MCMC Supervision for People Reidentification in Nonoverlapping Cameras

Boris Meden[1]
boris.meden@cea.fr
Frédéric Lerasle[2]
lerasle.laas.fr
Patrick Sayd[1]
patrick.sayd@cea.fr

[1] CEA, LIST,
Laboratoire Vision et Ingénierie des Contenus,
BP 94, F-91191 Gif-sur-Yvette, France

[2] CNRS ; LAAS ;
Université de Toulouse ; UPS, LAAS ;
F-31077 Toulouse Cedex 4, France

Abstract

We present a pedestrian tracking system that uses re-identification to monitor non-overlapping cameras. As tracking, re-identification is an assignment problem, the difficulties being to generate an accurate representation and to prune unlikely pairings. The assignments are realised in two stages. First, a Markovian multi-target tracking-by-detection framework which includes identification in the search space is run in the cameras. This generates tracks in the cameras and a first assignment between them thanks to the local identification. This solution is then optimized globally by a network supervisor benefiting from coarse topology knowledge over a sliding window with MCMC sampling. The tracking results obtained on a large ground-truthed dataset demonstrate the effectiveness of the approach.

1 Introduction

In this paper, we present a novel approach to perform multiple objects tracking (MOT) and re-identification on-the-fly, to monitor Non Overlapping Fields Of View networks (abbreviated NOFOV networks in the following).

Pedestrian tracking using a distributed NOFOV network offers the following advantages: (i) larger areas can be covered with few sensors; (ii) existing camera infrastructure can be exploited. The goal of MOT is then to cope with these discontinuities and to still guarantee spatio-temporal consistency in the whole camera network. The problem becomes twofold: beyond the image plane multi-target tracking, the system should be able to re-identify the targets when they appear in a new camera and thus achieve a "handover" of identity from one camera to another. This is called re-identification. The goal of fusing re-identification and MOT is to be able to retrieve *global identity* trajectories in the NOFOV network.

An overview of the whole algorithm is given in Section 3 while the trackers and the supervisor are detailed in Section 4 and Section 5 respectively. Experimental results are reported in Section 6 and our conclusions are given in Section 7.

2 Related Work

Pionneering works on MOT in the Computer Vision community started with sequential logic and first order markovian model. Particle filtering algorithms' interest for tracking (CONDENSATION) have been established by Isard and Blake, notably for multiple targets in [11]. Then, since [16], *tracking-by-detection* has emerged. The increased robustness of deferred logic methods has been proven by Xing *et al*. in [18]. Among the current *state-of-the-art* methods in terms of image plane multi-target mono-camera tracking we can cite Benfold and Reid in [1] for tracklets temporal optimization and Breitenstein *et al*. in [3] for Markovian methods. Both approaches are system-based and rely on a pedestrian detector such as HOG [5].

In terms of pure re-identification, recent works have sought to build a good representation for pedestrians to yield a reliable similarity function. In that vein, [8] and [7] were the first to propose to formulate it as a ranking problem, evaluating it with Cumulative Matching Curves. Thus, [8] propose to train a classifier on the invariant parts during a camera change, based on ground truth training pairs. The learned model focuses on the stable features through camera change. Besides, Zheng *et al*. [20] put the effort on the distance matching and proposed a probabilistic learned distance, trained to minimize the distance between correct pairs. Again, the training is performed on ground truth pairs, *i.e.* assume to have solved the problem for some samples. In opposition to these approaches, Farenzena *et al*. [6] adopt an unsupervised approach, with no learning. They propose a robust fixed signature based on symmetry and asymmetry of the appearance and well positionned colorimetric features, added to mean color blob matching extracted by Maximal Stable Color Region, and local texture patches. However, these methods are not designed for online application because of their important computation time and the need of ground truthed pairs as training samples.

In a camera network, when targets' trajectories present discontinuities due to the lack of observability, *e.g*. between nonoverlapping cameras, applied pedestrian re-identification becomes a necessity. Huang and Russel [9] represent some early work for multi-camera tracking with non-overlapping fields of view. They formalized their car-reidentification problem using association in a probability space built on similar target sizes and mean color. Then Pasula *et al*. [17] proved the efficiency of MCMC sampling to explore the assignment space. Some extra knowledge on the network can be learned, such the spatio-temporal correspondances [14] or appearance relationship through brightness transfer functions [12].

Several recent works start to transfer the re-identification descriptors to application needing online identification. In terms of embedded system, Matei *et al*. in [15] are on the particular problem of vehicle tracking in NOFOV networks. The linear motion model and the constant speed allow them to draw out a novel Multiple Hypothesis Tracking formulation using kinematic and appearance features. On the contrary, Kuo *et al*. in [13] are concerned with pedestrian. They adopt a similar re-identification approach as [8]. Finally, re-identification is seen as a MAP problem and is solved by Hungarian Method enumerating every possibility at the end of the sequence. However the gathering of training samples relies on a weak constraint and it is unclear how the approach can scale to multiple sensors as the number of assignments to test grows exponentially.

From these insights, our two-layered *tracking-by-reidentification* approach exhibits the following contributions. First, target re-identification within the network is achieved on-line thanks to a mixed-state sampling exploring both image plane location and identity of the current target. Second, this local identification is used as prior to an assignment optimization

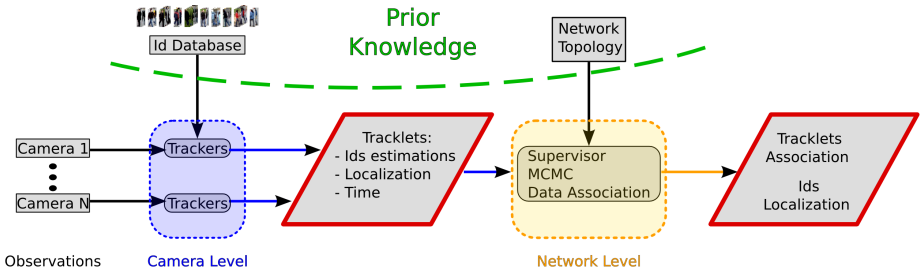

Figure 1: Synoptic diagram of the combination between markovian tracking-by-reidentification and MCMC data association at the network level.

process benefiting from topology knowledge and achieved through MCMC sampling. This final track assignment between cameras is performed in a deferred logic way on a sliding window to ensure robustness.

3 Overview of the Approach

The overall synoptic diagram for the centralized/decentralized tracking algorithm presented in this paper is described in figure (1). A NOFOV network is deployed to monitor some areas of interest. We assume to know the topology of the network, *i.e.*which cameras can exchange targets. We also assume to know from which areas (feeding areas) new people may enter in the network. Then, the main steps of our algorithm are:

- At the camera level , automatic distributed trackers based on HOG detections and following the approach of [3] track targets for each camera. Inspired by [6], the appearance model used is composed of horizontal stripes of HSV histograms weighted by their distances to the symmetry axis. The use of topology allows to instantiate new identities from the feeding areas in an identity database, which we compare with to perform re-identification. The mixed-state formalism [19] uses that database and sample in this identity space. That way the tracker produce a tracklet and re-identification probabilities in the database representing the belief of the tracker. The resulting tracklets are sent to the supervisor along with their probabilities of identity, their time of existence and their areas.

- At the network level, the supervisor resorts to deferred logic to optimize the assignment between the received tracklets using re-identification distributions and network topology information. The combinatorial space is efficiently explored through MCMC sampling. Tracks output by the supervisor are optimized to represent the activity of the same person.

4 Local Tracking-by-Reidentification at the Camera Level

4.1 Detections Integration

4.1.1 Associations to Detections

An association matrix is built between trackers and detections. The score of pair detection d vs. tracker tr given by equation (1), involves:

- the distance between the tracker's particles and the detection, evaluated under a gaussian kernel $p_{\mathcal{N}}(.) \sim \mathcal{N}(.,\sigma^2)$;
- the tracker's box area $\mathcal{A}(tr)$ relatively to the detection's one also evaluated under a gaussian kernel ;
- the tracker's appearance model score on the detection ($w_{App}(.)$) ;
- the identity models scores on the detection ($w_{Id}(.)$) weighted by the particle subsets Υ_j as defined in equation 3.

$$S(d,tr) = \underbrace{\sum_{p\in tr}^{N} p_{\mathcal{N}}(d-p)}_{\text{euclidean distance}} \times \underbrace{p_{\mathcal{N}}\left(\frac{|\mathcal{A}(tr)-\mathcal{A}(d)|}{\mathcal{A}(tr)}\right)}_{\text{relative size}} \times \underbrace{w_{App}(d,tr)}_{\text{appearance model}} \times \underbrace{\sum_{j=1}^{N_{id}} \Upsilon_j \cdot w_{Id}(d,j)}_{\text{identities distribution}} \quad (1)$$

Thus, tracker and detection should present simultaneously a similar position, a similar size, a similar colorimetric response, and the detection should resemble to the most likely identities for the mixed-state tracker. Maxima are extracted iteratively with the Hungarian Method [4].

4.1.2 Automatic Tracker Initializations / Terminations

Every temporally recurrent detection, which is not associated to any tracker, yields the instanciation of a new tracker. On a similar manner, every tracker which has not been associated with a detection for a time period longer than the suppression threshold is stopped.

4.2 Mixed-state based Particle Filtering

4.2.1 Prediction Model

Each target initialized on a detection is tracked by a particle filter. Given the identity database, we have extra reference descriptors to compare with. To do so, following [19], we use Mixed-State CONDENSATION filters, introduced in [10]. We aim to estimate a mixte state vector, composed of several continuous terms and a discrete one. $\mathbf{X} = (\mathbf{x}, id)^{\mathsf{T}}$, $\mathbf{x} \in \mathbb{R}^4$, $id \in \{1,\dots,N_{id}\}$ The continuous part of the state $\mathbf{x} = [x, y, v_x, v_y]^{\mathsf{T}}$ is composed of the position in the image plane $(x,y)^{\mathsf{T}}$ and of the speed vector $(v_x, v_y)^{\mathsf{T}}$. The integer part id refers to one of the N_{id} identities in the database. The tracking is conducted in the image plane, and tracking box dimension is updated on the associated detections. The appearance model is also updated on the associated detection. Given this extended state vector, the density of sampling process at image t can be decomposed [10]:

where $T_{ij}(\mathbf{x}_{t-1})$ is the transition probability from identity i to j, applied to the discrete identity parameter, and $p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the sampling applied to the continuous part. The

transition matrix $T = [T_{ij}]$ is built over the set of key-frames. The element T_{ij} is the similarity $w_{id}(.)$ between identities i and j of the database, computed between the most different key-frames. Particles are propagated according to a first order motion model.

4.2.2 Observation Model Integrating Detections

The weight $w_{tr}^{(p)}$ associated with the p-th particle of tracker tr is computed integrating the distance to the associated detection d^*, the colorimetric similarity to the appearance model $w_{App}(.)$ and the colorimetric similarity to the identity of the particle $w_{Id}(.)$. $Id(p)$ represents the identity taken by particle p. This is the discrete parameter of p.

$$w_{tr}^{(p)} = \underbrace{\alpha \cdot \mathcal{I}(tr) \cdot p_{\mathcal{N}}(d^* - p)}_{\text{distance to the detection}} + \underbrace{\beta \cdot w_{App}(d, tr)}_{\text{appearance model}} + \underbrace{\gamma \cdot w_{Id}(d, id(p))}_{\text{identity}} \tag{2}$$

where α, β and γ are weighting coefficients empirically set, and $\mathcal{I}(tr)$ is a boolean signifying the existence or not of an associated detection to the tracker. As in [19], the introduction of similarity relative to the identity in the particle weighting drives the particle cloud towards the most likely identities given the received observations. In that way, each tracker maintains a discrete distribution over the *global identities*, the modes of that distribution being the most likely identities.

The state estimation is a two-stage process. First we compute the Maximum A Posteriori over the discrete parameter relatively to the current observation $\mathbf{Z}_t$ with equation (3), *i.e.* the most likely identity at time step t.

$$\hat{id}_t = \arg\max_j P(id_t = j|\mathbf{Z}_t) = \arg\max_j \sum_{p \in \Upsilon_j} w_{tr}^{(p)}(t), \text{where } \Upsilon_j = \left\{ p | \mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, j) \right\} \tag{3}$$

Then, the continuous components are estimated over the subset of particles $\hat{\Upsilon}$ which have that most likely identity, following equation (4).

$$\hat{\mathbf{x}}_t = \sum_{p \in \hat{\Upsilon}} w_{tr}^{(p)}(t) \cdot \mathbf{x}_t^{(p)} / \sum_{p \in \hat{\Upsilon}} w_{tr}^{(p)}(t), \text{where } \hat{\Upsilon} = \{ p | \mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, \hat{y}_t)^\top \}$$

That way, on top of target image position estimation, each filter provides a discrete identity distribution for its target.

5 Global tracklet association at the network level

5.1 Problem Formulation

Let $Y = \{y_k = (ids_k, t_k^{in}, t_k^{out}, a_k^{in}, a_k^{out}), k = 1, \ldots, K\}$ denote the set of K tracklets generated by the mixed-state filters, where ids_k is the identity distribution, t_k^{in} and t_k^{out} are time of appearance and disappearance and a_k^{in} and a_k^{out} are the areas of appearance and disappearance. Unlike [15], we track pedestrians, *i.e.*without a priori motion, yielding completely unordered duration of visibility. That is why, instead of fixing an unrelevant duration for the sliding window, we wait for the supervisor to have gathered K tracklets before performing the data association search.

Here we define the problem as given the observations Y, inferring N tracks at the network level as composition of tracklets, where N is the number of identities wandering in the

network, known from counting targets coming from feeding areas. Equation (4) summarizes that, where τ_0 is the set of false alarms, τ_n is the n^{th} network track.

$$H = \{\tau_0, \tau_1, \ldots, \tau_N\} \tag{4}$$

Each τ_n in H is defined as a collection of camera tracklets. In our framework, the tracking problem is formulated as maximizing a posterior (MAP) of a tracklet assignment given the set of observations Y :

$$H^* = \arg\max_H (p(H|Y)) \text{ , where } H \sim p(H|Y) \propto p(Y|H)p(H) \tag{5}$$

5.2 Likelihood Model

The likelihood model we propose $p(Y|H)$ is composed by two terms: a topological part and a mixed-state distribution result: $p(Y|H) = \mathcal{P}_{Topo}(Y|H) \cdot \mathcal{P}_{MSR}(Y|H)$

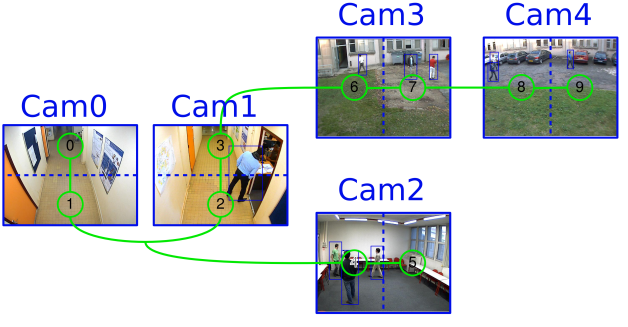

Figure 2: Topological graph of the testing network.

Topological Likelihood: Shortest paths on the graph are computed with Dijsktra algorithm. As the cameras are static, the topology is fixed and these distances can be precomputed offline and stored in a database. Figure (2) displays our private network topology.

$$\mathcal{P}_{Topo}(Y|H) = \prod_{i=1}^{|\tau_n|-1} p_{\mathcal{N}}(d_{topo}(a_{i-1}^{out}, a_i^{in})), \tag{6}$$

where $d_{topo}(.)$ is the distance between two nodes of the topological graph, $a_i^{in/out}$ are the area of beginning (*resp.* ending) of the i-th tracklet, $p_{\mathcal{N}}(.)$ is a gaussian kernel to transform the distance into a similarity between 0 and 1 and $|\tau_n|$ is the cardinal of the tracklet set τ_n.

Identities Distributions: In addition to topologic constraints we add appearance features. However comparing directly descriptors yields an homogeneity problem. Indeed, descriptors taken from the same camera may be more similar than from others, even if the target is different. At that point some papers resort to color calibration, a heavy process to project descriptors from different sensors in the same subspace. We use instead the mixed-state

trackers belief on the tracklet identity, resulting from online comparison with the database generated from the feeding area.

$$\mathcal{P}_{MSR}(Y|H) = \prod_{i=1}^{|\tau_n|-1} ids_i(id), \tag{7}$$

where ids_i is the discrete probability distribution over the identity database for the i-th tracklet. That way $ids_i(id)$ represents the probability that tracklet i has the identity id.

5.3 Topologic and Appearance driven MCMC Data Association

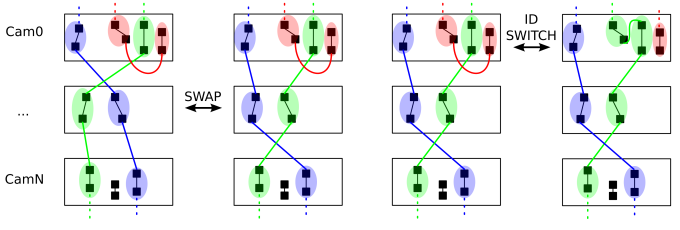

Figure 3: Examples of the Swap and Id Switch moves used for MCMCDA between multiple cameras (temporality is not represented).

We resort to the Metropolis Hastings algorithm to sample from equation (5). By design, a series of reversible proposal moves yields a Markov chain that is irreductible, aperiodic, and that converges to a stationnary distribution by the ergodic theorem. In our case, the stationnary distribution $\pi(H)$ is defined by equation (5.2), and the acceptance ratio for the j-th iteration is computed as

$$p(H_j \leftarrow H^*) = \min\left(\frac{\pi(H^*)q(H_{j-1}|H^*)}{\pi(H_{j-1})q(H^*|H_{j-1})}, 1\right) \tag{8}$$

The proposal distributions $q(H,H')$ consist of two pairs of reversible moves as illustrated in figure (3).

Id Switch Move: In an Id Switch move, one tracklet y_{switch} and one track τ_{new} (different from the track of y_{switch}) are chosen *u.a.r.* That way, y_{switch} goes from one track to another, changing their lengths.

Swap Move: In a Swap move, two tracklets y_i and y_j from different tracks are chosen *u.a.r* and are swapped.

6 Experimental Results

Considering the lack of public datasets in terms of NOFOV network, we evaluate the system first on synthetic data to validate the supervisor part and then the whole system on a real sequence on a NOFOV camera network.

6.1 Tracking Performances

Table 1 presents quantitative results on the PETS'09 sequence. First, we validate our partial implementation of [3] (without HOG + ISM detector, detector confidence use in the observation model, and Boosting Online based appearance model).

However, our approach presents an extra modality with the notion of *global identity*. We show first that the introduction of mixed-state particle filtering does not decrease much tracking performances. To do so, we compare MOTP and MOTA for our implementation without and with the reidentification module activated. Then, this extra modality allow us to compute TRR for the sequence. Finally, we compare the reidentification results of the distributed mixed-state filters alone against the supervised ones. There, exclusivity constraints (section 5) yield improved results. The stochastic aspect of particle filtering has been taken into account in our experiences: table 1 shows results averaged over ten repetitions of tracking.

Table 1: CLEAR MOT metrics tracking results [2] and true reidentification rates on the monocamera sequence PETS'09 S2L1. We give here Multi-Object Tracking Precision (MOTP), Multi-Object Tracking Accuracy (MOTA), and True Reidentification Rate (TRR).

Sequence PETS'09	MOTP	MOTA	TRR
Tracking-by-detection [3]	**56.3%**	**79.7%**	-
Tracking-by-detection implemented	42.7%	77.9%	-
Tracking-by-Reidentification	42.5%	77.7%	59.7%
Tracking-by-Reidentification supervised	42.4%	75.9%	**64%**

6.2 Synthetic Data

We build a network graph and simulate targets random exploration of that graph. At each intersection, the target chooses its destination with equiprobability between every possible destination. The identity vectors are generated such that

$$ids(i) = \begin{cases} \max(1 - abs(\varepsilon_i), 0) & \text{if } i = id_{GT}, \\ \min(abs(\varepsilon_i), 1) & \text{else.} \end{cases}$$

In both cases $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. That way, we build a pool of tracklets $Y_{synth} = \{y_k = (ids_k, t_k^{in}, t_k^{out}, a_k^{in}, a_k^{out}), k = 1, \ldots, K\}$ and we optimize it with our MCMC routine.

Figure 4 presents the results of a synthetic network with 30 cameras, and 10 to 40 identities wandering in it and 1000000 iterations of the Metropolis Hastings algorithm. Each identity appears at most 20 times in the network, yielding at most 20 tracklets.

This part validates the supervisor task. Indeed, with wrong measurements coming from the mixed-state filters (maximum value not on the ground truth identity), topology knowledge allows to correct these errors.

6.3 Real Data issued from our NOFOV Network

The NOFOV sequence presents a total of 12 pedestrians wandering between 5 cameras. Figure (5) gives an overview of the network. The dataset is 4000 frame-long and we plan

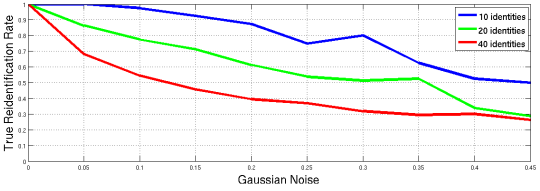

Figure 4: Evolution of the True Reidentification Rate on the MCMC filtering over the synthetic data in function of the gaussian noise on the identity vectors.

to release it publicly. The identity database is built at runtime using models of the trackers starting in Cam0, entry area0. We compare here the method based only on colorimetric information and particle filtering inspired by [19], with the supervised system we propose in Section 5 which optimizes the tracklets with topological constraints.

Table 2 presents true re-identification rates of the supervisor applied to real data optimizing the output of our *tracking-by-reidentification* module. Fusing topological constraints with identities distributions allows our MCMC formulation to increase the assignment quality.

Table 2: True Reidentification Rates for each camera of the sequence NOFOVNetwork: comparison of the approaches without, and with supervisor on the network.

NOFOV Sequence	cam0	cam1	cam2	cam3	cam4
Tracking-by-Reidentification	88.7%	65.3%	58.5%	54.6%	54%
Tracking-by-Reidentification supervised	90.6%	76.2%	68.2%	63.8%	62%

Unlike Kuo *et al*. [13] who perform a Hungarian Method to solve the tracklet assignment at the end of the sequence, our approach provides reidentification at each time-step through the mixed-state framework. This information is optimized when sufficient data is gathered. Where they only precise how to treat a pair of cameras, our approach is directly scalable to multiple sensors. Finally, figure 5 gives an overview of our system output. Left, the cameras of the network display current tracks, and down right the reidentification are displayed.

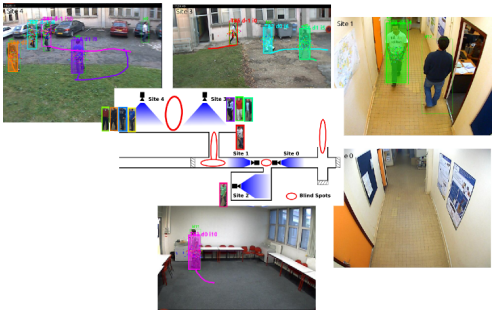

Figure 5: Overview of the monitored network (see the video enclosed as *supplementary material*).

7 Conclusion

In this paper, we present a novel system to perform re-identification on-the-fly while tracking, to monitor NOFOV camera networks. We base our approach on a markovian *tracking-by-detection* and we extend its distributed particle filters to also include a discrete identity parameter in their search space and so re-identify the targeted person. The identity points towards the database of targets present in the network, built from an entry area.

We derive a novel MCMC data association dedicated to NOFOV camera networks. The local identification of the distributed mixed-state filters is fused with coarse topology knowledge to yield the likelihood function of the MCMC data association. This paper is among the first ones to propose a tracking dedicated appearance-based method for person re-identification embedded in the tracking framework to monitor NOFOV camera networks.

References

[1] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.

[2] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*, 2008.

[3] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. *Pattern Analysis and Machine Intelligence*, 2010.

[4] F. Burgeois and J.-C. Lasalle. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 1971.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[7] N. Gheissari, TB Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[9] Timothy Huang and Stuart Russell. Object identification in a bayesian context. In *Int. Joint Conference on Artificial Intelligence*, 1997.

[10] M. Isard and A. Blake. A mixed-state CONDENSATION tracker with automatic model-switching. In *ICCV*, 1998.

[11] M. Isard and A. Blake. BraMBLe: a Bayesian multiple blob tracker. In *ICCV*, 2001.

[12] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *CVPR*, 2005.

[13] C.H. Kuo, C. Huang, and R. Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *ECCV*, 2010.

[14] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *CVPR*, 2004.

[15] Bogdan C. Matei, Harpreet S. Sawhney, and Supun Samarasekera. Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In *CVPR*, 2011.

[16] K. Okuma, A. Taleghani, N. De Freitas, J. Little, and D. Lowe. A boosted particle filter: multitarget detection and tracking. In *ECCV*, 2004.

[17] Hanna Pasula, Stuart J. Russell, Michael Ostland, and Yaacov Ritov. Tracking many objects with many sensors. In *Int. Joint Conference on Artificial Intelligence*, 1999.

[18] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, pages 1200–1207. IEEE, 2009.

[19] XXX. (Supplementary Material). 2011.

[20] W.S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.