Appearance-based Tracking of Persons with an Omnidirectional Vision Sensor

Grzegorz Cielniak[1], Mihajlo Miladinovic[1], Daniel Hammarin[1],
Linus Göranson[1], Achim Lilienthal[2] and Tom Duckett[1]

[1]Dept. of Technology, AASS, Örebro University, SE-70182 Örebro, Sweden
http://www.aass.oru.se
[2]W.-Schickard-Inst. for Comp. Science, University of Tübingen, D-72076 Tübingen, Germany
lilien@informatik.uni-tuebingen.de

Abstract

This paper addresses the problem of tracking a moving person with a single, omnidirectional camera. An appearance-based tracking system is described which uses a self-acquired appearance model and a Kalman filter to estimate the position of the person. Features corresponding to "depth cues" are first extracted from the panoramic images, then an artificial neural network is trained to estimate the distance of the person from the camera. The estimates are combined using a discrete Kalman filter to track the position of the person over time. The ground truth information required for training the neural network and the experimental analysis was obtained from another vision system, which uses multiple webcams and triangulation to calculate the true position of the person. Experimental results show that the tracking system is accurate and reliable, and that its performance can be further improved by learning multiple, person-specific appearance models.

1 Introduction

The ability to interact with people is an important requirement for robots which operate in populated environments. In tasks such as cleaning, housekeeping, rehabilitation, entertainment, inspection and surveillance, so-called service robots need to communicate and cooperate with people. To enable this interaction, the robot needs to know how many people there are in the neighbourhood, their position, and who they are (the three fundamental problems of people recognition, tracking and identification). In this paper, we focus on the problem of people tracking.

Sensory information about humans can be obtained by the robot in different ways. The most common sensors used today are range-finder sensors (e.g., sonar, laser), sound detectors (e.g., speech recognition) and, with increasing popularity, vision sensors (e.g., single camera, stereo vision). This paper investigates the use of omnidirectional vision for people tracking by autonomous robots.

In contrast to previous methods that use multiple cameras, our method is based on a single omni-camera mounted on top of a mobile robot (see Fig. 1). The use of a single camera means that we cannot use geometric triangulation methods to estimate the position of the person. Instead, we extract a number of simple statistical features from the images that correspond to "depth cues" indicating the apparent position of the person relative the robot. These features are presented in the input vector to an artificial neural network, which learns an "appearance model" that estimates the distance of the person from the robot. In the experiments presented here, the robot was stationary throughout, though we discuss the problems of implementing the method on a moving robot in future works.

To train the neural network, and also to obtain the ground truth information needed for the experimental analysis, some external measurement of the actual position of the person is required. In the experiments presented here, this information was obtained from another, independent vision system that uses multiple webcams located around the room and triangulation to calculate the true position of the person (see Section 3). Our results show that it is possible to train the neural networks in the tracking system using the position information from the external measurement system.

We then describe how to construct an appearance model that can be used to estimate the position of a moving person in the nearby environment (Section 4). From the panoramic images taken by the omni-camera we extract a set of features that capture information about the distance and direction of the person from the robot. An artificial neural network is then used to estimate the distance to the person. The results obtained with the learned appearance model are improved by using a discrete Kalman filter to track the position of the person over time (Section 5). In addition, we

show that performance can be further improved by learning different appearance models for different people using multiple neural networks (Section 6). In the experiments presented, we show that the performance of the system using person-specific appearance models is significantly better than that obtained with a general appearance model.

The are several reasons why using an artificial neural network to learn the appearance model is advantageous for the intended application of people tracking. First, the method is self-calibrating, meaning also that we do not need to design a model of the omni-camera by hand: the appearance model captures statistical properties of both the sensor and the relative position (depth) of the person in the images. Second, the method is appearance-based, and does not require any structural model of a human being. All necessary parameters are acquired from data during the training phase. Third, the method uses multiple features (depth cues) to recover information about the relative position of the person. This means that it is more robust in handling effects such as shadows and lighting variations, and should be more tolerant to additional noise when the robot itself is moving. Fourth, different appearance models can be learned for different people in order to further improve performance, since people come in different shapes and sizes.

2 Related Work

Omnidirectional cameras have become popular in computer vision, especially in applications like surveillance systems [1] and automated meeting recording [7]. In robotics, omni-cameras are used mostly for navigation and localization of mobile robots (see e.g., [11],[4]). A people tracking system using multiple, stationary omni-cameras was presented by Sogo et al. [9]. However, for a mobile robot this system would not be so useful, since it requires several omni-cameras at different positions.

Some very advanced methods for vision-based people tracking using regular or stereo cameras have been developed. For example, in Pfinder [12] a 3D description of the person is recovered. Also the W^4 system tracking body parts proposed by Haritaoglu [3]. To locate and track people these systems use information such as colour cues, shape analysis and robust tracking techniques. Use of these methods with an omni-camera sensor is limited (e.g., we don't have information about the whole person), although some of the vision processing and tracking techniques could be used in our future work.

A good example of mobile robots designed to operate in populated environments is the museum tourguide system RHINO [2] tested in Deutsches Museum Bonn and its successor MINERVA [10] which operates at the Smithsonian's National Museum of American History. RHINO and MINERVA use information from laser range finder and sonar to detect people. Recently a laser-based tracking system for mobile robots was proposed [8] which can track multiple persons using Joint Probabilistic Data Association Filters (JPDAF). A benefit of this approach is that the JPDAF can represent multi-modal distributions, compared to a Kalman filter which assumes a Gaussian distribution.

Figure 1: Omnidirectional camera mounted on the top of the Nomad 200 mobile robot.

3 External Measurement System

In order to carry out learning of the appearance model and to evaluate results, it was necessary to acquire information about the true position of the person (ground truth). Therefore, an external positioning system was developed to measure the real position of the person. To achieve this aim while keeping down costs, web-cameras were used to track a distinctly coloured object (the green "hat" worn by the person shown in Fig. 2). The system was developed so that it can operate with an arbitrary number of cameras ($N \geq 2$). Here, four Philips PCVC 740K web-cameras (resolution 320×240), connected by a 4×USB port to a Pentium III PC, were mounted in the corners of the 10×5 m area of the robotics lab at our institute (see Fig. 3). The orientation and position of the cameras was adjusted to cover the area of interest with as many cameras as possible.

Each camera first computes an estimate of the angle φ_i to the centre of the coloured object. For each combination of two cameras that can actually sense the whole coloured object, an estimate of the position $\vec{p}_{ij}$ is then calculated by

Figure 2: Pictures from measurement system with the person tracked.

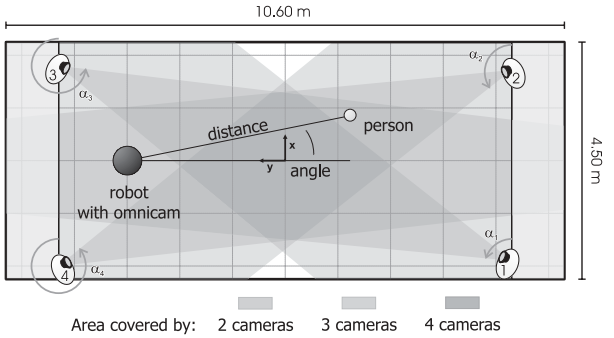


Figure 3: Absolute positioning system with 4 cameras. The figure shows a floor plan of the laboratory room and the placement of the webcams and the robot with the omnicam. Also plotted are the fields of view for each camera, shaded according to the number of cameras which can sense a particular region.

triangulation. With N cameras up to $N(N-1)/2$ valid position estimates $\vec{p}_{ij}$ are produced at each time interval, which are then combined to determine a final position estimate $\vec{p}$ in room coordinates.

The parameters of the cameras (heading α_i, coordinates X_i, Y_i and angular range $\Delta\alpha_i$) were determined by an initial calibration process that minimizes the average distance $\bar{d}$ between measured and known positions of several locations at which the coloured obect is placed. The calibration process is crucial because the positioning performance heavily depends on the accuracy of the camera parameters.

In the experiments presented here, the person taking part wears a coloured hat, which can be tracked by the measuring system but cannot be seen by the omni-camera. During a calibration procedure, the person stands at a number of fixed positions. Despite the comparatively poor resolution a good accuracy in the order of just a few centimeters ($\bar{d} \approx 1$ cm) could be achieved in this way.

The robot with the omni-camera was placed in the middle of one side of the experimental area, so that the performance of the system could be assessed over the largest possible range of distances (see Fig. 3).

4 Learned Appearance Model

In order to obtain useful information from the omni-camera, an appearance model is required. This could be derived by analytical methods, but in the case of non-linearities and noise this process can be difficult. Learning techniques can help either to find unknown parameters, or to learn the whole model of the sensor. In our work, we used an artificial neural network to estimate the distance of the person to the robot from a set of features extracted from the omni-camera images. The angle to the person can be calculated directly from the horizontal position in the panoramic image (see below), so we only need to consider learning of the distance.

4.1 Camera Set-up

The vision sensor was built from a CCD camera (Hitachi KP-D50) with a conical mirror attached above. The sensor is mounted on top of a Nomad 200 mobile robot, though in this work we have assumed that the robot was not moving. The total height of the robot with the omni-cam was about 1.7 m (see Fig. 1). This meant that the sensor could not see the whole person, but just a lower part of the body and legs (see Fig. 4). However, this was enough for our experiments.

4.2 Pre-processing

The omni-camera produces a circular image of its surroundings, so to use it in a convenient way, all coordinates were first changed from cartesian to polar. After unwrapping the picture to polar coordinates, the person can be detected and localized by using the following steps:

- *Background subtraction:* for every frame, the difference with the background is calculated. The background was recorded earlier with no moving person in the picture (taking the average of five pictures). This method can only be used under the assumption that the robot is not moving.
- *Segmentation of the person:* a histogram of difference data in both vertical and horizontal directions is created (see figure 4.b). Data which has a value higher

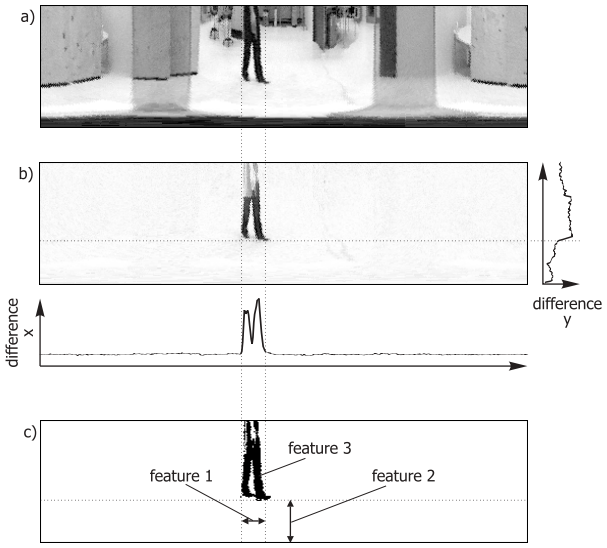
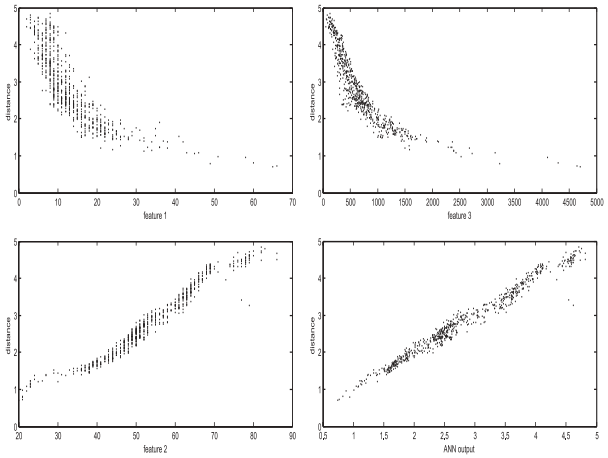


Figure 4: Feature extraction: a) original image (resolution 480×120 pixels) b) background subtraction and histograms along vertical and horizontal directions c) resulting image



Feature	r
1. Person width	-0.778
2. Apparent distance	0.973
3. Total no. pixels	-0.803
ANN output	0.987

Figure 5: Correlation between ground truth distance and every feature (top-left, bottom-left, top-right) and appearance model output (bottom-right). In the table, r is the linear correlation coefficient [6].

then a certain threshold (learned during background acquisition) is used for localization of the person in the image (4.c).

The angular position of the person can be obtained directly from the horizontal histogram (using the position of the mean value of this data). The average angle error value was about 2.01 ± 1.60 degrees, so there was no need to learn to estimate the angle.

4.3 Feature Extraction

We decided to use three features that can be extracted from the processed image:

- *Feature 1 - person width:* this is obtained from the distance between the limits of the horizontal histogram. If the person is closer to the omni-camera, their width tends to be bigger, however this can vary depending on the size and orientation of the person.
- *Feature 2 - apparent distance:* this is obtained from the distance between the lower limit of the vertical histogram and the bottom edge of the picture. This is the most useful feature, increasing with the true distance of the person from the camera, although shadows can be a problem.
- *Feature 3 - total number of pixels:* this is obtained from the number of pixels with intensity above a certain threshold (learned during background acquisition). Again, this can vary with the size and orientation of the person.

The quality of these features depends on several factors. The most important are the resolution of the omnicamera and quality of the converted polar images. Disturbances in the environment such as light conditions, shadows or unexpected movements can also be a problem. In order to assess the quality of our feature data, we measured the linear correlation coefficient [6] for each feature compared to the true distance of the person. The results are shown in Fig. 5.

4.4 Artificial Neural Network

An artificial neural network (ANN) was used to map the extracted features onto distance values. We used a multi-layer feedforward neural network (MLFF) with three inputs, one hidden layer and one output. During training, the distance information from the external measuring system was used to provide the target outputs for the ANN.

In our experiments, we used 684 images collected at a frequency of 3 Hz. Two different people took part in the experiment, one in each half of the data. After feature extraction, 30% of the data was used for training and 70% for testing the MLFF network. The best results were obtained with 4 units in the hidden layer and a learning rate of 0.3.

The results in Fig. 5 show that the ANN improves on the correlation of the input features with the ground truth distance.

5 Kalman Filter

The appearance model provides information about the distance and angle to the person. To improve these results, a Kalman filter can be used [5]. The Kalman filter uses all of the available knowledge about the process to produce the best estimate of the person's position (the errors are minimized statistically). The filtering procedure consists of two basic steps: prediction and correction. The estimated velocity of the person is used to predict their next position. This prediction is then combined with the next observation obtained from the appearance model.

Let $\boldsymbol{x} \in \mathbb{R}^2$ be a position of the person. At a given time k it can be expressed by the difference equation

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \boldsymbol{u}_k + \boldsymbol{w}_{k-1}, \tag{1}$$

where $\boldsymbol{u} \in \mathbb{R}^2$ is the nominal velocity of the person and $\boldsymbol{w} \in \mathbb{R}^2$ velocity disturbances.
The information obtained from the sensor is a measurement

$$\boldsymbol{z}_k = \boldsymbol{x}_k + \boldsymbol{v}_k, \tag{2}$$

where $\boldsymbol{v} \in \mathbb{R}^2$ represents measurement noise. Random variables $\boldsymbol{w}$ and $\boldsymbol{v}$ are assumed to be independent and are modelled as a white noise with normal probability distribution with covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$.

If $\hat{\boldsymbol{x}}_k^- \in \mathbb{R}^2$ is a prediction of the position then the estimate error can be defined as

$$\boldsymbol{e}_k^- = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-, \tag{3}$$

and its covariance matrix as

$$\boldsymbol{P}_k^- = E[\boldsymbol{e}_k^- \boldsymbol{e}_k^{-T}]. \tag{4}$$

In every prediction step, estimates of the position and error covariance matrix are updated

$$\hat{\boldsymbol{x}}_k^- = \hat{\boldsymbol{x}}_{k-1} + \boldsymbol{u}_k, \tag{5}$$
$$\boldsymbol{P}_k^- = \boldsymbol{P}_{k-1} + \boldsymbol{Q}. \tag{6}$$

Then the correction procedure is applied

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^- + \boldsymbol{K}_k(\boldsymbol{z}_k - \hat{\boldsymbol{x}}_k^-), \tag{7}$$
$$\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{K}_k)\boldsymbol{P}_k^-, \tag{8}$$

where

$$\boldsymbol{K}_k = \boldsymbol{P}_k^-(\boldsymbol{P}_k^- + \boldsymbol{R})^{-1}. \tag{9}$$

Filtering was applied to data expressed in room coordinates. All the initial conditions for the Kalman filter were obtained during the training phase. In our experiments:

$$\boldsymbol{Q} = \begin{bmatrix} 0.024 & -0.001 \\ -0.001 & 0.043 \end{bmatrix}, \boldsymbol{R} = \begin{bmatrix} 0.011 & -0.004 \\ -0.004 & 0.020 \end{bmatrix}. \tag{10}$$

6 Experimental Results

6.1 Appearance Model

The artificial neural network was tested with 70% of all collected data. We repeated the training and testing procedure 10 times, where the data for training were randomly chosen from whole sample set. The results in the following table show the average distance error with standard deviation.

Results	Avg. error in distance / m
Average	0.126 ± 0.167
Best	0.110 ± 0.099
Worst	0.163 ± 0.325

We also tested with different appearance models for each person. Training data was chosen individually from the set belonging to the given person. The results in the following table show the average distance error with standard deviation.

Test Subject	Appearance model trained for Person 1	Person 2
Person 1	0.096 ± 0.074	0.133 ± 0.126
Person 2	0.231 ± 0.276	0.117 ± 0.111

The results show that the performance of the person-specific appearance models is significantly better than that of the general appearance model (at the 99% confidence level, using Student's t-test for unpaired samples [6]), provided that the person has been identified correctly.

6.2 Kalman Filter

The results obtained by tracking with the Kalman filter are shown in Fig. 6 and the following table.

Tracking method	Avg. position error / m
Appearance model	0.154 ± 0.094
With Kalman filter	0.145 ± 0.092

The results show that the performance of tracking with the Kalman filter is significantly better than that of the appearance model alone (at the 99% confidence level, unpaired t-test).

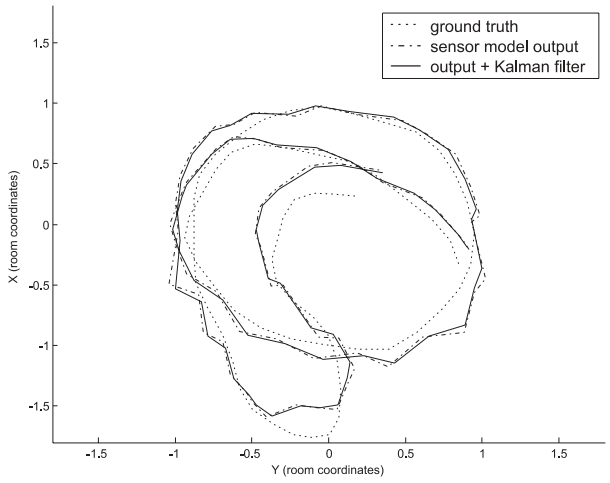


Figure 6: Fragment of tracked path presented in room coordinates.

7 Conclusions and Future Work

In this paper, we have presented an appearance-based algorithm for tracking a human using an artificial neural network to learn the appearance model together with a Kalman filter. Possible extensions to the system are discussed as follows:

- *Motion model:* to obtain a better velocity estimate in tracking, a more sophisticated motion model could be developed, or such a model could be learned from data.
- *Multi-person tracking:* the system should be extended to track more than one person at the same time. To achieve this, we would need to be able to represent multi-modal distributions, and also deal with possible occlusions.
- *Tracking on a moving robot:* in order to use the system on a moving robot, a more sophisticated algorithm for background-object extraction is required. Possible methods would include correlation methods to minimise the difference between successive images from the omni-camera. This ability is required so that the robot can learn tasks such as following, finding or guiding people.
- *Integration with a people identification system:* Our experiments show that more accurate tracking is possible if the person being tracked can be identified. It would be possible with our system to use the general appearance model first and then switch to the person-specific appearance model when the person has been identified with high certainty. In ongoing experiments, we are investigating integration of methods for people recognition, tracking and identification.

References

[1] T. Boult, A. Erkin, P. Lewis, R. Michaels, C. Power, C. Qian, and W. Yin. Frame-rate multi-body tracking for surveillance. In *Proc. DARPA IUW*, 1998.

[2] W. Burgard, A.B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1-2):3–55, 1999.

[3] I. Haritaoglu, D. Harwood, and L. Davis. Who, when, where, what: A real time system for detecting and tracking people. In *Proc. Third Face and Gesture Recognition Conference*, pages 222–227, 1998.

[4] M. Jogan and A. Leonardis. Robust localization using panoramic view-based recognition. In *Proc. 15th Int. Conf. on Pattern Recognition (ICPR'00)*, pages 136–139. IEEE Computer Society, 2000.

[5] P.S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, 1979.

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

[7] Y. Rui, A. Gupta, and J.J. Cadiz. Viewing meeting captured by an omni-directional camera. In *CHI*, pages 450–457, 2001.

[8] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. Tracking multiple moving objects with a mobile robot. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[9] T. Sogo, H. Ishiguro, and M. Trivedi. N-ocular stereo for real-time human tracking. In R. Benosman and S.B. Kang, editors, *Panoramic Vision: Sensors, Theory, and Applications*. Springer, 2000.

[10] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C.R. Rosenberg, N.Roy, J. Schulte, and D. Schulz. MINERVA: A tour-guide robot that learns. In *KI - Kunstliche Intelligenz*, pages 14–26, 1999.

[11] N. Winters, J. Gaspar, G. Lacey, and J. Santos-Victor. Omni-directional vision for robot navigation. In *Proc. IEEE Workshop on Omnidirectional Vision - Omnivis00*, 2000.

[12] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.