Multi-View Object Tracking Using Sequential Belief Propagation

Wei Du and Justus Piater

University of Liege, Department of Electrical Engineering and Computer Science,
Institut Montefiore, B28, Sart Tilman Campus, B-4000 Liege, Belgium
weidu@montefiore.ulg.ac.be, justus.piater@ulg.ac.be

Abstract. Multiple cameras and collaboration between them make possible the integration of information available from multiple views and reduce the uncertainty due to occlusions. This paper presents a novel method for integrating and tracking multi-view observations using bidirectional belief propagation. The method is based on a fully connected graphical model where target states at different views are represented as different but correlated random variables, and image observations at a given view are only associated with the target states at the same view. The tracking processes at different views collaborate with each other by exchanging information using a message passing scheme, which largely avoids propagating wrong information. An efficient sequential belief propagation algorithm is adopted to perform the collaboration and to infer the multi-view target states. We demonstrate the effectiveness of our method on video-surveillance sequences.

1 Introduction

Visual tracking involves object detection and recursive inference of the target states in time. A popular approach is to generate target hypotheses and then to verify them by matching with a pre-learned reference model. However, a drawback of this approach is that if the target is occluded, a partial or – even worse – complete target observation is missing, making the comparison with the reference model impossible. The problem is particularly severe in the context of single-camera tracking in crowded scenes such as surveillance and team sports. However, the use of multiple cameras and collaboration between them make possible the integration of multi-view information and reduce the uncertainty due to occlusions.

A potential problem of conventional multi-view tracking is that wrong information may be integrated and propagated from one view to other views. To solve this problem, this paper presents a novel method for integrating and tracking multi-view observations using bidirectional belief propagation. The method is based on a dynamic graphical model where target states at different views are represented as different but correlated random variables, and image observations at one view are only associated

This work has been sponsored by the Région Wallonne under DGTRE/WIST contract 031/5439.

with the target states at the same view. As all views are correlated with each other, the graphical model is fully connected. The tracking processes at different views exchange information using a message passing scheme, which largely avoids propagating wrong information. Hua and Wu introduced an efficient Sequential Belief Propagation (SBP) algorithm to perform the multi-scale visual tracking [1]. In the present paper, we adapt the approach to our multi-view tracking task and our specific graphical model. In particular, we apply SBP to integrate individual trackers at different views so that the multi-view target states are inferred based on the multi-view observations.

Similar multi-camera tracking frameworks have been presented, e.g. in the context of video surveillance [2] or soccer player tracking [3]. Targets are tracked by individual trackers at different views, and the results are fused by a fusion module. To prevent wrong information integration, uncertainties of individual trackers are computed and used during fusion. However, with no interaction between individual trackers, the multi-view information is not fully exploited and the robustness of these systems is limited.

Particle filters are popular in multi-view tracking [4, 5]. Both cited approaches are based on the best-view-selection strategy: the target states are estimated using mainly those views that contain the most likely information. The problem is that the targets of interest may not be sufficiently distinctive from clutter and as a result, the wrong selection of the best view will cause the complete loss of tracks.

Different from previous work, our approach involves both recursive inference of target states using particle filters [6], making the system capable of coping with non-Gaussian clutter and non-linear dynamics, and exchanging information across views using belief propagation [7, 8], making the system robust to occlusions. Belief propagation provides a systematic solution for propagating uncertainties in a graphical model. The specific flavor of belief-propagation that we use, sequential belief propagation, enables us to reduce the risk of wrong information propagation. A fully connected graphical model for multi-view tracking is proposed based on a multi-view target state representation. We demonstrate the effectiveness of our method on video-surveillance sequences.

Section 2 describes the multi-view representation and the graphical models. Sequential Belief Propagation is introduced in Section 3. Section 4 introduces the SBP-based multi-view tracking algorithm. Results on sequences of video surveillance from PETS2001 datasets [9] are illustrated in Section 5.

2 A Graphical Model for Multi-View Tracking

The target state at each view is denoted by x_i, where $i = 1, \ldots, L$ is the view index. Putting all states at different views together results in a multi-view representation for the target, denoted by $X = \{x_1, \ldots, x_L\}$. The benefit of this representation is that the multi-view target model makes possible the integration of multi-view image observations, which helps overcome the occlusion problem if the target is not occluded in all views. The image observation associated with x_i in the same view is denoted by z_i, and $Z = \{z_1, \ldots, z_L\}$.

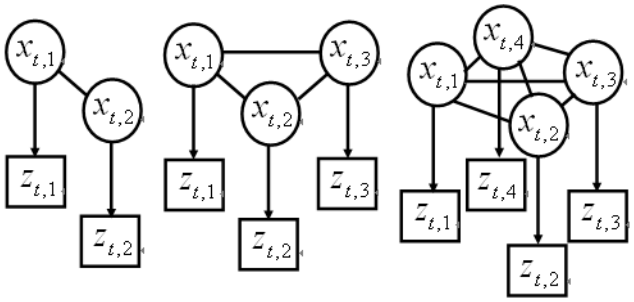


Fig. 1. Graphical models for the multi-view states at a time instant. From left to right, 2, 3 and 4 views are used respectively.

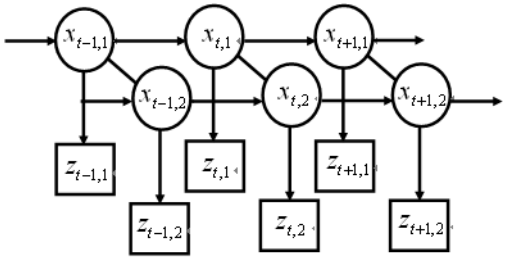


Fig. 2. Dynamic graphical model for the multi-view states. Only a two-view model is illustrated as an example.

Given the above definitions, our approach performs bidirectional belief propagation in a graphical model shown in Figure 1, and recursively infers the multi-view target states in a dynamic graphical model shown in Figure 2.

In both figures, the undirected link between $x_{t,i}$ and $x_{t,j}$ describes the mutual influence of multiple views and is associated with a potential function $\psi^t_{i,j}(x_{t,i}, x_{t,j})$, and the directed link from $x_{t,j}$ to $z_{t,j}$ represents the image observation process and is associated with an image likelihood function $p_j(z_{t,j}|x_{t,j})$. In Figure 2, the directed link from $x_{t-1,i}$ to $x_{t,i}$ represents the prior dynamics and is associated with a dynamic model $p(x_{t,j}|x_{t-1,j})$.

According to Bayes' rule, the recursive inference of the posterior distribution of the multi-view state $p(X_t|Z^t)$ is formulated as

$$\begin{aligned} \textbf{predict:}\ & P(X_t|Z^{t-1}) = \int P(X_t|X_{t-1})\, P(X_{t-1}|Z^{t-1})\, \mathrm{d}\, X_{t-1} \\ \textbf{update:}\ & P(X_t|Z^t) \quad \propto P(Z_t|X_t)\, P(X_t|Z^{t-1}) \end{aligned}$$

X_t is the multi-view state at time t, and $Z^t = \{Z_1, \ldots, Z_t\}$ are the image observations up to time t.

The inference of the joint multi-view state is difficult due to the lack of a closed-form solution. In practice, we infer the posterior of the single-view states $P(x_{t,j}|Z^t)$,

$j = 1, \ldots, L$. We show in the following sections how the inference is done using sequential belief propagation.

3 Sequential Belief Propagation

Sequential belief propagation, a non-parametric and sequential version of belief propagation, was first introduced by Hua and Wu [1]. We borrow the idea and apply it to our specific task.

The basic idea of the multi-view tracking algorithm is to calculate the inference of multi-view states through a message passing process. The local message passed from view i to view j in the graphical model in Figure 1 is

$$m_{ji}(x_{t,j}) \leftarrow \int \left(\prod_{k \in N(x_{t,i}) \backslash j} m_{ik}(x_{t,i}) \right) p_i(z_{t,i}|x_{t,i})\, \psi^t_{i,j}(x_{t,i}, x_{t,j})\, \mathrm{d}\, x_{t,i}, \quad (1)$$

where $N(x_{t,i})$ denotes the set of views connected to $x_{t,i}$ through an undirected link, and $N(x_{t,i})\backslash j$ means the neighboring views of $x_{t,i}$ except $x_{t,j}$. The first part of the right side of Equation 1 is the message that view i receives from its neighbors except j, the second part is the information of the image likelihood in view i, and the last part is the potential function mapping these information from view i to view j.

To infer $P(x_{t,j}|Z^t)$, $j = 1, \ldots, L$ based on the dynamic graphical model in Figure 2, we take into consideration the message from the previous time instants.

We assume independent dynamic models at each view,

$$P(X_t|X_{t-1}) = \prod_j p(x_{t,j}|x_{t-1,j}). \quad (2)$$

Given the posterior at the previous time instant $P(x_{t-1,j}|Z^{t-1})$, $j = 1, \ldots, L$, Equation 1 is updated as

$$\begin{aligned} m_{ji}(x_{t,j}) \leftarrow \int \Bigg[& \int p(x_{t,i}|x_{t-1,i})\, P(x_{t-1,i}|Z^{t-1})\, \mathrm{d}\, x_{t-1,i} \\ & \left(\prod_{k \in N(x_{t,i}) \backslash j} m_{ik}(x_{t,i}) \right) p_i(z_{t,i}|x_{t,i})\, \psi^t_{i,j}(x_{t,i}, x_{t,j}) \Bigg] \mathrm{d}\, x_{t,i}. \end{aligned} \quad (3)$$

Actually, only the information from the previous time instant is integrated into the new message passing process in the graphical model in Figure 2.

Thus, the marginal posterior $P(x_{t,j}|Z^t)$ is given by

$$\begin{aligned} & P(x_{t,j}|Z^t) \propto \\ & p_i(z_{t,j}|x_{t,j}) \left(\prod_{i \in N(x_{t,j})} m_{ji}(x_{t,j}) \right) \int p(x_{t,j}|x_{t-1,j})\, P(x_{t-1,j}|Z^{t-1})\, \mathrm{d}\, x_{t-1,j}. \end{aligned} \quad (4)$$

In fact, the new marginal posterior of Equation 4 is the traditional version plus a message passing process that integrates information from other views.

In practice, the SBP algorithm, implemented using sequential Monte Carlo methods, iterates the message passing process until convergence. Consult Hua and Wu [1] for the details of SBP and its Monte Carlo implementation.

4 Multi-View Tracking using SBP

Our goal is to solve the occlusion problem in single-view tracking by exploiting multi-view information. The SBP algorithm introduced above is well suited for the task.

4.1 The Monte Carlo Implementation

The key of the approach is to propagate the marginal posterior $P(x_{t,j}|Z^t)$ in time using Equation 2, 3, 4. Both the posterior and the messages are represented by weighted particles,

$$m_{ji}(x_{t,j}) \sim \{s_{t,j}^{(n)}, \omega_{t,j}^{(i,n)}\}_{n=1}^{N}, i \in N(x_{t,j}),$$
$$P(x_{t,j}|Z^t) \sim \{s_{t,j}^{(n)}, \pi_{t,j}^{(n)}\}_{n=1}^{N}, \quad j = 1, \ldots, L,$$

where $s_{t,j}^{(n)}$ is the particles sampled at view j, $\omega_{t,j}^{(i,n)}$ is the weight of the message received from view i, and $\pi_{t,j}^{(n)}$ is the belief of the particle based on the observations at all the views. N is the number of particles. Note that the same particle set is used to represent the message and the posterior distribution. The Monte Carlo implementation of the algorithm is described in Algorithm 1.

It is easy to see that the occlusion problem can be effectively solved by the proposed algorithm unless the target is occluded in all the views. Our approach is superior to the best view selection strategy proposed in [4, 5] in that the full information at all the views is taken into consideration during tracking. Even a view in which the target is completely occluded "contributes" to the tracking results by propagating uniformly distributed belief to other views. Although the view isn't informative, it will not affect the inference of the target states at other views. As a result, wrong information propagation is avoided.

Algorithm 1 is similar to the one proposed by Hua and Wu [1]. We extend the original algorithm by adding a fusion module to infer the target states on the ground plane and by modifying the potential function to fit the mulit-view tracking task.

4.2 The Potential Function

An issue in Algorithm 1 is the potential function that describes the spatial relation between the states at two different views. To simplify the problem, we model the target in a view as a rectangle so that the view state x_j is a 4D vector (u_j, v_j, h_j, w_j), where (u_j, v_j) is the middle point of the bottom of the bounding box and (h_j, w_j) is the 2D size.

We assume that the targets of interest always move on a calibrated ground plane, which is usual in video surveillance and team sports senarios, so that the positions

Algorithm 1 SBP based Multi-view Tracking

Require: Given $\{s_{t-1,j}^{(n)}, \pi_{t-1,j}^{(n)}\}_{n=1}^{N}, j = 1, \ldots, L$

Ensure: Generate $\{s_{t,j}^{(n)}, \pi_{t,j}^{(n)}\}_{n=1}^{N}, j = 1, \ldots, L$

1. INITIALIZATION: $k \longleftarrow 1$, for $j = 1, \ldots, L$

1.1 *Resampling*: resample $\{s_{t-1,j}^{(n)}, \pi_{t-1,j}^{(n)}\}_{n=1}^{N}$ to get $\{s_{t-1,j}^{(n)}, 1/N\}_{n=1}^{N}$

1.2 *Prediction*: generate $\{s_{t,j,k}^{(n)}\}_{n=1}^{N}$ from $p(x_{t,j}|x_{t-1,j})$

1.3 *Belief Initialization*: for $n = 1, \ldots, N$

$$\pi_{t,j,k}^{(n)} = p_j(z_{t,j,k}^{(n)}|s_{t,j,k}^{(n)})$$

1.4 *Message Initialization*: for $n = 1, \ldots, N, i \in N(j)$

$$\omega_{t,j,k}^{(i,n)} = \frac{1}{N} \quad \text{(uniformly distributed)}$$

2. ITERATION: SBP

2.1 *Importance Sampling*: Sample $\{s_{t,j,k+1}^{(n)}\}_{n=1}^{N}$ from $P(x_{t,j}|x_{t-1,j})$

2.2 *Message Reweighting*: for $n = 1, \ldots, N, i \in N(j)$

$$\omega_{t,j,k+1}^{(i,n)} \;=\; G_{t,j}^{(i)}(s_{t,j,k+1}^{(n)}) \;\Bigg/\; \left(\frac{1}{N}\sum_{r=1}^{N} p(s_{t,j,k+1}^{(n)}|s_{t-1,j}^{(r)})\right),$$

where

$$G_{t,j}^{(i)}(s_{t,j,k+1}^{(n)}) = \sum_{m=1}^{N}\Bigg[\pi_{t,i,k}^{(m)}\, p_i(z_{t,i,k}^{(m)}|s_{t,i,k}^{(m)}) \left(\prod_{l\in N(i)\setminus j} \omega_{t,i,k}^{(l,m)}\right) \left(\frac{1}{N}\sum_{r=1}^{N} p(s_{t,i,k}^{(m)}|s_{t-1,i}^{(r)})\right) \psi_{i,j}(s_{t,i,k}^{(m)}, s_{t,j,k+1}^{(n)})\Bigg].$$

Normalize so that $\sum_n \omega_{t,j,k+1}^{(i,n)} = 1$.

2.3 *Belief Reweighting*: for $n = 1, \ldots, N$

$$\pi_{t,j,k+1}^{(n)} \;=\; p_j(z_{t,j,k+1}^{(n)}|s_{t,j,k+1}^{(n)}) \left(\prod_{l\in N(j)} \omega_{t,j,k+1}^{(l,n)}\right) \left(\frac{1}{N}\sum_{r=1}^{N} p(s_{t,j,k+1}^{(n)}|s_{t-1,j}^{(r)})\right)$$

Normalize so that $\sum_n \pi_{t,j,k+1}^{(n)} = 1$.

2.4 *Iteration*: $k \longleftarrow k + 1$, iterate until convergence.

3. INFERENCE ESTIMATION:

$$p(x_{t,j}|Z_t) \sim \{s_{t,j,k}^{(n)}, \pi_{t,j,k}^{(n)}\}_{n=1}^{N}, \quad j = 1, \ldots, L$$

.

4. FUSION: The target states in 3D are estimated by fusing the individual view states.

(u_j, v_j) at different views are related to each other by a homography between each pair of views [3]. The propagation of the target sizes between views is a little more difficult because we need the full camera calibration information, by which we can infer the real target sizes in 3D and then project to other views. Fortunately, this information is available in most video surveillance applications where still cameras are used.

Fig. 3. Uncertainty propagation. The red circle in each view is the uncertainty of the target position in the current view (we assume constant and diagonal gaussian noise), and the white ellipse is the uncertainty propagated from the other view. It is clear that the transformation from the right view to the left is more certain.

Therefore, the potential function $\psi_{i,j}$ is defined as

$$\psi_{i,j}(x_i, x_j) \propto \lambda N(x_i; u_{x_i}, \Lambda_i) + (1-\lambda) N\left(x_i; \Pi_j(x_j), \Sigma_j(x_j)\right), \tag{5}$$

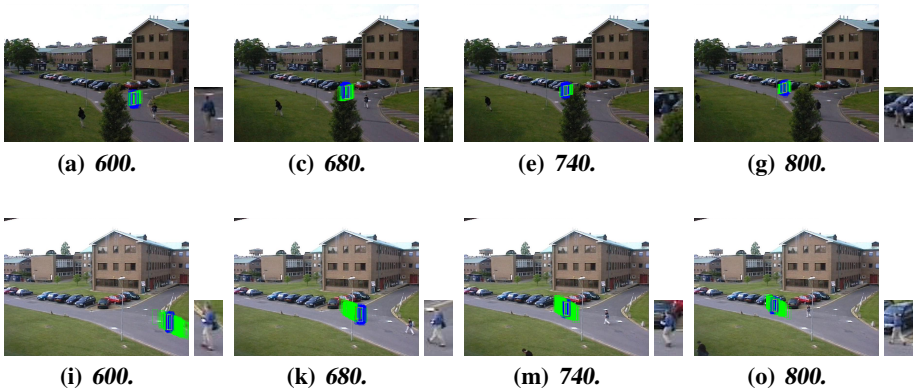
where the first term is the standard Gaussian outlier process, Π_j is a function that transforms the view state x_j to view i, and Σ_j is a function that propagates the uncertainty of x_j to view i using techniques from perturbation theory [10], see Fig. 3.

5 Results

Since 2-view data are most readily available, a SBP-based 2-view tracker was developed. The same principles apply when three or more views are used, although loops exist in the graphical model. For such situations, it was shown that loopy BP typically still yields good approximate results [8].

As described in Section 4.2, the target state $x_{t,j}$ is defined as a 4D vector with two coordinates for the position and the other two for the size to handle the scale changes. The motion model $p(x_{t,j}|x_{t-1,j})$ at each view is the standard constant-velocity model.

Following Perez et al. [11], a classical color observation model based on HSV color histograms is adopted which has the advantage of being insensitive to illumination effects. Thus, the observation process is to match the color histogram in a candidate region, a particle, with a pre-learned reference model, where the Bhattacharyya similarity coefficient is computed to measure the distance. The effectiveness of this model has been shown previously [11, 12, 4] and is confirmed by this work. In all the experiments, we manually initialize the regions of targets of interest at the first frame of each camera and learn the reference color models.



(a) *600.* (c) *680.* (e) *740.* (g) *800.*

(i) *600.* (k) *680.* (m) *740.* (o) *800.*

Fig. 4. Result of tracking a pedestrian. Blue rectangles are the particles sampled in the current view, while the green ones are the particles mapped from the other view using the homography between the two views. The white rectangles are the estimated target states of the view, whereas the red rectangles are the target states of the other view which are mapped to the view using the same homography.

5.1 Video Surveillance

PETS2001 Dataset Two contains sequences taken from two calibrated cameras and is used to evaluate the above algorithm. Figure 4 shows the result of tracking a pedestrian in subsequences of Camera 1 and Camera 2 from Frame 600 to Frame 800. The pedestrian is completely occluded by a tree in Camera 1 but is visible all the time in Camera 2. Thus, the algorithm successfully tracks the target during the occlusion in Camera1 by receiving messages from Camera 2. Although the result is a little biased due to the uncertainty propagation, it is corrected when the target reappears after the occlusion.

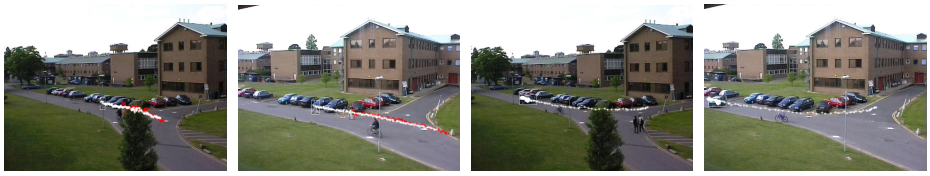
Figure 5 shows the result of tracking the same pedestrian from Frame 775 to Frame 850 and the comparison with Condensation [13]. Since we learn a simple color model of the target from only one frame, sometimes it is not very distinctive from the background. As a result, Condensation fails at the 805th frame of Camera 1 and at the 819th frame of Camera 2. However, our multi-view tracking algorithm keeps tracking by exchanging information across views. We agree that the problem may be solved by learning a better model or using another color space, but the problem still exists: if the target is not distinctive from clutter, it is difficult to maintain the target distribution with a small and fixed number of particles. By integrating multi-view observations, the algorithm is capable of dealing with unstable appearance in one view if stable appearance can be obtained in another view.

Other tracking results can be seen in Figure 6. Note that the two targets that are tracked from Frame 600 to Frame 850 in Figure 6 (a) and (b) are lost afterwards because they become too small to track using only color information. The tracking of a car from Frame 1300 to Frame 1800 is shown in Figure 6 (c) and (d). Since the car turns around in the subsequences of both cameras causing significant appearance changes, it

is impossible to learn the reference color model from only one frame. To solve this problem, we sample particles in this experiment from both the motion prior and a proposal distribution obtained from a change detection process based on a background model [14].

(a) ***SBPMVT-805.*** **(b)** ***Condensation-805.*** **(c)** ***SBPMVT-819.*** **(d)** ***Condensation-819.***

Fig. 5. Comparison of the SBP-based multi-view tracker (SBPMVT) with Condensation.



(a) ***Camera 1 (600-850).*** **(b)** ***Camera 2 (600-850).*** **(c)** ***Camera 1 (1300-1800).*** **(d)** ***Camera 2 (1300-1800).***

Fig. 6. Results of tracking several different targets.

Thus, the importance sampling function is

$$(1-\alpha)p(x_{t,j}|x_{t-1,j}) + \alpha p(x_{t,j}|B_{t,j}), \tag{6}$$

where $B_{t,j}$ is the observation of the foreground.

5.2 Discussion

We find that the potential function $\psi_{i,j}(x_i, x_j)$ is critical to the success of the multi-view tracking algorithm. As is described in Section 4.2, the target states at one view are transformed to another view by a homography under the assumption that the targets move on the ground plane. However, the transformation has large uncertainty if the camera view direction is highly oblique, for instance, when the camera position is close to the ground plane. In this case, more particles are needed to model the target distribution. This motivates the use of more views (> 2) which will reduce the uncertainty.

The extra fusion module that combines results at each view can be removed by adding a node representing the target states in 3D (2D ground) in the graphical model in Figure 2. The addition of this global node does not only change the current, fully-connected graphical model to a two-level, tree-structured graphical model, making the system more scalable and flexible to varying numbers of cameras, but also enables us to infer the 3D target states inside the SBP algorithm.

6 Conclusion and Future Work

This paper presents a novel multi-view tracking method that addresses the occlusion problem using bidirectional belief propagation. The strength of the method relies on the fact that information is integrated and exchanged across views so that a collaborative tracking scheme is formed. Technically, the tracking processes at different views perform the inference of the target states separately but based on the multi-view observations. A sequential and purely non-parametric belief propagation algorithm is adopted to allow individual trackers to collaborate in each view, which largely avoids the problem of propagating wrong information. As demonstrated, the method is robust and capable of dealing with occlusions as long as the targets of interest are visible in at least one view.

We are currently extending this work by adding one node in the graphical model representing the 3D target states. Another extension which is also ongoing is to track multiple targets simultaneously, which will broaden the applicability of the system.

References

1. Hua, G., Wu, Y.: Multi-scale visual tracking by sequential belief propagation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2004)
2. Black, J., Ellis, T.: Multi camera image tracking. In: the Second IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Hawaii, USA (2001)
3. Hayet, J.B., Mathes, T., Czyk, J., Piater, J., Verly, J., Macq, B.: A modular multi-camera framework for team sports tracking. In: International Conference on Advanced Video and Signal based Surveillance, Como, Italy (2005)
4. Nummiaro, K., Koller-Meier, E., Svoboda, T., Roth, D., Van Gool, L.: Color-based object tracking in multi-cameras environment. In: 25th Pattern Recognition Symposium, DAGM. (2003)
5. Wang, Y., Wu, J., Kassim, A.: Multiple cameras tracking using particle filtering. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Breckenridge, USA (2005)
6. Doucet, A., de Freitas, N., Gordon, N.: sequnetial Monte Carlo methods in practice. Springer-Verlag, New York (2001)
7. Sudderth, E., Ihler, A., Freeman, W., Willsky, A.: Nonparametric belief propagation. In: IEEE Conference on Computer Vision and Pattern Recognition, Madison, USA (2003) 605–612
8. Freeman, W., Pasztor, E.: Learning low-level vision. In: International Conference on Computer Vision, Greece (1999)
9. Ferryman, J.: (Pets websites: http://visualsurveillance.org/pets2001)

10. Criminisi, A., Reid, I., Zisserman, A.: A plane measuring device. In: British Machine Vision Conference. (1997)
11. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: European Conference on Computer Vision. Volume 1. (2002) 661–675
12. Okuma, K., Taleghani, A., Freitas, N., Little, J., Lowe, D.: A boosted particle filter: multitarget detection and tracking. In: European Conference on Computer Vision. (2004) 28–39
13. Isard, M., Blake, A.: Condensation-conditional density propagation for visual tracking. International Journal of Computer Vision **29** (1998) 5–28
14. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition. (1999)