Data Association and Occlusion Handling for Vision-Based People Tracking by Mobile Robots

Grzegorz Cielniak[a], Tom Duckett[a], Achim J. Lilienthal[b]

[a] *Department of Computing and Informatics, University of Lincoln, LN6 7TS Lincoln, United Kingdom*
[b] *Centre for Applied Autonomous Sensor Systems, Örebro University, SE-701 82 Örebro, Sweden*

Abstract

This paper presents an approach for tracking multiple persons on a mobile robot with a combination of colour and thermal vision sensors, using several new techniques. First, an adaptive colour model is incorporated into the measurement model of the tracker. Second, a new approach for detecting occlusions is introduced, using a machine learning classifier for pairwise comparison of persons (classifying which one is in front of the other). Third, explicit occlusion handling is then incorporated into the tracker. The paper presents a comprehensive, quantitative evaluation of the whole system and its different components using several real world data sets.

Key words: AdaBoost, Occlusion Detection, Thermal Vision, Colour Vision, Bayesian estimation

1. Introduction

This paper addresses the problem of people detection and tracking by mobile robots in indoor environments. A system that can detect and recognise people is an essential part of any mobile robot that is designed to operate in populated environments. Information about the presence and location of persons in the robots surroundings is necessary to enable interaction with the human operator, and also for ensuring the safety of people near the robot.

Email addresses: gcielniak@lincoln.ac.uk (Grzegorz Cielniak), tduckett@lincoln.ac.uk (Tom Duckett), achim.lilienthal@tech.oru.se (Achim J. Lilienthal)

The presented people tracking system uses a combination of thermal and colour information to robustly track persons. The thermal camera simplifies the detection problem, which is especially difficult on a mobile platform. The system is based on a fast and efficient sample-based tracking method that enables tracking of people in real-time. The measurement model using gradient information from the thermal image is fast to calculate and allows detection and tracking of persons under different views. An explicit model of the human silhouette effectively distinguishes persons from other objects in the scene. Moreover the process of detection and localisation is performed simultaneously so that measurements are incorporated directly into the tracking framework without thresholding of observations. With this approach persons can be detected independently from current light conditions and in situations where other popular detection methods based on skin colour would fail.

A very challenging situation for a tracking system occurs when multiple persons are present on the scene. The tracking system has to estimate the number and position of all persons in the vicinity of the robot. Tracking of multiple persons in the presented system is realised by an efficient algorithm that mitigates the problems of combinatorial explosion common to other known algorithms. A sequential detector initialises an independent tracking filter for each new person appearing in the image, using thermal information. A single filter is automatically deleted when it stops tracking a person.

While thermal vision is good for detecting people, it can be very difficult to maintain the correct association between different observations and persons, especially where they occlude one another, due to the unpredictable appearance and social behaviour of humans. To address these problems the presented tracking system uses additional information from the colour camera, introducing several techniques for improving data association and occlusion handling.

First, an adaptive colour model is incorporated into the measurement model of the tracker to improve data association. For this purpose an efficient integral image based method is used to maintain the real-time performance of the tracker.

Second, to deal with occlusions the system uses an explicit method that first detects situations where people occlude each other. This is realised by a new approach based on a machine learning classifier for pairwise comparison of persons that uses both thermal and colour features provided by the tracker. Our approach uses the AdaBoost algorithm [1] to build the classifier from the available thermal and colour features.

Third, the information from the occlusion detector is then incorporated into the tracker for occlusion handling and to resolve situations where persons reappear in a scene.

Further to our previously published results [2], this paper presents a comprehensive, quantitative evaluation of the whole system and its different components using several real world data sets recorded in an office environment. We analyse the relative influence of different visual features for occlusion handling, and further demonstrate the robustness and efficiency of the approach.

1.1. Related Work

Many approaches for people tracking on mobile platforms are based on skin colour and face recognition (e.g., [3, 4]). However these methods require persons to be close to and facing the robot so that their hands or faces are visible. Stereo vision provides extra range information that makes the segmentation of persons easier, allowing for detection of both standing and moving people regardless their orientation [5]. However the provided depth information is very coarse and therefore limits the number of possible applications. Our system makes use of thermal vision that takes advantage of the fact that humans have a distinctive thermal profile compared to nonliving objects. Moreover thermal information is not influenced by changing lighting conditions and allows detection of people even in darkness. Infrared sensors have been applied to detect pedestrians in driving assistance systems (e.g. [6],[7]) but their use in robotic applications is limited, probably due to the high price of the sensors.

Other people tracking systems are based on range-finder sensors such as laser scanner and sonar that are very popular sensors in mobile robotics for navigation and localisation tasks. The system in [8] uses a laser scanner sensor to track multiple persons. It is based on a particle filter and JPDAF data association, uses a global representation of the environment, requires thresholded sensor data and deals with occlusions of non-interacting persons only. In contrast, our system uses sensor coordinates, incorporates unthresholded data and can reason about occlusions of interacting persons. The work of Zajdel at al. [9] presents a robotic system that tracks and re-identifies persons when they re-appear on the scene. However the tracking procedure is realised by a Bayesian network that grows rapidly and requires storage of all data. It is therefore of limited use for on-line applications.

Classical tracking algorithms usually handle the detection and tracking tasks separately in order to simplify the whole problem [10, 11]. However, such an architecture can cause loss of information between these steps, in addition to the computational cost of detection by exhaustive search of all possible object states [12]. The alternative approach considers these problems simultaneously (track-before-detect, also called unified tracking [13]). The presented system is designed in this latter spirit, using a track-before-detect technique.

To deal with problems of occlusions several authors proposed solutions that use special sensors or their special arrangement. One example system uses a camera placed above the observed scene [14]. Persons observed from such a view-point cannot occlude each other. Another example is a multi-camera system [15] where ambiguities caused by occlusion are resolved by combining information from different cameras placed in different places. All these solutions can be used only in a few, very controlled scenarios and their use in mobile applications would be especially troublesome if not impossible.

In the majority of people tracking systems the problem of occlusion is solved within the tracking framework. Possible approaches handle occlusions either implicitly without reasoning, or model them explicitly. Implicit solutions use kinematic information as well as dedicated measurement models [16, 17, 18]. However the behaviour of people tends to be highly unpredictable in general, and they may or may not interact. Therefore implicit approaches can deal only with specific cases, i.e., short-term occlusions. The proposed system uses an explicit approach to deal with occlusions. This reasoning requires domain specific knowledge, i.e., detection of situations when persons appear to merge and split, and making decisions about their behaviour during occlusion (see for example [19, 20, 21]). We use colour as additional information that helps to detect occluded persons and resolve occlusions when occluded persons appear again on the scene.

In the next section we introduce the experimental platform. Section 3 presents the basic tracker using gradient information from the thermal camera. The next sections describe the techniques developed to maintain the correct associations between observations and persons, by exploiting the combination of thermal and colour vision: incorporation of colour information into the measurement model (Section 4), an occlusion detector based on the machine learning algorithm AdaBoost (Section 5) and the occlusion handling procedure (Section 6). Experimental results are presented in Section 7, followed by conclusions and suggestions for future work.

Figure 1: ActivMedia PeopleBot robot equipped with a thermal camera and a standard camera (left). Example of an image from the colour camera (right-top) and thermal camera (right-bottom).

2. Experimental Set-up

We used an ActivMedia PeopleBot robot (Fig. 1) equipped with different sensors, including a colour pan-tilt-zoom camera (VC-C4R, Canon) and, a thermal camera (Thermal Tracer TS7302, NEC), and an Intel Pentium III processor (850 MHz). The colour and thermal camera are mounted close to each other which simplifies the calibration procedure between the two cameras (see Section 4.1).

The robot was operated in an indoor environment (a corridor and lab room). Persons taking part in the experiments were asked to walk in front of the robot while it performed a corridor following behaviour or while the robot was stationary. At the same time, image data were collected with a frequency of 15Hz. The resolution of both thermal and colour images was 320×240 pixels. In our set-up the visible range on the grey-scale thermal image was equivalent to the temperature range from 24 to 36 °C.

3. Basic Tracker Using Thermal Vision

3.1. Particle-based Tracking of a Single Person

To reliably estimate the location and movement of persons it is necessary to apply a tracking procedure. Our system uses a particle filter to

provide an efficient solution to this problem despite the high dimensionality of the state space. The particle filter performs both detection and tracking simultaneously without exhaustive search of the state space. Moreover the measurements are incorporated directly into the tracking framework without any preprocessing such as thresholding that could cause loss of information.

The posterior probability $p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t})$ of the system being in state $\boldsymbol{x}_t$ given a history of measurements $\boldsymbol{z}_{1:t}$ is approximated by a set of N weighted samples such that

$$p(\boldsymbol{x}_t|\boldsymbol{z}_{1:t}) \approx \sum_{i=1}^{N} w_i \delta(\boldsymbol{x}_t - \boldsymbol{x}_t^i). \tag{1}$$

Each $\boldsymbol{x}_t^i$ describes a possible state together with a weight w_t^i which is proportional to the likelihood that the system is in this state. We use a standard Sampling Importance Resampling (SIR) filter [22] starting with a uniform initial distribution. The dynamic model used in the particle filter is random walk with drift. The measurement model used to calculate new weights for particles is presented later in Section 3.3. The resampling step was implemented using the systematic resampling algorithm [23].

3.2. Tracking Multiple Persons

The above method is extended to the multi-person case by detecting new persons incrementally as they appear while maintaining existing tracks of persons. This system uses a set of independent particle filters to track different persons. To assign new filters to new persons we use a sequential detector consisting of a set of N randomly initialised particles. These particles are used to "catch" a new person entering the scene. To avoid multiple detections in the same or similar regions, the weight of detection particles is penalised by a factor $\psi_d < 1$ in cases where particles cross already detected areas. The weight update equation for the i^{th} detection particle is modified to $w_t^i \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t = \boldsymbol{x}_t^i)\psi$, where $\psi = \psi_d$ if particle i overlaps with other detected regions and $\psi = 1$ otherwise. Thus already existing filters naturally limit the search space for the detector. Detection occurs when the average fitness of the particles exceeds a certain threshold for a few consecutive frames (3 in our experiments). Then the particles from the detector are used to initialise a new tracker before being re-initialised for detection of the next new person.

A solution based on independent tracking filters is computationally inexpensive and appropriate for on-line applications, but suffers in cases when tracked persons are too close to each other. To reduce these problems we

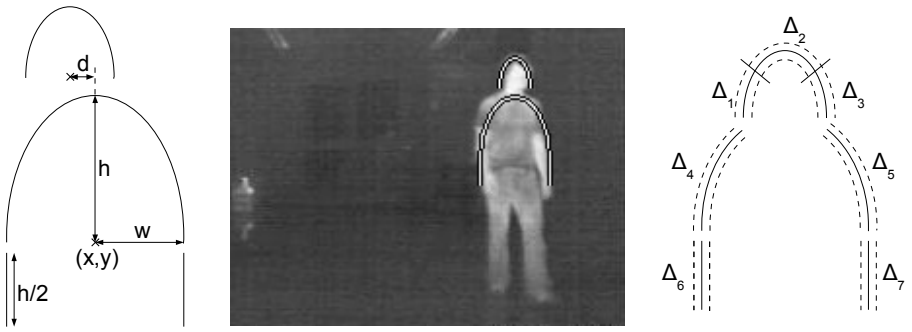


Figure 2: The elliptic measurement model for thermal images. Model parameters are shown on the left. Division of ellipses into 7 regions is shown on the right.

explicitly model interactions between persons by penalising the weights of particles that intersect with other detected regions. The weight update equation for established tracking filters is changed to $w_t^i \propto p(\boldsymbol{z}_t|\boldsymbol{x}_t = \boldsymbol{x}_t^i)\psi$, where $\psi = e^{(-\rho g_{ij})}$ and g_{ij} expresses the amount of overlap between particle i and already detected region tracked by filter j, which is multiplied by a factor ρ in the exponent of the penalty term. The penalty factor ρ allows for specifying the "strength" of interactions between persons and the amount of handled partial occlusions. This solution is similar to the interaction model proposed by [24], where the authors propose a Random Markov Field using a joint state space representation. The treatment of interactions in both approaches has the drawback that in the case of occlusions weaker filters disappear. Motion information could help here only in specific situations where persons are just passing by each other at sufficient speeds. However this is not the case in situations where people stop to talk, shake hands, walk in groups, etc.

3.3. Elliptic Contour Model

The measurement model used by our thermal tracker is a contour model consisting of two ellipses: one describes the position of the body part and the other measures the position of the head part (Fig. 2). Thus we obtain a 9-dimensional state vector: $\boldsymbol{x}_t = (x, y, w, h, d, v_x, v_y, v_w, v_h)$ where (x, y) is the mid-point of the body ellipse with width w and height h. The height of the head is calculated by dividing h by a constant factor. The displacement of the middle of the head part from the middle of the body ellipse is described by d. We also model velocities of the body part as (v_x, v_y, v_w, v_h). The

velocity of the d component has very noisy characteristics and is therefore not considered in the state vector. To calculate the importance weight w_t^i of a sample i with state $\boldsymbol{x}_t^i$ we divide the ellipses into $m = 7$ different regions (see Fig. 2) and for each region j the image gradient Δ_j^i between pixels in the inner and outer parts of the ellipse is calculated. The gradient is maximal if the ellipses fit the contour of a person in the image data. A fitness value f^i for each sample i is then calculated as the sum of all gradients multiplied with individual weights α_j for each region: $f^i = \sum_{j=1}^{m} \alpha_j \Delta_j^i$. The weights α_j sum to one and are chosen such that the shoulder parts have lower weight to minimise the measurement error that occurs due to different arm positions. The fitness value is finally scaled to values in $[0, 1]$ in order to represent a likelihood:

$$p_g(\boldsymbol{z}_t|\boldsymbol{x}_t^i) = \frac{\exp(\kappa \cdot (f^i - \theta))}{\exp(\kappa \cdot (f^i - \theta)) + \exp(\kappa \cdot (\theta - f^i))}, \tag{2}$$

where θ denotes a fitness threshold and the value of κ defines the slope of the likelihood function. This kind of likelihood function was also used in [25] for visual tracking of objects.

When the mean gradient value from Eq. 2 is greater than 0.5 then a person is considered to be detected. We also check the uncertainty of the estimate [26] to avoid detections in wrong regions when the posterior is multi-modal (e.g. for multiple persons).

This approach is similar to the work by Isard and Blake [27] for tracking people in a greyscale image. However, they use a spline model of the head and shoulder contour which cannot be applied in situations where the person is far away or visible in a side view, because there will be no recognisable head-shoulder contour. The elliptic contour model used here is able to cope with these situations.

4. ADAPTIVE COLOUR MODEL

4.1. Colour representation

Since the baseline between the cameras is small compared to the distance to persons, it is possible to align the thermal and colour images by affine transformation. We then use an efficient colour representation proposed in [28] based on the first three moments (mean, variance and skewness) of the colour distribution. This representation was shown to be more effective than histogram methods (e.g., [29]) in the domain of image indexing. To include

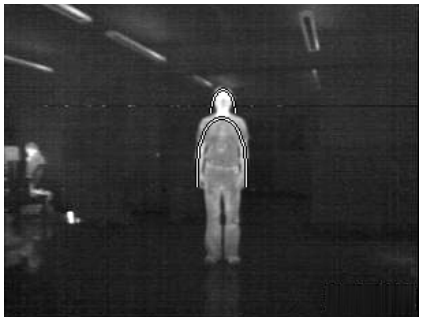
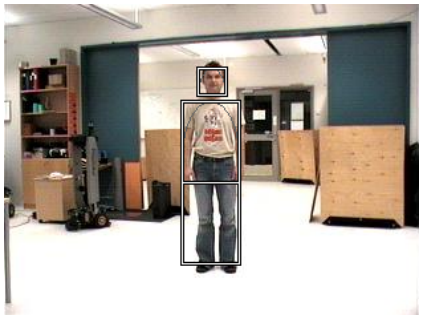
a)  b) 

Figure 3: Rectangular features: a) thermal image b) colour image with regions corresponding to different body parts from which colour information is extracted.

information about the spatial layout of the colour we divided the region corresponding to a person's body into rectangular sub-areas from which we calculate the colour statistics (see Fig. 3b). The position and size of these regions are determined from the information provided by the elliptic contour model.

4.2. Colour likelihood

The appearance model based on colour moments is created every time a new detection occurs, i.e. a new track is initialised in the thermal image. By using the affine transformation we are able to determine the region corresponding to a person on the colour image (see Fig. 3). From three rectangular regions corresponding to the person's head, torso and legs we collect colour statistics $\boldsymbol{c}_t$ of the first three moments (m_1, m_2, m_3) for three colour channels (R, G, B). Finally we obtain a feature vector $\boldsymbol{c}_t$ of size $3 \times 3 \times 3 = 27$. To make the model more robust to changing light conditions we adapt it while a person is tracked. In our implementation we store colour statistics from the last n_k frames and calculate their mean value. The parameter n_k influences the robustness and adaptivity of the colour model. In our experiments $n_k = 10$ corresponding to 0.7 s. We use Euclidean distance to measure the similarity between the model $\boldsymbol{c}_t^\star$ and region of interest $\boldsymbol{c}_t$. Finally, the likelihood model for colour information is

$$p_c(\boldsymbol{z}_t|\boldsymbol{x}_t) = exp\left(-\lambda d_t^2\right), \tag{3}$$

where λ is a parameter that determines the shape of the colour likelihood. Since λ scales the distance, higher values of λ mean that the colour-based

likelihood model is more peaked, thus having more importance when combined with the gradient information from the ellipse model.

4.3. Rapid rectangular features

The colour moments can be rapidly calculated using an integral image representation [30]. The estimators for the first three moments of the colour distribution can be obtained by means of k-statistics calculated using sums of the rth powers of the colour data:

$$S_r = \sum_{i=x}^{x+w} \sum_{j=y}^{y+h} I^r(i,j), \tag{4}$$

where $I(i,j)$ is a pixel value of the colour image selected from the rectangular region specified by coordinates $\{x, y, x+w, y+h\}$. Each S_r can be quickly calculated using the integral image representation. The first three k-statistics are obtained as

$$k_1 = S_1/n, \tag{5}$$

$$k_2 = \frac{nS_2 - S_1^2}{n(n-1)}, \tag{6}$$

$$k_3 = \frac{2S_1^3 - 3nS_1S_2 + n^2S_3}{n(n-1)(n-2)}, \tag{7}$$

where $n = w \times h$. Finally the normalised values of estimators for mean m_1, variance m_2 and skewness m_3 can be obtained as $m_1 = k_1$, $m_2 = k_2/k_1$ and $m_3 = k_3/k_2^{\frac{3}{2}}$. The normalisation is performed to balance the influence of each moment on the final score.

4.4. Combining thermal and colour information

If we assume that the likelihoods for the gradient model $p_g(\boldsymbol{z}_t|\boldsymbol{x}_t)$ (Eq. 2) and colour model $p_c(\boldsymbol{z}_t|\boldsymbol{x}_t)$ (Eq. 3) are independent then the data fusion can be realised by taking a product of these two likelihoods

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t) = p_g(\boldsymbol{z}_t|\boldsymbol{x}_t)p_c(\boldsymbol{z}_t|\boldsymbol{x}_t). \tag{8}$$

The parameters κ, θ (gradient model) and λ (colour model) specify the shape of the gradient and colour likelihood functions, thus specifying the importance of the respective features. The influence of possible correlations between colour and thermal distributions should be investigated more thoroughly in future work.

When a person is not detected, a colour model cannot be built and only gradient information can be used to update the weight of the particles of a single tracking filter as $w_t^i = p_g(\boldsymbol{z}_t|\boldsymbol{x}_t^i)\psi$. However as soon as a person is detected the colour model can be created and the weight update equation changes to:

$$w_t^i = p_g(\boldsymbol{z}_t|\boldsymbol{x}_t^i)p_c(\boldsymbol{z}_t|\boldsymbol{x}_t^i)\psi,\ i = 1, \ldots, N. \tag{9}$$

Note that the sequential detector relies only on gradient information from the thermal image.

5. OCCLUSION DETECTION WITH ADABOOST

To detect occlusions we propose an approach that sorts the order of all persons in the image according to pairwise comparisons. The proposed occlusion detector specifies which one of two overlapping persons is in front of the other. The order of the persons from front-to-back is then determined by a sort procedure requiring $M_O \cdot log(M_O)$ comparisons where M_O specifies the number of overlapping persons.

There are several features that could indicate the correct order of two overlapping persons in the image, from which we have chosen a set of three thermal and three colour features:

- The strength (i.e., mean gradient value) of a tracking filter, since a person for which the corresponding tracker indicates a higher confidence is more likely to be in the front. This feature is, however, very noisy and is affected by many factors such as movement of the camera, temperature of the environment, etc.

- The top and bottom of the elliptic model can also indicate the depth of a person since closer persons appear taller and closer to the upper and bottom border of the image. However the bottom part can be cut when persons stand too close to the camera. The top of a person's head is a more reliable feature, though it is affected by the different height of persons.

- Another set of features is the colour similarity of the region corresponding to a person. We have chosen three such regions including the overlapping, non-overlapping and whole areas of a person. Occluded persons should have lower similarity values.

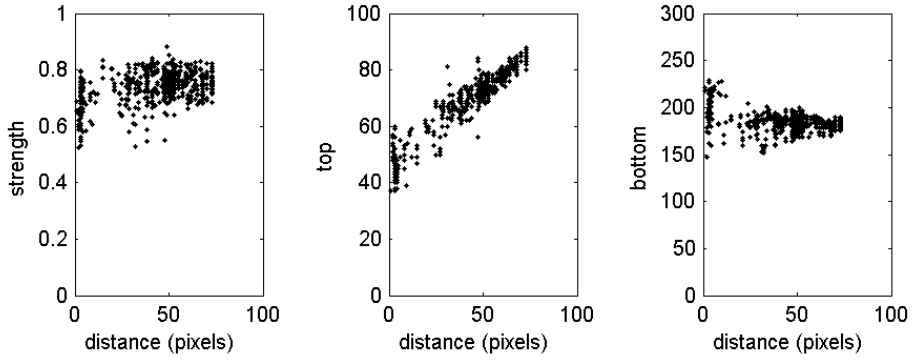


Figure 4: Relationship of the different thermal features to the apparent distance of a person taken from the ground truth data.

Since a single feature cannot easily determine the right order of the persons we use a boosting algorithm [1] to weight and combine a number of "weak classifiers" built from these features, resulting in a strong classifier with much improved occlusion detection accuracy.

To give an impression of the discriminative power of the thermal features used, we present a graphical representation of their relationship to the apparent distance of a person taken from the ground truth data (see Fig. 4). This distance uniquely determines the order of the persons. Note that range information from a laser scanner could also be used to simplify this problem. However in this work we consider an exclusively vision-based system. (It would not be meaningful to provide a similar visualisation for the colour features, since these features are based on comparisons of two tracked persons rather than a single tracked person as in the thermal case.)

We use the AdaBoost (Adaptive Boosting) classification algorithm [1] for selecting the best combination of features to detect occlusions. AdaBoost combines results from so-called "weak" classifiers $h_t(x)$ into one "strong" classifier $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$, where T is the number of weak classifiers and α_t is an importance weight given to each "weak" classifier $h_t(x)$ according to the performance during the iterative learning process (see [30] for details). During learning focus is put on the training examples which were most difficult to classify (this process is called "boosting"). As a result we obtain a final classifier that performs better than any of the weak classifiers alone.

Following [30] we use simple weak classifiers based on a single-valued feature $f_j(x)$

$$h_j(x) = \begin{cases} 1: & p_j f_j(x) < p_j \theta_j \\ 0: & \text{otherwise,} \end{cases} \tag{10}$$

where θ_j is a threshold and $p_j = \{-1, 1\}$ is a parity indicator determining the direction of the inequality sign. During the training procedure optimal values of θ_j and p_j are determined by minimising the number of misclassified training examples.

In addition, we use weak classifiers based on a weighted combination of features $f_j(x) = \sum_{i=1}^{G} \alpha_i f_i(x)$, where α_i specifies the weight for an input feature $f_i(x)$ ($G = 2$ in our experiments). We discretise possible weight values α_i from the range $\{-1, 1\}$ into N_f fractions. As a result we obtain a sufficient number of different weak classifiers for selection by the boosting algorithm.

6. OCCLUSION HANDLING

The learned occlusion detector can be used to improve tracking performance during occlusion. It is used in two different ways: first, to alter the penalising policy between the trackers (as described in Section 3), and second, to re-identify occluded persons when they reappear.

Our interaction model for tracking multiple persons allows tracking of people that overlap to a certain degree. This is achieved by modifying the interaction factor ρ to prevent target fetching (i.e., to prevent two filters in close proximity from collapsing around the same tracked object). The proposed pairwise occlusion detector is used to determine which of the tracking filters is occluded. We consider two possible situations: partial occlusion and total occlusion. During partial occlusion, some part of a person is still visible. However, the gradient along the contour is disturbed, which can cause a quick disappearance of the tracker. To avoid this we change the penalty equation to $\psi = e^{(-\rho_o g_{ij})}$, where the penalty term $\rho_o < \rho$ is used to model interactions between the partially occluded tracking filters. Interaction with other filters (non-overlapping with this pair) remains unchanged.

A modified update procedure for the tracker with improved occlusion handling is presented in Algorithm 1.

When the head contour of a person becomes occluded the corresponding tracker is considered to be totally occluded. This means that we can only

Algorithm 1 A modified update procedure for tracking filters (not totally occluded).

for each filter
  measure(thermal,colour)

handle occlusions():
- determine overlaps between filters
- determine occlusions between overlapping filters:
  - detect occlusions using the AdaBoost detector
  - assign occluded/occluding filters
  - assign partial/total occlusions
- adjust the penalty term ρ for each filter according to the type of occlusion

for each filter
  penalise()
  calculate estimates()

guess the true position of this person. We assume that the state of the occluded person is the same as the state of the occluding person. No penalty is considered for the occluded tracker. We keep particles of the totally occluded tracker for a short time (we use a value of 8 frames here) in situations when quick occlusions occur and the velocity of particles may allow resolution of this occlusion. However after this time has elapsed the particles of the tracker are removed and the only information kept is the colour model. When a new person is detected this information is used to match the colour model to all occluded trackers. If the colour model is most similar to the closest occluded tracker then the detected person is considered to be an occluded one. Otherwise the person is considered to be a new person. To avoid situations where the occluded tracker stays forever behind the occluding one, we also specify a maximum duration of occlusion (in our case 10 s). This minimises errors in the case where an occluded person disappears from the scene in some other way (e.g., through a door or a corridor behind an occluding person) or in cases of missed assignments to newly detected persons.

7. Experiments

7.1. Evaluation

Our system was tested on the data collected by the robot during several runs. We collected 11 tracks using corridor following behaviour and 42 tracks

	detection	localisation
recall	$\frac{N_R}{N_T}$	$\frac{\lvert A_T \cap A_R \rvert}{\lvert A_T \rvert}$
precision	$\frac{N_R}{N_C}$	$\frac{\lvert A_T \cap A_R \rvert}{\lvert A_C \rvert}$
accuracy	$\frac{2 \cdot N_R}{N_T + N_C}$	$\frac{2 \cdot \lvert A_T \cap A_R \rvert}{\lvert A_T \rvert + \lvert A_C \rvert}$

Table 1: Detection and localisation metrics.

with a stationary robot resulting in 53 different tracks including 12 different persons (5607 images containing at least one person and 6769 images in total). To obtain the ground truth data we used a flood-fill segmentation algorithm corrected afterwards by hand using the ViPER-GT tool [31]. We considered only a bounding box around a person. The top and bottom edges were determined from the contours of the head and feet while the sides were specified by the maximum width of the torso (without arms). The cases when persons appeared too close ($< 3m$) to or too far ($> 10m$) from the robot were not taken into account. The size of the bounding box was specified as $2 \cdot \mathit{width}$ and $3.5 \cdot \mathit{height}$ of the elliptic contour model, an approximation to the proportions of the human body. Bounding boxes from the ground truth data are referred to as *targets* and those from the tracker as *candidates*.

We use two kinds of metrics that indicate the quality of the tracking procedure: detection metrics (counting persons) and localisation metrics (area matching). Each type of metric is further divided into three statistics: recall, precision and accuracy. Recall indicates true positives ("hits"), precision indicates false alarms, and accuracy is a combination of both recall and precision (see Table 1). These metrics allow thorough testing of the properties and performance of the tracker as in [31] and [32].

A candidate is considered to be correctly detected if the overlap ratio between candidate and target bounding boxes is greater than 50%. Detection metrics take into account the number of correctly detected candidates N_R in one frame and compare it with the number of targets N_T and number of all candidates N_C. The final result is a weighted average of all frames. Localisation metrics express relations between areas corresponding to correctly detected candidates A_R, all candidates A_C and targets A_T. The final result is a weighted average of all frames. All of the metrics are normalised to give percentages.

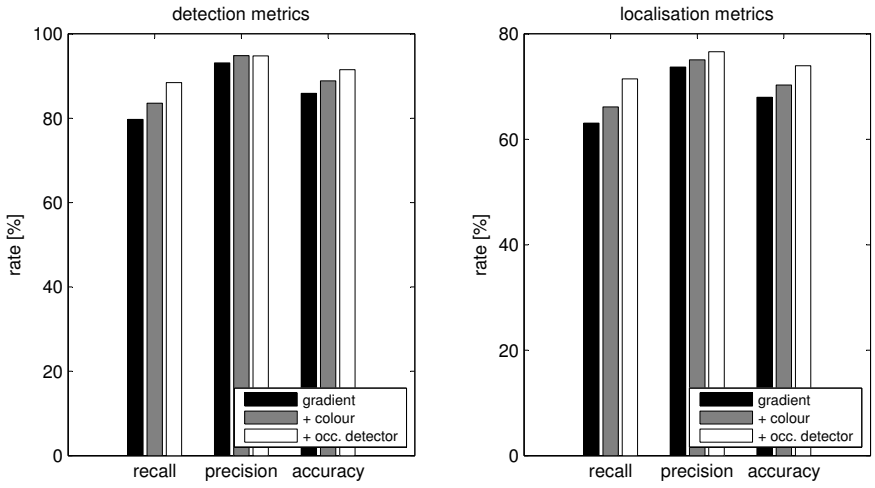


Figure 5: Detection and localisation metrics for tracking multiple persons without and with colour information and with occlusion handling procedure.

7.2. Training of the AdaBoost Classifier

To train the AdaBoost classifier the described thermal and colour features were extracted from the collected data. The only cases considered were situations when two or more people were overlapping. Moreover since the behaviour of the tracker without proper occlusion handling is unpredictable after a total occlusion occurs, only those examples that preceded the moment of the total occlusion were selected. During the occlusions, the colour models of the respective persons were not updated. In this way we obtained 121 positive and 121 negative examples giving a total of 242 examples.

We created additional weak classifiers based on weighted sums of pairs of features with 20 fractions giving, in the case of all six thermal and colour features used, 1200 new weak classifiers. We used 60% of randomly selected input examples as a training set and the remaining part as a test set. Each training procedure was repeated 10 times.

7.3. Results

7.3.1. Tracking Results

Fig. 5 shows the tracking performance using only thermal gradient information, with additional colour information, and with both colour information

Figure 6: Selected thermal images from the sequence showing the output from the tracker before, during and after the occlusion of three simultaneously tracked persons. The bounding boxes corresponding to occluded persons are marked by a dotted line.

and explicit occlusion handling. Each experiment was repeated 10 times with different random variations in the particle filter for each trial using $N = 1000$ particles per filter. The system parameters were optimised individually using an area accuracy metric as the performance criterion. Both detection and localisation metrics indicate a significant improvement when using additional colour information ($p < 0.01$). This leads to more precise estimates and decreases the number of cases where the tracker loses track of a person. However the overall accuracy (84.2% in detection and 68.7% in localisation) is affected by low recall values. Adding the occlusion detector gives an increase of 6.8% in area recall metrics and 3.1% in area accuracy metrics. Examples of the output from the tracker can be seen in Fig. 6.

7.3.2. Occlusion Detector

The strong classifier learned from the combination of thermal and colour features was able to predict occlusions correctly in around 89% of all cases (see Table 2). This gives a significant advantage over the results obtained when thermal and colour features were used separately ($p < 0.01$). Thermal features provided significantly better results than colour features alone.

Table 3 shows results for different methods of combining features into weak classifiers. The comparatively bad results when using single features

Feature type	**Results** [%]
thermal	76.39 ± 4.49
colour	69.07 ± 1.94
both	89.38 ± 2.48

Table 2: Classification results for different feature types.

Combination of features	**Results** [%]	T total
single	74.94 ± 4.88	6
weighted pairs	89.36 ± 2.48	1206
weighted triplets	89.38 ± 1.82	129206

Table 3: Classification results for different combination of features to create weak classifiers (resulting in T *total* weak classifiers).

are caused by the low number of weak classifiers. The proposed method of using a weighted combination of pairs of features increased the performance of the final classifier by around 15%. We also made tests with weighted triplets of features for comparison. Despite the much higher number of possible weak classifiers the difference in performance compared to weighted pairs was not found to be significant (based on a paired t-test with confidence level $p = 0.01$).

From the results presented in Table 5 we can get an impression about how much information is provided by a single feature. The most reliable features are the top of a person's head, colour similarity of the whole region and of the non-overlapping area. Weak classifiers based on combinations of these features had the highest importance (see Table 4). Other features also contributed to the final classifier (e.g., the position of the bottom of the elliptic model) even though their individual performance was relatively poor.

7.3.3. Processing Time

Table 6 presents the average processing time needed for calculation of 1000 samples when using different colour representations. It takes about two times longer to calculate one step of the tracking procedure when using

Place	Weight of a feature					
	strength	top	bottom	colour	colour_o	colour_no
1	-0.05	-	-	-	-	1.00
2	-	-0.05	-	1.00	-	-
3	-	-1.00	0.45	-	-	-
4	-	-0.75	1.00	-	-	-
5	-	-	0.05	-	-	1.00
6	-	-0.80	1.00	-	-	-
7	-	-	0.10	-	-	1.00
8	-0.55	1.00	-	-	-	-
9	-1.00	0.05	-	-	-	-
10	-	-	-0.05	1.00	-	-

Table 4: 10 best weak classifiers with their respective weights (colour_o and colour_no labels stand for colour similarity of the overlapping and non-overlapping area respectively).

all three moments compared to the tracker based on thermal information only (around 30Hz on a 2.00 GHz processor when using 1000 samples). A good trade-off between time requirements and performance of the tracker for our set-up is a representation using just the first moment of the colour distribution (46% more time compared to the gradient based tracker). The overall performance of the tracker based on this representation is about 2% lower than the variant using the three colour moments. When tracking multiple persons, additional processing time is required for calculation of penalty terms for the detector and individual tracking filters. In our case tracking one person required around 8% extra time for the detector and in the case of four persons around 36% extra time is needed for calculation of penalty terms between the trackers.

8. Conclusions and Future Work

We presented a people tracking system that uses a combination of thermal and colour information to robustly track persons. While thermal vision is good for detecting people, it can be very difficult to keep track of which observation corresponds to which person, due to the unpredictable appearance and social behaviour of humans. To address these problems the presented

Single feature	**Results** [%]
strength	50.10 ± 4.91
top	72.99 ± 3.85
bottom	56.49 ± 4.25
colour	67.62 ± 3.04
colour_o	45.56 ± 2.69
colour_no	67.42 ± 2.92

Table 5: Classification results for single features (colour_o and colour_no labels stand for colour of the overlapping and non-overlapping area respectively).

Platform		**Model**		
		gradient [ms]	colour I [ms]	colour III [ms]
robot	int. image	-	5.12	16.09
0.85 GHz	1000 samples	33.37	50.24	68.79
modern PC	int. image	-	2.09	4.90
2.00 GHz	1000 samples	13.52	17.66	25.89

Table 6: Average processing time needed to calculate 1000 samples using different measurements models. Label "colour I" and "colour III" correspond to a colour representation using the first moment and the first three moments respectively.

tracking system uses additional information from the colour camera. An adaptive colour model is incorporated into the measurement model of the tracker to improve data association. For this purpose an efficient integral image based method is used to maintain the real-time performance of the tracker.

To deal with occlusions, the system uses an explicit method that first detects situations where people occlude each other. This is realised by a new approach based on a machine learning classifier for pairwise comparison of persons that uses both thermal and colour features provided by the tracker. This information is then incorporated into the tracker for occlusion handling and to resolve situations where persons reappear in a scene.

We believe that the question of how to handle occlusions is impossible to

answer in a general way, i.e. independent of a particular application. However our solution demonstrates that it is plausible to deal with occlusions to some extent and through experiments we showed that this increases the overall performance of the tracker. Such a solution has obvious pitfalls that should be considered in future work such as proper handling of misclassification errors, wrong assignments after occlusions, uniformly dressed people, etc. A mobile robot itself could be used to check if the occluded person is really behind another person by taking appropriate actions. Recognition of human behaviour could also help to solve this kind of problem.

References

[1] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: Computational Learning Theory: Eurocolt, Springer-Verlag, 1995, pp. 23–37.

[2] G. Cielniak, T. Duckett, A. J. Lilienthal, Improved data association and occlusion handling for vision-based people tracking by mobile robots, in: Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Diego, CA, USA, 2007, pp. 3436–3441.

[3] T. Wilhelm, H. J. Böhme, H. M. Gross, Sensor fusion for vision and sonar based people tracking on a mobile service robot, in: Int. Workshop on Dynamic Perception, Bohum, Germany, 2002, pp. 315–320.

[4] L. Brèthes, P. Menezes, F. Lerasle, J. Hayet, Face tracking and hand gesture recognition for human-robot interaction, in: Proc. IEEE ICRA, New Orleans, LA, USA, 2004, pp. 1901–1906.

[5] R. Muñoz Salinas, E. Aguirre, M. García-Silvente, People detection and tracking using stereo vision and color, Image and Vision Computing 25 (6) (2007) 995–1007. doi:10.1016/j.imavis.2006.07.012.
URL `http://dx.doi.org/10.1016/j.imavis.2006.07.012`

[6] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, M. Meinecke, Pedestrian detection in infrared images, in: Proc. of the IEEE Intelligent Vehicles Symposium, Columbus, USA, 2003, pp. 662–667.

[7] H. Nanda, L. Davis, Probabilistic template based pedestrian detection in infrared videos, in: IEEE Intelligent Vehicle Symposium, Versailles, France, 2002.

[8] D. Schulz, W. Burgard, D. Fox, A. B. Cremers, Tracking multiple moving objects with a mobile robot, in: Proc. IEEE CVPR, 2001.

[9] W. Zajdel, Z. Zivkovic, B. J. A. Kröse, Keeping track of humans: Have i seen this person before?, in: Proc. IEEE ICRA, Barcelona, Spain, 2005.

[10] D. B. Reid, An algorithm for tracking multiple targets, in: Proc. IEEE Trans. Autom. Control, Vol. 6, 1979, pp. 843–854.

[11] Y. Bar-Shalom, T. Fortmann, Tracking and Data Association, Academic Press, 1988.

[12] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, D. G. Lowe, A boosted particle filter: Multitarget detection and tracking, in: Proc. ECCV, Vol. 1, 2004, pp. 28–39.

[13] L. D. Stone, T. L. Corwin, C. A. Barlow, Bayesian Multiple Target Tracking, Artech House, 1999.

[14] S. S. Intille, J. Davis, A. Bobick, Real-time closed-world tracking, in: Proc. IEEE CVPR, 1997, pp. 697–703.

[15] A. Mittal, L. S. Davis, M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo, in: Proc. IEEE CVPR, 2002, pp. 18–36.

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

[17] S. Khan, M. Shah, Tracking people in presence of occlusion, in: Asian Conference on Computer Vision, Taipei, Taiwan, 2000.

[18] M. Isard, J. MacCormick, Bramble: A Bayesian multiple-blob tracker., in: Proc. of the International Conference on Computer Vision, Vol. 2, Vancouver, British Columbia, Canada, 2001, pp. 34–41.

[19] A. Elgammal, L. S. Davis, Probabilistic framework for segmenting people under occlusion, in: Proc. of the International Conference on Computer Vision, Vancouver, Canada, 2001.

[20] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, R. Bolle, Appearance models for occlusion handling, in: Proceedings of the 2nd IEEE Workshop on Performance Evaluation of Tracking and Surveillance, Kauai, Hawaii, USA, 2001.

[21] S. Mckenna, S. Jabri, Z. Duric, A. Rosenfeld, Tracking groups of people, Computer Vision and Image Understanding 1 (80) (2000) 42–56.

[22] N. J. Gordon, D. J. Salmond, A. F. M. Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation, Proc. Inst. Elect. Eng. F 140 (2) (1993) 107–113.

[23] B. Ristic, S. Arulampalam, N. Gordon, Beyond the Kalman Filter - Particle Filters for Tracking Applications, Artech House, Boston, 2004.

[24] Z. Khan, T. Balch, F. Dellaert, An MCMC-based particle filter for tracking multiple interacting targets, in: Proc. ECCV, 2004.

[25] P. Li, H. Wang, Probabilistic object tracking based on machine learning and importance sampling, in: Proc. of the Iberian Conference on Pattern Recognition and Image Analysis, Vol. 1, 2005, pp. 161–167.

[26] R. Karlsson, F. Gustafsson, Monte Carlo data association for multiple target tracking, in: In IEEE Target tracking: Algorithms and applications, The Netherlands, 2001.

[27] M. Isard, A. Blake, Condensation – conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.

[28] M. A. Stricker, M. Orengo, Similarity of color images, in: Storage and Retrieval for Image and Video Databases, 1995, pp. 381–392.

[29] M. Swain, D. Ballard, Color indexing, International Journal of Computer Vision 7 (1991) 11–32.

[30] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. IEEE CVPR, 2001.

[31] D. S. Doermann, D. Mihalcik, Tools and techniques for video performance evaluation, in: Proc. ICPR, Vol. 4, Barcelona, Spain, 2000, pp. 4167–4170.

[32] K. Smith, D. Gatica-Perez, J. M. Odobez, S. Ba, Evaluating multi-object tracking, in: Workshop on Empirical Evaluation Methods in Computer Vision, San Diego, CA, USA, 2005.