Noname manuscript No.
(will be inserted by the editor)

A Bank of Unscented Kalman Filters for Multimodal Human Perception with Mobile Service Robots

Nicola Bellotto · Huosheng Hu

Received: date / Accepted: date

Abstract A new generation of mobile service robots could be ready soon to operate in human environments if they can robustly estimate position and identity of surrounding people. Researchers in this field face a number of challenging problems, among which sensor uncertainties and real-time constraints. In this paper, we propose a novel and efficient solution for simultaneous tracking and recognition of people within the observation range of a mobile robot. Multisensor techniques for legs and face detection are fused in a robust probabilistic framework to height, clothes and face recognition algorithms. The system is based on an efficient bank of Unscented Kalman Filters that keeps a multi-hypothesis estimate of the person being tracked, including the case where the latter is unknown to the robot. Several experiments with real mobile robots are presented to validate the proposed approach. They show that our solutions can improve the robot's perception and recognition of humans, providing a useful contribution for the future application of service robotics.

Keywords Robot Perception · Human Tracking and Recognition · Bayesian Estimation · Service Robotics

1 Introduction

People tracking algorithms try to estimate the position of humans in the environment from noisy measurements. However, service robots must also distinguish and recognize different persons, which are otherwise treated as simple moving objects. Without recognition, robots would not be able to deal with the actual user needs and they could not deliver high-quality services.

Mobile service robots acting in human environments and provided with interaction skills (e.g. tour-guide or security robots) are more effective if they can distinguish between new and former users, or between public visitors and staff members. Besides these applications, the capability to identify people gives robots a certain grade of "social intelligence" [19] as they can better adapt to individual human behaviours.

Most of existing robotic systems provided with vision-based human recognition operate in two separate steps: first, a frame is selected where the subjects satisfies some criteria, like pose, size or number of visible features; then, some standard recognition algorithm is applied versus a fixed database of known people [2, 15, 30]. Unfortunately, this approach ignores important clues like time and spatial evolution of the subject to be identified. This information can be provided by human trackers and used to improve the robot's recognition system, while the latter can increase the robustness of the tracking process itself.

In this paper we propose a novel solution for simultaneous tracking and identification of humans with mobile service robots, which integrates several detection and recognition algorithms in a robust probabilistic framework, making use of different data sources and a bank of Bayesian filters. A new histogram-based recognition algorithm for human clothes is presented, which takes into account the uncertainty of the human position to select the image region where the histogram match is best. The success rate of human identification is increased with a simple face recognition algorithm, which makes use of an improved method for fast face alignment and scaling. Finally, these recognition

N. Bellotto
School of Computer Science
University of Lincoln
Lincoln LN6 7TS, United Kingdom
E-mail: nbellotto@lincoln.ac.uk

H. Hu
School of Computer Science and Electronic Engineering
University of Essex
Colchester CO4 3SQ, United Kingdom

algorithms are fused with human height and 2D spatial information, thanks to a modularized architecture that keeps multi-hypothesis estimates of the subject being tracked.

The remainder of the paper is organized as follows. Section 2 presents and overview of related research work. An overview of the multisensor human detection is introduced in Section 3, while the algorithms for vision-based recognition are illustrated in Section 4. The following Section 5 introduces a bank of Bayesian filters and describes the architecture implemented for simultaneous people tracking and recognition. Several experimental results with mobile robots are presented and discussed in Section 6. Finally, Section 7 concludes the paper with a summary of the progresses achieved and future research directions.

2 Related Work

In the context of social robotics [19, 22], human detection, tracking and recognition are essential prerequisites for successful applications. Before starting close interactions, indeed, mobile social robot have to detect people and approach them. This is also necessary for robots to attract human attention and explicitly ask for help in case they need assistance [34].

In literature, different solutions for tracking humans with mobile robots are reported. Many applications use robots' on-board cameras to detect people, often searching for faces [30] or other body parts [12, 42]. Several approaches instead make use of range sensors, considering people as moving entities [29, 9].

Human recognition is another broad research field that includes detection and interpretation of biometric features [26]. Most solutions are vision-based and use algorithms for face recognition [40, 43, 44], or in some case gait and full body analysis [18, 33]. However, just a few recognition systems are actually implemented on real mobile robots, as their perception capabilities are limited by sensor uncertainty, motion and changes in the environment [2, 8, 15].

Humans have articulated body postures and motion behaviours, difficult to be modelled and observed from a robot platform. In order to reduce uncertainty and improve redundancy, two or more sensors can be integrated for human detection and recognition. Heuristic approaches are in some cases used to estimate the position of a person with mobile robots equipped with laser and camera [38]. These sensors can be combined to reduce the searching area during human detection, using the laser to find the direction of possible targets and applying face detection only to small portions of the current frame [13]. Thermal and CCD cameras are used in [16] for human recognition, segmenting people's contours on the thermal image and applying a Bayesian classifier to the relative region on the colour image.

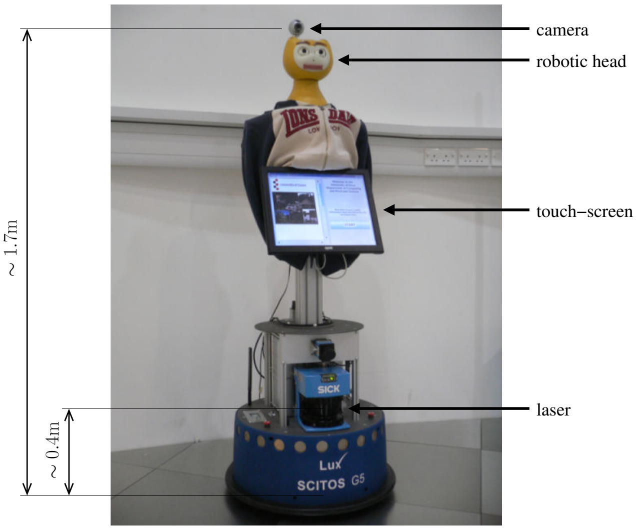


Fig. 1 The Scitos G5 robot with laser, camera and interaction devices.

Probabilistic methods for robot sensor fusion have also been proposed. The system implemented in [21] and [7] adopts Kalman filters to track people using lasers and cameras mounted on mobile robots. In [39], the authors illustrate a robot equipped with two laser range sensors that can track several people using a combination of particle filters and probabilistic data association. Another solution based on particle filter is proposed in [14], which integrates laser data and visual information from a panoramic camera. A covariance intersection method, using sonar, laser and visual data, is implemented in [31] for tracking multiple people. The last two implementations, however, are evaluated only with static robot platforms.

3 Multisensor Human Detection

In this work we consider mobile robots equipped with SICK laser range sensors and colour cameras. The two robotic platforms used in our research are a Scitos G5, shown in Fig. 1, and a Pioneer 2 DX with on-board PCs and similar sensor configurations, illustrated in Fig. 2.

3.1 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. Most of the existing legs detection algorithms are based on the search of local minima [10, 39], motion detection [14, 29] or machine learning techniques [1].

The legs detection algorithm implemented in this work is based on the recognition of typical legs patterns extracted from a single laser scan. These patterns correspond to three possible postures: legs apart, forward straddle and two legs

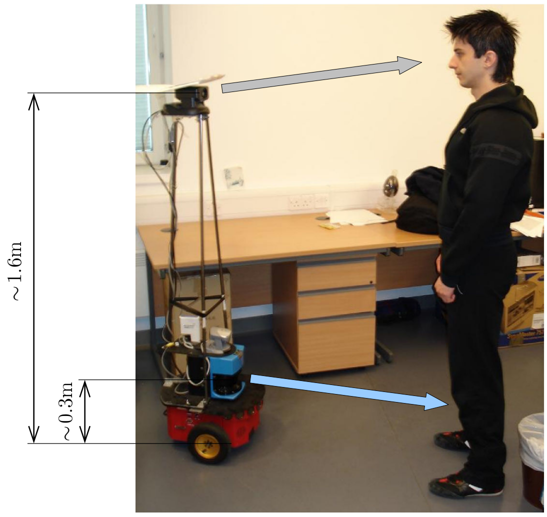


Fig. 2 The Pioneer 2 DX robot with laser and camera, detecting legs and face respectively.

together (or single leg). An example of legs detection is illustrated in Fig. 3. Briefly, the algorithm initially filters laser data in order to smooth the readings, then detects all the edges lying on the directions of the laser scans. Groups of adjacent edges, possibly corresponding to legs, are identified according to simple geometric relations and spatial constraints. Fig. 4 shows a scan of the three different legs postures with the angle of the laser beam in the abscissa and the measured range in the ordinate. Direction and distance of each legs pattern is computed from the midpoint (red cross) between the extremes (black circles) of the outer vertical edges. The method is quite robust even in case of cluttered environments and, besides being computationally efficient, it is not influenced by the robot's motion. Further details and comparisons to other techniques can be found in [7].

3.2 Face Detection

The camera on the robot can be used to detect faces and recognize people. Some of the most popular techniques to perform real-time face detection are based on the color segmentation of skin regions [23], but these are usually prone to errors due to light variations, shadows and skin tones. Like in our previous work [7], the face detection algorithm implemented in the current system is based on the solution of Viola & Jones [41], which offers a good balance between detection performance and computational efficiency. Fig. 3 shows an example of face detection with the robot's camera.

4 Vision-based Recognition

The algorithms described next are used for clothes and faces recognition. Although the former alone cannot provide a biometric measure for robust human identification, it can greatly improve the system performance when combined with height and face recognition, as later shown in our experiments.

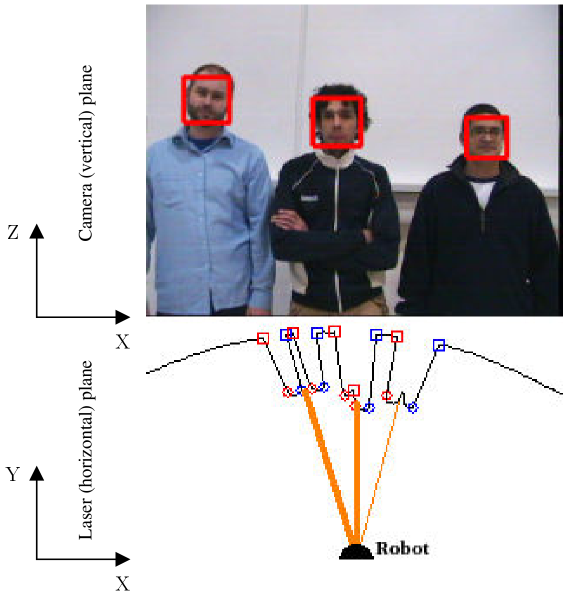


Fig. 3 Face and legs detection. On the bottom, from left to right, three different legs postures can be noted from the laser scan: legs apart, forward straddle and two legs together.

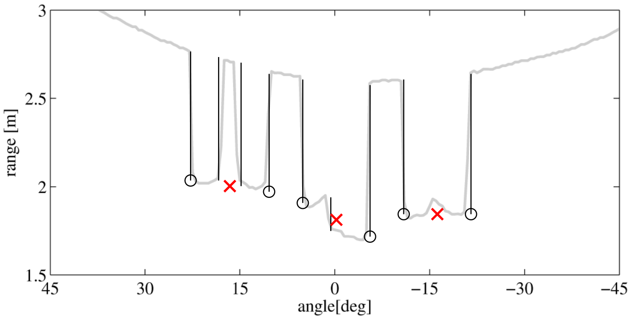


Fig. 4 Legs patterns and relative midpoints for measuring their direction and distance.

4.1 Clothes Recognition

Clothes recognition is performed using an improved version of the color histogram comparison described in [5]. Since the main task of the robot is to have close interactions with humans, it is generally not possible to consider the histogram of the whole body, so the region of interest (ROI) is limited to the human torso, which is the only part always visible from the camera when the robot is at a minimum distance from the person (at least 2m in our case).

An efficient measure to compare color histograms is the one adopted for the mean-shift tracking algorithm [17], which is based on the sample estimate of the Bhattacharyya coefficient. Given a discrete normalized density of reference $\mathbf{q} = \{q_u\}_{u=1\ldots m}$ (i.e. an m-bin histogram) and the one to be compared $\mathbf{p}(R) = \{p_u(R)\}_{u=1\ldots m}$, where R is the ROI, the sample estimate of the Bhattacharyya coefficient is the following:

$$\rho\left[\mathbf{p}(R), \mathbf{q}\right] = \sum_{u=1}^{m} \sqrt{p_u(R)\, q_u} \tag{1}$$

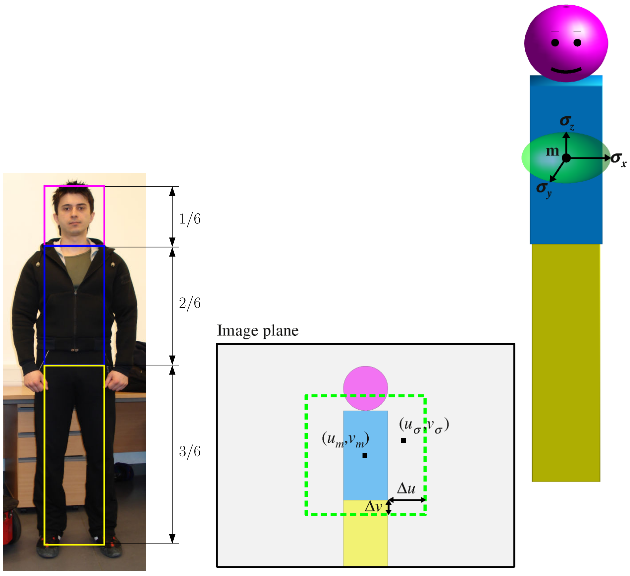
Using (1), the distance between the two distributions is defined as follows:

$$d\,(R) \equiv d\left[\mathbf{p}(R), \mathbf{q}\right] = \sqrt{1 - \rho\left[\mathbf{p}(R), \mathbf{q}\right]} \tag{2}$$

Since based on discrete densities, this distance is scale invariant and is normalized between 0 and 1. From empirical tests on a number of subjects, it showed also to be very discriminative, yet quite robust to different human poses.

Some of the histogram-based techniques for human recognition rely on a precise calibration between camera and laser range finder to select the body region on the current frame [8]. In case the laser is not available, motion detection techniques are used to highlight the ROI [42]. It is very difficult however to select precisely the correct body region on the current frame, in particular when the robot and the person are moving. If this selection is not accurate, the histogram considered might be completely wrong. A fixed scale factor to increase the size of the ROI does not solve the problem, as the measure could be seriously influenced by other objects on the background. Differently from other works [8, 15, 33], the algorithm described next explicitly considers the uncertainty from tracking, and therefore does not need an accurate calibration between camera and laser. The selection procedure is also independent from lighting conditions and related problems that usually affect video segmentation techniques.

The distance between color histograms is calculated using a selection procedure of the ROI that takes into account the uncertainty of the current human estimate. The considered body proportions are those proposed in [15], and illustrated in Fig. 5(a), where the torso is $2/6$ of the total height. Given the current 3D estimate $[x_k, y_k, z_k]^T$ of the face position, the centre of the torso $\mathbf{m} = [x_k, y_k, \eta\, z_k]^T$ is initially determined, where $\eta = 8/11$ is a constant value calculated considering the abovementioned human proportions. The point $\mathbf{m}$ is projected onto the image plane, obtaining the relative pixel (u_m, v_m). This is the centre of the initial torso region R_t, the size of which is also set according to the given body proportions (i.e. blue rectangle in Fig. 5(a)). Note that clothes recognition can be applied even when the person is not facing the camera because R_t is proportional to the estimated height, which is kept in the state as long as the subject is being tracked.



(a) Body proportions. (b) Procedure for histogram region selection.

Fig. 5 Selection of the region for clothes recognition.

Given the vector of standard deviations $\boldsymbol{\sigma} = [\sigma_x, \sigma_y, \sigma_z]^T$ from the covariance matrix of the current human estimate and a scale factor s, the point $(\mathbf{m} + s\boldsymbol{\sigma})$ is also projected on the relative pixel (u_σ, v_σ). The differences $\Delta u = |u_\sigma - u_m|$ and $\Delta v = |v_\sigma - v_m|$ are the quantities used to extend the initial ROI and get the new R_σ, as illustrated in Fig. 5(b). A scale factor $s = 2$ guarantees a region sufficiently large to include the targets torso in most of the situations.

In a way similar to standard template matching, the histogram of reference $\mathbf{q}$ is compared to the histograms of all the sub-regions R, with the same size of R_t, inside the considered region R_σ. The histogram of reference is provided by a fixed database of known subjects. This is manually initialized before operation, although in the future it would be desirable to include an automatic initialization and update of the database. In order to limit the influence of light variations, histograms are calculated in the HSV color space from the Hue and Saturation components.

The sub-region R^* for which the distance $d^* \equiv d(R^*)$ is minimum is where the histogram of reference matches best. The centre (u^*, v^*) of R^* can be used to calculate the direction of the human target with respect to the camera. The whole procedure is briefly described in Algorithm 1.

4.2 Face Recognition

During the last decades, many solutions have been proposed for the challenging task of face recognition [43]. Most of the initial work concentrates on recognition from still images, but the recent availability of fast algorithms for real-time face detection made possible the recognition on video sequences. However, a number of problems, like head pose, lighting condition and low resolution cameras, makes face

Algorithm 1 Histogram-based Detection

```
Input: estimated $x_k$, $y_k$ and $z_k$, with relative $\sigma_x$, $\sigma_y$ and $\sigma_z$
Output: position $(u^*, v^*)$ and distance $d^*$

$\mathbf{m} \Leftarrow [x_k, y_k, \eta\, z_k]^T$                                ▷ initialize ROI
$\boldsymbol{\sigma} \Leftarrow [\sigma_x, \sigma_y, \sigma_z]^T$
project $\mathbf{m}$ and $(\mathbf{m} + s\boldsymbol{\sigma})$ on image plane to obtain, respectively,
    $(u_m, v_m)$ and $(u_\sigma, v_\sigma)$
$\Delta u \Leftarrow |u_\sigma - u_m|$
$\Delta v \Leftarrow |v_\sigma - v_m|$
select initial ROI $R_t$ from the torso, centered in $\mathbf{m}$, as shown in
    Fig. 5(a)
calculate new ROI $R_\sigma$ increasing $R_t$ by $2\Delta u$ and $2\Delta v$, as shown
    in Fig. 5(b)
$d^* \Leftarrow \infty$                                          ▷ initialize histogram match
$(u^*, v^*) \Leftarrow (u_m, v_m)$
get histogram of reference $\mathbf{q}$
for all $R \in R_\sigma$ with $R$ centred in $(u_\mathrm{R}, v_\mathrm{R})$ and having the same
    size of $R_t$ do
    if $d(R) < d^*$ then                                         ▷ new best match found
        $(u^*, v^*) \Leftarrow (u_\mathrm{R}, v_\mathrm{R})$                    ▷ memorize match position
        $d^* \Leftarrow d(R)$                                     ▷ memorize histogram distance
    end if
end for
```

recognition in video more difficult. On the other hand, videos contain important time and spatial information, not available otherwise from still images [44].

In general, the whole process of face recognition consists of three main steps: detect a face from the current frame, process the relative image region and finally apply some recognition algorithm. Processing the considered image region is one of the most crucial part of every identification system. To align a face horizontally, a common technique is to locate the eyes' position and rotate the face image, so that their final inclination is null. The distance between the eyes is used to resize the face to a pre-determined value. The processed face can be compared to a reference template, that in our system is the canonical face representation proposed in [25] and shown in Fig. 6.

A fast algorithm for eye detection has been proposed in [28] and is based on the extraction of histogram minima within sub-regions containing the eyes. The method relies on the assumption that the iris' color is darker than the surrounding region, which is not always true. The method illustrated in [20] makes use of a more robust probabilistic approach that takes into account the uncertainty of face detection. Unfortunately the implementation of the latter could not work in real-time on the available robots.

The solution developed for our system is a fast, color-independent procedure based on the same algorithm used for face detection [41, 37]. Using two classifiers, one trained for right eyes and another for left eyes, two independent local searches are performed on specific sub-regions of the face bounding box. With reference to Fig. 6, the regions scanned are the $2w \times 2w$ top-left and top-right areas. If more than one eye is found within the considered region (false positives), the detection closest to its centre is chosen.

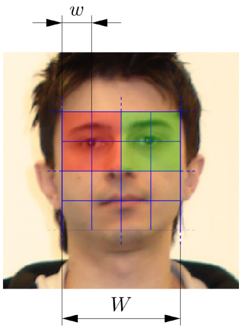


Fig. 6 Canonical model for face processing and recognition.

In order to align it horizontally, the face is rotated of an angle α_{RL} calculated from the positions $(u_\mathrm{R}, v_\mathrm{R})$ and $(u_\mathrm{L}, v_\mathrm{L})$ of the right and left eye respectively. From the model in Fig. 6, where the distance $d_{RL} = 2w$ between eyes is half the size of the face, the scaling factor to obtain a face of size $W \times W$ is $s = W/(2\ d_{RL})$. Rotation and scaling of the face, centred on the right eye, are performed with an affine transformation as follows:

$$
\begin{aligned}
a &= s\ \cos\left(\alpha_{RL}\right) \\
b &= s\ \sin\left(\alpha_{\mathrm{RL}}\right) \\
\begin{bmatrix} u' \\ v' \end{bmatrix} &= \begin{bmatrix} a & b & (1-a)\, u_\mathrm{R} - b\, v_\mathrm{R} \\ -b & a & b\, u_\mathrm{R} + (1-a)\, v_\mathrm{R} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}
\end{aligned}
\tag{3}
$$

where (u, v) is a pixel of the source image and (u', v') of the destination. Note that, in order to avoid possible outliers of the rotated face image, the affine transformation is actually applied to a sub-region slightly bigger than the original face bounding box, so a correction term for the offset is added to the coordinates $(u_\mathrm{R}, v_\mathrm{R})$ of the right eye. Eventually, the resulting face will be $W \times W$ pixels, with the right and left eyes centred in (w, w) and $(3w, w)$ respectively.

The last step includes cropping the face's area with an elliptical mask. This reduces the influence of hair and background pixels on the four corners of the rectangular region. Then, considering only the area within the ellipse, the face is equalized and normalized so that the distribution of the pixels intensity has zero mean and standard deviation one. An example of face processing, from detection to normalization, is shown in Fig. 7.

Finally, the popular *Eigenface* recognition algorithm [40] is applied to the normalized face versus a database of known faces. The difference between current and reference face is a quantity ξ, between -1 and 0, given by the standard Mahalanobis cosine [11].

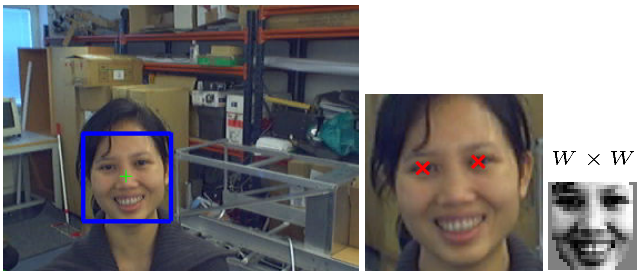


Fig. 7 Image processing before recognition. The detected face is shown on the left, the position of the eyes in the middle and the final result on the right.

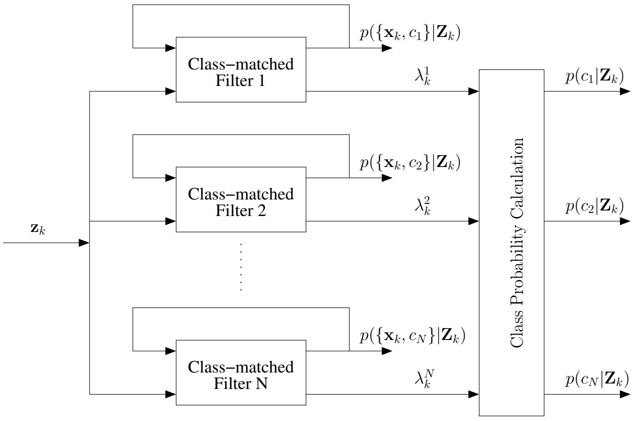


Fig. 8 Schematic representation of a generic bank of filters.

5 Simultaneous Human Tracking and Recognition

A part for military applications, where most of the results are available only in simulation, little work has been done so far for joint tracking and classification of objects and humans [32, 44]. Current solutions are generally based on the use of single sensor data and are often constrained by computation resources, which make them unfeasible for robot applications. The proposed approach, instead, is an effective solution that uses sensor fusion to track and recognize humans simultaneously and in real-time.

5.1 Bank of Filters

The estimate of a particular target at time t_k can be expressed by the joint state $\{\mathbf{x}_k, c_i\}$, where $\mathbf{x}_k \in \mathbb{R}^n$ is a vector containing information like position and velocity of the target, and c_i (with $i = 1, \ldots, N$) is a time-invariant attribute, or target class. Given the sequence of measurements $\mathbf{Z}_k = \{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$, the prior distribution of the i^{th} joint state can be written as follows [24]:

$$p(\{\mathbf{x}_k, c_i\}|\mathbf{Z}_{k-1}) \\ = \int_{\mathbb{R}^n} p(\mathbf{x}_k|\{\mathbf{x}_{k-1}, c_i\})\, p(\{\mathbf{x}_{k-1}, c_i\}|\mathbf{Z}_{k-1})\, d\mathbf{x}_{k-1} \quad (4)$$

Applying Bayes' rule, the (normalized) posterior is calculated using the following equation:

$$p(\{\mathbf{x}_k, c_i\}|\mathbf{Z}_k) \propto p(\mathbf{z}_k|\{\mathbf{x}_k, c_i\})\, p(\{\mathbf{x}_k, c_i\}|\mathbf{Z}_{k-1}) \quad (5)$$

Equations (4) and (5) form a recursive estimation for tracking the i^{th} target.

It is also possible to write a recursive update of the class probability with the following expression:

$$p(c_i|\mathbf{Z}_k) \propto \lambda_k^i\, p(c_i|\mathbf{Z}_{k-1}) \quad (6)$$

where $\lambda_k^i = p(\mathbf{z}_k|\mathbf{Z}_{k-1}, c_i)$ is the likelihood function of class i.

The optimal solution for joint target tracking and classification is a bank of class-matched filters that, in absence of specific feature observations, is characterized by different prediction models [24, 36]. Any combination of Bayesian estimators can be used: for example, a BoF can be built using both Kalman and particle filters, as long as each of them provides a class likelihood. The advantage is that the designer can choose the most appropriate filters for a particular class of targets, depending on state evolution (i.e. prediction model) and sensor used (i.e. observation model). At every time step k, each filter outputs the likelihood λ_k^i of the relative class c_i which is used to update recursively the class probabilities with (6). A typical bank of filters (BoF) is schematically represented in Fig. 8.

Given a BoF with standard Bayesian estimators, the only unknown quantities are the class likelihoods $\lambda_k^1, \ldots, \lambda_k^N$, which must be provided by the filters at each time step. For Kalman filters, this quantity corresponds to the *mode likelihood function* [4, 36], which under linear-Gaussian assumptions is a zero-centred Gaussian function:

$$\lambda_k^i = \mathcal{N}\left(\nu_k; \mathbf{0}, \mathbf{S}_k^i\right) \quad (7)$$

where ν_k is the innovation term of the Kalman filter, i.e. the difference between real and expected observation, and $\mathbf{S}_k^i$ is the relative covariance. The same expression is also used as an approximation when the linear-Gaussian assumptions do not hold.

5.2 System Architecture

The joint state in our system contains the vector $\mathbf{x}_k$, which consists of position (x_k, y_k), height z_k, orientation ϕ_k and velocity v_k, plus the attribute c_i representing the identity of the human target (label).

Previous target classification solutions based on BoF use a different prediction model for each estimator, chosen to best fit the relative motion behaviour. In the current system, human recognition is performed both at prediction and observation level. Every estimator of our BoF differs from each other on the z_k component of the prediction model, so to reflect the expected height of a known person. Each filter is

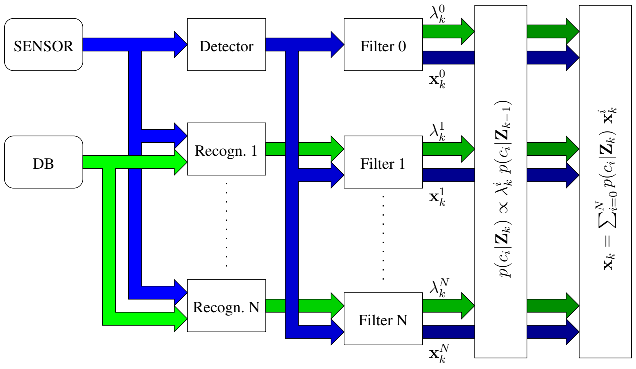


Fig. 9 Bank of filters for joint people tracking and identification.

then updated with target-specific measures of the person's identity given by face and clothes recognition.

The implementation adopts a modular approach where *detectors*, used to measure the human position (i.e. legs detector and face detector), are integrated with *recognizers*, which measure the similarity between current human features and information stored in a database (i.e. clothes recognizer and face recognizer). The system is schematically illustrated in Fig. 9, where $\mathbf{x}_k^i$ is the state estimated by the i^{th} filter, and λ_k^i is the relative class likelihood. In case the identity information is unavailable, our system includes an additional estimator, called *zero-filter*, that outputs the likelihood λ_k^0 of a person to be unknown. This is discussed later in Section 5.3.5.

Raw sensor data from laser and camera are processed by the detectors. The extracted information, which is described by opportune observation models, is sent to all the filters. The current implementation includes a laser-based legs detector and a vision-based face detector, but the system could be easily extended to include other sensors (e.g. sonar, microphone, etc.) and detection algorithms (e.g. motion, sound, etc.).

Sensor data are also used by recognizers, each one specifically designed to identify a particular person. These recognizers, described too by observation models, can access a database of known subjects and provide the filters with identity information. Unlike detectors however, each recognizer can only serve one estimator, as shown in Fig. 9. In our system, the database contains height, color histogram (of the torso) and face of each subject. The first one is used in the prediction model of each filter, the other two during the update step of the estimation. Other recognizers could be added in case more sensors and recognition algorithms were available (e.g. voice, gait, etc.).

At every time step k, all the filters are updated with the current observations, and the identity probabilities are computed with (6). The actual output of the BoF is a mixture of probability densities, not necessarily Gaussian, the mean and covariance of which are calculated as follows [4]:

$$\mathbf{x}_k = \sum_{i=0}^{N} p(c_i|\mathbf{Z}_k) \, \mathbf{x}_k^i \tag{8}$$

$$\mathbf{P}_k = \sum_{i=0}^{N} p(c_i|\mathbf{Z}_k) \left[\mathbf{P}_k^i + \left(\mathbf{x}_k^i - \mathbf{x}_k\right) \left(\mathbf{x}_k^i - \mathbf{x}_k\right)^T \right] \tag{9}$$

Equations (8) and (9) are the current state of the human target and its covariance. In case of multiple people, $\mathbf{x}_k$ and $\mathbf{P}_k$ are estimated for each person being tracked and also used to assign new observations to the proper targets using Nearest Neighbour data association [3, 7].

Since the number of estimators and recognizers needed is equal to the number of subjects in the database, the maximum size of a BoF depends on the available computing resources. Using the Unscented Kalman Filter (UKF), which in [6] showed to provide fast and accurate estimations for people tracking, our system can run in real-time on a PIII 800 MHz (on-board PC of the Pioneer robot), tracking and identifying several people at the same time. The proposed architecture could accommodate other Bayesian estimators, including particle filters, and deal with a large database of known people, provided sufficient computing resources are available.

5.3 Implementation

The choice of the best Bayesian estimator for the BoF is fundamental. The standard Kalman filter [3] provides an efficient way to integrate different sensor data and, in case of linear systems with Gaussian noise, it is known to be optimal. An Extended Kalman Filter (EKF) can be used to provide approximate solutions in case of non-linearities, although most of the recent approaches for tracking people are based on particle filters [14, 39] because they are not constrained by any linear or Gaussian assumption. Unfortunately, in terms of computational cost, particle filters can be very demanding and pose serious constraints in case of BoFs or multiple people tracking.

5.3.1 Unscented Kalman Filter

The estimator adopted for our system is the UKF [27]. Instead of the first-order linearization used by the EKF, the UKF captures mean and covariance of the probability distributions with carefully chosen weighted points. Differently from particle filters, these points are not randomly sampled and their weights do not have to sum up to one. Also, the number of points used by the UKF is small enough (twice the size of the state vector, plus one) to make this estimator particularly suitable for real-time applications of mobile robots with limited hardware resources.

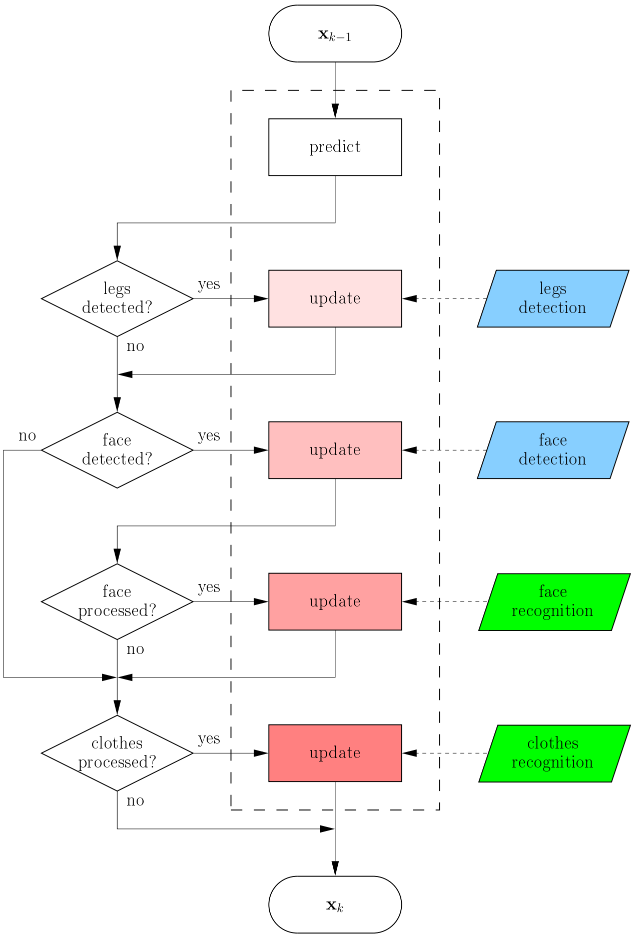


Fig. 10 Sequential update of the UKF.

In case of asynchronous and uncorrelated measurements, the UKF can be updated sequentially using only the observations that are currently available. If all of them are ready at the same time, a sequential update of the filter, starting from the most to the least accurate sensor, gives a better estimate for non-linear systems and is computationally more efficient [3]. A diagram showing the sequential estimation process of a single UKF is shown in Fig. 10.

The UKF's state vector is $\mathbf{x}_k = [x_k, y_k, z_k, \phi_k, v_k]^T$, already defined in Section 5.2. The observation space is constituted by bearing b_k and range r_k from legs detection, bearing α_k and elevation β_k from face detection, bearing φ_k and histogram distance d_k from clothes recognition, and difference ξ_k from face recognition. The models described next consider human motion relative to the local coordinate frame of the robot, the position of which is given by odometry. Note that in order to estimate the absolute human position, the robot should be provided with an accurate localization system. However, since in this case only the robot's frame of reference is considered, tracking is not affected by the cumulative error of odometry. Furthermore, the odometry error between two consecutive estimations is usually very small and can be safely included in the noises of the observation models [7].

5.3.2 Prediction Model

The model used to described the motion of a walking person is a variant of the standard constant-velocity model and consists of the following equations [7]:

$$\begin{cases} x_k &= x_{k-1} + x_{k-1}\,\Delta t_k\,\cos\phi_{k-1} \\ y_k &= y_{k-1} + v_{k-1}\,\Delta t_k\,\sin\phi_{k-1} \\ z_k &= z(c_i) + n^z_{k-1} \\ \phi_k &= \phi_{k-1} + n^\phi_{k-1} \\ v_k &= |v_{k-1}| + n^v_{k-1} \end{cases} \tag{10}$$

where $\Delta t_k = t_k - t_{k-1}$ is the time interval between two consecutive predictions. Supposing a person can only walk forward, the velocity v_k is assumed to be always positive. The noises n^z_{k-1}, n^ϕ_{k-1} and n^v_{k-1} are all zero-mean Gaussians.

As shown later in the experiments, height information can improve the recognition rate. In order to recognize people from their height, the prediction models of the estimators for the BoF differ from each other on the z_k component. The predicted z_k is a constant $z(c_i)$ available from the database (plus a noise term), which is the known height of the i^{th} subject. Except in case the heights of two or more people are the same, this equation makes every prediction model unique, in a way conceptually similar to target classification with different motion models.

Note that height z_k does not evolve over time and therefore could be simply modeled in the likelihood function for recognition purposes. However, in the current implementation z_k is part of the state vector for consistency with our previous work [7] and, most of all, with the zero-filter explained in Section 5.3.5, where the height does actually evolve over time. This choice simplifies also the software implementation of the ROI selection for clothes recognition.

5.3.3 Observation Models of the Detectors

The measurements provided by our laser-based legs detection are bearing b_k and range r_k. The legs observation model can be therefore written as follows:

$$\begin{cases} b_k = \tan^{-1}\left[\dfrac{y_k - l^y_k}{x_k - l^x_k}\right] - l^\phi_k + n^b_k \\ r_k = \sqrt{(x_k - l^x_k)^2 + (y_k - l^y_k)^2} + n^r_k \end{cases} \tag{11}$$

where the noises n^b_k and n^r_k are zero-mean Gaussians. The quantities l^x_k, l^y_k and l^ϕ_k are correction terms taking into account the current position and orientation of the robot from odometry, as well as the displacement of the laser device with respect to its frame of reference [7].

The face observation model takes into account the angles ψ_k and θ_k of the camera's pan and tilt respectively. The relative equations can be written as follows:

$$\begin{cases} \alpha_k = \tan^{-1}\left[\dfrac{y_k - c_k^y}{x_k - c_k^x}\right] - c_k^\phi - \psi_k + n_k^\alpha \\ \beta_k = -\tan^{-1}\left[\dfrac{z_k - c_k^z}{\sqrt{(x_k - c_k^x)^2 + (y_k - c_k^y)^2}}\right] - \theta_k + n_k^\beta \end{cases} \tag{12}$$

Even in this case, the noises n_k^α and n_k^β are zero-mean Gaussians, while c_k^x, c_k^y, c_k^z and c_k^ϕ are correction terms depending on the robot and camera's position. Further details on the observation models (11) and (12) can be found in [7].

5.3.4 Observation Models of the Recognizers

To integrate the class likelihood (7) relative to face and clothes recognition, the histogram distance d_k and the Eigenface difference ξ_k are included as noisy constants in the following observation models, thus providing additional information to update the identity probability. This solution could be improved implementing likelihood functions where histogram and face recognition errors are modeled from the current state vector [35, 44].

Besides the value of the histogram distance, the procedure for clothes recognition provides the person's direction with respect to the camera. The relative observation model includes therefore the bearing φ_k of the torso centre and the distance d_k of its color histogram, modeled as follows:

$$\begin{cases} \varphi_k = \tan^{-1}\left(\dfrac{y_k - c_k^y}{x_k - c_k^x}\right) - c_k^\phi + n_k^\varphi \\ d_k = d(c_i) + n_k^d \end{cases} \tag{13}$$

where $d(c_i)$ is the histogram distance for the i^{th} subject in the database, null in case of perfect match ($d(c_i) = 0$), and the noises n_k^φ and n_k^d are zero-mean Gaussians with parameters empirically determined. The quantities c_k^x, c_k^y and c_k^ϕ are the same correction terms used in (12).

Note that in (13) the elevation angle has not been included, although it could be calculated from the best match position (u^*, v^*) of the histogram detection. The reason is that, when close to a person, our robots can observe only the top part of the torso, from the chest up to the head. In this case, the elevation measure would inevitably fall on a location higher than the actual centre of the torso. The problem would influence the z_k estimate (and hence the height recognition), so we prefer to rely exclusively on the elevation provided by face detection.

The best performance of the Eigenface algorithm can only be achieved under controlled conditions. Unfortunately, in real applications faces have various postures and expressions that make their recognition very difficult. Therefore, in the current implementation the main purpose of face recognition is to enhance the identification performance, without actually reducing the uncertainty of single UKFs. Note however that face recognition does influence the output of the BoF, since the final estimate in (8) and (9) is weighted by identity probabilities.

The face recognition measure from Eigenface is modeled as follows:

$$\xi_k = \xi(c_i) + n_k^\xi \tag{14}$$

where $\xi(c_i)$ is the value obtained for the i^{th} face in the database in case of perfect match ($\xi(c_i) = -1$). For simplicity, the noise n_k^ξ is assumed to be normally distributed, with standard deviation determined by a number of tests with different faces.

Because both the recognition algorithms depends on face detection, the noises n_k^d and n_k^ξ are actually correlated. However, this is true only for the initialization stages, while the subsequent refinements (best histogram match and face normalization) are performed independently. Therefore, in our opinion, the assumption of uncorrelated measures is a justified simplification. This accommodates also the use of Kalman estimators and proved to work reasonably well for our application.

5.3.5 Zero-filter

The main purpose of our system is to track and recognize a certain number of known people. However, there are situations in which the robot has to deal with uncertain or missing human data, in particular:

- the database has no information about the person currently being tracked;
- the information about this person is incomplete or incorrect;
- the person cannot be recognized because outside the camera's field of view.

Without considering the probability of a subject to be unknown, the BoF would assign wrong identity probabilities to the N available subjects. Our system includes therefore an additional estimator, the zero-filter, which has the function to track and identify unrecognized subjects.

Obviously, the (nearly) constant z_k component in (10) does not apply to the zero-filter, but uses instead the following height prediction:

$$z_k = z_{k-1} + n_{k-1}^z \tag{15}$$

Like all the other estimators, the zero-filter is corrected by anonymous legs and face detections. However, since it does not hold any histogram information, clothes observations are substituted by "virtual" measurements including the predicted bearing $\hat{\varphi}_k$ and a constant histogram distance

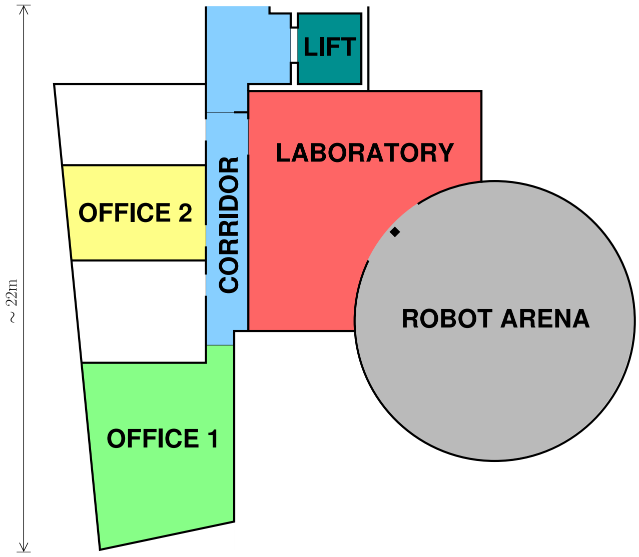


Fig. 11 Floor plan of the test environment.

$\hat{d} = 2\,\sigma_d$ (i.e. twice the standard deviation of the noise n_k^d). This sets an adaptive threshold on the clothes observation, assuring that only good histogram detections influence the identity probability. A similar approach is used for face recognition, the measure of which is set to $\hat{\xi} = 2\,\sigma_\xi$ (i.e. twice the standard deviation of the noise n_k^ξ) for the relative update of the zero-filter.

6 Experimental Results

The system has been implemented in C++ and runs in real-time on the embedded PCs of two mobile robots, a Pioneer 2 DX and a Scitos G5, both equipped with a camera on the top and a laser sensor on the bottom. The experimental scenario includes a typical office environment with cluttered rooms and a narrow corridor, which are illustrated in Fig. 11. The data used for the current evaluation have been collected with our robots following and approaching one or more persons in different rooms. The first part of the experiments present tracking and recognition results using only height and clothes observations. The integration of face recognition is evaluated in the last part.

Since we are particularly interested in recognizing humans who are willing to interact with the robot, most of the cases illustrated next refer to people facing its camera. Note however that a variety of situations has been covered during the experiments, with people often moving randomly in the environment. In general, the face had to be visible only a few instants for the BoF's recognition to converge successfully. Once recognized, people were correctly identified as long as they were tracked, even when outside the camera's field of view.

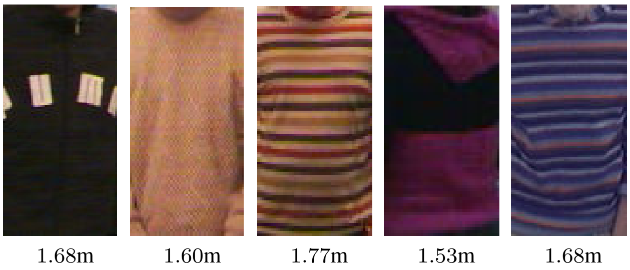


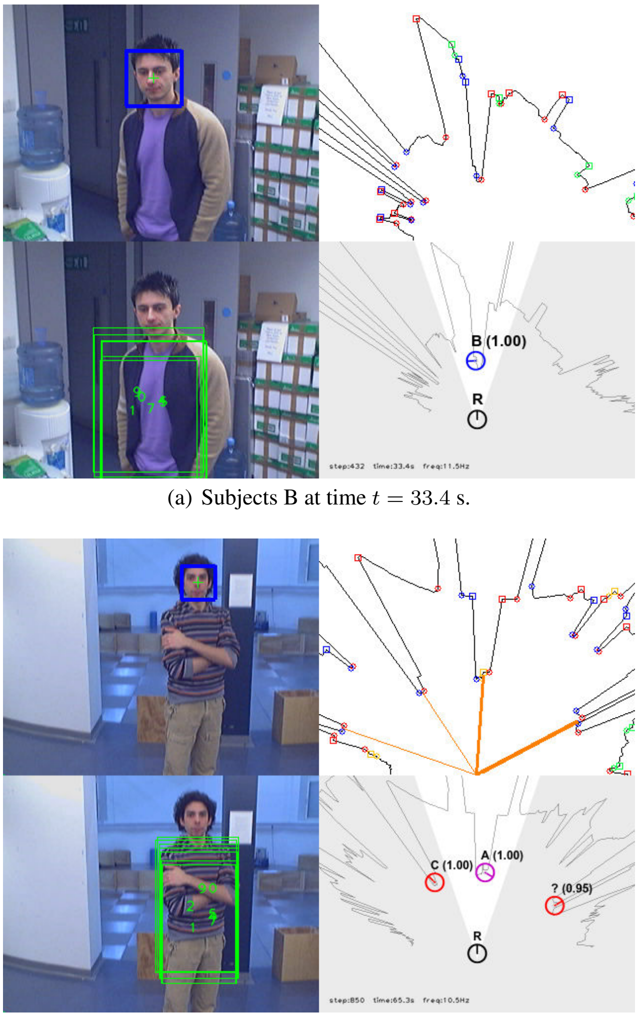
Fig. 12 Some examples of clothes and heights from the database of known subjects.

6.1 Simultaneous People Tracking and Recognition

The situation described next illustrates a typical case of simultaneous people tracking and recognition, executed in real-time with the Scitos robot. Besides legs and face detectors, height and clothes recognizers were used for human identification. The database has been created manually including histograms and height information relative to 13 different subjects, so the BoF consists of 14 UKFs (one is the zero-filter). Some of the clothes worn by people during the experiment are shown in Fig. 12, together with their relative height information. Note that some of the subjects had similar clothes and heights, which made the recognition particularly challenging.

The snapshots in Fig. 13 illustrate a few moments of the experiment, with the robot approaching subjects A and B. Each figures includes, from the robot's point of view, face and legs detection on the top, together with clothes recognition and position estimates on the bottom. In this particular case, the robot was programmed to move autonomously and perform simple interactions. If there were no people, the robot simply wandered in the environment avoiding obstacles. If one or more persons were detected, the robot tracked and approached them; once in proximity, it stopped to engage them in audio-visual interactions.

The graphs in Fig. 14 show the temporal evolution of the identity probabilities for subjects A and B. From Fig. 14(a), it can be noted that subject B was correctly recognized, but not immediately. In this case, the person was detected and tracked with the laser for a few seconds, before being actually observed by the camera. The identity probability of the zero-filter was therefore the highest one until $t \simeq 32$ s, when eventually the probability of subject B took over. The graph in Fig. 14(b) shows instead that subject A, visible by the camera since his first detection, was promptly recognized by the robot at time $t \simeq 63$ s. The identity probability of subject A approached immediately 1, while the unknown identity of the zero-filter dropped to a minimum threshold value close to zero ($p_{min} = e^{-100}$). A video of the experiment is available at the following address: *http://robots.lincoln.ac.uk/users/nbellotto/videos/soro.mpg* .


B (1.00)
R
step:432 time:33.4s freq:11.5Hz
C (1.00)
A (1.00)
? (0.95)
R
step:850 time:65.3s freq:10.5Hz

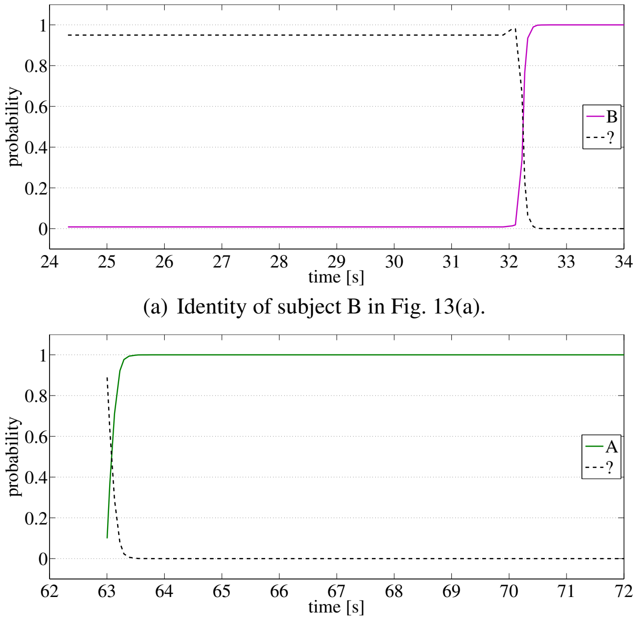
(a) Subjects B at time $t = 33.4$ s.

(b) Subject A at time $t = 65.3$ s.

Fig. 13 Simultaneous people tracking and identification. A, B, and C are the identities of the subjects being tracked by the robot R.

6.2 Evaluation of Height and Clothes Recognition

The success of height and clothes recognition depends on the quality of the tracking estimate. Several recognition experiments have been therefore conducted with people and robot moving across different rooms. The results in Fig. 15 show the recognition performance for approximately 10 minutes of recorded data, during which 13 different people have been followed and approached by the robot in 30 different occasions. The chart indicates the correct and wrong recognition rates, computed by the number of successful and error cases out of the total number of observed people. A subject was considered recognized when the relative identity probability reached at least 0.9. The case where the person was identified as unknown is also reported. The recognition performances have been compared first using height, then clothes and finally the combination of both. From the results, it can be seen that their integration led to a more reliable estimation of the human identity, with a significant improvement on the number of successful recognitions and, at the same time, a substantial reduction of error cases.



(a) Identity of subject B in Fig. 13(a).

(b) Identity of subject A in Fig. 13(b).

Fig. 14 Identity probabilities. The thick line is the probability of the subject's identity, the dashed line is the probability of being unknown from the zero-filter. The identity probabilities from the other filters are almost always null and omitted for clarity.

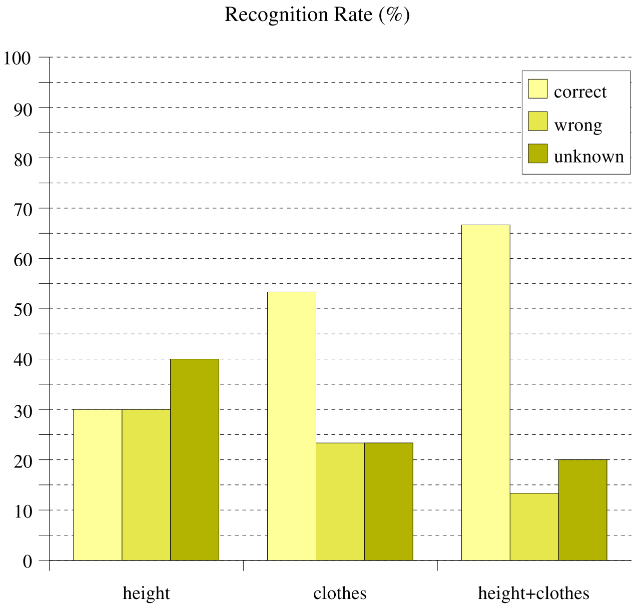
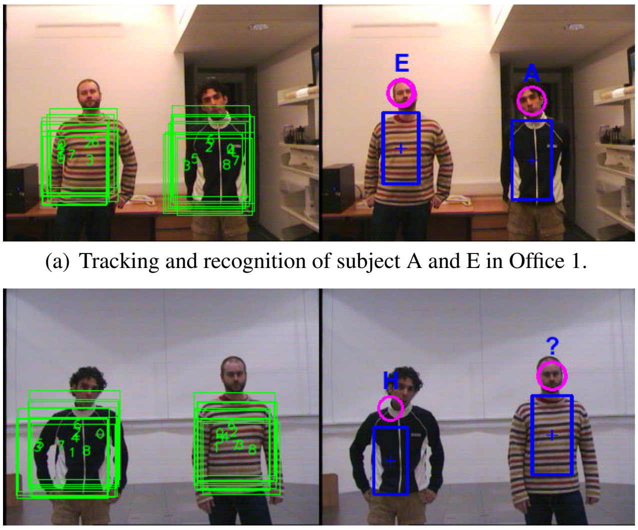


Fig. 15 Recognition performance using height and clothes information.

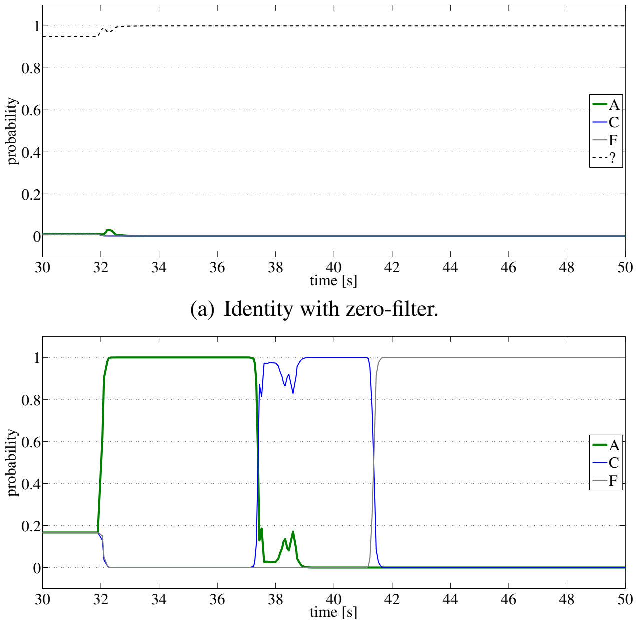
Note in particular the increment of approximately 15% in the recognition rate when both clothes and height were used, compared to the results of clothes only. Although height



(a) Tracking and recognition of subject A and E in Office 1.

(b) Recognition error in the robot arena.

Fig. 16 Recognition failure due to different lighting conditions.



(a) Identity with zero-filter.

(b) Identity without zero-filter.

Fig. 17 Identity probability of an unknown person (subject B) with and without zero-filter.

was not informative per se, the relative bars in Fig. 15 show there was a majority of indecision cases (unknown person) where the zero-filter prevailed. Many of these cases went eventually in favor of the correct recognition once "boosted" by clothes recognition.

Most of the recognition errors were due to considerable light variations during the experiments, like the situation illustrated in Fig. 16. A couple of persons, subject A and E, were initially tracked and correctly recognized starting inside Office 1, as shown in Fig. 16(a). Unfortunately, a few minutes later in the Robot Arena, the system failed to recognize the same people. This was because the particular halogen lamps of the Robot Arena modified significantly the color histograms of the clothes, making their recognition very unreliable. As shown in Fig. 16(b), subject A was identified as a completely different person (i.e. subject H). Thanks to the zero-filter, instead, subject E was identified as unknown, an error that can be considered more acceptable than the previous one. This problem suggests however that further improvements are needed to make clothes recognition more robust to lighting conditions.

6.3 Identification of Unknown People

The main task of the zero-filter is to track and identify an unknown person when there is not sufficient information to recognize him/her. Without this additional estimator, the subject would be otherwise confused with the most similar person in the robot's database. An example of recognition with and without zero-filter is shown in Fig. 17. The graph reports the identity probability estimated for a subject B, during the experiment in Section 6.1, removing the relative height and clothes information from the database. The first graph refers to the case with zero-filter and shows that the BoF can identify correctly subject B as unknown. The second graph shows instead that without zero-filter the maximum identity probability switches erroneously between three different subjects (A, C and F).

The convergence speed of the probability could be "tuned" increasing or decreasing the noises of the observation models. However, the use of the zero-filter permits to have the desirable property of fast, yet correct convergence. This is a necessary condition for real-world robot applications, in particular when a few seconds delay can completely spoil human-robot interactions.

The effectiveness of the zero-filter was also evaluated using the same data recorded for the experiments in Section 6.2. The same trial has been repeated several times, every time removing one of the 13 people from the database, and checking if the missing subject was actually identified as unknown. The error in this case was less than 7%, that is, the robot confused the unknown subjects with someone else in the database only in 2 occasions out of 30.

6.4 Tracking Errors

Recognition can improve human tracking reducing the uncertainty of the estimation and recovering from occasional data association errors. In this experiment, a quantitative evaluation of the tracking robustness has been conducted comparing the number of errors occurred with and without human recognition, using respectively BoFs and sim-

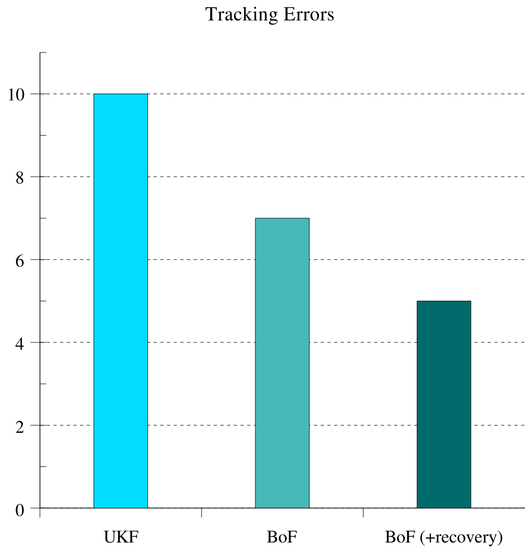


Fig. 18 Comparison of tracking errors with normal UKF (no recognition) and BoF (with recognition).

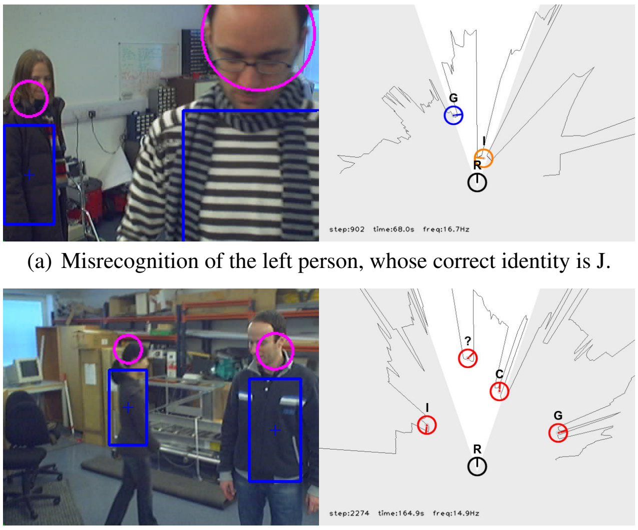
ple UKFs (one for each target). The parameter used is the total number or tracking errors, each one corresponding to one of the following cases: a) the track hypothesis deviates from the correct trajectory of the human target and is eventually deleted by the system; b) the track "jumps" to an adjacent object due to a false positive in the human detection; c) the track shifts to another person (close to the original one) because of data association errors. We considered approximately 20 minutes of data recorded with our robots on a number of trials. The test scenario included 13 people in 6 different locations, illustrated in Fig. 11, with various lighting conditions and levels of clutter.

The chart in Fig. 18 shows that, compared to the standard UKF, the number of tracking failures decreased by 30% with our new solution. We consider also the situation where two track hypotheses switched because of data association errors, but the BoF promptly recovered their correct identity labels. Without counting these cases, automatically corrected by the system, the number of actual failures drops even further, as shown by the last column of the chart.

6.5 Integration of Face Recognition

In this section we analyze the identification performance using also face recognition. The experiments carried out are similar to the previous ones, but in this case only the Scitos robot was used, due to the computational burden of face recognition. The results described next refer to approximately 5 minutes of data, simultaneously tracking and recognizing up to 8 different people. Some of the faces contained in the robot's database, scaled to 24 × 24 pixels and normalized, are shown in Fig. 19.

Fig. 19 Some of the faces in the database of known people.


G
I
R
step:902 time:68.0s freq:16.7Hz
?
C
I
G
R
step:2274 time:164.9s freq:14.9Hz

(a) Misrecognition of the left person, whose correct identity is J.

(b) Person on the top-left not yet identified (correct identity is G).

Fig. 20 Some examples of tracking and recognition with BoFs.

Despite face recognition, in a few cases the robot failed to identify some people wearing similar clothes. During the situation depicted in Fig. 20, for example, two persons were sometimes misrecognized because they had similar (brown) jackets. This happened in particular when the face of the subject was not clearly visible by the robot.

Nevertheless, including face recognition, the performance of the system was generally better than the previous case (only height and clothes), especially on the number of successful identifications. The chart in Fig. 21 shows indeed that the amount of successful identifications increased combining the three modalities (height, clothes and face), while the number of errors remained unchanged. The performance of the system using only height and face recognition, instead, was not very reliable because the poor performance of Eigenface on low-resolution images was often worsened by occlusions and particular head postures.

7 Conclusions and Future Work

In this paper, an original solution for multimodal human perception with mobile robots has been presented using multisensor detection and data fusion. A robust histogram comparison for clothes recognition has been illustrated, which takes into account the uncertainty of the target estimate to maximizes the histogram match and determine the position

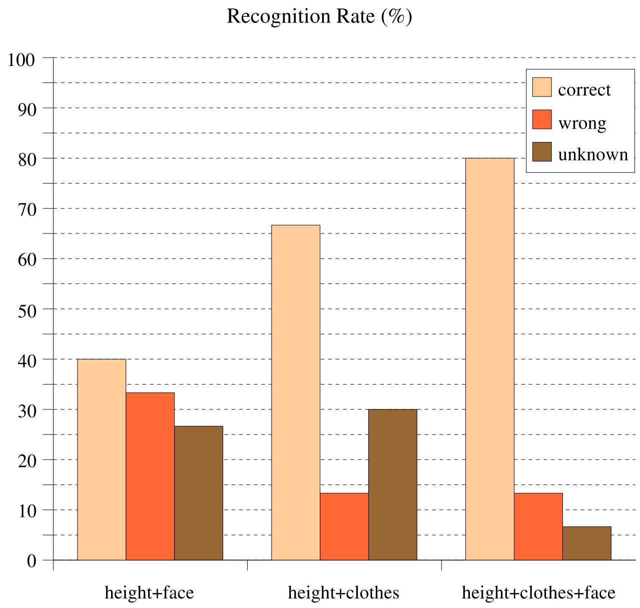


Fig. 21 Recognition performance using face, clothes and height.

of the human torso. A fast technique to process and normalize images for face recognition has also been proposed.

The major contribution of this work lies in the new architecture for simultaneous human tracking and recognition, which is based on a bank of UKFs to combine different algorithms and sensing modalities, and includes a dedicated filter for the identification of unknown persons. Experiments in real scenarios prove the effectiveness of our solutions and show that the robot perception of humans can be improved fusing tracking to height, clothes and face recognition.

The proposed system could be ameliorated in a number of ways, in particular with a robust face recognition algorithm and a solution to deal with the scalability issue. Our future research will focus also on the integration of sensor data using higher representation levels, providing service robots with semantic information about human appearance and behaviours. The objective is to make these robots capable of "understanding" what (and who) they are perceiving by means of real-time AI techniques.

References

1. Arras, K. O., Mozos, O. M., and Burgard, W. (2007). Using boosted features for the detection of people in 2d range data. In *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3402–3407, Rome, Italy.
2. Asoh, H., Vlassis, N., Motomura, Y., Asano, F., Hara, I., Hayamizu, S., Ito, K., Kurita, T., Matsui, T., Bunschoten, R., and Kröse, B. (2001). Jijo-2: An office robot that communicates and learns. *IEEE Intelligent Systems*, 16(5):46–55.
3. Bar-Shalom, Y. and Li, X. R. (1995). *Multitarget-Multisensor Tracking: Principles and Techniques*. Y. Bar-Shalom.
4. Bar-Shalom, Y., Li, X. R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. Wiley.
5. Bellotto, N. and Hu, H. (2007a). Multisensor data fusion for joint people tracking and identification with a service robot. In *Proc. of IEEE Int. Conf. on Robotics and Biomimetics (ROBIO)*, pages 1494–1499, Sanya, China.
6. Bellotto, N. and Hu, H. (2010). Computationally Efficient Solutions for Tracking People with a Mobile Robot: an Experimental Evaluation of Bayesian Filters. *Autonomous Robots*, 28(4):425–438.
7. Bellotto, N. and Hu, H. (2009). Multisensor-based human detection and tracking for mobile service robots. *IEEE Trans. on Systems, Man, and Cybernetics – Part B*, 39(1):167–181.
8. Bennewitz, M., Burgard, W., Cielniak, G., and Thrun, S. (2005). Learning motion patterns of people for compliant robot motion. *The Int. Journal of Robotics Research*, 24(1):31–48.
9. Bennewitz, M., Burgard, W., and Thrun, S. (2002). Learning motion patterns of persons for mobile service robots. In *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3601–3606, Washington, DC, USA.
10. 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.
11. Beveridge, R., Bolme, D., Teixeira, M., and Draper, B. (2003). *The CSU Face Identification Evaluation System User's Guide: Version 5.0*. Computer Science Department, Colorado State University.
12. Beymer, D. and Konolige, K. (2001). Tracking people from a mobile platform. In *IJCAI Workshop on Reasoning with Uncertainty in Robotics*, Seattle, WA, USA.
13. Blanco, J., Burgard, W., Sanz, R., and Fernánez, J. (2003). Fast face detection for mobile robots by integrating laser range data with vision. In *Proc. of the Int. Conf. on Advanced Robotics (ICAR)*, volume 2, pages 953–958, Coimbra, Portugal.
14. 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.
15. 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.
16. Cielniak, G. and Duckett, T. (2004). People recognition by mobile robots. In *Proc. of AILS 2nd Joint SAIS/SSLS*

Workshop, Lund, Sweden.

17. 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, SC, USA.
18. Cunado, D., Nixon, M. S., and Carter, J. N. (2003). Automatic extraction and description of human gait models for recognition purposes. *Comput. Vis. Image Underst.*, 90(1):1–41.
19. Dautenhahn, K. (1995). Getting to Know Each Other - Artificial Social Intelligence for Autonomous Robots. *Robotics and Autonomous Systems*, 16:333–356.
20. Fasel, I., Fortenberry, B., and Movellan, J. (2005). A generative framework for realtime object detection and classification. *Computer Vision and Image Understanding*, 98:182–210.
21. Feyrer, S. and Zell, A. (2000). Robust real-time pursuit of persons with a mobile robot using multisensor fusion. In *Proc. of the 6th Int. Conf. on Intelligent Autonomous Systems (IAS)*, pages 710–715, Venice, Italy.
22. Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4):143–166.
23. 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.
24. Gordon, N. J., Maskell, S., and Kirubarajan, T. (2002). Efficient particle filters for joint tracking and classification. In *Proc. of Signal and Data Processing of Small Targets (SPIE)*, pages 439–449, FL, USA.
25. Gorodnichy, D. (2003). Facial recognition in video. In *Proc. of Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 505–514, Guildford, United Kingdom.
26. Jain, A. K., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(1):4–20.
27. Julier, S. J. and Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proc. of the IEEE*, 92(3):401–422.
28. Li, G., Cai, X., Li, X., and Liu, Y. (2006). An efficient face normalization algorithm based on eyes detection. In *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3843–3848, Beijing, China.
29. 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.
30. Liu, J. N. K., Wang, M., and Feng, B. (2005). iBotGuard: an internet-based intelligent robot security system using invariant face recognition against intruder. *IEEE Trans. on Systems, Man, and Cybernetics (Part C)*, 35(1):97–105.
31. Martin, C., Schaffernicht, E., Scheidig, A., and Gross, H.-M. (2005). Sensor fusion using a probabilistic aggregation scheme for people detection and tracking. In *Proc. of the 2nd European Conference on Mobile Robots (ECMR)*, pages 176–181, Ancona, Italy.
32. Minvielle, P., Marrs, A., Maskell, S., and Doucet, A. (2005). Joint Target Tracking and Identification - Part II: Shape video computing. In *Proc. of the 8th Int. Conf. on Information Fusion*, Philadelphia, PA, USA.
33. Nakajima, C., Pontil, M., Heisele, B., and Poggio, T. (2003). Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006.
34. Nourbakhsh, I., Kunz, C., and Willeke, T. (2003). The mobot museum robot installations: a five year experiment. In *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3636–3641.
35. Pérez, P., Vermaak, J., and Blake, A. (2004). Data fusion for visual tracking with particles. *Proc. of IEEE*, 92(3):495–513.
36. Ristic, B., Gordon, N., and Bessell, A. (2004). On target classification using kinematic data. *Information Fusion*, 5(1):15–21.
37. Santana, M. C., Suarez, O. D., Canalis, L. A., and Navarro, J. L. (2008). Face and facial feature detection evaluation. In *Proc. of the 3rd Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, pages 167–172.
38. Scheutz, M., McRaven, J., and Cserey, G. (2004). Fast, reliable, adaptive, bimodal people tracking for indoor environments. In *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 2, pages 1347–1352, Sendai, Japan.
39. 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.
40. Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):72–86.
41. Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154.
42. Zajdel, W., Zivkovic, Z., and Kr̃ose, 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.
43. Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458.
44. Zhou, S. and Chellappa, R. (2002). Probabilistic human recognition from video. In *Proc. of the 7th European Conference on Computer Vision (ECCV)*, pages 681–697, London, UK. Springer-Verlag.

Nicola Bellotto is a Lecturer in the School of Computer Science at the University of Lincoln and a member of the Lincoln Robotics Group. He holds a PhD in Computer Science from the University of Essex and a Laurea in Electronic Engineering from the University of Padua. Before joining the University of Lincoln, he was a research assistant in Cognitive Computer Vision at the University of Oxford. His research interests range from mobile robotics to cognitive perception, including sensor fusion, Bayesian estimation, robot vision and localization. He gained also several years of professional experience in the industry as software developer and embedded system programmer.

Huosheng Hu is a Professor in the School of Computer Science and Electronic Engineering, University of Essex, UK, leading the Human Centred Robotics Group. His research interests include autonomous mobile robots, human-robot interaction, evolutionary robotics, multi-robot collaboration, embedded systems, pervasive computing, sensor integration, intelligent control and networked robotics. He has published over 250 papers in journals, books and conferences, and received a number of best paper awards. He has been a founding member of Networked Robots within the IEEE Robotics and Automation Technical Committee since 2001. He was a General Co-Chair of IEEE International Conference on Mechatronics and Automation, Harbin, China, 2007, Publication Chair of IEEE International Conference on Networking, Sensing and Control, London, 2007; Co-Chair of Special & Organised Sessions of IEEE International Conference on Robotics and Biomimetics, Sanya, China, 2007; Chair for Special and Organised Sessions, IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Xi'an, China, 2008, etc. Prof. Hu is currently Editor-in-chief for the International Journal of Automation and Computing. He is a reviewer for a number of international journals such as IEEE Transactions on Robotics, SMC-Part B, Automatic Control, Neural Networks and International Journal of Robotics Research. Since 2000 he has been a Visiting Professor at 6 universities in China - namely Central South University, Shanghai University, Wuhan University of Science and Engineering, Kunming University of Science and Technology, Chongqing University of Post & Telecommunication, and Northeast Normal University. Prof. Hu is a Chartered Engineer, a senior member of IEEE and ACM, and a member of IET and IAS.