Proceedings of the 2008 IEEE
International Conference on Robotics and Biomimetics
Bangkok, Thailand, February 21 - 26, 2009

Multimodal Perception and Recognition of Humans with a Mobile 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— **Mobile service robots are becoming more and more popular, both in public and private places, so their perception capabilities must be adequate to detect and recognize people. One of the methods to accomplish the last task is face recognition, but this is unfortunately very challenging because of the human and robot motion. Also, besides the identities of individuals, the system should be able to distinguish between known and unknown people, and deal with this information accordingly. These challenging problems can be solved with a bank of Bayesian filters that simultaneously track and recognize the person of interest using laser and visual data. The paper extends this solution and proposes an improved version, which combines face recognition with human clothes and height identification, and that can also distinguish unknown people. The effectiveness of the system is demonstrated by several experiments with a mobile service robot.**

Index Terms— **Robot perception, human recognition, tracking, sensor fusion, service robotics.**

I. Introduction

In recent years, a growing number of mobile service robots has been employed in human environments for entertainment or other applications, for example for security tasks, elderly and patient care or museum guides [1, 2]. In many cases, the kind of service provided is rather anonymous, that is, the robot does not really know who are the persons being served. However, a service robot that can track people, but cannot distinguish and recognize them, is suitable only for a limited number of applications.

A successful service robot, indeed, should be aware of human presence and needs. Even basic interactions performed by a tour-guide robot should be user-oriented, distinguishing between new and former visitors, or between them and staff members. Knowing the identity of a person, indeed, could mean anticipating his/her possible destination, and this could effectively improve the tracking, as shown in some recent work [3].

For human recognition, several biometric features can be observed using different sensors. Vision-based solutions are obviously the most common, and include human identification from face, iris, ear shape, gait or clothes [4, 5]. Other sensors are also used for the recognition of fingerprints, voices, odors, etc. [6]. The task however becomes extremely challenging if these sensors are mounted on a mobile platform, in particular if the people to be recognized behave and assume poses in a way to spoil the observation of their biometric features.

Our approach is based on a bank of Bayesian estimators that integrates the information provided by two robot sensors, laser and camera, to track people in the surroundings [5]. At the same time, this generates also identity probabilities for the recognition of the human currently being tracked. The solution presented in this paper improves and extends our previous work as follows:

- besides height and clothes recognition, the current system includes a fast procedure for eye detection and face recognition, which improve the perception capabilities of the robot;
- the final architecture integrates a new estimator to deal with an unknown subject, so the robot can correctly track and recognize people who are not present in its database.

The remainder of this paper is organized as follows. Section II introduces the multisensor implementation of human detection and recognition with our mobile robot. Section III describe the method implemented to detect eyes and perform face recognition. Section IV presents our new Bayesian architecture for simultaneous human tracking and recognition. The performance of the system are evaluated with experiments in Section V. Finally, Section VI summarizes the progresses achieved and gives directions for future research.

II. Multisensor Perception

People can be observed using different robot sensors. In our case, the robot is equipped with two common devices, a laser range finder and a colour camera. These are used to detect and recognize humans in the neighbourhood as explained below.

A. Face and Legs Detection

Faces are detected by a camera mounted on the top of the robot, approximately 1.70m high, using a popular algorithm for object detection [7]. Using a standard pin-hole camera

978-1-4244-2679-9/08/$25.00 ©2008 IEEE

model, both direction and size of a face contribute to the estimation of the relative person's position and height.

The robot is equipped also with a laser device, mounted a few decimeters from the floor. An efficient algorithm for legs detection measures distance and direction of people in the range ±90° covered by the laser [8].

B. Clothes and Height Recognition

A common way to distinguish people is histogram comparison, that is, a region of the current camera's frame is selected which contains all or part of the human body. Besides difficulties caused by varying light conditions, a challenging problem is also the correct selection of the region of interest, in particular when both robot and people are moving. If this region is not accurate, the histogram considered might be completely wrong and cause a recognition failure. In [5] we implemented then a selection procedure that takes into account the uncertainty of the human position estimate and extracts the best matching histogram, together with directional information useful for tracking.

Combining range and elevation measurements, provided respectively by the laser and the camera of the robot, it is also possible to determine the height of the person being tracked (more precisely, the height of his face centre). This is also an important biometric information that can be used in conjunction with clothes and face recognition.

III. Face Recognition

In general, the whole process of face recognition consists of three main steps:

- detect and select a face from the current image;
- process the selected image region;
- apply a recognition algorithm.

The first part, in our case, is already accomplished by the face detection procedure mentioned in Section II-A. The selected face is the one that has been assigned to the considered track by the procedure for data association [8]. The second part, instead, consists in aligning and scaling the face according to the position of the eyes. The image region is also converted to grey levels and equalized, then cropped with an elliptic mask and normalized. Finally, in the third part, the recognition algorithm compares the processed image to a set of prerecorded faces, all belonging to the same subject, to find the best match. This algorithm usually returns also a measure of similarity between the current face and the reference template. The last two steps are discussed in details next.

A. Image Processing

One of the most crucial part of every face recognition system is the pre-processing of the considered image region. To align a face horizontally, a common technique consists in locating the eyes and then rotating the selected image so that their inclination is null. Also, the distance between the eyes is used to resize the same image to a given value of width and height. In this way, every new detected face can be oriented and resized as the reference template, that in our system has been set to 24×24 pixels.

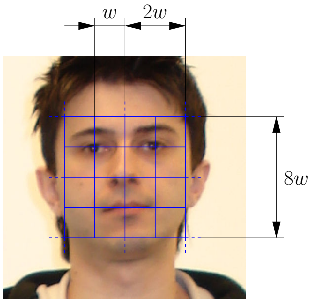


Fig. 1. Canonical face model.

A fast algorithm for eye detection has been proposed in [9] and is based on the extraction of histogram minima within sub-regions containing the eyes. The method relies on the assumption that the eye's centre is darker than the surrounding area, which in general is not true, since in many cases the color of the eyes might be brighter than the skin. The method illustrated in [10], instead, makes use of a more robust probabilistic approach that takes into account the uncertainty of face detection, and integrates also the capability to distinguish blinking eyes. Unfortunately, from a number of tests performed with our robot, the latter method showed to be computationally expensive and not feasible for real-time tasks.

The solution we implemented, instead, is a fast, color-independent procedure based on the same algorithm used for face detection [7]. Using two classifiers, one trained for right eyes and another for left eyes [11], we run two independent local search on specific sub-regions of the face bounding box. With reference to the example in Fig. 1, which illustrates the canonical face representation proposed in [4], the regions scanned are the $2w \times 2w$ top-left area, for one eye, and the top-right area, for the other. It is possible in some case that more than one eye is detected within the considered region. If so, we just choose the detection closer to its centre.

Once the positions $(u_{\mathrm{R}}, v_{\mathrm{R}})$ and $(u_{\mathrm{L}}, v_{\mathrm{L}})$, of the right and left eye respectively, have been determined, we simply calculate the angle and the distance between them as follows:

$$\alpha_{RL} = tan^{-1}\left(\frac{v_{\mathrm{L}} - v_{\mathrm{R}}}{u_{\mathrm{L}} - u_{\mathrm{R}}}\right) \tag{1}$$

$$d_{RL} = \sqrt{(u_{\mathrm{L}} - u_{\mathrm{R}})^2 + (v_{\mathrm{L}} - v_{\mathrm{R}})^2} \tag{2}$$

The face is then rotated of an angle $-\alpha_{RL}$ in order to align the eyes, and then resized to 24×24 pixels, i.e. the size of the templates used for comparison. From the model in Fig. 1, where the distance between the eyes is half the size of the face, this means scaling the image by a factor $s = 12/d_{RL}$.

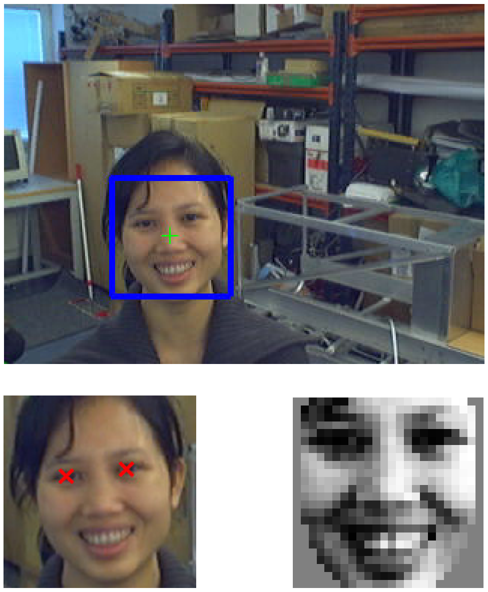


Fig. 2. Image processing before face recognition. The face is aligned and resized according to the position of the eyes, then converted to grey levels, cropped with an elliptical mask and finally equalized.

Rotation and scaling, centred on the right eye $(u_\mathrm{R}, v_\mathrm{R})$, are performed with a simple affine transformation as follows:

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

where (u, v) are the pixel coordinates of the source image and (u', v') those of the destination. Note that, in order to avoid possible outliers in the final image after rotation, the affine transformation is actually applied to a sub-region slightly bigger than the original face bounding box. The coordinates $(u_\mathrm{R}, v_\mathrm{R})$ of the right eye contain obviously a correction term for the offset. In practice, after the transformation, the resulting face will be 24×24 pixels, with the right and left eyes centred, respectively, in $(6, 6)$ and $(18, 6)$.

The final step of the image processing consists in cropping the face's area with an elliptical mask. This reduces the influence of hair and background pixels at the four corners of the rectangular region. Then, considering only the area within the ellipse, the face is equalized and finally normalized, so that the distribution of the pixels intensity has zero mean and standard deviation of one. An example of face processing, from detection to normalization, is shown in Fig. 2.

B. Eigenfaces

After the image processing described above, we apply one of the most popular techniques for face recognition, called *Eigenfaces*[12]. The procedure consists in measuring the similarity between a new face image and a reference one, projecting both the images into an eigenspace, previously created by training, and calculating the distance between the projections.

Several methods have been proposed to measuring this similarity. The most common one, which is considered to give the best performance, is the Mahalanobis Cosine distance [13]. As the name suggests, this is given by the cosine of the angle θ_{ij} between the two face projections $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ in the Mahalanobis space:

$$
\xi\left(\boldsymbol{f}_i, \boldsymbol{f}_j\right) = -\frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_j}{||\boldsymbol{f}_i||\,||\boldsymbol{f}_j||} = -\cos\theta_{ij} \qquad (4)
$$

where the minus sign is introduced just to produce a "distance like" metric.

IV. Simultaneous People Tracking and Recognition

In our system, the human state at time k is described by the vector $\mathbf{x}_k = [x, y, z, \phi, v]_k^T$, where the first two components are the coordinates of the 2D position, z is the face's height, ϕ the orientation and v the velocity. This can be estimated using the anonymous measurements provided by the legs and the face detection, and a Bayesian filter like UKF [14].

For simultaneous tracking and recognition, we consider the joint state $\mathbf{x}_k^i = \{\mathbf{x}_k, c_i\}$, where c_i is a time-invariant feature that depends on the identity of the person. In this case, a solution based on a bank of filters (BoF) can be adopted [5, 8], which keeps track of a moving person and, at the same time, estimates his/her identity. In particular, given the set of observations $\mathbf{Z}_k = \{\mathbf{z}_0, \ldots, \mathbf{z}_k\}$, the probability of the latter can be updated recursively as follows:

$$
p(c_i|\mathbf{Z}_k) \;=\; \frac{\lambda_k^i\ p(c_i|\mathbf{Z}_{k-1})}{\sum_{i=0}^{N} \lambda_k^i\ p(c_i|\mathbf{Z}_{k-1})} \qquad (5)
$$

The likelihood $\lambda_k^i = p(\mathbf{z}_k|\mathbf{Z}_{k-1}, c_i)$, for Kalman filters, is a zero-mean Gaussian:

$$
\lambda_k^i = \mathcal{N}\left(\nu_k^i; \mathbf{0}, \mathbf{S}_k^i\right) \qquad (6)
$$

where ν_k^i is the innovation and $\mathbf{S}_k^i$ its covariance. Although it normally requires linear-Gaussian assumptions, (6) is used in practice even when these do not hold.

A. Implementation

The implementation of the system for joint tracking and recognition is illustrated in Fig. 3. Sensor data are processed by a legs and a face detector to provide positional information to all the filters of the BoF. Each one of these filters corresponds to a possible human identity, and receives therefore information from its dedicated face and clothes recognizer. The information about human height, instead, is embedded on the prediction model of each filter.

At each time step k, all the filters are updated by the current observations. The identity probabilities are calculated

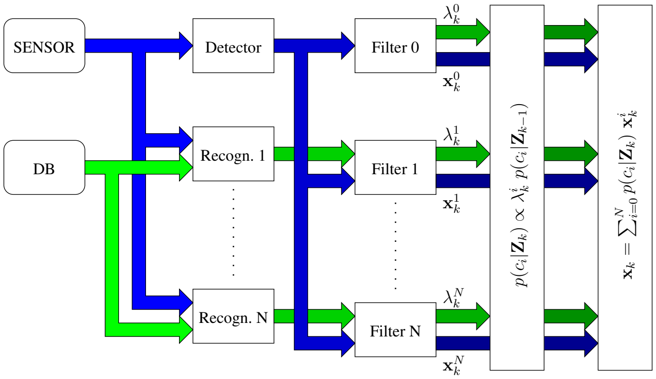


Fig. 3. BoF for joint people tracking and recognition. Filter 0 is the zero-filter used to track and identify unknown persons.

using (5) and the likelihood λ_k^i of each filter. The actual output of the BoF is a mixture of probability densities, not necessarily Gaussian, the mean and covariance of which are computed as follows [15]:

$$\mathbf{x}_k = \sum_{i=0}^{N} p(c_i|\mathbf{Z}_k)\ \mathbf{x}_k^i \qquad (7)$$

$$\mathbf{P}_k = \sum_{i=0}^{N} p(c_i|\mathbf{Z}_k) \left[\mathbf{P}_k^i + \left(\mathbf{x}_k^i - \mathbf{x}_k\right)\left(\mathbf{x}_k^i - \mathbf{x}_k\right)^T\right] \qquad (8)$$

These are basically the current position of the targeted human and its covariance. Also, since a BoF is used for each person being tracked, $\mathbf{x}_k$ and $\mathbf{P}_k$ are both necessary for data association in case of multiple people [8].

B. Face Recognition Model

The face recognition measure, introduced in (4), can be modeled then by the following expression:

$$\xi_k = n_k^{\xi} - 1 \qquad (9)$$

For simplicity, the noise n_k^{ξ} is assumed to have a zero-centred normal distribution with standard deviation $\sigma_{\xi} = 0.15$, empirically determined by a number of tests on different faces. Although incorrect, this approximation makes the model easy to use and showed to work well in practice. Equation (9) accommodates also the fact that the face recognition measure, based on the Mahalanobis Cosine, is -1 in case of best match.

Note that, in the current implementation, face recognition is integrated with the only purpose of enhancing the identifycation performance, without actually reducing the uncertainty of the filter's estimate. This is mainly due to the difficulty of achieving good performance on real video frames with Eigenfaces, where faces have various postures or expressions, and cannot be recognized under controlled conditions. However, face recognition does influence the final estimate, since the output of the BoF, as shown in (7), is weighted by identity probabilities.

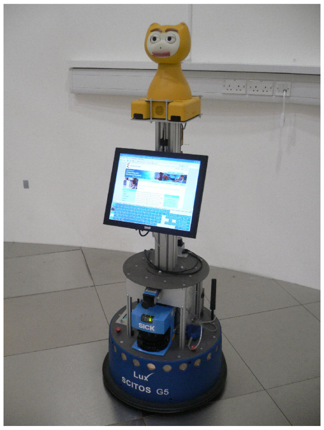


Fig. 4. Interactive mobile robot.

C. Zero-filter

The zero-filter is an additional estimator of the BoF, which has the function to keep track and identify unknown subjects. This is useful in the following situations:

- the database does not contain information about the person currently being tracked;
- the person's information, recorded in the database, is incomplete or not correct;
- the person cannot be recognized because out of the camera's field of view.

In all these cases, if we do not consider the probability $p_0 = 1 - \sum_{i=1}^{N} p(c_i|\mathbf{Z}_k)$ of a subject to be unknown, the BoF would just assign one of the available identities, chosen more or less randomly from those recorded in the database. The zero-filter then, which is not updated by any clothes or face recognition, generates the necessary p_0 and, at the same time, keeps track of the human target.

As shown in Fig. 3, this filter is corrected by the anonymous legs and face detections like all the other estimators. However, since it does not hold any identity information, its face recognition is updated instead by a "virtual" measurement given by a constant distance $\xi^* = 2\ \sigma_{\xi}$, i.e. twice the standard deviation of the noise n_k^{ξ} in (9). This sets a threshold on the face recognition, assuring that only good face observations influence the identity probability. A similar approach is used also for clothes recognition.

V. Experimental Results

Several experiments have been conducted with a mobile robot in real situations and on recorded data. Fig. 4 shows our interactive mobile robot, which is based on a Scitos G5 platform provided with SICK laser, color camera, touch screen and iCat on the top. The latters are used for interaction, providing visual information and generating speech and facial expressions. The on-board PC is a Core Duo 1.6GHz running Linux OS.

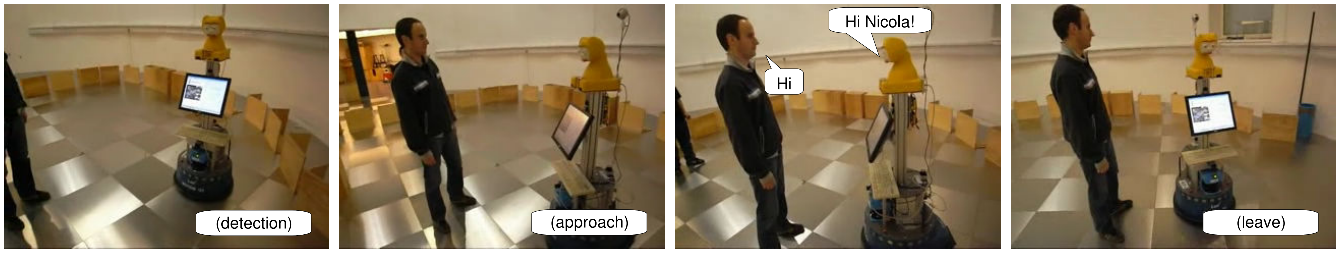


Fig. 6. The robot approaches a person, initially unknown, but avoid the interaction as soon as it recognizes him as staff member.

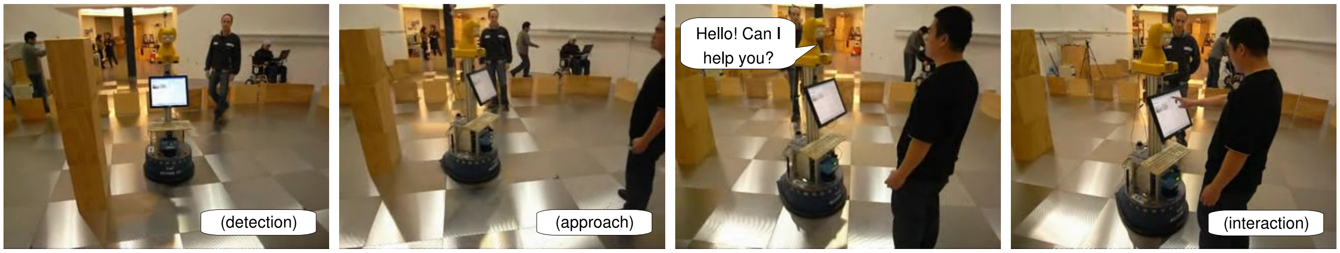


Fig. 7. The robot approaches a person, who is identified as (unknown) visitor, and stops in front of him for interaction..

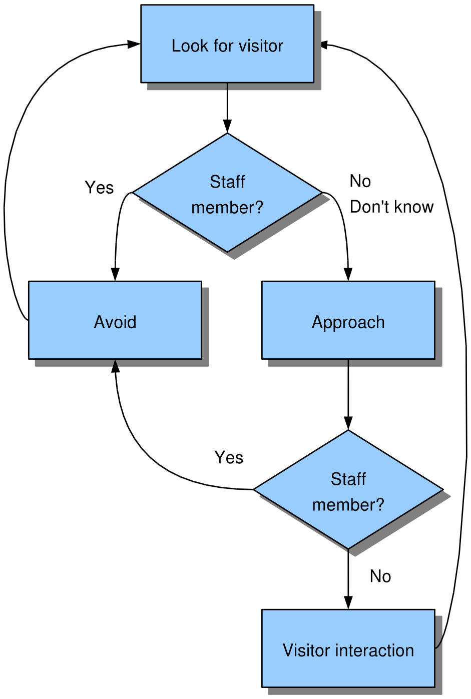


Fig. 5. Interaction scheme to approach visitors and avoid staff members.

A. *Identification of Unknown People*

The main task of the zero-filter is to classify a person as unknown whenever this is not present in the database or there is not sufficient evidence to recognize him/her. Without this filter, the subject would be eventually recognized as the most similar person in the database. This feature becomes useful also in many practical situations of human-robot interaction, in particular in all those cases where the robot's behaviour needs to adapt to the identity of the users.

Our robot is frequently used for demonstrations with young students coming to visit the university. The robot is required to approach some visitors and establish a short interaction, which consists in an initial greeting, followed by an invite to use the touch-screen and get more information about the university. Since during these events an operator has to stand beside the robot, we implemented a simple but effective behaviours scheme, as illustrated in Fig. 5, that enables the robot to approach only visitors and avoid the operator.

Two demonstrative cases are illustrated by the sequences in Fig. 6 and Fig. 7. The first one illustrates the robot that approaches the operator, who is initially too far to be recognized. However, as soon as the robot identifies him correctly, it moves away, avoiding the interaction with a polite greeting. The second sequence, instead, shows the robot approaching a visitor, stopping in his proximity and starting an interaction. During this steps, the robot classifies the user as "unknown" (i.e. anonymous visitor) and provides the necessary information, both vocal and visual, as long as the person stays in front of it.

The effectiveness of the zero-filter has also been measured on a number of tests with the robot approaching 13 different people. Tracking and recognition performances have been observed several times on 10 minutes of recorded data, every time removing one of the 13 people in the database, and checking if the missing subject was actually identified as unknown. In only 2 cases, out of 30, the zero-filter was not able to threshold the other identities, that is, the misrecognition error of unknown people was less than 7%.

B. Recognition Performance

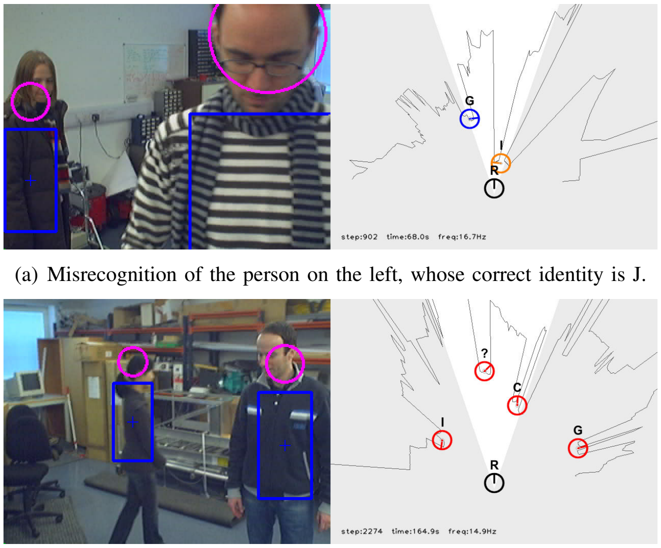
Similar experiments have been carried out using the robot. In some case the recognition failed because of similar clothes worn by the people. For example, during the experiment illustrated in Fig. 8, the two girls were sometimes misrecognized because of the brownish color of both their jackets, especially when their faces were not clearly visible by the robot.

The recognition performance is explained with a chart in Fig. 9, reporting the percentage of correct recognitions and errors observed on 5 minutes of data with 8 different people, who were moving around the robot and approached by it. The chart shows the results obtained from three different combinations of identity features, considering subjects fully recognized only when their relative identity probabilities reached at least 0.9. It shows also that, unfortunately, the performance of the system using only height and face recognition was not very reliable. The main reason was the rather poor performance of the Eigenface algorithm on low-resolution face images, often disturbed also by occlusions and head orientation (e.g. hair, nodding, etc.). However, the overall recognition performance, when face, clothes and height are considered, is particularly good, with 80% of correct identifications and only 14% of errors. The remaining 6% were cases of non-identification handled by the zero-filter.

VI. Conclusions and Future Work

This paper presented an improved bank of Bayesian filters for the simultaneous tracking and recognition of people with a mobile service robot. The approach adopted is a multimodal solution that integrates face, clothes and height recognition to identify known and unknown persons, which has been shown experimentally to perform well and to be feasible for real applications of interactive robots.

In the future, we would like to improve the face recognition with a more robust algorithm, less sensitive to light variations and head orientation. This would be particularly useful to implement an automatic update of the clothes information.


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 person on the left, whose correct identity is J.

(b) The walking persons on the left is the original subject G.

Fig. 8. Some examples of tracking and recognition with BoFs. The right part shows the laser scan, where R is the robot and the others are people.

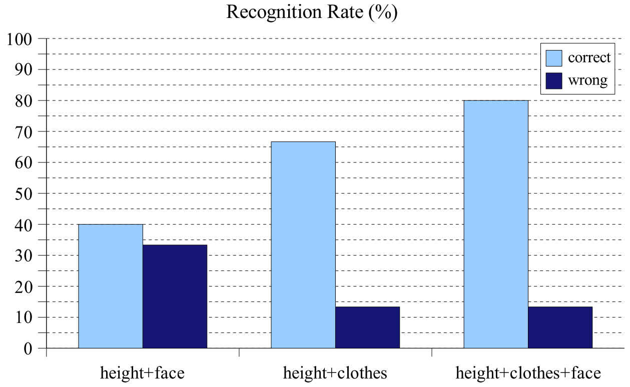


Fig. 9. Recognition performance.

References

[1] M. Hans, B. Graf, and R. Schraf, "Robotic home assistant Care-O-bot: Past – present – future," in *Proc. of the IEEE Int. Workshop on Robot and Human Interactive Communication*, Germany, 2002, pp. 380–385.

[2] A. Chella and I. Macaluso, "Sensations and Perceptions in Cicerobot, a Museum Guide Robot," in *Proc. of BICS*, Greece, 2006.

[3] S. Thompson, T. Horiuchi, and S. Kagami, "An environment driven model of human navigation intention for mobile robots," in *Proc. of IASTED Int. C. on Robotics and Appl.*, Germany, 2007, pp. 119–125.

[4] D. Gorodnichy, "Facial recognition in video," in *Proc. of Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, Guildford, United Kingdom, 2003, pp. 505–514.

[5] N. Bellotto and H. Hu, "Multisensor data fusion for joint people tracking and identification with a service robot," in *Proc. of IEEE Int. Conf. on Robotics and Biomimetics*, China, 2007, pp. 1494–1499.

[6] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

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

[8] N. Bellotto and H. Hu, "Simultaneous people tracking and recognition for interactive service robots," in *New Research on Mobile Robots*, E. V. Gaines and L. W. Peskov, Eds. Nova Science, 2008, (in press).

[9] G. Li, X. Cai, X. Li, and Y. Liu, "An efficient face normalization algorithm based on eyes detection," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, China, 2006, pp. 3843–3848.

[10] Fasel, Fortenberry, and Movellan, "A generative framework for real-time object detection and classification," *Computer Vision and Image Understanding*, vol. 98, pp. 182–210, 2005.

[11] M. C. Santana, O. D. Suarez, L. A. Canalis, and J. L. Navarro, "Face and facial feature detection evaluation," in *Proc. of the 3rd Int. Conf. on Computer Vision Theory and Applications*, 2008, pp. 167–172.

[12] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 72–86, 1991.

[13] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, *The CSU Face Identification Evaluation System User's Guide: Version 5.0*, Computer Science Department, Colorado State University, May 2003.

[14] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[15] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. Wiley, 2001.