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

Grzegorz Cielniak†, Tom Duckett† and Achim J. Lilienthal*

† Department of Computing and Informatics
University of Lincoln
LN6 7TS Lincoln, United Kingdom
gcielniak@lincoln.ac.uk,
tduckett@lincoln.ac.uk

* Centre for Applied Autonomous Sensor Systems
Örebro University
SE-701 82 Örebro, Sweden
achim.lilienthal@tech.oru.se

***Abstract*— This paper presents an approach for tracking multiple persons using a combination of colour and thermal vision sensors on a mobile robot. 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.**

I. INTRODUCTION

This paper presents a vision-based people tracking system allowing a mobile robot to detect and localise people in its surroundings, which uses a combination of thermal and colour information (see [2] for further details). The approach is based on an existing tracking system for thermal images [13]. 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. To further improve tracking of multiple persons, this paper introduces three main improvements to the system:

- incorporation of an adaptive colour model into the measurement model of the tracker to improve data association, using the integral image representation to speed up processing,
- explicit detection of occlusions, using a machine learning algorithm AdaBoost for pairwise comparison of persons (classifying which one is in front of the other), and
- integration of occlusion handling into the particle filter.

Many approaches for people tracking on mobile platform are based on skin colour and face recognition (e.g., [15], [1]). However these methods require persons to be close to and facing the robot so that their hands or faces are visible. The system in [9] uses a laser sensor to track multiple persons. It is based on a particle filter and JPDAF data association. It 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 [16] 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 Baysian network that grows rapidly and requires storage of all data, and is therefore limited for use in on-line applications.

II. 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 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 to allow for easy combination of the information (see Section IV-A). In our set-up the visible range on the grey-scale thermal image was equivalent to the temperature range from 24 to 36 °C.

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.

III. BASIC TRACKER USING THERMAL VISION

A. Tracking a Single Person

Our system uses a particle filter to provide an efficient solution to the estimation 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 $S_t = \{\boldsymbol{x}_t^i, w_t^i\},\ i = 1, \ldots, N$. 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 [5] starting with a uniform initial distribution. The resampling step was

Fig. 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).

implemented using the systematic resampling algorithm. The dynamic model used in the particle filter is a movement with constant velocity plus small random changes.

B. 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 ith 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 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 $\psi = e^{(-\rho g_{im})}$ and g_{im} expresses the amount of overlap between particle i and region m, which is multiplied by a factor ρ in the exponent of the penalty term. This solution is similar to the interaction model proposed by [8], 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.

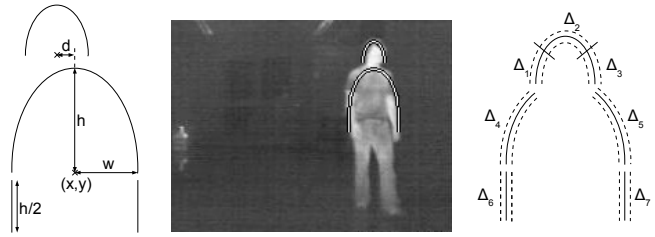

Fig. 2. The elliptic measurement model for thermal images. Model parameters are shown on the left. Division of ellipses into 7 regions is shown on the right.

C. 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 taken into account. 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 minimize 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:

$$p_g(\boldsymbol{z}_t|\boldsymbol{x}_t^i) = \frac{\exp(\kappa \cdot (f^i - \theta))}{\exp(\kappa \cdot (f^i - \theta)) + \exp(\kappa \cdot (\theta - f^i))}, \quad (1)$$

where θ denotes a fitness threshold and the value of κ defines the slope of the likelihood function.

When the mean gradient value from Eq. 1 is greater than 0.5 then a person is considered to be detected. We also check the uncertainty of the estimate [7] 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 [6] 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

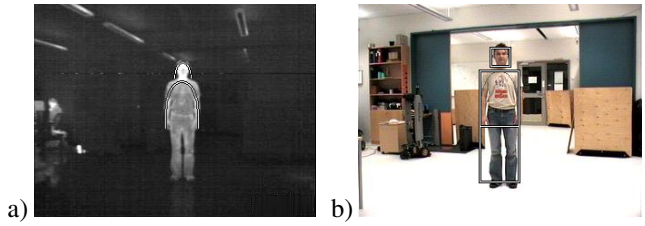

Fig. 3. Rectangular features: a) thermal image b) colour image with regions corresponding to different body parts from which colour information is extracted.

will be no recognisable head-shoulder contour. The elliptic contour model used here is able to cope with these situations.

IV. ADAPTIVE COLOUR MODEL

A. Colour representation

Since the baseline between 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 [11] 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., [12]) in the domain of image indexing. To include 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 is determined from the information provided by the elliptic contour model.

B. 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

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

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.

C. Rapid rectangular features

The simple features based on the colour moments can be rapidly calculated using an integral image representation [14]. 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:

$$S_r = \sum_{i=x}^{x+w} \sum_{j=y}^{y+h} I^r(i,j), \tag{3}$$

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

$$k_1 = S_1/n, \tag{4}$$

$$k_2 = \frac{nS_2 - S_1^2}{n(n-1)}, \tag{5}$$

$$k_3 = \frac{2S_1^3 - 3nS_1S_2 + n^2S_3}{n(n-1)(n-2)}, \tag{6}$$

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.

D. Combining thermal and colour information

If we assume that the likelihoods for the gradient model $p_g(\boldsymbol{z}_t|\boldsymbol{x}_t)$ (Eq. 1) and colour model $p_c(\boldsymbol{z}_t|\boldsymbol{x}_t)$ (Eq. 2) are independent then the data fusion can be realised by taking a product of these two likelihoods

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t) = p_g(\boldsymbol{z}_t|\boldsymbol{x}_t)p_c(\boldsymbol{z}_t|\boldsymbol{x}_t). \tag{7}$$

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:

$$w_t^i = p_g(\boldsymbol{z}_t|\boldsymbol{x}_t^i)p_c(\boldsymbol{z}_t|\boldsymbol{x}_t^i)\psi, \; i = 1, \ldots, N. \tag{8}$$

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

V. 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. 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 order of overlapping persons in the image, from which we have chosen a set of three thermal and three colour features. The first feature chosen is 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 front. This feature is, however, very noisy and affected by many factors such as movement of the camera, ambient temperature, 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.

We use the AdaBoost (Adaptive Boosting) classification algorithm [4] 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 $H(x) = sign(f(x))$ as $f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$, 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 [14] 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 [14] we use simple weak classifiers based on a single-valued feature $f_j(x)$

$$h_j(x) = \begin{cases} 1: & p_j f_j(x) < p_j \theta_j \\ 0: & \text{otherwise,} \end{cases} \tag{9}$$

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.

VI. 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 III), 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 $\rho_o < \rho$. Interaction with other filters (non-overlapping with this pair) remains unchanged. When the head contour of a person becomes occluded the corresponding tracker is considered to be totally occluded. This means that we can only 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.

VII. EXPERIMENTS

A. Evaluation

Our system was tested on the data collected by the robot during several runs. In total we collected 11 tracks using corridor following and 42 tracks with a stationary robot. In total we obtained 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 [3]. 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 $2 \cdot width$ and $3.5 \cdot height$ of the elliptic contour

	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 I

DETECTION AND LOCALISATION METRICS.

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 the level of false alarms, and accuracy is a combination of both recall and precision (see Table I). These metrics allow thorough testing of the properties and performance of the tracker as in [3] and [10].

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.

B. Training of the AdaBoost classifier

We extracted the described thermal and colour features from the collected data. We considered only cases 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, we took only those examples that preceded the moment of the total occlusion. 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.

C. Results

Fig. 4 shows the tracking performance using only thermal gradient information, with additional colour information, and with both colour information and explicit occlusion handling. Each experiment was repeated 10 times with different

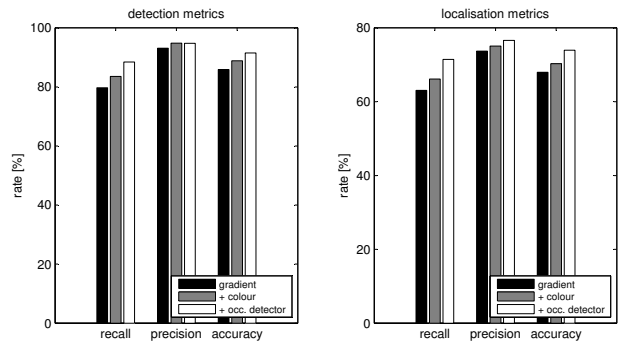

Fig. 4. Detection and localisation metrics for tracking multiple persons without and with colour information and with occlusion handling procedure.

Fig. 5. 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.

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. The overall accuracy (84.2% in detection and 68.7% in localisation) however 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. The output from the tracker can be seen in Fig. 5.

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

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

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

TABLE II

CLASSIFICATION RESULTS FOR DIFFERENT FEATURE TYPES.

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

TABLE III

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.

these features had the highest importance. 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.

Table III 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 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.

VIII. CONCLUSIONS AND FUTURE WORK

From the viewpoint of a typical service robot, 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. 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] L. Brèthes, P. Menezes, F. Lerasle, and J. Hayet, "Face tracking and hand gesture recognition for human-robot interaction," in *Proc. IEEE ICRA*, New Orleans, LA, USA, 2004, pp. 1901–1906.

[2] G. Cielniak, "People tracking by mobile robots using thermal and colour vision," Ph.D. dissertation, Örebro University, April 2007.

[3] D. S. Doermann and D. Mihalcik, "Tools and techniques for video performance evaluation," in *Proc. ICPR*, vol. 4, Barcelona, Spain, 2000, pp. 4167–4170.

[4] Y. Freund and 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.

[5] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. Inst. Elect. Eng. F*, vol. 140, no. 2, pp. 107–113, April 1993.

[6] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

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

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

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

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

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

[12] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[13] A. Treptow, G. Cielniak, and T. Duckett, "Real-time people tracking for mobile robots using thermal vision," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 729–739, 2006.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE CVPR*, 2001.

[15] T. Wilhelm, H. J. Böhme, and 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.

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