International Conference on Robotics and Biomimetics
December 15 -18, 2007, Sanya, China

Multisensor Data Fusion for Joint People Tracking and Identification with a Service Robot

Nicola Bellotto and Huosheng Hu
Dept. of Computing and Electronic Systems, University of Essex
Colchester, CO4 3SQ, United Kingdom
{nbello, hhu}@essex.ac.uk

Abstract— **Tracking and recognizing people are essential skills modern service robots have to be provided with. The two tasks are generally performed independently, using ad-hoc solutions that first estimate the location of humans and then proceed with their identification. The solution presented in this paper, instead, is a general framework for tracking and recognizing people simultaneously with a mobile robot, where the estimates of the human location and identity are fused using probabilistic techniques. Our approach takes inspiration from recent implementations of joint tracking and classification, where the considered targets are mainly vehicles and aircrafts in military and civilian applications. We illustrate how people can be robustly tracked and recognized with a service robot using an improved histogram-based detection and multisensor data fusion. Some experiments in real challenging scenarios show the good performance of our solution.**

Index Terms— *People Tracking and Identification, Histogram-based Detection, Multisensor Data Fusion, Service Robotics.*

I. Introduction

The important developments of the last years in the robotic field are very encouraging and make us believe that the day when service robots will populate our lives is not too far. Since these will be placed in our environment, the study of the so called Human-Centred Robotics becomes essential for a synergistic cooperation between men and intelligent robots.

One of the first skills necessary for service robots is the ability to track and recognize people in the surrounding. This is essential for a mobile robot to perform some simple tasks, like following a person in a crowded environment, and more complicated behaviours, such as locating and identifying people to interact with. Many real applications can be found in literature where people tracking and recognition with a mobile robot play an important role. This is the case for example of the system described in [1], which uses a thermal and a normal camera to detect, track and recognize a person with a security robot. There is also a diffuse interest towards applications in public places for entertainment purposes, like the tour-guide robot presented in [2] that tracks visitors using two laser-range finders. In the new research area of socially assistive robotics [3], human tracking and recognition is also necessary for a robot to maintain an appropriate spatial distance from people and to engage in social interactions depending on their identity.

Vision-based tracking systems have shown to work well in simple situations where the direction of human targets is sufficient to accomplish the robot's task, or where the subject being tracked is a single person [4, 5]. However, in most of the cases, an accurate estimation of the targets location is necessary, for example if the robot has to keep a fixed distance from a moving person or interact with several people sparse in the environment. In situations like these, good performances can be achieved using multisensor data fusion, so to increase the detection range of the robot and, at the same time, reduce the uncertainty due to sensor data inaccuracy. Recent works report for example the integration of laser range finders with a unidirectional or omnidirectional cameras [6, 7].

Very often, however, tracking people is not enough for an interacting service robot. For example, if the latter works as receptionist for an exhibit, it would be preferable the robot welcomes only new visitors and avoids people already met before. This task requires at least a basic system for human recognition. The most common solutions are vision-based and often utilise color histograms to classify people wearing different clothes [8, 6].

Normally, people tracking and identification are two distinct processes executed sequentially without sharing information. The mutual benefits deriving from the combination of spatial and identity information have been shown with the mobile robot application of [4] and the simultaneous face tracking and recognition of [9].

With regards to the considerations above, two major contributions are given by the current work:

- An improved histogram comparison for human identification, which is a robust extension of the previous solution adopted in [6]. The method compares, in real-time, a template histogram with sub-regions of the image containing a person's torso. The best histogram match is located and used to update the target estimate.
- An original implementation of joint people tracking and identification for a service robot using multisensor data fusion. Instead of a general target classification based on different motion models, our human recognition is mainly done at observation level using the histogram information and is completely integrated in the Bayesian estimation performed with a bank of filters.

The paper is organized as follows. Section II illustrates the new histogram-based identification. An overview of our multisensor tracking is given in Section III. Section IV then explains the solution developed to perform joint people

978-1-4244-1758-2/08/$25.00 © 2008 IEEE.

tracking and identification. Some experiments with a mobile robot are described in Section V. Finally, the paper terminates with our conclusions and future work in Section VI.

II. Human Identification

People recognition is performed using color histogram comparison of the human torso, that is, the color of the clothes. The method explained in [6], which used a distance between histograms based on the Bhattacharyyas coefficient, has been extended and improved so to give a more reliable measure of similarity and additional input about the location of the best match. The information provided by the new histogram comparison is integrated in the probabilistic framework of the same Bayesian estimator used for tracking. Details are given in the following sections.

A. Histogram Comparison

The histogram comparison explained in [6] made use of a distance based on the Bhattacharyya's coefficient that showed to be very discriminative, yet quite robust to different human poses. The weak point of our first implementation, however, was the difficulty of selecting the proper image region that was supposed to contain the human body, in particular when the robot and/or the person were moving. If this selection was not accurate, the histogram considered could be completely different from the real one of the person. Also, such a distance was measured only with a single comparison of the whole region's histogram against that one of reference. In the approach here illustrated, instead, the distance between color histograms is calculated in a more robust way.

First of all, from the current image, a region of interest is selected according to the most recent estimate of the target's position and its uncertainty. To do this, we consider the 3D points $\mathbf{m} = [x, y, z]^T$ and $\boldsymbol{\sigma} = [\sigma_x, \sigma_y, \sigma_z]^T$, given respectively by the human torso coordinates and their standard deviations. Chosen a scale factor s, we project then $(\mathbf{m} - s\boldsymbol{\sigma})$ and $(\mathbf{m} + s\boldsymbol{\sigma})$ into the image plane, obtaining the relative pixels (u_-, v_-) and (u_+, v_+) that are the corners of the rectangular region to select. In our experiments, we found that a scale factor $s = 2$ guarantees a region sufficiently large to include always the target's torso. Fig. 1 shows an example of region selection.

Then, in a way similar to standard template matching techniques, the histogram of reference is compared to the histograms of all the sub-regions (with fixed size) inside the considered region. Instead of proceeding pixel by pixel, the sub-regions are selected using wider steps to reduce the computational cost and perform in real-time. In order also to limit the influence of light variations, histograms are calculated in the HSV color space from the Hue and Saturation components. The result of each comparison between sub-region and reference's histogram is stored in a matrix $\mathbf{D}$ with values between 0 and 1, the size of which depends on the selected region, sub-regions and pixel-step used.

The element $d \in \mathbf{D}$ with the smallest value is the minimum histogram distance, and its location inside the matrix indicates the sub-region of the image where the histogram of reference matches best. Eventually, the centre (u_c, v_c) of this sub-region can be used to calculate the direction of the target with respect to the camera.

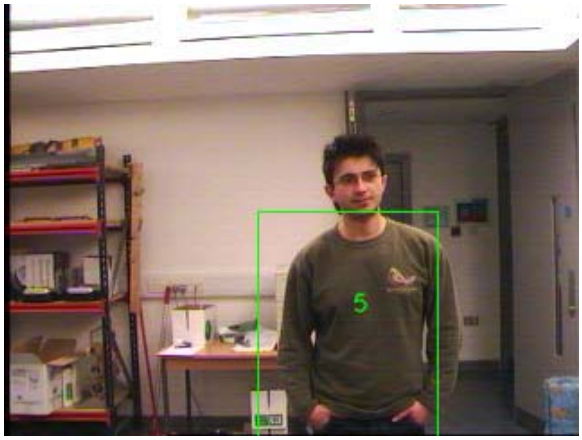


Fig. 1. Region from where histograms are extracted and then compared to that one in the database. The number 5, which indicates the record inside the current database, is centred on the sub-region where the histogram matches best, namely at the pixel location (u_c, v_c).

B. Clothes Detection

The histogram comparison explained above is the core of the clothes detection. This is basically giving the direction of the clothes (color histogram) for a selected subject of the database. Bearing and elevation are calculated from the sub-region's centre (u_c, v_c) using a pinhole-camera model and simple geometric transformations, as we already did for the face detection [6]. The histogram distance is also provided.

At time step k, the observation model of the clothes detection takes into account the current absolute position (x_k^c, y_k^c, z_k^c) of the camera and orientation ϕ_k^R of the robot, as given by the odometry, in order to get bearing α_k and elevation β_k of the torso centre, plus the distance d_k of the its color histogram:

$$\begin{cases} \alpha_k = \tan^{-1}\left(\dfrac{y_k - y_k^c}{x_k - x_k^c}\right) - \phi_k^R + n_k^\alpha \\ \beta_k = \tan^{-1}\left[\dfrac{z_k^c - \mu\, z_k}{\sqrt{(x_k - x_k^c)^2 + (y_k - y_k^c)^2}}\right] + n_k^\beta \\ d_k = n_k^d \end{cases} \quad (1)$$

The costant μ in the second member of (1) is chosen so that the product $\mu\, z_k$ corresponds approximately to the height of the torso centre. The noises n_k^α, n_k^β and n_k^d are zero-mean Gaussians with $\sigma_\alpha = \sigma_\beta = \frac{\pi}{30}$rad and $\sigma_d = 0.3$. For the latter, the assumption of Gaussianity is motivated also by the empirical results of [10].

This observation model is used by the clothes recognizers of the Bayesian estimators inside the bank of filters, as explained in Section IV, so to perform a new correction step whenever a tracked person is visible by the robot's camera.

III. Multisensor People Tracking

Most of the recent systems for people tracking rely on cameras or laser range sensors to detect humans [2, 4, 6, 7]. In our robot, both the devices are used, the first to detect faces and clothes, the second to find legs. After a data association step to assign new readings to the proper track, sensory information is fused using a recursive Bayesian estimation.

A. Legs and Face Detection

Legs are detected from a single laser scan using the same procedure described in [6]. Briefly, the algorithm recognizes patterns for three typical legs postures: legs apart, forward straddle and two legs together (or single leg). The method is computationally inexpensive and works well even in cluttered environments.

The face detection is based on the real-time solution of [11]. With simple geometric transformations, the direction of the face is calculated from its location inside the image [12].

B. Data Association and Tracks Creation

We adopt Nearest Neighbour (NN) data association to handle multiple targets [13]. False readings are excluded using a validation gate and the same similarity measure adopted in [12] is used for the creation of the association matrices.

Sensor readings discarded by the gating or the assignment procedure are used to create new candidate tracks. A track is eventually deleted from the database if not updated for more than a certain time or if the uncertainty of its position is too large.

C. State Estimation

For the prediction of the human motion, we adopt an extension of the constant velocity model, where the state includes position (x_k, y_k), height z_k, orientation ϕ_k and velocity v_k of the human target. The observation models of the legs and face detection take into account the current position of the robot, as provided by odometry. Details are given in [12].

The estimation can be theoretically performed with any recursive Bayesian filter. However, for this specific task, the Unscented Kalman Filter (UKF) and the SIR Particle Filter (with at least 500 samples) showed to perform particularly well and give similar results [12]. Using the relative observation models, the estimation is updated whenever new data are available from the laser or the camera.

IV. Joint Tracking and Identification

In recent years, the problem of tracking and classifying targets simultaneously has been studied in particular for military and civilian applications, where the targets of interest are mainly vehicles and aircrafts [14, 15]. Out of this context, for which most of the results are available only in simulation, only little work has been done for object classification [16], face recognition [9] or people tracking and identification [17]. These are generally limited to the case of a single target, with a single sensor or from stationary positions.

A. Bank of Filters

The optimal solution for joint target tracking and classification consists of a bank of class-matched filters [14] characterized by different motion (or prediction) models. At every time step k, each filter outputs the likelihood $\lambda_k^i = p(\mathbf{z}_k|\mathbf{Z}_{k-1}, c_i)$ of the class c_i, with $\mathbf{Z}_k = \{\mathbf{z}_m\}_{m=0}^k$, which is used to update recursively the class probabilities as follows:

$$p\left(c_i|\mathbf{Z}_k\right) \propto \lambda_k^i \; p\left(c_i|\mathbf{Z}_{k-1}\right) \tag{2}$$

Normally, for Kalman filters, the considered likelihood has a zero-mean normal distribution [18]:

$$\lambda_k^i = \mathcal{N}\left(\nu_k; \mathbf{0}, \mathbf{S}_k^i\right) \tag{3}$$

where ν_k is the innovation term and $\mathbf{S}_k^i$ the relative covariance. Particle filters, instead, provide a class likelihood that is implicit in the normalization step of the estimation [14] and is calculated as follows:

$$\lambda_k^i = \sum_{i=1}^{M} \tilde{w}_k^i \tag{4}$$

where $\{\tilde{w}_k^i\}_{i=1}^M$ are the weights of the M samples. If c_i is the identity of the target (or some feature that characterizes it), the system can perform simultaneous tracking and identification.

B. System Architecture

Instead of classifying targets according only to their prediction models, in our system the differentiation is mainly done at observation level, that is, each filter is updated with a target-specific input, which gives a "measure" of the target's identity. This is implemented using a modular approach where the "detectors", used to measure the human position (i.e. laser-based legs detector, vision-based face detector), are accompanied by "recognizers", which measure the similarity between the current observation and the information stored in a database (e.g. clothes recognizer, face recognizer, voice recognizer, etc.). The system is schematically illustrated in Fig. 2. At each time step, the considered estimate of the current track is that of the n^{th} filter for which the identity probability $p(n|z_k)$ is maximal.

The database contains pre-recorded information about known subjects. In the current implementation, for each person the database contains his height and the color histogram of his torso. The bank of filters, then, has one estimator for each subject. With respect to our previous implementation [12], besides the histogram observation model (1), these filters differ on the human height component of the prediction model. In absence of identity information, this was modelled as $z_k = z_{k-1} + w_{k-1}$, where the last term was Gaussian noise. Since the database now contains height information, the new model is the following one:

$$z_k = H_n + w_{k-1} \tag{5}$$

where H_n is the known height of the n^{th} subject. The concept is similar to a target classification based on different motion

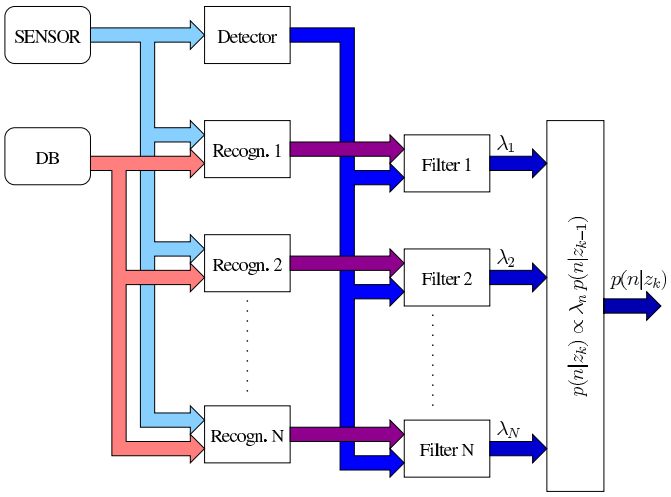


Fig. 2. Bank of filters for joint people tracking and identification. Each filter corresponds to a subject stored in the database.

Fig. 3. Mobile robot with laser and camera.

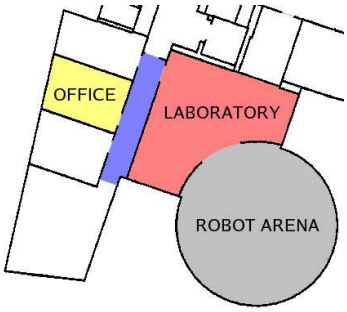


Fig. 4. Plan of the environment used for experiments.

models, as explained in IV-A, and has a direct influence on the prediction of the face and clothes observations.

Since the number N of filters (and recognizers) inside a bank is equal to the number of subjects in the database, its maximum size depends on the available computational power. Using UKF, which provides fast and accurate estimations [12], our system could work in real-time, on a slow PC, with $N = 10$ and more. Particle filters and larger databases can be used with faster hardware.

V. Experimental Results

To test our system, we recorded sensory data using the mobile robot shown in Fig. 3 and performed several off-line trials. The robot was remotely controlled to approach some people and to follow them across different rooms. In the experiments described next, UKFs were used for the bank of filters, the good accuracy of which was already shown in our previous work [12]. The database was created manually and included information relative to 10 different subjects, each record containing the height of a person and a histogram of his torso. Note also that some of the subjects had similar clothes or height, making the identification even more difficult.

A. Identity Probability

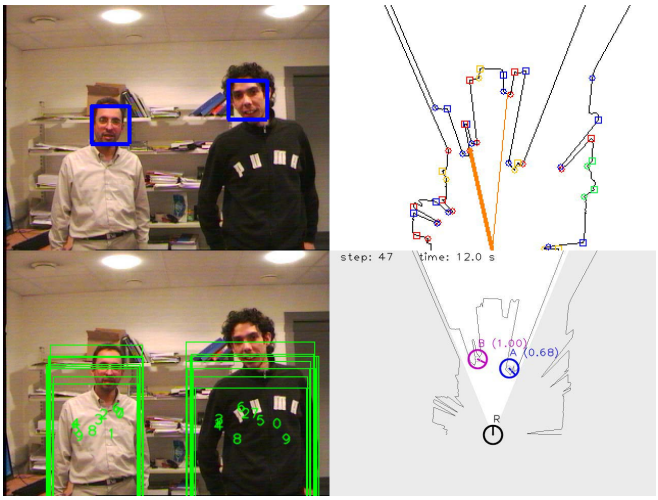
The system was tested in different rooms and with several people in a typical office environment. To show the performance of the human identification while tracking, we report a case where the robot had to follow and recognize people in a possible situation for a service robot. The reconstructed scenario was that of an instructor (person A) introducing the robot to some other people (B, C and D) located in different rooms. The action started in an office and continued across a corridor to reach other people in an adjacent laboratory (see Fig. 4). The approximate length of the experiment was 70s.

The first two people, A and B, are shown in Fig. 5(a), together with the relative vision and laser-based detections. Please note that possible false positives in the legs or face detection did not influence the correctness of the estimation since they were properly discarded by the gating procedure. The positional estimates are also illustrated and labelled according to their probability identities, the evolutions of which are graphically reported in Fig. 6(a) and 6(b).

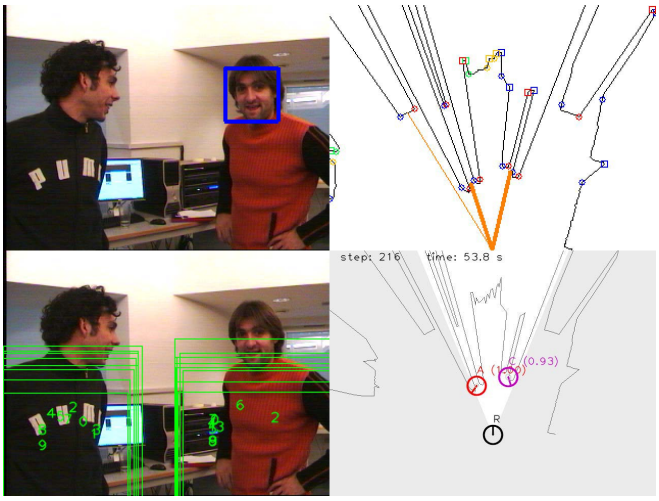
From the office, the instructor A moved then to the laboratory, always followed and tracked by the robot, to meet the other persons C and D, who are shown in Fig. 5(b) and 5(c). Approaching the last one, while temporarily out of the camera's field of view, the track associated to A "jumped" erroneously to the adjacent person D. This is the reason why, in the last part of Fig. 6(a), the estimated identity switched from A to D, updating correctly the label of the current track. The A identity was correctly assigned then to the last track, as shown by the graph in Fig. 6(d). In practice, the system is able to promptly recover from possible tracking errors as soon as visual information is available.

B. Improvement of the Tracking Robustness

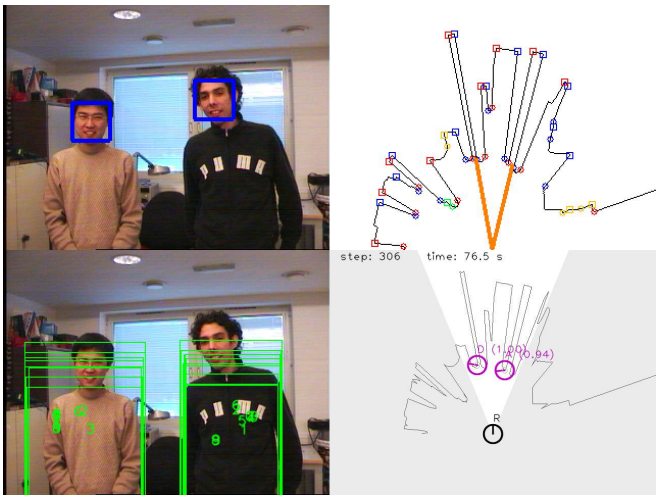
Another important feature of the system, that derives from the integration of the histogram-based identification, is the robust tracking in challenging situations. There are cases where people are walking very close and it is difficult to distinguish them properly using only range information, like the situation depicted in Fig. 7. Since human motion is very



(a) Targets A and B, respectively on the right and left, in the office.

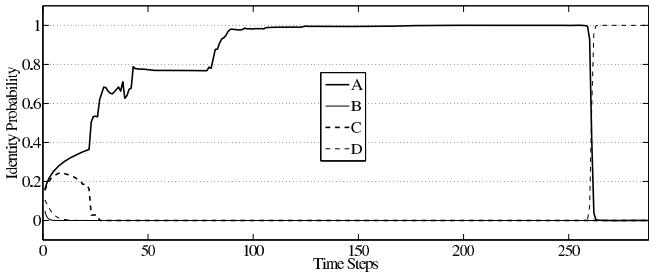


(b) Targets A and C in the laboratory.

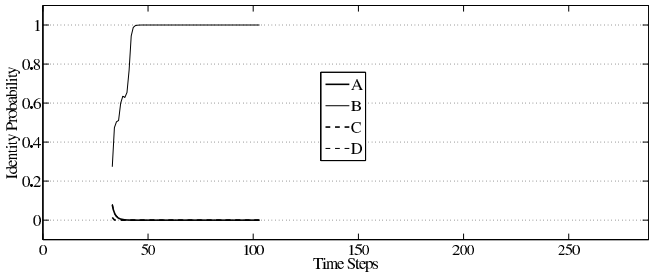


(c) Targets A and D in the laboratory.

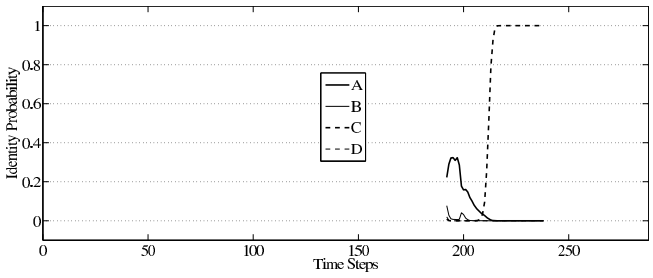
Fig. 5. Joint Tracking and Identification. For each figure, the two frames on the top-left and top-right show respectively the outputs of the face and legs detection (as seen by the robot). The bottom-left, instead, is the histogram-based clothes recognition (a different region is selected for each subject of the database). Finally, the bottom-right frame illustrates the robot R, the track estimates with their labels and the current probability of the relative identity.



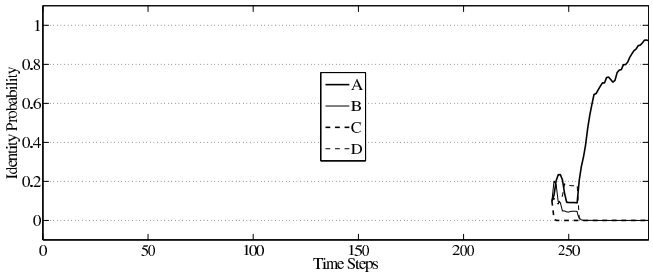
(a) Identity of the first track.



(b) Identity of the second track.



(c) Identity of the third track.



(d) Identity of the forth track.

Fig. 6. Identity probabilities. The start points of the graphs are different since they are relative to tracks created in different moments. Note that the identity of the first and the last track swapped soon after the time step 250.

unpredictable, adjacent tracks could often swap and keeping them fixed to the same targets becomes a hard task.

The sequence illustrated in Fig. 8 shows the different performance of two single estimators using only legs and face detection (left column), and the result of the joint tracking and identification including clothes recognition (right column). In the first case, the tracks E and F swapped due to their proximity and current uncertainty of the estimates. In the second case, instead, the two banks of filters were successfully updated by the additional information of the clothes detection modules (bearing, elevation and histogram distance of the torso), so the tracks were kept on the proper sides.

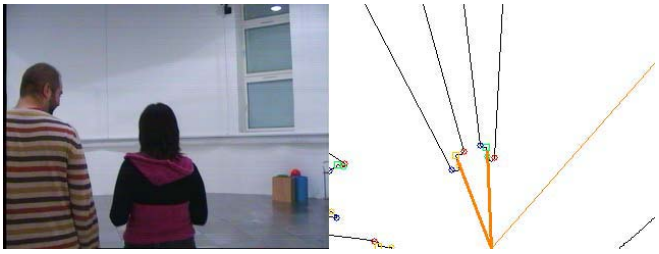


Fig. 7. Case of adjacent people walking very close to each other. Since they are not facing the camera, the tracking can rely only on the legs detection and clothes recognition.

VI. Conclusions and Future Work

The paper presented an improved histogram comparison for the detection and recognition of human clothes. The information provided have been integrated in a system designed to perform joint people tracking and identification. A solution based on a bank of filters, which uses sensor fusion to combine visual and laser range data, has been illustrated. The effectiveness and robustness of the tracking and identification have been tested with several experiments. The results show the high performance of the system and its strong potential for service robotics applications.

Further experiments in various conditions, such as colorful clothes with some patterns, should be also carried out, and the performance compared with conventional techniques. The next implementation of our system, which is the first part of a more complex project for a receptionist robot, will include face recognition and automatic update of the human database. This will help to handle the case where people change their clothes and to distinguish between known and unknown users.

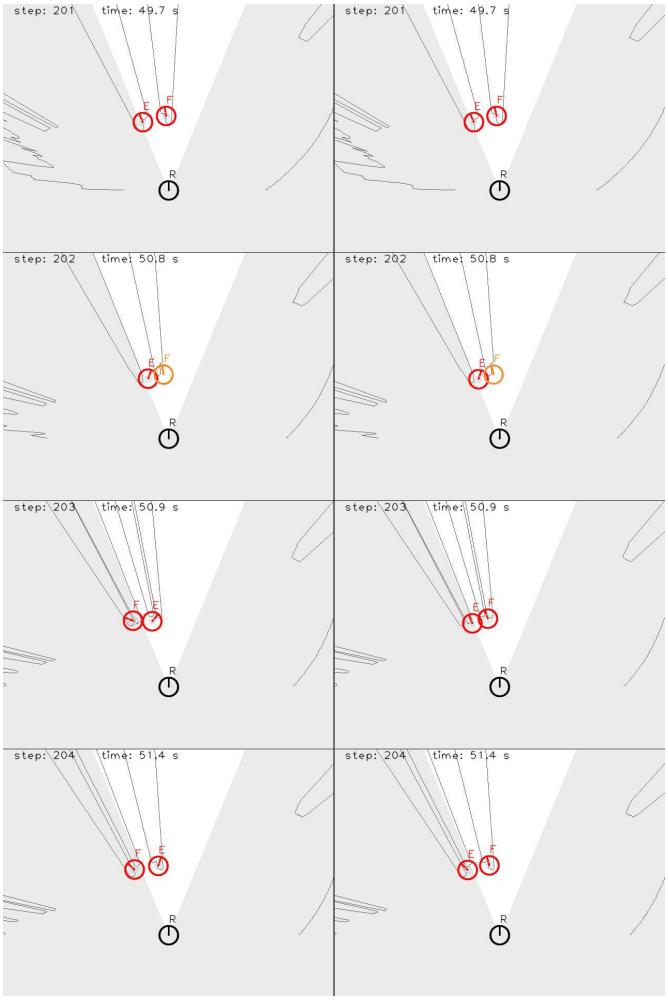


Fig. 8. Sequence of people being tracked by normal UKFs (on the left) and by banks of filters (on the right). Tracks E and F swapped erroneously in the first case, but kept the proper side in the second.

References

[1] A. Treptow, G. Cielniak, and T. Duckett, "Active people recognition using thermal and grey images on a mobile security robot," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Canada, 2005, pp. 2103–2108.

[2] W. Burgard, P. Trahanias, D. Hähnel, M. Moors, D. Schulz, H. Baltzakis, and A. A., "TOURBOT and WebFAIR: Web-Operated Mobile Robots for Tele-Presence in Populated Exhibitions," in *Proc. of the IROS 2002 Workshop on Robots in Exhibitions*, 2002.

[3] A. Tapus, M. J. Mataric, and B. Scasselati, "Socially assistive robotics," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 35–42, 2007.

[4] W. Zajdel, Z. Zivkovic, and B. J. A. Krõse, "Keeping track of humans: Have I seen this person before?" in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, Barcelona, Spain, 2005, pp. 2093–2098.

[5] D. Calisi, L. Iocchi, and G. R. Leone, "Person following through appearance models and stereo vision using a mobile robot," in *Proc.of Int. Workshop on Robot Vision*, 2007.

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

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

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

[9] S. Zhou and R. Chellappa, "Probabilistic human recognition from video," in *Proc. of the 7th European Conference on Computer Vision (ECCV)*. London, UK: Springer-Verlag, 2002, pp. 681–697.

[10] P. Pérez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. of IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[12] N. Bellotto and H. Hu, "People Tracking with a Mobile Robot: a Comparison of Kalman and Particle Filters," in *Proc. of the 13th IASTED Int. Conf. on Robotics and Applications*, Würzburg, Germany, 2007, pp. 388–393.

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

[14] N. J. Gordon, S. Maskell, and T. Kirubarajan, "Efficient particle filters for joint tracking and classification," in *Proc. of Signal and Data Processing of Small Targets (SPIE)*, FL, USA, 2002, pp. 439–449.

[15] P. Smets and B. Ristic, "Kalman filter and joint tracking and classification in the TBM framework," in *Proc. of the 7th Int. Conf. on Information Fusion*, Stockholm, Sweden, 2004, pp. 46–53.

[16] P. Minvielle, A. Marrs, S. Maskell, and A. Doucet, "Joint target tracking and identification - part ii: Shape video computing," in *Proc. of the 8th Int. Conf. on Information Fusion*, Philadelphia, PA, USA, 2005.

[17] D. Schulz, D. Fox, and J. Hightower, "People tracking with anonymous and id-sensors using rao-blackwellised particle filters," in *Proc. of the Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Acapulco, Mexico, 2003, pp. 921–926.

[18] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: particle filters for tracking applications*. Artech House, 2004.