Real-Time Monocular People Tracking by Sequential Monte-Carlo Filtering

Christian A.L. Waechter
Fachgebiet Augmented Reality
Boltzmannstr. 3
85748 Garching b. München
waechter@cs.tum.edu

Daniel Pustka
Fachgebiet Augmented Reality
Boltzmannstr. 3
85748 Garching b. München

Gudrun J. Klinker
München
Boltzmannstr. 3
85748 Garching b. München
klinker@cs.tum.edu

ABSTRACT

We present a solution to the people tracking problem using a monocular vision approach from a bird's eye view and Sequential Monte-Carlo Filtering. Each tracked human is represented by an individual Particle Filter using spheroids as a three-dimensional approximation to the shape of the upstanding human body. We use the bearings-only model as the state update function for the particles. Our measurement likelihood function to estimate the probability of each particle is imitating the image formation process. This involves also partial occlusion by dynamic movements from other humans within neighbored areas. Due to algorithmic optimization the system is real-time capable and therefore not only limited to surveillance or human motion analysis. It could rather be used for Human-Computer-Interaction (HCI) and indoor location. To demonstrate this capabilities we evaluated the accuracy of the system and show the robustness in different levels of difficulty.

Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Sensor Fusion, Tracking—*complexity measures, performance measures*; I.5.4 [**Pattern Recognition**]: Applications, Implementation—*Computer Vision, Interactive Systems*

General Terms

Algorithms

Keywords

People Tracking, Realtime, Sensor Fusion, Particle Filter

1. INTRODUCTION

It is still a challenge to track a person in real-time over a longer period of time in the presence of several other people – and especially when there are no restrictions on the behavior of humans regarding interaction, entering & leaving the scenery or type of movements, etc. Such movements can frequently cause partial or full occlusions of persons. These are a major problem in people tracking as they result in swapping or loss of identities. In the fields of human-computer interaction, security, entertainment or motion analysis the tracking of individual persons is essential for higher level applications. Therefore objects should be trackable independently from each other.

We present an approach to this task using a single overhead camera that is mounted at the ceiling of the room. To account for the frequent partial occlusions between persons in this bird's eye view and to provide continued identification despite such occlusions, our system uses a three-dimensional Cartesian space as its underlying model of the observed scenery and uses individual Sequential Monte Carlo Filtering to maintain a large number of parallel hypotheses about human positions and to reason about occlusions within the scenery. The system uses separate particle filters to track different persons, and it uses the three-dimensional model and an image formation model to generate synthetic images according to the hypothesized human positions of individual particles. The model is the basis for the measurement likelihood function of the particle filter that compares camera images, processed by simple background subtraction and ramp-thresholding, with artificial views of the scenery generated for each hypothesis for each human. In doing so, it focuses on image subareas according to a hypothesized human position and takes recent positions of other tracked persons into account in order to explicitly discard borderline image regions that bear the potential of belonging to another person – thereby reducing the risk of unwanted fusion or swapping of identities. The system assumes that the subjects move on a planar floor. Each subject that enters the observed area is recognized by the system and represented by a particle filter. Since prediction of human walking behavior is unfeasible and can only be given under high uncertainty we apply a standardized, linear movement model to spread the hypotheses of the particle filters around the positions of humans. A rough appearance model is used to approximate the peoples' shape and to estimate an individual measurement likelihood of the particle filter. Furthermore, we apply algorithmic optimizations to lower the overall computational cost for all particles of each filter.

MIRAGE '13, June 06 - 07 2013 Berlin, Germany

ACM 978-1-4503-2023-8/13/06 ...$15.00.

2. RELATED WORK

2.1 Monocular Solutions

Isard *et al*[8] showed a single-camera real-time surveillance system that uses the so called Condensation Algorithm from Isard and Blake [7] to estimate a global likelihood of the belief about all available objects. For the tracking of a single person 500 particles are used and eight parameters, two for position and six for appearance, are used to describe the state of each object. In contrast, our approach uses a smaller state description of four positional and velocity parameters plus three globally fixed shape parameters. This lowers the required number of particles to represent the whole scenery.

Smith *et al*[11] also present a single camera approach to track multiple persons by using a pixel-wise binary distinction between fore- and background of the image and a color model of the foreground objects to assign the extracted information to the objects. A single Trans Dimensional Particle Filter represents the states of all objects and reasons about them using interaction potentials. These especially are important to distinguish between people crossing each other. The approach still shows swapping of identities which should be avoided. We avoid the problem of swapping the identity of passing objects by reasoning about object states in a three-dimensional representation of the scene rather than in the two-dimensional image space.

Zhao and Nevatia [13] showed a promising monocular approach toward people tracking in crowded scenes. They use three-dimensional representations of humans consisting of three ellipsoids for the head body and the legs, and color histograms in combination with a mean-shift algorithm to estimate correspondences of features to targets. In contrast to their joint likelihood that uses color information our method uses separate particle filters for every object and handles occluding interactions between objects by an image mask that integrates possible occlusions before estimating the likelihood of an object. This results in a single likelihood of one objects, reducing computational costs. We also show our method to differentiate between objects of similar color, as e.g. common in office environments with many persons wearing dark suits.

2.2 Stereo Vision / Multi-View Solutions

Many people tracking systems rely on stereo vision or multi-view camera systems to directly obtain 3D data of the objects that are being tracked. Most systems rely on these inter-camera relationships to use epipolar geometry or triangulation. Thus, the algorithms cannot be transferred to monocular vision approaches.

The system of Du *et al*[3] uses individual particle filters to track subjects that have been chosen manually in advance. For each attached camera view, a particle filter for each object is used to reason about the principal axes of the objects. The trajectories of the individual objects are tracked on the ground plane, fusing the principal axes for each object across all views.

A Bayesian multiple camera tracking approach is given by Qu *et al*[10] to avoid computational complexity. Their collaborative multiple-camera model uses epipolar geometry without using 3D coordinates of the targets.

Heath and Guibas [6] present an approach to people tracking using a particle filter system that represent single objects. Their approach uses multiple stereo sensors to individually estimate the 3D trajectories of salient feature points on moving objects.

Fleuret *et al*[4] use a probabilistic occupancy map that provides robust estimation of the occupancy on the ground plane. They apply a global optimization to all the trajectories of each detected individuum over a certain number of time frames. They showed reliable tracking for up to six people using a four camera setup. Yet, the solution suffers from a four-second lag that is needed to estimate the data is robustly. It is thus impossible to use the system for real-time human-computer interaction.

The method of Osawa *et al*[9] uses a three-dimensional representation of humans to track them in a cluttered office environment. Their concept of using a likelihood function inspired the approach that we introduce in this paper. In contrast to the stereo vision approach, our solution is a monocular camera system, mounted at the ceiling of a room. It uses intensity values instead of binary images to better integrate the slight illumination differences of a person from the background subtraction, making it more robust against image noise. The bird's eye view is optimal for a single camera system, since a complete occlusion of one person by another is unlikely to happen in normal office environments. This also enables tracking more than two objects at the same time with partial occlusion.

3. PARTICLE FILTER FOR MULTI HUMAN STATE ESTIMATION

We use a Sequential Monte Carlo (SMC) Filter technique [2] to estimate the state of multiple tracked people. The technique consists of two procedures to compute the prior and posterior distribution functions. These two procedures, which are also often referred to as the *motion* and *observation model*, are used to predict the state of an object and to validate this prediction. In contrast to other Bayesian filter mechanisms, (e.g. the Kalman filter) the SMC Filter has superior ability in maintaining a discrete number of distinct hypotheses of an object state. It provides the means to reason about the position of an object in complex situations, e.g., when being partially occluded or when emerging in a camera's field of view.

3.1 Human States and Dynamics

The state of an observed 3D scene is represented at each time step t by a set of I particle filters $M_t = \{P_t^1, \ldots, P_t^I\}$, where I denotes the number of currently tracked persons. Each particle filter P_t^i is a belief about a person's state and can be approximated by X_t^i, the weighted average of the m samples $\{\mathrm{x}_t^{i,j}, w_t^{i,j}\}$, with $j = 1, \ldots, m$. The weight $w_t^{i,j}$ is the calculated probability of a hypothesis $\mathrm{x}_t^{i,j}$. This design allows for easy parallelization of the particle filters as they can be processed mostly independently from each other: particles from a particle filter i that are evaluated at time step t only depend on the average states $X_{t-1}^{i'}$ of the other particle filters $i' \neq i$ at time step $t-1$. Under real-time conditions, the difference in movement is small enough. Subsequently, we refer to the particle filters of individual persons i without explicitly using the index i for each filter.

The description of the humans is kept simple in order to avoid high dimensionality. For a minimal configuration of the state we take a hypothetical x- and y-position of the person and the velocities in x and y-direction. The state

can then be described as the quadruple $\mathrm{x}_t^j = \langle x_t^j, y_t^j, \dot{x}_t^j, \dot{y}_t^j \rangle$. The transition of $p(\mathrm{x}_t^j|\mathrm{x}_{t-1}^j)$ is described by a first order linear system $p(\mathrm{x}_t^j|\mathrm{x}_{t-1}^j) = \Phi \mathrm{x}_{t-1}^j + \Gamma \mathrm{w}_t^j$ with w_t^j as a zero mean additive Gaussian noise $\mathcal{N}(0, \mu)$ for the position and a zero mean additive Gaussian noise $\mathcal{N}(0, \eta)$ for the change in velocities. Φ is a 4×4 transition matrix and Γ a 4×2 matrix. Since we use a well-known state description and dynamics further details are also explained by Chang and Bar-Shalom's JPDA approach [1].

3.2 Observation Model

The perceptive model is based on a projective pin-hole model similar to the process of picture generation within a normal camera. To evaluate the sample weight we generate an artificial view of the scenery from the viewpoint of the camera and compare this to the background-subtracted and ramp-thresholded image taken at the same time step. The resulting images z_t are used to estimate the likelihood of a sample as $p(z_t|\mathrm{x}_t^j)$.

3.2.1 Shape Description

We apply a three-dimensional shape model in Cartesian coordinates to all detected persons in the observed area. This approach minimizes the computational costs but applies only to humans of similar shape. As a simple three-dimensional shape description we use spheroids, prolate ellipsoids where the two minor semi-axes are equal. In our case the major semi-axis a is set to 0.9 meters and the minor semi-axes b and c are set to 0.25 as we track persons of an approximate height of 1.8 meters. The height of the center is also set to 0.9 such that the south pole of the spheroids is on the ground level. The three dimensional model can be described by a quadric Q, modeled as a 4×4 matrix. This quadric can be projected onto the image plane by multiplying Q with the camera-specific 3×4 projection matrix P. This process is illustrated in figure 1 and details are given by Hartley and Zisserman [5]. The resulting 3×3 matrix C^* describes the spheroid on the image plane as a line-conic, the dual to the point-conic. We use the inverse of matrix C^* to calculate the upper and lower bounds of the conic on the image plane. In between these two borders we determine for each y value, the corresponding x-values as the left and right border of the conic at the specific y.

3.2.2 Measurement Likelihood Function

The particle filter described by [9] uses the following measurement function to evaluate the likelihood of a generated virtual scene :

$$C_j(V_{t,j}) \;=\; \frac{\sum_{x,y} B_t(x,y) \cap V_j(x,y)}{\sum_{x,y} B_t(x,y) \cup V_j(x,y)} \qquad (1)$$

where B_t is the thresholded background-subtracted image and $V_{t,j}$ is a virtual binary image, generated for the particle filter hypothesis j. Unfortunately, the hard threshold approach has the drawback of being very sensitive to lighting and noise. A small illumination change can quickly make a person undetectable.

To improve the detection quality, we replaced the set operations on binary images by arithmetic operations on grayscale images. The intersection of binary regions was replaced by the product of grayscale intensity values, and the union operation by the maximum of two intensity values. The new evaluation function is shown in equation 2.

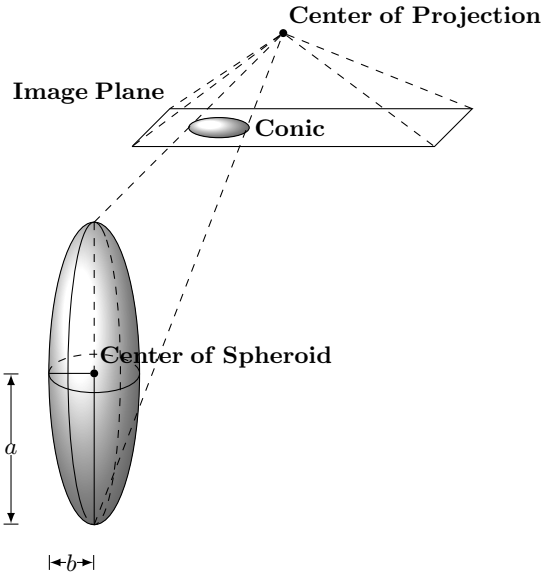


Figure 1: The projection of the spheroid appears as a conic on the image plane. The major semi-axis of the spheroid is marked as a, the minor semi-axes are equal so only b is shown.

$$C_j(V_{t,j}) \;=\; \frac{\sum_{x,y} B_t(x,y) V_j(x,y)}{\sum_{x,y} \max(B_t(x,y), V_j(x,y))} \qquad (2)$$

Instead of the hard threshold operation on the background-subtracted image, a softer ramp-mapping of intensity values is used.

Region of Evaluation (RoE)

To reduce the influence of image noise and of information from distant parts of the image, the evaluation of the measurement likelihood function is restricted to a local region of the background subtracted image (yellow and pale blue regions in Fig. 2b).

The shape of the RoE is a function of the three-dimensional shape description in section 3.2.1. The RoE is a conic corresponding to the projection of an enlarged version of the spheroid of the current particle. The semi-axes of the enlarged spheroid are scaled by a value α_e. The height h_e of the center is adjusted such that the larger spheroid touches the ground. The x- and y-positions are unchanged. We estimated reasonable values of $\alpha_e \approx 2.0$ and $h_e = \frac{1}{2} \cdot a$ for defining the RoE.

For a robust estimation of the particle weight, it is essential to take the potential presence of other nearby persons into account. The shape of other persons i', according to their last average position $\mathrm{x}_{t-1}^{i'}$, is masked out in the current RoE since this pixel information is ambiguous. This makes the evaluation more robust and reduces mixups between persons occluding each other. This RoE also increases the speed of the algorithm as the computational costs are lowered extremely in comparison to a complete image analysis.

$$C_{t,j}(V_{t,j}) = \frac{\sum_y (S(e_r(y), y) - S(e_l(y), y))}{\sum_y (S(e_l(y), y) - S(0, y) + e_r(y) - e_l(y) + S(x_{\max}, y) - S(e_r(y), y))} \quad (3)$$

Performance Optimizations

The computation of equation 2 can be simplified by assuming that the virtual image $V_{t,j}$ is a binary image, i.e. $V_{t,j}(x, y) \in \{0, 1\}$. In this case, the numerator is the sum of all pixels of $B_t(x, y)$ where $V_{t,j}(x, y) = 1$. The denominator is just the number of pixels where $V_{t,j}(x, y) = 1$.

As the virtual image contains exactly one convex ellipsoid, we can represent $V_{t,j}$ by just describing the left and right ellipsoid edges $e(y) = (e_l(y), e_r(y))$ for each line y of the image. Also the repeated evaluation of pixel sums in the image can be improved by using a pre-computed sum image $S_{t,j}(x, y) = \sum_{i=0}^{x} B_{t,j}(i, y)$. This optimization makes sense, as typically 300 hypotheses are evaluated per image and person. The optimized likelihood computation is expressed in equation 3.

Reasoning in 3D

As our people tracking maintains the position of the humans in 3D we can use this knowledge to reason about the particle weights. If the Euclidean distance of particle $\mathrm{x}_t^{k,j}$ of person k to the average position x_{t-1}^{l} of another person l is less than twice the length of the semi-axes (b or c) of their spheroid, the particle weight is set to zero. As we set the semi-axes b and c to 0.25 meters, this makes a minimum distance of 0.5 meters between the centers of the spheroids. If prior 3D knowledge of the environment is available, this information can be used to reason further about particle weights.

3.3 Instantiation and Deletion of Particle Filters

To robustly estimate the positions of the people within the scenery, the system needs to know about every human entering or leaving the observed area. These mechanisms are kept simple and are briefly introduced here.

Instantiation

At spaced time intervals, the system recognizes new objects within the scenery by a blob detection algorithm on the background subtracted images. Blobs that pass some early checks on constraints, *e.g.*object size, are used for instantiation. But only if no already tracked person is near the blob a new Particle Filter will be instantiated at the Blobs rough position in respect to the ground plane.

Deletion

The covariance of the particles can be interpreted as the belief of a person's state with respect to the mean value. If samples of a particle filter are widely distributed in the environment, this is taken as an indicator that the filter does not represent the person any more, and the filter is deleted. This mechanism cares for objects that are either occluded or have left the scenery. It results in the deletion of the associated particle filter.

4. EXPERIMENTAL RESULTS

4.1 Experiment Setup

We have evaluated our tracking system using a single camera (FOV ≈ 35, framerate $\approx 15Hz$, resolution $= 640 \times 480$) mounted at the ceiling in our offices at a height of ca. 3.60 meters. The camera's sensor plane is nearly parallel to the observed ground floor (3.60×4.70 meters). The camera images were directly undistorted using Zhangs method [12], such that the conics could be projected onto the image plane without an additional correction for radial distortion. The people tracking system ran on a standard Intel Core 2 Duo platform, running at 2.5 Ghz, all images could be processed without loss.

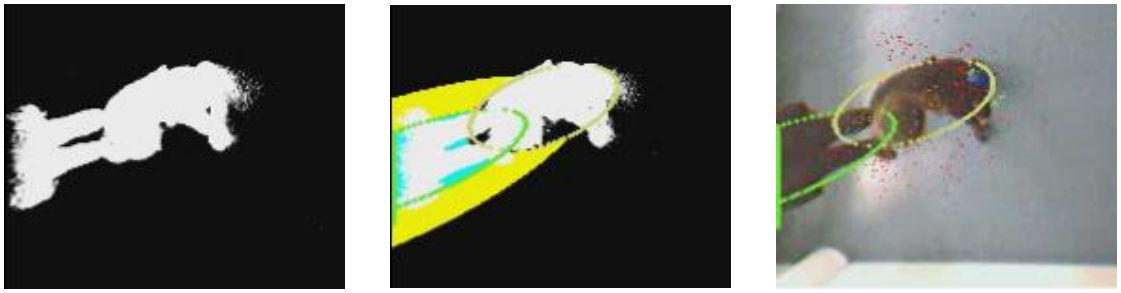
4.2 Accuracy

We carried out three different experiments in order to evaluate the accuracy of our tracking system. For all experiments we measured various positions of different difficulty for the tracking system within the observed area in advance. By placing a marker on the ground plane we could measure the various positions in respect to the common coordinate system of the ground plane, also defined by placing a marker. These positions were than marked on the floor and given different labels ranging from A1 to E2. See figure 3 for the various locations and visibility of the person. The position starting with the same latter were used for estimation under occlusion, were the 1 indicates the outer position, were a person is generally worse visible.

In the first experiment a persons had to step at the ten marked positions such that the spot was roughly centered under the point of highest gravity of the person. We than recorded the position of the person at this spot for a period of 30 seconds ($\sim$ 450 frames) without any intentional movement of the person. From this data we calculated the mean error and the standard deviation of our tracking system, which is listed in table 1 in the upper row. The overall archived error of the people tracking system under no movement can be stated as 0.09 ± 0.04 meters and demonstrate the capabilities this monocular tracking system.

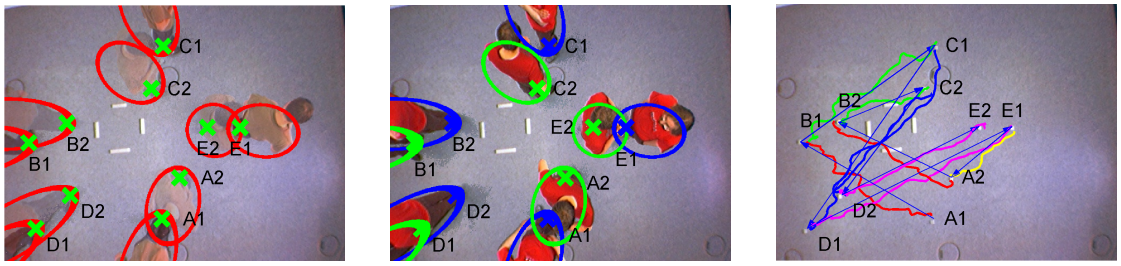
In a second experiment we recorded the positions of two persons standing near to each other. The persons again positioned themselves over the marked areas and their location was recorded. This time we were especially interested in the position accuracy of the occluded person. And estimated the position error for this constellation. As it can be seen from the results the position of the person which is occluded from the other person has less accuracy. The final result with an error of 0.15 ± 0.09 meters for the outer person and 0.12 ± 0.09 meters for the inner person is little higher than for a single person but still within the range of a human bodies circumference. With this error our system is also in the range of Fleuret's *et al*[4] multi-camera system, where they used a slightly different evaluation method.

In the final experiment we estimated the tracking accuracy of the system for a moving person. The person was advised to walk from one marker to the another in a direct way. First the subject had to visit all outer marker



(a) Background Subtraction (b) Region of Evaluation (c) Particles

Figure 2: Tracking result dealing with partial occlusion between two person.Subfigure 2(a) shows the background subtracted and thresholded image that is used as an input. Subfigure 2(b) shows the yellow region of evaluation of one person. Note that pixels corresponding to the last position of the other person are masked out within this region. Subfigure 2(c) shows the original camera output with colored ellipses for the detected persons and the particles for each persons, blue particles represent a high weight and red is for low weights.



(a) One Person (b) Two Persons (c) Trajectory

Figure 3: The different evaluation positions on the floor are marked with crosses. One can easily observe the different visibility levels of the person at these positions.

starting at $A1$ finally ending up in $E2$. So the order was $A1 \rightarrow B1 \rightarrow \ldots \rightarrow E1 \rightarrow A2 \rightarrow \ldots \rightarrow E2$. The trajectory of this movement was recorded and can be seen in figure 3(c). At each timestep we estimated the error by constructing the perpendicular line from the measured position to the direct path and calculating the euclidean distance of this perpendicular line. The results can be seen in table 2. Interestingly the error 0.0813 ± 0.0712 meters from the moving person is similar to the error from no movements at all Error. (See table 2)

4.3 Robustness

To evaluate the robustness of our tracking system we recorded three sequences with 5 difficult situations each. We paid attention that the persons were similar dressed (red top and brown trousers) to show the advantage of our method compared to color based methods, which we expect to fail in the same task. We asked three experts, not belonging to the authorship of this paper, to judge about the general level of difficulty of these events for a people tracking system. The results can be seen in table 3. All sequences rated as 1 or 2 have been processed without problems but one. In one sequence, rated as 2, the identities of two persons were swapped. We believe that this happened from our simple background subtraction method that does not respect reflections and shadows on the ground plane, which were visible on the floor here. Three of the hardest and second hardest sequences could be solved without problems. At two of these sequences (one rated as 4, the other as 5) the tracking dropped one person due to heavy occlusion. Later on, the person was detected again and given a new id, when the person was well observable again. In one of the medium sequences the tracking performed well. Especially in the case where one hardly visible person was not tracked at all, but at the same time it did not confuse the tracking of the other person. We provided two additional videos for the experts, each one with one longer difficult situation. In these sequences all tracked persons wore black suits, similar to common office environments. The first video with five persons was always rated as 4, the second video was two times rated 4 and once 5. There was no problem in the first sequence but in the second sequence the identities of two persons were swapped at the end, when one person sneaked through the other two persons.

	Sequence 1				Sequence 2				Sequence 3			
Expert	1	2	3	∅	1	2	3	∅	1	2	3	∅
Situation 1	3	2	1	**2**	5	5	5	**5**	4	4	3	**4**
Situation 2	2	3	1	**2**	5	5	4	**5**	5	5	4	**5**
Situation 3	3	1	2	**2**	2	5	2	**3**	4	3	2	**3**
Situation 4	1	1	1	**1**	3	5	2	**3**	4	4.5	1	**3**
Situation 5	1	2	1	**1**	3	2	2	**2**	5	3.5	4	**4**

Table 3: Levels of difficulty of the sequences. The mean values are already rounded to the nearest integer value.

Position	A1	B1	C1	D1	E1	A2	B2	C2	D2	E2
Mean [cm]	5	11	8	13	7	8	10	10	5	10
Std. Dev.	± 2	± 3	± 2	± 4	± 3	± 3	± 6	± 2	± 3	± 2
Mean [cm]	8	16	8	31	13	8	26	5	18	5
Std. Dev.	± 4	± 4	± 4	± 2	± 3	± 2	± 5	± 2	± 5	± 2

Table 1: The upper row shows the accuracy of the tracking for a single person being tracked at the different locations. The lower row shows the accuracy at the different locations while two persons are being tracked. At position D1 the occluded person could only be tracked for 108 frames. The person was then deleted from the system due to heavy occlusion.

Position	A1B1	B1C1	C1D1	D1E1	E1A2	A2B2	B2C2	C2D2	D2E2
Mean [cm]	10	7	15	7	7	11	5	9	4
Std. Dev.	± 8	± 5	± 13	± 4	± 3	± 4	± 3	± 4	± 3

Table 2: The table shows the error from the estimated trajectory to the shortest path from one location to the other.

We and one expert also noticed a major advantage of our tracking approach in case of hardly visible persons. It is capable of tracking a previously detected person robustly even if only small parts of the body, e.g. the shoes, are visible to the camera.

In case of one person occluding another completely there is no reliable forecast which identity will be lost as the particle filters are non optimal and deterministic in their computation of the position. But it can be assumed that most likely the identity of the occluding person will remain as there is still more image information from that person visible within the camera. This behavior could also be seen in our two cases mentioned earlier.

5. CONCLUSION

Our system concentrates on imitating the processes of picture generation within a camera to reliably estimate the likelihood of hypotheses in an intuitive way. With the projective approach introduced in this paper as much information as possible is kept until the likelihood of the measurement is calculated, as we interpret a complete image as a measurement and avoid the early extraction of information which can result in errors at a much prior stage in the data acquisition. The additional optimization step we introduce allows to use multiple Particle Filters with many hypotheses at the same time, meeting real-time conditions. Rather than optimizing the prior distribution function this seems to be promising as the prediction of human's movements are always hard to model in an optimal way even in certain environments when knowledge about human movements can be retrieved (e.g. sports arena).

In general it is possible to track more than five persons as the computational costs rise linearly with the number of persons, such that the system can easily be scaled to an arbitrary number of persons. This results from the fact that we do not compute a joint likelihood for the states of all particles. As long as the objects to track can be approximated by simple geometric shapes, as in our case a spheroid, the computational costs do not exceed the performance of present end user systems (see section 4). This is important for satisfying real-time requirements, e.g. with respect to HCI and Augmented Reality applications.

6. REFERENCES

[1] K. Chang and Y. Bar-Shalom. Joint probabilistic data association for multitarget tracking with possibly unresolved measurements and maneuvers. *Automatic Control, IEEE Transactions on*, 29(7):585–594, 1984.

[2] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[3] W. Du, J. Hayet, J. Verly, and J. Piater. Ground-Target Tracking in Multiple Cameras Using Collaborative Particle Filters and Principal Axis-Based Integration. *IPSJ Transactions on Computer Vision and Applications*, 1(0):58–71, 2009.

[4] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, February 2008.

[5] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[6] K. Heath and L. Guibas. Multi-person tracking from sparse 3D trajectories in a camera sensor network. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–9, 2008.

[7] M. Isard and A. Blake. CondensationŮconditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28, 1998.

[8] M. Isard, J. MacCormick, C. Center, and P. Alto. BraMBLe: A Bayesian multiple-blob tracker. In *Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings*, volume 2, 2001.

[9] T. Osawa, X. Wu, K. Wakabayashi, and T. Yasuno. Human tracking by particle filtering using full 3d model of both target and environment. In *Proceedings of the 18th International Conference on Pattern Recognition-Volume 02*, pages 25–28. IEEE Computer Society Washington, DC, USA, 2006.

[10] W. Qu. Distributed Bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. *EURASIP Journal on Advances in Signal Processing*, 2007:1–15, 2007.

[11] K. Smith, D. Gatica-Perez, and J. Odobez. Using particles to track varying numbers of interacting people. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 1, 2005.

[12] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330–1334, 2000.

[13] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, 2004.