Mixed-State Particle Filtering for Simultaneous Tracking and Re-identification in Non-overlapping Camera Networks

Boris Meden[1], Patrick Sayd[1], and Frédéric Lerasle[2,3]

[1] CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Point Courrier 94, F-91191 Gif-sur-Yvette, France
[2] CNRS; LAAS; 7 avenue du Colonel Roche, F-31077 Toulouse Cedex 4, France
[3] Université de Toulouse; UPS, INSA, INP, ISAE; UT1, UTM, LAAS; F-31077 Toulouse Cedex 4, France
{boris.meden,patrick.sayd}@cea.fr, lerasle@laas.fr

Abstract. This article presents a novel approach to person tracking within large-scale indoor environments monitored by non-overlapping field-of-view camera networks. We address the image-based tracking problem with distributed particle filters using a hierarchical color model. The novelty of our approach resides in the embedding of an already-seen-people database in the particle filter framework. Doing so, the filter performs not only position estimation but also does establish identity probabilities for the current targets in the network. Thus we use online person re-identification as a way to introduce continuity to track people in disjoint camera networks. No calibration stage is required. We demonstrate the performances of our approach on a 5 camera-disjoint network and a 16-person database.

Keywords: re-identification, tracking, camera network, non-overlapping fields of view, particle filtering.

1 Introduction

The problem of estimating the trajectory of an object as it moves in an area of interest, known as tracking, is one of the major topics of research in computer vision (see a comprehensive survey in [14]). That becomes even more challenging with multiple objects tracking (MOT), aiming to maintain identities of tracks. MOT has been tackled by supervised approaches [13], but also with distributed particle filters [12] [1]. However, it is usually not feasible to completely cover large areas with cameras having overlapping views due to economic and/or computational reasons. Thus, in realistic scenarii, the system should be able to handle multiple cameras with non-overlapping fields of view (NOFOV). Beyond the intra-camera tracking problem, the crucial difficulty resides in the transitions between cameras and the problem of maintaining targets' identities at the network level. The differences in target appearances are mainly due to different poses of cameras and different colorimetric responses.

A. Heyden and F. Kahl (Eds.): SCIA 2011, LNCS 6688, pp. 124–133, 2011.

That jump between cameras, known as the re-identification problem, can be seen as twofold with on the one hand the robustness of the descriptor and on the other hand the specific strategy to match identities. As most approaches, Gray *et al.* [4] focus on the target descriptor to achieve the best frame to frame re-identification rate. They propose the VIPeR dataset, composed of pedestrian images taken for two cameras with different viewpoints and illuminations. Prosser *et al.* used a similar approach in [11]. Rather than choosing which cues to use, these works let a meta-algorithm provide a descriptor highlighting the invariant cues of pedestrian silhouette relatively to a learning database. The limitations here are the great number of samples needed. Other works, also on the descriptor level, try to project their color descriptors on the same subspace, putting the focus on the color consistency issue and resorting to color calibration. Thus, Javed *et al.* in [6] compute a subspace based color brightness transfer function. That transfer function is estimated over a set of training samples seen in the network. Bowden *et al.* in [3] go further into the learning of that transfer function as they compute it incrementally. Again here, the limitations reside in the training phase which takes processing time and is biased as the sample target set cannot be exhaustive.

The different aforementioned approaches consider all a frame to frame comparison strategy. Cong *et al.* in [2] propose a more enhanced process in matching not only single images of the target but whole tracking sequences. They perform a spectral analysis of the graph Laplacian of the matrix of two sequences. The number of clusters in the matrix (*i.e.* the number of similar descriptors) is directly linked to the eigenvalues. With SVM classification in the reduced eigenspace, the decide whether the sequences belong to the same target or not. The inter-camera re-identification is achieved through the RGB Greyworld normalization and an elaborated strategy over the sequences. However the comparison is only done for two sequences at the same time. To bridge the gap between cameras, Makris *et al.* in [8] learn spatiotemporal transitions to infer re-appearance time of targets evolving in a blind spot of the network. With the same goal, Lim *et al.* in [7] propose a two-behavior particle filter: when the tracked target is visible, usual tracking is performed, when it is in a blind spot, particles evolve in the metric map of the building, dividing the current group in two at each intersection. Re-identification is achieved when a detection in a camera coincide with particles in the metric map. Limitations of these approaches resides clearly in the multiple objects configurations.

In this paper, we see the tracking in NOFOV networks as an extension of the multiple objects tracking (MOT) in mono- or multi-ocular sequences [13] [12]. To do so, we propose to embed re-identification within the particle filter framework. Thus we estimate not only relative position in a given camera but also the identity of the target in respect to an *a priori* learned person database (people that have entered the building where the network is set). Moreover, we use distributed filters, which implies low complexity and enable the approach to be extended to large networks. This strategy is an enhanced frame to frame comparison as the filter introduce temporality, but still produce a re-identification

result at each new frame. As there is no public dataset treating about extended camera networks, we tested our algorithms on our private 5-camera network composed of a 34-meter long corridor, a meeting room and a building outdoor entrance with a total 16 pedestrians are wandering in it.

In the following, Section 2 first presents how we learn the identities that we will track in the other cameras and build a target-database. Section 3 details the particle filter adaptation. Section 4 introduces a supervisor notion for the network monitoring. And finally Section 5 presents the way we evaluated the approach.

2 Learning Identities to Re-identify and Track

2.1 Target Representation

To avoid any camera geometrical calibration problem, the tracking is conducted in the image plane. We use a rectangular geometric model as Region Of Interest (ROI). The descriptor is hierarchical: the ROI is sliced into regular horizontal stripes and each stripe is described by its color distribution. Color histograms have proven to be robust to appearance changes [9] with their global aspect. The addition of spatial constraints in the signature localizes the colors and increases the discrimination power. It has been successfully used for tracking purpose by Pérez *et al.* [10] as well as for re-identification purpose by Cong *et al.* [2]. Moreover that type of descriptor (termed Hand Localized Histogram) was part of the evaluations conducted by Gray *et al.* in [4] and also achieved good results in frame to frame comparison. We use color histograms in the RGB color space with 8 bins per channel for tracking computing time, and we tuned the number of stripes using [4] evaluation. As they did, we computed Cumulative Matching Characteristic (CMC) curves over the VIPeR dataset, for different numbers of stripes from 1 to 30 and kept the best curve, corresponding to the 5-stripe descriptor. Associated to a normalization process explained in subsection 3.2, large bins allow us to handle color discrepancy between cameras.

2.2 Reducing the Database to Key Frames

Before recognizing people we have to see them a first time. We propose to do a learning of identities in a first camera seen as an entrance point in the network (*e.g.* the hall of a building, and Site 0 in the figure 4). Then we will treat the network as a closed system with a collection of identities walking in it. First we run a traditional CONDENSATION[1] particle filter on the learning sequence. We extract a view of the target for each frame. Then, we reduce offline this collection of descriptors to key ones. To select an appropriate number of key frame to retain the same amount of variation for each identity, we perform a spectral analysis of the tracking sequences. Our approach is inspired by [2] but here limited to a single person. Thus we focus on the variations of the target descriptor for the

[1] For Conditional Density Propagation.

same target to extract the main representative descriptors of the sequence. To do so, we build the similarity matrix of each tracking sequence taken from the learning camera as $W_{ij} = \exp(-K \cdot \sum_{k=1}^{N_c} d^2(s_i(k), s_j(k)))$, where $d(.,.)$ is the discrete Bhattacharyya Distance, $s_i(k)$ (*resp.* $s_j(k)$) is k-th color distribution of target i (*resp.* j), N_c the number of color distributions per target and K a normalization constant. We apply spectral clustering method to that similarity matrix calculating its un-normalized Graph Laplacian $\Delta = D - W$ where D is the diagonal matrix of the horizontal sums of W elements: $D_{ii} = \sum_j W_{ij}$. We then diagonalize the Graph Laplacian. The eigenvalues present an eigengap when the number of clusters is reached [2]. In a one person sequence, the gap may not be obvious, so just put a threshold on the eigenvalues and perform k-means clustering in the reduced space of the k first eigenvectors, k being the number of eigenvalues lesser than the threshold. Thus we summarize a tracking sequence to key frames that retain the main variability in terms of appearance. Figure 1 shows some of the tracking boxes and the chosen key frames. Out of 100 tracking boxes, we extract between 4 and 10 key frames.

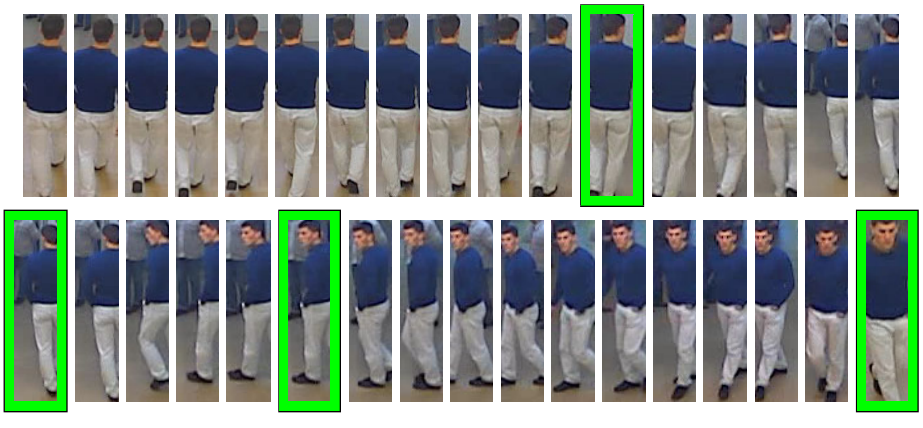

Fig. 1. Some boxes part of a tracking sequence (scaled to have the same height for design purpose). The green frame highlights the four key frames selected by the k-means algorithm to represent that identity. These key frames capture the biggest variation of the target appearance in the tracking sequence.

3 Embedding Re-identification into the Tracking Process

3.1 Particle Filter Framework

A bayesian tracking filtering process begins with the choice of a reference region in an image, and then proceed to a recursive search of similar regions in the remaining of the sequence. Given the identity database, we have got here another reference descriptor to compare with. We use the Mixed State CONDENSATION particle filter framework [5], to estimate our Mixed State vector composed by

continuous parameters (the target's image coordinates $\mathbf{x}$) and also a discrete parameter (the target's identity y) in the filter loop, namely

$$\mathbf{X} = (\mathbf{x}, y)^{\mathsf{T}},\ \mathbf{x} \in \mathbb{R}^4,\ y \in \{1, \dots, N_{id}\}$$

In our case, we track in the image plane with a rectangular geometric model. We have $\mathbf{x} = [x_c, y_c, h_x, r]^{\mathsf{T}}$, where $(x_c, y_c)^{\mathsf{T}}$ are the coordinates of the box center, h_x is the half width of the box, and r is the width-height ratio which is assumed constant, and where N_{id} is the cardinal of the identity database, and N the number of particles. Given that extended state, the sampling process density at frame t can be written as in [5]:

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = T(\mathbf{X}_t, \mathbf{X}_{t-1}) \cdot p(\mathbf{x}_t|\mathbf{x}_{t-1})$$

where $T(\mathbf{X}_t, \mathbf{X}_{t-1})$ is a transition probability matrix which will sample the discrete ID parameter, and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the sampling on the continuous part of the state. The transition matrix $T = [t_{ij}]$ is built over the key frame set. The element t_{ij} is the similarity between identities i and j in the database, computed using equation (1) between the most distant key frames of each identity.

The difference with [5] resides in the discrete parameter meaning. They used it to include different motion models into the filter and to have it decide which one fits the best. For us, and this is the main novelty of that paper, this parameter refers to an identity in our already-seen-person database and allows us to perform simultaneous tracking and re-identification. To the best of our knowledge this has not been done before.

3.2 Estimating the Identity and the Position

After the sampling stage, the new positions of the particles are evaluated relatively to the new image $\mathbf{Z}_t$. The traditional temporal likelihood $p(\mathbf{Z}_t|\mathbf{x}_t^{(n)})$ is estimated as:

$$w_{Temp}^{(n)}(t) = \exp\{-K \cdot \sum_{j=1}^{N_c} d^2\left(s_t^{(n)}(j), s_{model}(j)\right)\},\ \forall\ n = 1, \dots, N$$

where N_c is as previously the number of color distributions per target, $s_{model}(.)$ the set of color distributions of the tracking reference model (*i.e.* the initial box of a tracking process, that we do not update during the process), $s_t^{(n)}(.)$ the set of color distributions of the current particle, and N is the number of particles.

The Mixed-State CONDENSATION framework adapted to re-identification provides an additional likelihood, weighting the particle relatively to its identity, $p(\mathbf{Z}_t|\mathbf{x}_t^{(n)}, y_t^{(n)})$:

$$w_{Id}^{(n)}(t) = \exp\{-K \cdot \min_{i \in N_y} \sum_{j=1}^{N_c} d^2\left(s_t^{(n)}(j), s_{identity}(j, y_t^{(n)}, i)\right)\},\ \forall\ n = 1, \dots, N \quad (1)$$

where N_y is the cardinal of the key frame class of identity $y_t^{(n)}$ ($y_t^{(n)}$ being the identity assigned to the n-th particle at time t), N_c the number of color distributions per target, $s_{identity}(., y_t^{(n)}, i)$ is the set of color distributions of the i-th keyframe of identity $y_t^{(n)}$ in the database, $s_t^{(n)}(.)$ is the set of color distributions of the current particle, and N is the number of particles. Figure 2 sum up the principle of these two likelihoods per particle. Each particle is evaluated relatively to the reference of tracking (w_{Temp}), but also (w_{Id}) relatively to its identity (described by a collection of key frames).

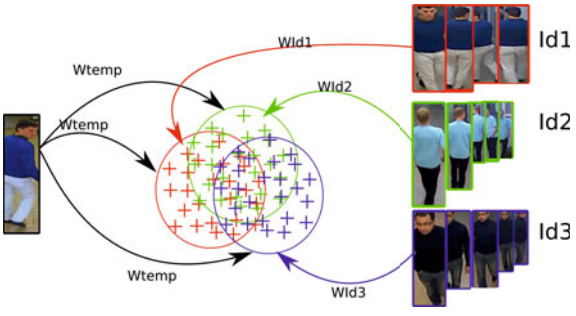

Fig. 2. Illustration of our mixed state particle filter in the case of a database of cardinal of 3. The particle cloud is divided into three subcloud, identically distributed at the initialisation of the filter (as displayed in the figure). Then the strongest identity will take the lead, because of the combined likelihood and the transition matrix T. Mixed state particles share the same temporal tracking reference (left), but a different identity in the database (right, with the key frames).

As these two types of likelihood do not share the same order of magnitude, we normalize them over the set of particles before the resampling stage. That way, we guarantee the jump between two cameras, known as re-identification. Here, we do not favor any bins in the histograms, as [6] and [3] do with the computation of a transfer function. We assume a linear transformation between the cameras colorimetric responses, adopt a rather large bin quantization to absorb that transformation and apply that normalization. Moreover, unlike color calibration, this approach is independent of the pair of cameras considered. We note the normalized likelihood w_{Id}^* and w_{Temp}^*. If w_{Temp}^* is greater than a threshold (*i.e.* if the particle is relevant, otherwise we just use the low temporal similarity as the combined one), we combine both of these similarities to obtain our likelihood formulation which will be injected into the particle weighting stage:

$$\pi_t^{(n)} = \alpha \cdot w_{Temp}^{*(n)}(t) + (1-\alpha) \cdot w_{Id}^{*(n)}(t), \ \forall\ n = 1, \ldots, N.$$

Doing so, we give weight to the particles that moved into the right place, assuming that they hold the right identity. The state estimation is then a two-stage

process. First we need to compute the MAP on the discrete parameter, *i.e.* a partial re-identification.

$$\hat{y}_t = \arg\max_j P(y_t = j|\mathbf{Z}_t) = \arg\max_j \sum_{n \in \boldsymbol{\Upsilon}_j} \pi_t^{(n)}, \text{ where } \boldsymbol{\Upsilon}_j = \left\{ n | s_t^{(n)} = (\mathbf{x}_t^{(n)}, j) \right\} \quad (2)$$

The continuous state components are then estimated on the subset of particles that have the strongest identity (equation (3)).

$$\hat{\mathbf{x}}_t = \sum_{n \in \hat{\boldsymbol{\Upsilon}}} \pi_t^{(n)} \cdot \mathbf{x}_t^{(n)} / \sum_{n \in \hat{\boldsymbol{\Upsilon}}} \pi_t^{(n)}, \text{ where } \hat{\boldsymbol{\Upsilon}} = \{ n | s_t^{(n)} = (\mathbf{x}_t^{(n)}, \hat{y}_t) \} \quad (3)$$

4 The Non-ubiquity Constraint

Our distributed approach provides a strategy for re-identification. Instead of comparing one query image to every entries in the database, we let our mixed-state particle filter perform the decision, allowing identity concurrency in the process. The drawback of the approach resides in the fact that there is no interactions between filters, which means that nothing constrain filters from choosing the same identity.

Thus we add to the approach a light supervising procedure that gather re-identification probabilities thanks to the online identity characterization and assign each filter its most likely identity respectively to the other filters. For a multiple targets configuration supervised, equation 2 transforms to equation 4.

$$\hat{y}_t(f) = \arg\max_j P(y_t = j|\mathbf{Z}_t, f) = \arg\max_j \sum_{n \in \boldsymbol{\Upsilon}_j(f)} \pi_t^{(n)}(f), \quad (4)$$

$$\text{where } \boldsymbol{\Upsilon}_j(f) = \left\{ n | s_t^{(n)}(f) = (\mathbf{x}_t^{(n)}(f), j) \right\}, \ \forall f = 1 \dots N_{filters}$$

where $(s_t^{(n)}(f), \pi_t^{(n)}(f))$ is the n-th particle and its likelihood, of the f-th filter, and $N_{filters}$ is the number of particle filters currently running. When a filter receive an identity, this identity becomes no more available for the remaining filters. That way, we avoid having the same identity for two separate targets.

5 Evaluations

5.1 Evaluation Network Setup

We used a five-camera network presenting non-overlapping fields of view (Figure 4). The camera 0 is the one we used to learn offline the database. Then, we let 16 pedestrians wandering in the network. Figure 3 shows a key frame per identity of the database.

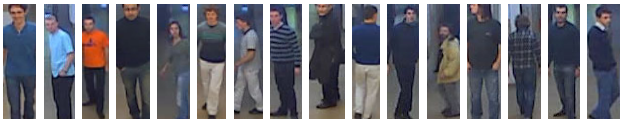

Fig. 3. Our private 16-pedestrian database for experiments

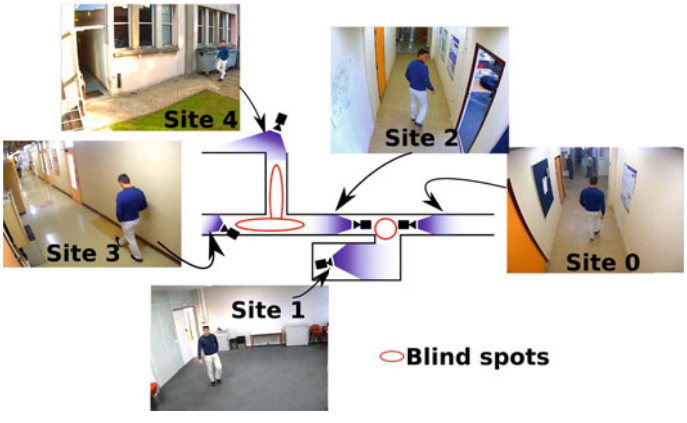

Fig. 4. Overview of the testing network composed of one 34-meter long corridor, one meeting room and one outside area

5.2 Re-Identification Efficiency

As we explained in Section 1 , our mixed-state approach provides a new strategy for the re-identification problem. First we provide a thorough comparison of that strategy to the state of the art one, considering the case of the single target tracking. We compute re-identification results of all the 16 database identities for all cameras, for a frame to frame strategy and for ours. In both cases we use the same descriptor (as we evaluate only the strategy), and initializations of tracking are provided by a configuration file hand-made. A complete system would resort to a detector. For the frame to frame, we run the tracking process with no identity feedback, and compare the estimated position to each entry in the database, at each time step. Evaluated with identity ground truth, both strategies produce binary answers at each time step for each target. For each camera, we sum the results, which true gives re-identification rates per frame, and then we average them over the overall sequence. The rates are also averaged over five runs of each tracking sequence due to the stochastic nature of Particle Filters. Table 1 summarize these results.

We observe different re-identification rates depending on the camera considered. The site #0 is where the identities have been learned, so descriptors are really similar, hence the almost 100% rate. However, sites #1 and #4 are rather different in terms of camera pose and background colors (site #4 being moreover

Table 1. Re-identification rates for camera to camera comparison of the trivial approach and Mixed State one

Approach	Site #0 to #0	Site #0 to #1	Site #0 to #2	Site #0 to #3	Site #0 to #4
Track then ID	0.96	0.40	0.66	0.65	0.30
Track + ID	0.98	0.46	0.81	0.71	0.34

outside). The descriptor chosen use no background subtraction, which is an explanation to the dropping rates. Still, for each camera, the simultaneous tracking and re-identification strategy performs better than the frame to frame one.

5.3 Multi-Target Re-identification

Figure 5 provides an illustration of the typical case where the non-ubiquity constraint is useful: multiple targets evolving in the network. Quantitative evaluations are being studied.

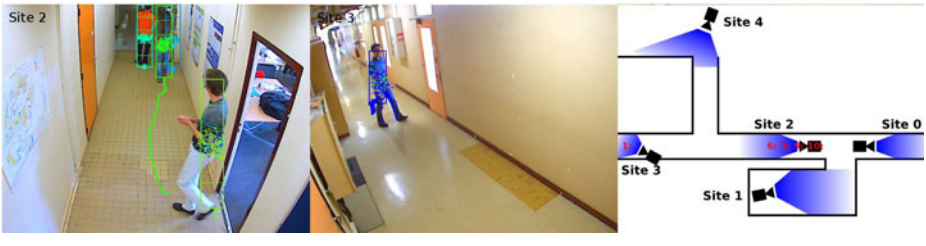

Fig. 5. Tracking 5 targets in the network: four in site #2 and one in site #3. Re-identification results are reported on a map of the network.

6 Conclusion and Perspectives

We have proposed a new approach for people tracking in NOFOV camera networks, which does not require any *a priori* knowledge on the network. Here we see person re-identification as a means to bring continuity between tracking sequences from different cameras. The main novelty of that paper is to embed re-identification into the particle filter framework to estimate simultaneously the target's position and its ID within the camera network. Rather than focusing on the descriptor, we propose here an enhanced matching strategy, introducing temporal filtering in the re-identification process. We have proved by a thourough comparison over every sites of our private network and every identities considered that our mixed-state particle filtering strategy outperforms the usual frame to frame comparison. Moreover, our approach is theoretically independent of the cameras number as the filters are distributed. And we also provide a way to constrain ubiquity of identities between the filters in case of multiple targets tracking.

Further work will investigate on an online construction and updating procedure of the identity database. Moreover, adding interaction forces between filters as proposed in [12] would reinforce the multi-targets mono-camera tracking. Finally, while our approach only uses 2D information, additional knowledge about the scene (*e.g.*, a ground plane to improve targets' size estimation), or about the network (*e.g.*, a topology map to infer some unlikely positions in the network) would be beneficial.

References

1. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: Proceedings of the International Conference on Computer Vision (2010)
2. Cong, D.N.T., Khoudour, L., Achard, C., Meurie, C., Lezoray, O.: People re-identification by spectral classification of silhouettes. Signal Processing (2009)
3. Gilbert, A., Bowden, R.: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 125–136. Springer, Heidelberg (2006)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
5. Isard, M., Blake, A.: A mixed-state condensation tracker with automatic model-switching. In: Proceedings of the International Conference on Computer Vision (1998)
6. Javed, O., Shafique, K., Shah, M.: Appearance modeling for tracking in multiple non-overlapping cameras. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2005)
7. Lim, F., Leoputra, W., Tan, T.: Non-overlapping distributed tracking system utilizing particle filter. The Journal of VLSI Signal Processing (2007)
8. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2004)
9. Nummiaro, K., Koller-Meier, E., Van Gool, L.: An adaptive color-based particle filter. Image and Vision Computing (2003)
10. Perez, P., Vermaak, J., Blake, A.: Data fusion for visual tracking with particles. Proceedings of the IEEE (2004)
11. Prosser, B., Zheng, W., Gong, S., Xiang, T., Mary, Q.: Person Re-Identification by Support Vector Ranking. In: Proceedings of the British Machine Vision Conference (2010)
12. Qu, W., Schonfeld, D., Mohamed, M.: Distributed bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. EURASIP J. Appl. Signal Process. (2007)
13. Smith, K., Gatica-Perez, D., Odobez, J.: Using particles to track varying numbers of interacting people. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2005)
14. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. Acm Computing Surveys (CSUR) 38(4), 13 (2006)