T U M

I N S T I T U T F Ü R I N F O R M A T I K

A Perspective Approach Toward Monocular Particle-based People Tracking

Christian Waechter Daniel Pustka Gudrun Klinker

TUM-I0923
August 09

T E C H N I S C H E U N I V E R S I T Ä T M Ü N C H E N

TUM-INFO-08-I0923-100/1.-FI

Druck: Institut für Informatik der
Technischen Universität München

A Perspective Approach Toward Monocular Particle-based People Tracking

Christian A. L. Waechter
Daniel Pustka
Gudrun J. Klinker

Fachgebiet Augmented Reality
Technische Universität München
Germany

August 31, 2009

We present a solution to the people tracking problem using a monocular vision approach from a bird's eye view and three-dimensional human representations maintained by Sequential Monte-Carlo Filtering. The system robustly estimates the image measurements' likelihoods imitating the image formation process and respecting partial occlusion by dynamic object movements. Due to algorithmic optimization the system is real-time capable and can be used for Human-Computer-Interaction (HCI) in combination with other sensors. The approach can also be extended to multi-view camera systems.

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 instead of a specialized adapted human motion model. 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.

2 Related Work

Various solutions to the problem of tracking multiple targets exist. Recent approaches mostly use particle filters for the general task, but differ in detail. The ability of particle filters to maintain many different hypotheses about possible object states are an advantage in handling occlusions in camera images or ambiguity in measurement data in general. Most recent solutions to the problem use a stereo vision approach to extract

features and use them as measurements in the posterior distribution function – but some single camera approaches exist as well. These are introduced next.

2.1 Monocular Solutions

Isard *et al.*[7] showed a single-camera real-time surveillance system that distinguishes between foreground and background objects. They use the so called Condensation Algorithm from Isard and Blake [6] to estimate the global likelihood of the belief about all available objects, such that it integrates various different hypotheses about different numbers of objects and their states. For the tracking of a single person 500 particles are used and the number of particles increases with additional objects that are being tracked. 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.*[10] 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 [11] 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 (BMCT) approach is given by Qu *et al.*[9] to avoid computational complexity. Their collaborative multiple-camera model uses epipolar geometry without using 3D coordinates of the targets. The additional cameras are activated in case of close proximity or present occlusion in a particular camera. They were able to track multiple targets in a crowded station environment using two cameras for tracking.

Heath and Guibas [5] 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.*[8] 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 $M_t = \left\{P_t^1, \ldots, P_t^I\right\}$, 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 $\left\{\mathrm{x}_t^{i,j}, w_t^{i,j}\right\}$, 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 $\mathrm{x}_t^j = \langle x_t^j, y_t^j, \dot{x}_t^j, \dot{y}_t^j \rangle$, an already well-known state description used by Chang and Bar-Shalom's JPDA [1]

The prior distribution function is used to estimate the probable position of each human. The particle state x_t^j is updated by its probable position $p(\mathrm{x}_t^j | \mathrm{x}_{t-1}^j)$ depending on the last time step $t-1$. The transition of $p(\mathrm{x}_t^j | \mathrm{x}_{t-1}^j)$ is described by a 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. The matrices for Φ and Γ are as follows:

$$\Phi = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

$$\Gamma = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2}$$

The functions for μ and η include a scale factor α_v to allow for a change in speed and direction of an object in a given time interval.

$$\mu = v_{t-1} \cdot \delta_t \tag{3}$$

$$\eta = \delta_t \cdot \alpha_v \tag{4}$$

The velocity at the last time step is v_{t-1} and δ_t denotes the time difference between the updates. We set $\alpha_v \in \{1.0, \ldots, 2.0\}$ which can be used to represent normal walking behavior of humans.

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. The description of a spheroid is similar to the ellipsoid by the quadric Q^*, a 4×4 matrix. The spheroid is described by the diagonal matrix D, a description of the unit ellipsoid and the homogeneous square-matrix Q with the lengths of the semi-axes a, b and c of the spheroid on the diagonal and the position of the center in the last column. Note that b and c are set equal as they are the two minor semi-axes of the spheroid. Equation 5 describes the quadric in the corresponding matrix formulation:

$$Q^* = Q \cdot D \cdot Q^T \tag{5}$$

Using equation 5, the quadric can be projected onto the image plane by multiplying Q with the camera-specific 4×4 projection matrix P. The resulting formula 6 describes the spheroid on the image plane as a 3×3 matrix. This process is illustrated in figure 3.2.1.

$$C^* = (P \cdot Q) \cdot D \cdot (P \cdot Q)^T . \tag{6}$$

The inverse of matrix C^* is positive semi-definite, allowing for a test whether a point in image coordinates is inside the conic or not. For each discrete homogenized pixel $\hat{x}_p$ we can use

$$\hat{x}_p^T \cdot C^{*-1} \cdot \hat{x}_p \leq 0 \tag{7}$$

to determine this property. 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 0.2 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.

3.2.2 Measurement Likelihood Function

The particle filter described by [8] uses the following measurement function to evaluate the likelihood of a generated virtual scene :

$$C_j(V_{t,j}) \;=\; \frac{\sum_{x,y} B_t(x,y) \cap V_j(x,y)}{\sum_{x,y} B_t(x,y) \cup V_j(x,y)} \tag{8}$$

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.

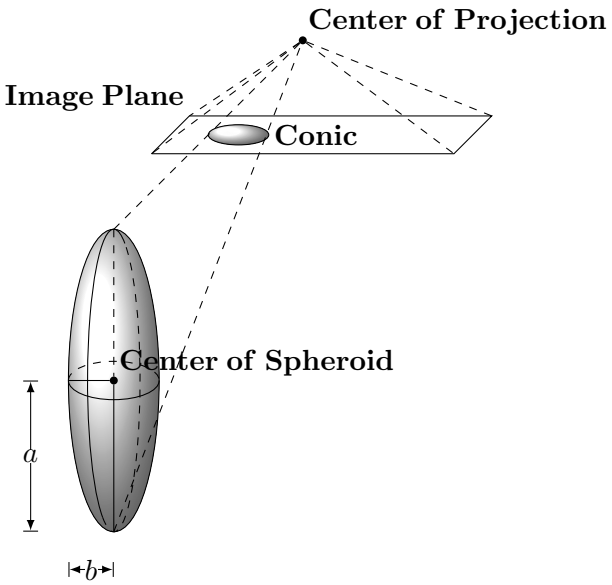


Figure 1: The projection of the spheroid appears as a conic on the image plane. The major semi-axis of the spheroid is marked as a, the minor semi-axes are equal so only b is shown.

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 9.

$$C_j(V_{t,j}) \;=\; \frac{\sum_{x,y} B_t(x,y) V_j(x,y)}{\sum_{x,y} \max(B_t(x,y), V_j(x,y))} \tag{9}$$

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.

Performance Optimizations The computation of equation 9 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 $\tilde{3}00$ hypotheses are evaluated per image and person. The optimized likelihood computation can now be expressed as:

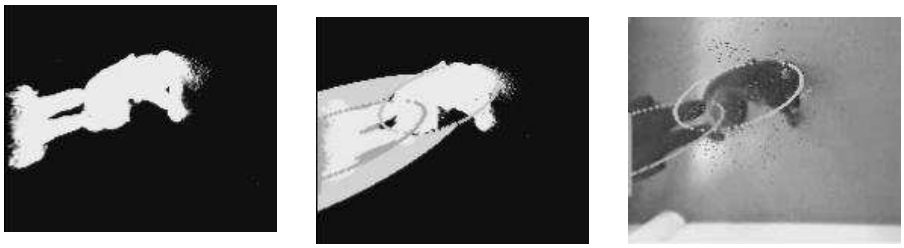
$$C_{t,j}(V_{t,j}) \;=\; \frac{\sum_y (S(e_r(y),y) - S(e_l(y),y))}{\sum_y (S(e_l(y),y) - S(0,y) + e_r(y) - e_l(y) + S(x_{\max},y) - S(e_r(y),y))} \quad (10)$$

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 0.2 meters, this makes a minimum distance of 0.4 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. The blob is approximated by an ellipse. The intersection of a line from the projection center of the camera through the blob center with the ellipse is used as a new x-, $y-$position. Objects with a distance to the previously detected objects that is greater as a certain threshold are considered as new and a homography maps the point coordinates on the image plane to the corresponding point on the x-,$y-$ ground plane of the three-dimensional environment map. The design of the instantiation allows to automatically detect new



(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.

objects and also to reinitialize objects that had been lost due to (nearly) total occlusion by other persons.

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 Evaluation

We have evaluated our tracking system using a single camera (FOV ≈ 35, framerate $\approx 15Hz$, resolution $= 320 \times 240$) mounted at the ceiling in a waiting area at a height of ca. 3.30 meters. The camera was mounted such that the camera plane was nearly parallel to the floor of the environment and 3.60×4.70 meters of the floor could be observed. The computation platform for the evaluation of the people tracking was an Intel Core 2 Duo, running at 2.5 Ghz. The images from the camera were undistorted, such that the conics could be projected into the image plane without adjustment. A background subtraction followed by a ramp-threshold was applied to process the image for the use of the likelihood function. The values for the threshold were determined by a histogram and were set to 30 and 60 for the lower and the higher threshold.

We applied our method in a single setup, yet with a varying number of persons. The persons were advised to wear black suits, similar to office environments, to show the advantage of our method compared to color based methods, which we expect to fail in

the same task. The persons were advised to enter the room one after the other since we cannot yet distinguish between persons that have not yet been recognized by the system.

The algorithm showed to be robust against swapping and loss of identities for two persons in the observed area. Identities were swapped or lost only when one person occluded the other one completely. The occluded person then was considered as a new person when it was redetected when the blobs were again far enough apart.

In tracking scenarios with three or more persons the system sometimes swapped or lost identities because situations with complete occlusions occurred more often. Yet the system still showed satisfying results in difficult situations. Our evaluation showed promising results of still differentiating between objects even when only little visual information was present. Especially, situations with three or more nearby persons in the very outer regions of the camera image resulted in reliable tracking as long as there was enough image information for the persons. An advantage of our tracking approach arises when only small pieces of a tracked person remain visible at the border of the image. If there is still intensity information, e.g. from the shoes, this information is enough to track the person if detected previously.

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.

The results were obtained at 15 Hz, but for 5 or more persons the preprocessed images were downsampled to half their size to still meet the real-time requirements. Otherwise the system would run at 7 Hz which is to slow for a usable prediction of the particles' positions.

5 Future Work

As future work we suggest using a statistical background detection algorithm to avoid problems with illumination and reflections in the scenery and shadows at walls or on the floor. The use of 3D knowledge of the environment would enable us to track persons also in a cluttered environment. In cases of pillars or other larger obstacles within a room we suggest making use of more hypotheses of a human's position than the single average of the particles. Instead of a global shape model for all humans, estimating the individual height and width of each human would enable the system to also track larger persons or small children. To overcome the problem of complete occlusions in the outer area we suggest a multi camera environment to observe areas where occlusions are most likely to appear. This will make the tracking more robust against loosing and swapping of identities. A multi camera system could also be used to observe a larger area if they are arranged in a way that the views overlap and gaps are avoided.

6 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.

References

[1] K.C. 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.B. 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] 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.

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

[7] M. Isard, J. MacCormick, C.S.R. 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.

[8] 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.

[9] 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.

[10] K. Smith, D. Gatica-Perez, and J.M. 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.

[11] 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.