Person Identification by Mobile Robots in Indoor Environments

Grzegorz Cielniak and Tom Duckett

Centre for Applied Autonomous Sensor Systems
Dept. of Technology, Örebro University
SE-70182 Örebro, Sweden
http://www.aass.oru.se

Abstract

This paper addresses the problem of identifying persons with a mobile robot. In the proposed system, people are first detected and then tracked with the robot's laser range-finder sensor, using an independent Kalman filter for each person. After segmentation, the rectangular region of the image containing the person is divided into regions corresponding to the person's head, torso and legs. Colour features are extracted from each region for input to a pattern recognition system. Five alternative classification methods were investigated, including experiments on a real robot and with a static camera system. The best identification performance was obtained with an ensemble of self-organizing maps (ESOM), where one self-organizing map is trained for each person in the robot's database. We also discuss how to incorporate the new method into a complete application of a robotic security guard.

1 Introduction

Recently a variety of so-called service robots have been developed. They have been designed to work in populated environments such as hospitals [11], museums [4], office buildings [1] and supermarkets [8], where they perform tasks such as cleaning, surveillance, entertainment, education and delivery. These robots must have the ability to cooperate with people. To enable this cooperation, a robot needs to know how many people there are in its surroundings, where they are and who they are (the three fundamental problems of people detection, tracking and identification). This paper concerns the problem of people identification.

Modern human identification systems use a variety of features including iris, face and speech patterns. The literature in this field is quite extensive and some special workshops confirm the general interest in this topic. Useful approaches for mobile robots are those that can be utilized from a distance, and should be able to operate in real-time under the extra noise and variations due to the motion of the robot itself. The ideal system should be able to recognise the humans in their natural environment, without requiring any special registration or scanning procedure. Possible techniques include face recognition [10], speaker identification, biometrics, etc. Recent work has focussed on how to combine different recognition techniques in order to improve identification accuracy [3]. However, most of the work to date has concentrated on static sensor systems. In this paper, we investigate an identification method for a mobile robot that is based on a learned colour model of the person's whole appearance, including face, hair, clothes, shoes, etc. We discuss the practical issues of integrating this method into a real world application in the conclusions.

An overview of our proposed system is given in Fig.1. To detect and track people in the surroundings of the robot, the built-in laser sensor is used (see Section 2). To identify people, colour images are used. Information from the laser-based tracker is first used to segment the area of the image containing a person. Then a set of colour features is extracted from the segmented area (Section 3). These feature values are classified using an ensemble of self-organizing maps (ESOM). Section 4 gives details of the various pattern classification algorithms investigated. An advantage of this particular method is that it can be trained incrementally, by adding a new self-organizing map for each new person as required, which is an important requirement for many applications of service robots. Section 5 gives experimental results showing the robustness of our method, including a comparison of ESOM with a number of other classification methods. This includes results on a real robot (with 3 persons) and also with a static camera system (with 9 persons) in order to test both the feasibility and scalability of the proposed method. Finally, in the concluding section, we discuss how to integrate the method with other techniques for enabling human-robot cooperation, including details of our proposed application, the *Robotic Security Guard*.

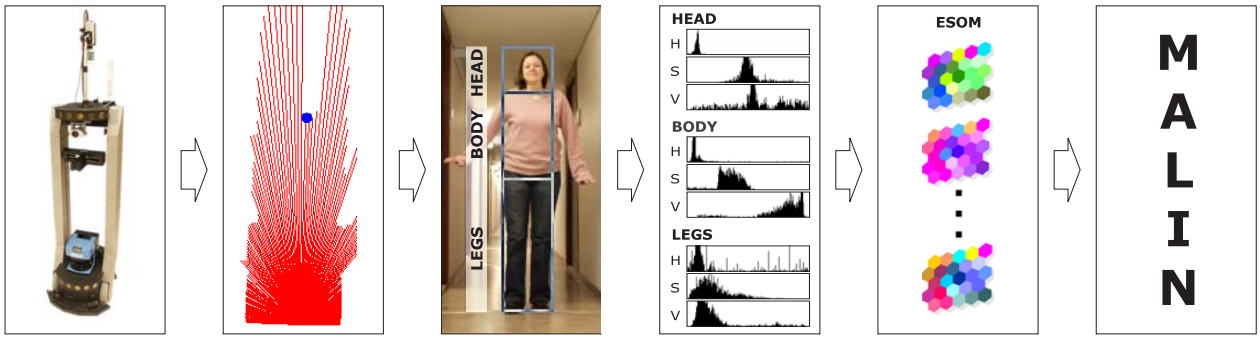

Figure 1: Overview of the proposed system: a) robot platform, b) information from the laser-based tracker, c) segmented image, d) colour distributions within the segmented areas, e) ensemble of SOMs, and f) classification result.

2 People Detection and Localisation

This section describes the pre-processing steps required to segment the people in the images, both on a real robot and with a static web-camera. The output of both of these different data collection methods is the rectangular region of an image containing a person.

2.1 Implementation on a B21 Robot

We used a SICK LMS 200 laser scanner mounted on an RWI B21 robot at Freiburg University, Germany, to detect and track people in the robot's immediate surroundings. The detection system first extracts local minima that correspond to the legs of persons. To increase the reliability of the system, consecutive scans are taken into account. To keep track of each person, we apply a separate Kalman filter for each person detected. The state x_r of a person at time step r is represented by a vector $[x, y, \delta x, \delta y]'$. Whereas x and y represent the position of the person, the terms δx and δy represent the velocity of the person in x- and y-direction.

Figure 2: Robot Albert (B21) tracking a person who is walking through the environment.

Accordingly, the prediction is carried out by the equation

$$x^-_{r+1} = \begin{vmatrix} 1 & 0 & t_r & 0 \\ 0 & 1 & 0 & t_r \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} x_r,$$

where t_r is the time elapsed between the measurement z_{r+1} and z_r. Since the laser range sensor does not provide the velocities δx and δy, which are also part of our state space, the measurement matrix projects onto the first two components of the state space. Accordingly, the predicted measurement at step $r+1$ is

$$z^-_{r+1} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{vmatrix} x^-_{r+1}.$$

To solve the data association problem, we apply the nearest neighbour approach.

To determine the area of the image corresponding to a person, as detected by the laser tracking system, we rely on an accurate calibration between the camera and the laser. We use a perspective projection to map the 3D position of the person in world coordinates to 2D image coordinates.

2.2 Experiments with a Web-camera

We used a web-camera Philips PCVC 740K (resolution 160×120 pixels) connected to a Pentium II PC to collect the data. The camera was placed in the corner of the robotics lab at our institute. The position and orientation of the camera were adjusted to cover the largest possible area of a 7×8 m room. Persons taking part in the experiments were asked to walk within a limited area of interest (see Fig. 3a). During this task, images from the web-camera were recorded with frequency 2Hz.

To localise and determine the area of the image corresponding to a person we used a vision-based approach.

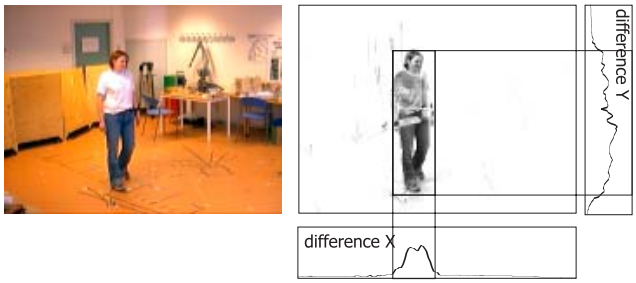

Figure 3: Web-camera: a) an example image with a walking person b) segmentation of the person from the picture.

Since the position of the camera was fixed, we could use a background extraction method. For every frame, the difference with the background was calculated. The background was recorded earlier with no moving person in the picture (taking the average of five pictures). Then a histogram of difference data in both vertical and horizontal directions was created. Data with a value higher then a certain threshold (learned during background acquisition) was used for detection of the person in the image (see Fig.3b).

3 Feature Extraction

The determined area of the image is first divided into three sub-areas corresponding to three human body parts: head, torso and legs. We used similar proportions to those proposed by Vitruvius [16]. In our approach, the proportions used were $\frac{1}{6}$ for the head, $\frac{2}{6}$ for the torso and $\frac{3}{6}$ for the legs (see Fig. 1c).

In the next step, we collect statistical information about the colour distribution within the segmented areas. Colour features are robust with respect to translation, rotation, scale and other kinds of geometric distortions, but very sensitive to varying lighting conditions. Therefore we used the HSV (Hue-Saturation-Value) colour space. In this colour model, the intensity factor can be easily separated and its influence reduced. We collected information about the first two moments (mean and variance) of the colour distribution for each segment, which gives $3 \times 3 \times 2 = 18$ features in total.

4 Pattern Classification

To identify persons, the pattern vector obtained after preprocessing (i.e., people detection and localisation) and feature extraction is presented to a pattern classification system. In this section, we describe the five different classification methods investigated.

4.1 Minimum Distance Classifier

A simple and very intuitive classification method is a minimum distance classifier (MDC). In this method, mean vectors calculated from the training data for each class are assumed to be ideal prototypes for the persons. To classify a new input vector, the Euclidean distance to each of the prototypes is calculated, and the vector is assigned to the class with the shortest distance.

Equivalently, the decision function for a minimum distance classifier can be written as

$$d_j(\mathbf{x}) = \mathbf{x}^T \mathbf{m}_j - \frac{1}{2}\mathbf{m}_j^T \mathbf{m}_j$$

where $\mathbf{x}$ is the pattern vector to be classified, and $\mathbf{m}_j$ is the mean vector of each class ω_j. Classification is then determined by the class that produces the highest decision value.

4.2 Bayes' Classifier

The classification results obtained by the minimum distance classifier can be improved by modelling the distribution of the data. For normally distributed classes, it it is shown in [6] that minimum-error-rate classification is achieved by using the decision function

$$d_j(\mathbf{x}) = \ln P(\omega_j) - \frac{1}{2}\ln|\mathbf{C}_j| - \frac{1}{2}([\mathbf{x}-\mathbf{m}_j]^T \mathbf{C}_j^{-1}[\mathbf{x}-\mathbf{m}_j])$$

where $\mathbf{m}_j$ is the mean vector and $\mathbf{C}_j$ is the covariance matrix of each class ω_j. Again, a given input vector is assigned to the class with the highest decision value.

4.3 Multi-layer Feedforward Network

Another way to perform a classification task is to use a multi-layer feedforward (MLFF) neural network. We decided to use a variant with one output unit for each class ω_j, which was trained with the 1-of-c coding [2], where c is the number of classes. During training, an input pattern is presented to the network together with a target output vector, in which the output corresponding to the correct class is set to 1 and all other ouputs are set to 0. During classification, the class corresponding to the output with the highest value is chosen as the classification result. In our experiments, we used a network with 18 input units, one hidden layer and c output units depending on the experiment ($c = 7$ for the B21 robot, and $c = 9$ for the web-camera data). The best performance was obtained with 15 units in the hidden layer and a learning rate of 0.3.

4.4 Simple Recurrent Network

If we use image sequences rather then a single image, the identification task can be treated as a dynamic process. In

this case, a simple recurrent neural network (SRN) can be used to improve classification performance by taking into account the "history" of the sequence of image patterns. In this type of network, the outputs of the neurons in the hidden layer are connected to another set of input units (called the context layer) in a feedback loop. These neurons play the role of a dynamic memory. The recurrent inputs at time t are taken from the outputs of the hidden units at time $t-1$. We used an Elman recurrent network [7] with a similar configuration to the MLFF (18 inputs, one hidden layer with 12 neurons, and c output neurons).

4.5 Ensemble of SOM Classifiers

The self-organizing map (SOM) is an unsupervised neural network that can be used for clustering sensor data [12]. The basic idea of this approach is to train one SOM for each person, and then during classification the SOM which gives the smallest distance error is chosen as the "winner". Such a structure can be called an ensemble of SOMs [15].

One self-organizing map consists of a set of neurons or cluster units that are arranged in a regular geometric pattern (in this paper, a hexagon was used). Each unit j has a weight vector $\mathbf{w}_j$ that acts as a prototype. When a pattern vector $\mathbf{x}$ is presented to the network, the best matching cluster unit is determined according to the smallest euclidean distance $\| \mathbf{x} - \mathbf{w}_j \|$. During training, the weight vector of the best matching unit is adapted to be more similar to the input vector. In addition, the weight vectors for the geometric neighbours of the winning unit are also adapted by a smaller amount, with the result that the network learns a topographic mapping from the input space onto the cluster space that preserves the underlying distribution of the training data. In other words, similar input vectors are mapped onto similar regions in the map of neurons. During testing, the distance to the best matching unit is used as a measure of the similarity of the presented pattern vector to the stored "signature" for that particular person.

The input to each SOM is the vector of extracted colour features (dimension 18) as described in the previous section. In our experiments, we used a basic structure consisting of 25 units arranged in a 2D square grid of 5×5 units, which provides a good compromise between classification performance and network complexity.

5 Experimental Results

To evaluate the performance of our method, we used data collected during several experiments with a mobile robot and with a web-camera. The data collected by the mobile robot consisted of 7 classes (3 different persons wearing 7 different sets of clothing), comprising 207 examples in total. 20% of these examples were used as a training set and the others for testing. There were 9 different persons taking part in the experiments with the web-camera, in which a total of 2438 examples were collected. We used 30% of this data for training and the remaining 70% for testing. The training-testing procedure was repeated 10 times with a different, randomly chosen training set for each repetition. Tables 1 and 2 present the average results with standard deviation for the classification methods described in the previous section.

Method	Results [%]
MDC	80.52 ± 2.66
MLFF	82.58 ± 3.16
ESOM	86.39 ± 2.44

Table 1: Results from the experiments with a mobile robot.

Method	Results [%]
MDC	69.06 ± 1.64
Bayes	80.01 ± 2.44
MLFF	81.28 ± 3.62
SRN	84.50 ± 1.86
ESOM	92.16 ± 1.12

Table 2: Results from the experiments with a web-camera.

The results show that the performance of the classifier based on an ensemble of SOMs is significantly better than that of the other classifiers ($p = 0.01$ using Student's t-test [14]). The image data obtained by the robot was very noisy (mostly due to inaccuracies in the synchronization between the laser and vision systems). This affected the performance of the classification procedure. The small amount of available data from the robot also meant that we were not able to perform classification with the Bayes' classifier (due to singularities in the calculated covariance matrices) or the simple recurrent network (because the length of the available image sequences was too short). However, the results obtained with the web-camera indicate that the SRN produced a small but significant improvement in performance over the MLFF network ($p = 0.05$), using image sequences of approximately 80 images per person.

6 Conclusions and Future Work

In this paper, we proposed a new method to identify people with mobile robots. Laser information is used to detect and track persons (segmentation), and vision information is used to identify the segmented persons. The identification

procedure uses an ensemble of SOMs rather than a single neural network. Our experiments demonstrated the robustness of this method compared to the other methods investigated. Possible topics for future work would include more sophisticated image segmentation routines and incorporation of motion patterns of persons to further improve identification performance.

This work represents a step towards a complete application, the Robotic Security Guard (RSG). The main goal of this application is to combine different aspects of learning, navigation, localisation, planning and interaction into one platform that is able to patrol an environment, guard valuable equipment, recognise known persons, discriminate intruders from known persons, and cooperate with human security staff. The required functionalities will also include learning algorithms for simultaneous localisation and mapping [9], acquisition of navigation behaviours from human demonstration [13], and learning of typical motion patterns of persons [5]. The project will be developed and tested on an Activmedia Peoplebot robot equipped with an array of heterogenous sensor types, including a pan-tilt-zoom camera, omni-directional camera, thermal infrared camera, laser range-finder sensor, ultrasonic range-finder sensors, stereo microphones and odometry.

There are a number of issues that must be addressed in order to integrate the proposed identification method with the intended robotic application. The learned color model has several useful properties, for example, it is recognisable from a wide range of distances, and is fairly invariant to different orientations of the persons. However, it has a number of obvious drawbacks, e.g., people often change their clothes on a regular basis! To overcome this problem, we intend to combine different recognition techniques with orthogonal strengths and weaknesses. Accurate (but not so robust) techniques such as face or speech recognition could be used at the start of each day to obtain a confident initial estimate of the identity of a person. Then the clothing model could be re-acquired or added to an existing database of clothes for that person, and used for general identification purposes later in the day, and in situations where faces and voices cannot be easily recognised. Rather than making crisp decisions, the different sources of sensory evidence should be combined within a framework that allows belief revision based on new information, e.g., by maintaining probability distributions over the location and identity of detected persons.

Acknowledgments

We would like to thank Maren Bennewitz and Wolfram Burgard at Freiburg University for their help with providing the B21 robot data during the Marie Curie fellowship of the first named author.

References

[1] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui. Socially embedded learning of office-conversant robot Jijo-2. In *Proc. of the Int. Joint Conference on Artificial Intelligence (IJCAI)*, 1997.

[2] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[3] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[4] W. Burgard, A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1-2), 1999.

[5] G. Cielniak, M. Bennewitz, and W. Burgard. Where is ...? Learning and utilizing motion patterns of persons with mobile robots. In *Proc. of the Int. Joint Conference on Artificial Intelligence (IJCAI)*, Acapulco, Mexico, August 9-15, 2003.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2nd edition, 2000.

[7] J.L. Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225, 1991.

[8] H. Endres, W. Feiten, and G. Lawitzky. Field test of a navigation system: Autonomous cleaning in supermarkets. In *Proc. of the IEEE Int. Conference on Robotics and Automation (ICRA)*, 1998.

[9] U. Frese and T. Duckett. A multigrid approach for accelerating relaxation-based SLAM. In *Proc. IJCAI Workshop on Reasoning with Uncertainty in Robotics*, Acapulco, Mexico, August 9, 2003.

[10] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, December 2001.

[11] S. King and C. Weiman. Helpmate autonomous mobile robot navigation system. In *Proc. of the SPIE Conference on Mobile Robots*, pages 190–198, Boston, MA, November 1990. Volume 2352.

[12] T. Kohonen. *Self-organizing Maps*. Springer, 1995.

[13] J. Li and T. Duckett. Learning robot behaviours with self-organizing maps and radial basis function networks. In *Proceedings of the Second Swedish Workshop on Autonomous Robotics*, Stockholm, Sweden, October 10-11, 2002.

[14] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C, 2nd. edition*. Cambridge University Press, 1992.

[15] A.J.C. Sharkey. On combining artificial neural nets. In *Connection Science*, volume 8, pages 299–314, 1996.

[16] P. Vitruvius and F. Granger (trans.). *Vitruvius: On Architecture*. Harvard University Press, 1934.