# Human Re-identification in Crowd Videos using Personal, Social and Environmental Constraints

Shayan Modiri Assari, Haroon Idrees, and Mubarak Shah

Center for Research in Computer Vision (CRCV),
University of Central Florida (UCF), Orlando
{smodiri, haroon, shah}@cs.ucf.edu

**Abstract.** This paper addresses the problem of human re-identification in videos of dense crowds. Re-identification in crowded scenes is a challenging problem due to large number of people and frequent occlusions, coupled with changes in their appearance due to different properties and exposure of cameras. To solve this problem, we model multiple Personal, Social and Environmental (PSE) constraints on human motion across cameras in crowded scenes. The personal constraints include appearance and preferred speed of each individual, while the social influences are modeled by grouping and collision avoidance. Finally, the environmental constraints model the transition probabilities between gates (entrances / exits). We incorporate these constraints into an energy minimization for solving human re-identification. Assigning $1-1$ correspondence while modeling PSE constraints is NP-hard. We optimize using a greedy local neighborhood search algorithm to restrict the search space of hypotheses. We evaluated the proposed approach on several thousand frames of PRID and Grand Central datasets, and obtained significantly better results compared to existing methods.

**Keywords:** Video Surveillance, Re-identification, Dense Crowds, Social Constraints, Multiple Cameras, Human Tracking

## 1 Introduction

Human re-identification is a fundamental and crucial problem for multi-camera surveillance systems [49, 17]. It involves re-identifying individuals after they leave field-of-view (FOV) of one camera and appear in FOV of another camera (see Fig 1(a)). The investigation process of the Boston Marathon bombing serves to highlight the importance of re-identification in crowded scenes. Authorities had to sift through a mountain of footage from government surveillance cameras, private security cameras and imagery shot by bystanders on smart phones [22]. Therefore, automatic re-identification in dense crowds will allow successful monitoring and analysis of crowded events.

Dense crowds are the most challenging scenario for human re-identification. For large number of people, appearance alone provides a weak cue. Often, people in crowds wear similar clothes that makes re-identification even harder (Fig. 1c). Unlike regular surveillance scenarios previously tackled in literature, we address this problem for thousands of people where at any 30 second interval, hundreds of people concurrently enter a single camera.

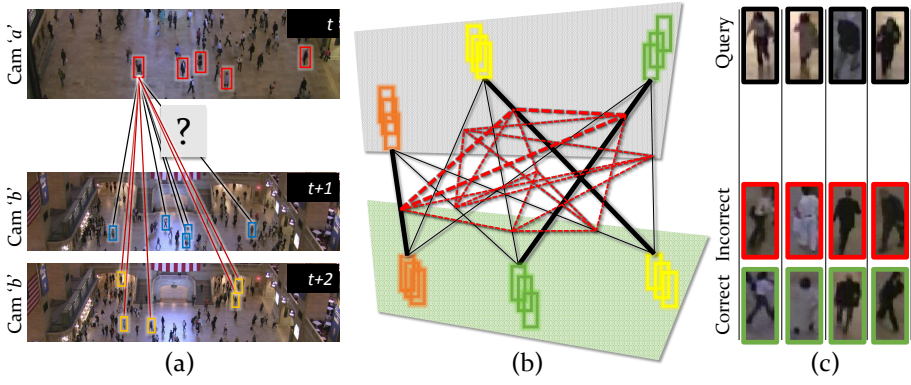



**Fig. 1.** (a) Our goal is to re-identify people leaving camera $a$ at time $t$ (top row) to when they appear in camera $b$ at some time $t+1, t+2, \ldots$ in the future. The invisible region between the cameras is not closed, which means people can leave one camera and never appear in the other camera. (b) We construct a graph *between individuals* in the two cameras, as shown with black lines. Some of the constraints are linear in nature (appearance, speed, destination) while others are quadratic (spatial and social grouping, collision avoidance). The quadratic constraints are shown in red and capture relationships *between matches*. In (c), the people in black boxes are from camera $a$, while the other two rows shows people with similar appearance from camera $b$. The red boxes indicate the best matches (using appearance) which are incorrect, and green boxes show the low-scoring correct matches. This highlights that crowded scenes make human re-identification across cameras significantly difficult.

Traditionally, re-identification has been primarily concerned with matching static snapshots of people from multiple cameras. Although there have been few works that modeled social effects for re-identification such as grouping behavior [58, 5, 4], they mostly deal with static images. In this paper, we study the use of time and video information for this task, and propose to consider the dynamic spatio-temporal context of individuals and the environment to improve the performance of human re-identification. We complement appearance with multiple personal, social and environmental (PSE) constraints, many of which are applicable without knowledge of camera topology. The PSE constraints include *preferred speed and destination*, as well as *social grouping* and *collision avoidance*. The environmental constraints are modeled by learning the repetitive patterns that occur in surveillance networks, as individuals exiting camera from a particular location (gate) are likely to enter another camera from another specific location. These happen both as soft (*spatial grouping*) and hard constraints (*transition probabilities*). The PSE constraints that are linear in nature, i.e. occur between objects, are shown with black lines in Fig. 1(b), while quadratic ones occur between matching hypotheses, i.e., pairs of objects, are shown with red lines in Fig. 1(b). Thus, if there are $N_a$ and $N_b$ number of people in two cameras, then the total number of possible matching hypotheses is $N_a N_b$, and there are $(N_a N_b)^2$ possible quadratic hypotheses. The time limits naturally reduce some of the hypotheses, nonetheless for large number of people these can be overwhelming. Since the proposed PSE constraints are both lin-

ear and quadratic in nature, we employ a greedy local neighborhood search algorithm to optimize the resulting objective function simultaneously for all people. Thus, in addition to producing rankings for different queries, our method also outputs the more useful $1-1$ correspondences for individuals.

To the best of our knowledge, this is the first paper to address human re-identification using personal, social and environmental constraints in dense crowds. The evaluation is performed on two datasets, PRID [19] and the challenging Grand Central dataset [53] which depicts dense crowds [1]. The rest of the paper is organized as follows. We discuss related work in Sec. 2, and present the proposed approach in Sec. 3. The results of our experiments are reported in Sec. 4, and we conclude with some directions for future research in Sec. 5.

## 2 Related Work

Our approach is at the crossroads of human re-identification in videos, dense crowd analysis and social force models. Next, we provide a brief literature review of each of these areas.

**Person Re-identification** is an active area of research in computer vision, with some of the recent works including [27, 28, 37, 7, 57, 54, 55, 1, 29] applicable to static images. In videos, several methods have been developed for handing over objects across cameras [49, 20, 45, 6, 10]. Most of them focus on non-crowd surveillance scenarios with emphasis on modeling color distortion and learning brightness transfer functions that relate different cameras [39, 40, 21, 16], others relate objects by developing illumination-tolerant representations [31] or comparing possible matches to a reference set [9]. Similarly, Kuo *et al.* [24] used Multiple Instance Learning to combine complementary appearance descriptors.

The spatio-temporal relationships across cameras [32, 46, 47] or prior knowledge about topology has been used for human re-identification. Chen *et al.* [8] make use of prior knowledge about camera topology to adaptively learn appearance and spatio-temporal relationships between cameras, while Mazzon *et al.* [34] use prior knowledge about relative locations of cameras to limit potential paths people can follow. Javed *et al.* [20] presented a two-phase approach where transition times and exit/entrance relationships are learned first, which are later used to improve object correspondences. Fleuret [14] predicted occlusions with a generative model and a probabilistic occupancy map. Dick and Brooks [11] used a stochastic transition matrix to model patterns of motion within and across cameras. These methods have been evaluated on non-crowd scenarios, where observations are sparse and appearance is distinctive. In crowded scenes, hundreds of people enter a camera simultaneously within a small window of few seconds, which makes learning transition times during an unsupervised training period virtually impossible. Furthermore, our approach is applicable whether or not the information about camera topology is available.

**Dense Crowds** studies [3, 59, 60] have shown that walking behavior of individuals in crowds is influenced by several constraints such as entrances, exits, boundaries, obstacles; as well as preferred speed and destination, along with interactions with other

[1] Data and ground truth available at: http://crcv.ucf.edu/projects/Crowd-Reidentification

pedestrians whether moving [35, 15] or stationary [53]. Wu *et al.* [51] proposed a two-stage network-flow framework for linking tracks interrupted by occlusions. Alahi *et al.* [2] identify origin-destination (OD) pairs using trajectory data of commuters which is similar to grouping. In contrast, we employ several PSE constraints besides social grouping.

**Social Force Models** have been used for improving tracking performance [25, 38, 52]. Pellegrini *et al.* [38] were the first to use social force models for tracking. They modeled collision avoidance, desired speed and destination and showed its application for tracking. Yamaguchi *et al.* [52] proposed a similar approach using a more sophisticated model that tries to predict destinations and groups based on features and classifiers trained on annotated sequences. Both methods use agent-based models and predict future locations using techniques similar to crowd simulations. They are not applicable to re-identification, as our goal is not to predict but to associate hypotheses. Therefore, we use social and contextual constraints for re-identification in an offline manner. Furthermore, both these methods require observations to be in metric coordinates, which for many real scenarios might be impractical.

For re-identification in static images, group context was used by Zheng *et al.* [58, 17], who proposed ratio-occurrence descriptors to capture groups. Cai *et al.* [5] use covariance descriptor to match groups of people, as it is invariant to illumination changes and rotations to a certain degree. For re-identifying players in group sports, Bialkowski *et al.* [4] aid appearance with group context where each person is assigned a role or position within the group structure of a team. In videos, Qin *et al.* [41] use grouping in non-crowded scenes to perform hand over of objects across cameras. They optimize track assignment and group detection in an alternative fashion. On the other hand, we refrain from optimizing over group detection, and use multiple PSE constraints (speed, destination, social grouping etc.) for hand over. We additionally use group context in space, i.e., objects that take the same amount of time between two gates are assigned a cost similar to grouping, when in reality they may not be traveling together in time. Mazzon and Cavallaro [33] presented a modified social force multi-camera tracker where individuals are attracted towards their goals, and repulsed by walls and barriers. They require a surveillance site model beforehand and do not use appearance. In contrast, our formulation avoids such assumptions and restrictions.

In summary, our approach does not require any prior knowledge about the scene nor any training phase to learn patterns of motion. Ours is the first work to incorporate multiple personal, social and environmental constraints simultaneously for the task of human re-identification in crowd videos.

## 3 Framework for Human Re-identification in Crowds

In this section, we present our approach to re-identify people using PSE constraints. Since transition probabilities between gates are not known a priori, we estimate correspondences and transition probabilities in an alternative fashion.

Let $O_{i_a}$ represent an observation of an object $i$ in camera $a$. Its trajectory is given by a set of points $[\mathbf{p}_{i_a}(t^{\eta}_{i_a}), \ldots, \mathbf{p}_{i_a}(t^{\chi}_{i_a})]$, where $t^{\eta}_{i_a}$ and $t^{\chi}_{i_a}$ represent the time it entered and exited the camera $a$, respectively. Given another observation of an object $j$ in camera $b$, $O_{j_b}$, a possible match between the two is denoted by $M^{j_b}_{i_a} = \langle O_{i_a}, O_{j_b} \rangle$. To

simplify notation, we drop the symbol for time $t$ and use it only when necessary, thus, $\mathbf{p}_{i_a}^{\chi} \equiv \mathbf{p}_{i_a}(t_{i_a}^{\chi})$ and $\mathbf{p}_{j_b}^{\eta} \equiv \mathbf{p}_{j_b}(t_{j_b}^{\eta})$.

The entrances and exits in each camera are divided into multiple gates. For the case of two cameras $a$ and $b$, the gates (locations) are given by $\mathbf{G}_{1_a}, \ldots, \mathbf{G}_{U_a}$ and $\mathbf{G}_{1_b}, \ldots, \mathbf{G}_{U_b}$, where $U_a$ and $U_b$ are the total number of gates in both cameras, respectively. Furthermore, we define a function $g(\mathbf{p}(t))$, which returns the nearest gate when given a point in the camera. For instance, for a person $i_a$, $g(\mathbf{p}_{i_a}^{\chi})$ returns the gate from which the person $i$ exited camera $a$, by computing the distance of $\mathbf{p}_{i_a}^{\chi}$ to each gate. Mathematically, this is given by:

$$g(\mathbf{p}_{i_a}^{\chi}) = \arg\min_{\mathbf{G}_{u_a}} \|\mathbf{G}_{u_a} - \mathbf{p}_{i_a}^{\chi}\|^2, \quad \forall u_a = 1, \ldots, U_a. \tag{1}$$

To compute appearance similarity, $\phi_{\text{app}}(O_{i_a}, O_{j_b})$, between observations $O_{i_a}$ and $O_{j_b}$, we use features from Convolutional Neural Networks [44]. In particular, we extract features from Relu6 and Fc7 layers, followed by homogenous kernel mapping [48] and linear kernel as the the similarity metric. Next, we describe the costs for different PSE constraints, $\phi(.)$, employed in our framework for re-identification. Since all costs have their respective ranges, we use a sigmoid function, $\hat{\phi}(.) = (1 + \exp(-\beta\phi(.)))^{-1}$, to balance them. Most of the constraints do not require knowledge about camera topology, and are described below.

### 3.1 PSE Constraints without Camera Topology

**Preferred Speed:** The walking speed of individuals has been estimated to be around 1.3 m/s [42]. Since, we do not assume the availability of metric rectification information, we cannot use this fact directly in our formulation. However, a consequence of this observation is that we can assume the walking speed of individuals, *on average*, in different cameras is constant. We assume a Normal distribution, $\mathcal{N}(.)$, on observed speeds in each camera. The variation in walking speeds of different individuals is captured by the variance of the Normal distribution. Let $\mathcal{N}(\mu_a, \sigma_a)$ and $\mathcal{N}(\mu_b, \sigma_b)$ denote the distribution modeled in the two cameras. Since a particular person is being assumed to walk with the same speed in different cameras, the cost for preferred speed using the exit speed of person $i_a$, $\dot{\mathbf{p}}_{i_a}^{\chi}$, and the entrance speed of person $j_b$, $\dot{\mathbf{p}}_{j_b}^{\eta}$ is given by:

$$\dot{\mathbf{p}}_{i_a}^{\chi} = \sigma_a^{-1}(\|\mathbf{p}_{i_a}^{\chi} - \mathbf{p}_{i_a}^{\chi-1}\| - \mu_a), \quad \dot{\mathbf{p}}_{j_b}^{\eta} = \sigma_b^{-1}(\|\mathbf{p}_{j_b}^{\eta+1} - \mathbf{p}_{j_b}^{\eta}\| - \mu_b), \tag{2}$$

$$\phi_{\text{spd}}(O_{i_a}, O_{j_b}) = |\dot{\mathbf{p}}_{i_a}^{\chi} - \dot{\mathbf{p}}_{j_b}^{\eta}|. \tag{3}$$

**Destination:** For re-identification in multiple cameras, the knowledge about destination gives a prior for an individual's location in another camera. Since individuals cannot be observed between cameras, we capture the common and frequent patterns of movement between gates in different cameras by modeling the transition probabilities between gates in those cameras. Assuming we have a set of putative matches $\{M_{i_a}^{j_b}\}$, we estimate the probability of transition between exit gate $G_{u_a}$ and entrance gate $G_{u_b}$ as:

$$p(G_{u_a}, G_{u_b}) = \frac{|g(\mathbf{p}_{i_a}^{\chi}) = G_{u_a} \wedge g(\mathbf{p}_{j_b}^{\eta}) = G_{u_b}|}{|g(\mathbf{p}_{i_a}^{\chi}) = G_{u_a} \wedge \sum_{u'_b, j'_b} g(\mathbf{p}_{j'_b}^{\eta}) = G_{u'_b}|}. \tag{4}$$

Thus, the cost for transition between gates for the match $\langle O_{i_a}, O_{j_b} \rangle$ is given by:

$$\phi_{\text{tr}}(O_{i_a}, O_{j_b}) = 1 - p\big(g(\mathbf{p}^{\chi}_{i_a}), g(\mathbf{p}^{\eta}_{j_b})\big). \tag{5}$$

**Spatial Grouping:** The distance traveled by different individuals between two points (or gates) across cameras should be the same. Since the camera topology is not available in this case, the distance can be implicitly computed as a product of velocity and time. This is a quadratic cost computed between every two possible matches, $M^{j_b}_{i_a}$ and $M^{j'_b}_{i'_a}$, given by:

$$\begin{aligned}\varphi_{\text{spt}}(M^{j_b}_{i_a}, M^{j'_b}_{i'_a}) = \exp(-|\mathbf{p}^{\chi}_{i_a} - \mathbf{p}^{\chi}_{i'_a}|) \cdot \exp(-|\mathbf{p}^{\eta}_{j_b} - \mathbf{p}^{\eta}_{j'_b}|) \\ \cdot |(\dot{\mathbf{p}}^{\chi}_{i_a} + \dot{\mathbf{p}}^{\eta}_{j_b})(t^{\eta}_{j_b} - t^{\chi}_{i_a}) - (\dot{\mathbf{p}}^{\chi}_{i'_a} + \dot{\mathbf{p}}^{\eta}_{j'_b})(t^{\eta}_{j'_b} - t^{\chi}_{i'_a})|.\end{aligned} \tag{6}$$

Effectively, if the exit and entrance locations are nearby (the first two terms in Eq. 6), then we compute the distance traveled by each match in the pair using the product of mean velocity and time required to travel between those locations (the third term). It is evident from Eq. 6 that the exponentiation in first two terms will allow this cost to take effect only when the entrance and exit locations are both proximal. If so, the third term will then measure the difference in distance traveled by the two possible matches (tracks), and penalize using that difference. If the distance is similar, the cost will be low suggesting both matches (tracks) should be included in the final solution. If the difference is distance is high, then at least one or both of the matches are incorrect.

**Social Grouping:** People tend to walk in groups. In our formulation, we reward individuals in a *social group* that exit and enter together from the same locations at the same times,

$$\varphi_{\text{grp}}(M^{j_b}_{i_a}, M^{j'_b}_{i'_a}) = \exp(-|\mathbf{p}^{\chi}_{i_a} - \mathbf{p}^{\chi}_{i'_a}| - |\mathbf{p}^{\eta}_{j_b} - \mathbf{p}^{\eta}_{j'_b}| - |t^{\eta}_{j_b} - t^{\eta}_{j'_b}| - |t^{\chi}_{i_a} - t^{\chi}_{i'_a}|). \tag{7}$$

Here, the first two terms capture the difference in exit and entrance locations, respectively, and the third and fourth terms capture the difference in exit and entrance times, respectively.

### 3.2 Optimization with PSE Constraints

In this subsection, we present the optimization technique which uses the aforementioned constraints. Let $z^{j_b}_{i_a}$ be the variable corresponding to a possible match $M^{j_b}_{i_a}$. Our goal is to optimize the following loss function over all possible matches, which is the weighted sum of linear and quadratic terms:

$$\begin{aligned}L = \sum_{i_a, j_b} z^{j_b}_{i_a} \underbrace{\big(\hat{\phi}_{\text{app}}(M^{j_b}_{i_a}) + \alpha_{\text{spd}}\hat{\phi}_{\text{spd}}(M^{j_b}_{i_a}) + \alpha_{\text{tr}}\hat{\phi}_{\text{tr}}(M^{j_b}_{i_a})\big)}_{\text{Linear Terms}} \\ + \sum_{\substack{i_a, j_b \\ i'_a, j'_b}} z^{j_b}_{i_a} z^{j'_b}_{i'_a} \underbrace{\Big(\alpha_{\text{spt}}\hat{\varphi}_{\text{spt}}(M^{j_b}_{i_a}, M^{j'_b}_{i'_a}) + \alpha_{\text{grp}}\hat{\varphi}_{\text{grp}}(M^{j_b}_{i_a}, M^{j'_b}_{i'_a})\Big)}_{\text{Quadratic Terms}},\end{aligned} \tag{8}$$

**Algorithm 1** : Algorithm to find $1-1$ correspondence between persons observed in different cameras using both linear and quadratic constraints.

**Input**: $O_{i_a}, O_{j_b}$ $\forall i_a, j_b, R$ (# steps)
**Output**: $L^*, \mathbf{z}^*$; $0 \leq |t^{\eta}_{j_b} - t^{\chi}_{i_a}| \leq \tau, \forall z^{j_b}_{i_a}$

**procedure** RE-IDENTIFY()
  Initialize $[L^*, \mathbf{z}^*]$ for Linear Constraints with MUNKRES [36] ▷ Initial solution
  **while** $L^*$ improves **do**
    **for** $r = 0$ to $R$ **do**
      $[L^-, \mathbf{z}^-] = \text{REMOVEMAT}(L^*, \mathbf{z}^*, r)$ ▷ Probabilistically remove $r$ matches
      $L' = L^-, \mathbf{z}' = \mathbf{z}^-$ ▷ Consider it the new solution
      **for** $s = r + 1$ to $1$ **do**
        $[L^+, \mathbf{z}^+] = \text{ADDMAT}(L', \mathbf{z}', s)$ ▷ Add $s$ new matches to the solution
        **if** $L' > L^+$ **then** ▷ Is the solution after adding new matches better?
          $L' = L^+, \mathbf{z}' = \mathbf{z}^+$ ▷ If so, update it as the new solution
        **end if**
      **end for**
      **if** $L^* > L'$ **then** ▷ Is the new solution better the best solution so far?
        $L^* = L', \mathbf{z}^* = \mathbf{z}'$ ▷ If so, update it as the best solution
      **end if**
    **end for**
  **end while**
**end procedure**

subject to the following conditions:

$$\sum_{i_a} z^{j_b}_{i_a} \leq 1, \forall j_b, \sum_{j_b} z^{j_b}_{i_a} \leq 1, \forall i_a, z^{j_b}_{i_a} \in \{0, 1\}. \tag{9}$$

Since the transition probabilities in Eq. 4 are not known in advance, we propose to use an EM-like approach that iterates between solving $1-1$ correspondences using the linear and quadratic constraints, and estimating transition information using those correspondences. Furthermore, due to the binary nature of variables, the problem of finding $1-1$ correspondences using PSE constraints is NP-hard. We use a local neighborhood search algorithm presented in Alg. 1 which optimizes Eq. 8 subject to the conditions in Eq. 9. The solution is initialized for linear constraints with Munkres [36]. The sub-procedure $\text{REMOVEMAT}(L, \mathbf{z}, r)$ removes $r$ hypotheses from the solution as well as their respective linear and quadratic costs by assigning probabilities (using respective costs) for each node in the current $z$. In contrast, the sub-procedure $\text{ADDMAT}(L, \mathbf{z}, s)$ adds new hypotheses to the solution using the following steps:

- Populate a list of matches for which $z^{j_b}_{i_a}$ can be 1 such that Eq. 9 is satisfied.
- Make combinations of $s$-lets using the list.
- Remove combinations which dissatisfy Eq. 9.
- Compute new $L$ in Eq. 8 for each combination. This is efficiently done by adding $|\mathbf{z}| * s$ quadratic values and $s$ linear values.
- Pick the combination with lowest loss $L$. Add $s$-let to $\mathbf{z}$ and return.

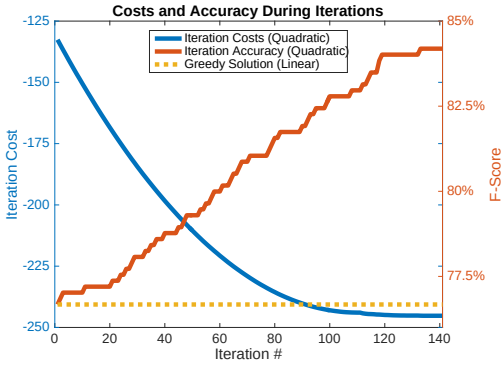



**Fig. 2.** The graph shows the performance of Algorithm 1 using both linear and quadratic constraints, compared against Hungarian Algorithm [36] using only the linear costs shown with orange dotted line. The loss function in Eq. 8 is shown in blue, whereas the accuracy is shown in red. Quadratic PSE constraints in conjunction with Alg. 1 yield an improvement of $\sim 8\%$ over linear constraints.

Algorithm 1 updates the solution when there is a decrease in the loss function in Eq. 8, as can be seen from Line 13. Once the change in loss is negligible, the algorithm stops and returns the best solution obtained. Fig. 2 shows the results quantified for our approach using Alg. 1. The $x$-axis is the step number, whereas the left $y$-axis shows the value of loss function in Eq. 8 (blue curve), and the right $y$-axis shows the F-Score in terms of correct matches (orange curve). We also show results of Hungarian Algorithm (Munkres) [36] in dotted orange line using linear constraints, which include appearance and speed similarity. These curves show that Alg. 1 simultaneously improves the loss function in Eq. 8 and the accuracy of the matches as the number of steps increases.

### 3.3 PSE Constraints with Camera Topology

The PSE constraints presented in the previous section are applicable when the spatial relations between the cameras are not known. However, if the inter-camera topology is available, then it can be used to infer the motion of people as they travel in the invisible or unobserved regions between the cameras. The quality of paths in the invisible region can be subject to constraints such as *preferred speed* or *direction of movement*, which can be quantified and introduced into the framework. Furthermore, collision avoidance is another social constraint that can only be applied when inter-camera topology is known.

Given two objects in cameras $a$ and $b$, $O_{i_a}$ and $O_{i_b}$, in the same reference of time, we predict the possible path between the objects. This is obtained by fitting a spline, given by $\boldsymbol{\gamma}_{i_a}^{j_b}$, in both $x$ and $y$ directions using cubic interpolation between the points $\mathbf{p}_{i_a}$ and $\mathbf{p}_{j_b}$ parameterized with their respective time stamps.

**Collision Avoidance:** Let the point of closest approach between two paths be given by:

$$d(\boldsymbol{\gamma}_{i_a}^{j_b}, \boldsymbol{\gamma}_{i'_a}^{j'_b}) = \min_{\max(t_{i_a}^{\chi}, t_{i'_a}^{\chi}), \ldots, \min(t_{j_b}^{\eta}, t_{j'_b}^{\eta})} \|\boldsymbol{\gamma}_{i_a}^{j_b}(t) - \boldsymbol{\gamma}_{i'_a}^{j'_b}(t)\|, \tag{10}$$

we quantify the collision avoidance as a quadratic cost between pairs of possible matches:

$$\phi_{\text{invColl}}(M_{i_a}^{j_b}, M_{i'_a}^{j'_b}) = \big(1 - \varphi_{\text{grp}}(M_{i_a}^{j_b}, M_{i'_a}^{j'_b})\big). \exp\big(-d(\boldsymbol{\gamma}_{i_a}^{j_b}, \boldsymbol{\gamma}_{i'_a}^{j'_b})\big). \tag{11}$$

Since people avoid collisions with others and change their paths, this is only applicable to trajectories of people who are not traveling in a group, i.e., the cost will be high if two people not walking in a group come very close to each other when traveling through the invisible region between the cameras.

**Speed in Invisible Region:** The second constraint we compute is an improved version of the *preferred speed* - a linear constraint which now also takes into account the direction is addition to speed of the person in the invisible region. If the velocity of a person within visible region in cameras and while traveling through the invisible region is similar, this cost would be low. However, for an incorrect match, the difference between speed in visible and invisible regions will be high. Let $\dot{\boldsymbol{\gamma}}$ denote the velocity at respective points in the path, both in the visible and invisible regions. Then, the difference of maximum and minimum speeds in the entire trajectory quantifies the quality of a match, given by,

$$\phi_{\text{invSpd}}(O_{i_a}, O_{j_b}) = |\max_{t_{i_a}^{\eta} \ldots t_{j_b}^{\chi}} \dot{\boldsymbol{\gamma}}_{i_a}^{j_b}(t) - \min_{t_{i_a}^{\eta} \ldots t_{j_b}^{\chi}} \dot{\boldsymbol{\gamma}}_{i_a}^{j_b}(t)|. \tag{12}$$

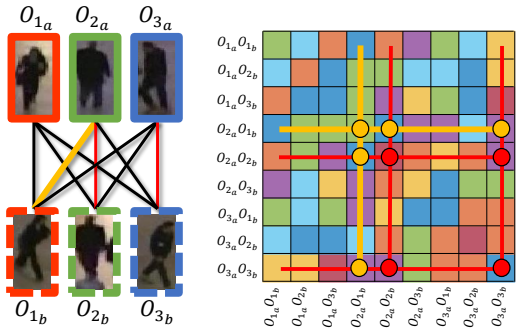
When the inter-camera topology is available, these constraints are added to the Eq. 8 and the method described in the Sec. 3.2 is used to re-identify people across cameras.

## 4 Experiments

Since PSE constraints depend on time and motion information in the videos, many commonly evaluated datasets such as VIPeR [18] and ETHZ [12] cannot be used for computing PSE constraints. We evaluate the proposed approach on the PRID dataset [19] and the challenging Grand Central Dataset [53]. First, we introduce the datasets and the ground truth that was generated for evaluation, followed by detailed analysis of our approach as well as contribution of different personal, social and environmental (PSE) constraints to the overall performance.

### 4.1 Datasets and Experimental Setup

**PRID 2011** is a camera network re-identification dataset containing 385 pedestrians in camera '$a$' and 749 pedestrians in camera '$b$'. The first 200 pedestrians from each camera form the ground truth pairs while the rest appear in one camera only. The most common evaluation method on this dataset is to match people from cam '$a$' to the ones in cam '$b$'. We used the video sequences and the bounding boxes provided by the authors of [19] so we can use the PSE constraints in our evaluation. Since the topology of the scene is unknown, we have used the constraints which do not need any prior knowledge about the camera locations. We evaluated on the entire one hour sequences and extract visual features in addition to various PSE constraints. In accordance with previous methods, we evaluate our approach by matching the 200 people in cam '$a$' to 749 people in cam '$b$' and quantify the ranking quality of matchings.




(a) Linear Costs (b) Quadratic Costs

**Fig. 3.** This figure illustrates the CMC evaluation procedure with quadratic constraints. Given object tracks in the two cameras $O_{1_a}, O_{2_a}, O_{3_a}$ and $O_{1_b}, O_{2_b}, O_{3_b}$, (a) the linear constraints are computed between objects, and (b) quadratic constraints between each possible pair of matches. Adding a new match (shown with amber) requires adding one linear value and number of quadratic values equal to the size of current solution.

**Grand Central** is a dense crowd dataset that is particularly challenging for the task of human re-identification. The dataset contains $120,000$ frames, with a resolution of $1920 \times 1080$ pixels. Recently, Yi *et al.* [53] used a portion of the dataset for detecting stationary crowd groups. They released annotations for trajectories of $12,684$ individuals for $6,000$ frames at $1.5$ fps. We rectified the perspective distortion from the camera and put bounding boxes at correct scales using the trajectories provided by [53]. However, location of annotated points were not consistent for any single person, or across different people. Consequently, we manually adjusted the bounding boxes for $1,500$ frames at $1.5$ fps, resulting in ground truth for $17$ minutes of video data.

We divide the scene into three horizontal sections, where two of them become separate cameras and the middle section is treated as invisible or unobserved region. The locations of people in each camera are in independent coordinate systems. The choice of dividing the scene in this way is meaningful, as both cameras have different illuminations due to external lighting effects, and the size of individuals is different due to perspective effects. Furthermore, due to the wide field of view in the scene, there are multiple entrances and exits in each camera, so that a person exiting the first camera at a particular location has the choice of entering from multiple different locations. Figure 1(c) shows real examples of individuals from the two cameras and elucidates the fact that due to the low resolution, change in brightness and scale, the incorrect nearest neighbors matches using the appearance features often rank much better than the correct ones for this dataset.

**Parameters:** Since there are multiple points / zones of entrances and exits, we divide the boundaries in each camera into $U_a = U_b = 11$ gates. The weights used in Eq. 8 are approximated using grid search on a separate set and then used for both datasets. They are $\alpha_{\text{spt}} = \alpha_{\text{invColl}} = .2$, $\alpha_{\text{tr}} = 1$, and $\alpha_{\text{spd}} = \alpha_{\text{invSpd}} = -\alpha_{\text{grp}} = 5$. Note that, social grouping is rewarded in our formulation, i.e. people who enter and exit together in space and time are more likely to be correct matches when re-identifying people across cameras.

**Table 1.** This table presents the quantitative results of the proposed approach and other methods on the **Grand Central Dataset**.

| Method | CMC | | | | | | F-Score |
|---|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | Rank-20 | Rank-50 | AUC (1:100) | (1-1) |
| Random | 1.83% | 5.48% | 11.36% | 21.91% | 54.36% | 51.00% | 6.90% |
| LOMO-XQDA [28] | 4.06% | 12.37% | 21.91% | 39.76% | 71.40% | 63.81% | 11.16% |
| SDALF [13] | 6.09% | 16.23% | 23.12% | 40.16% | 68.56% | 63.01% | 20.69% |
| SAM [2] | 6.09% | 27.18% | 42.60% | 51.72% | 74.44% | 69.60% | 26.98% |
| eSDC-knn [56] | 11.36% | 27.38% | 38.34% | 50.71% | 74.44% | 69.49% | 30.43% |
| Manifold Learning (Ln) [30] | 7.71% | 24.54% | 36.71% | 54.97% | 78.09% | 72.11% | 30.83% |
| Manifold Learning (Lu) [30] | 10.55% | 34.08% | 48.68% | 66.53% | 87.83% | 80.50% | 32.66% |
| CNN Features [44] | 12.98% | 32.45% | 44.62% | 62.07% | 83.77% | 77.79% | 41.99% |
| **CrowdPSE (w/o topology)** | **25.56%** | **81.54%** | **93.31%** | **97.57%** | **98.38%** | **95.80%** | **67.94%** |
| **CrowdPSE (w/ topology)** | **49.29%** | **95.13%** | **98.17%** | **98.17%** | **98.17%** | **97.31%** | **84.19%** |

### 4.2 Evaluation Measures

Cumulative Matching Characteristic (CMC) curves are typically used evaluating performance of re-identification methods. For each person, all the putative matches are ranked according to similarity scores, i.e. for each person $O_{i_a}$, the cost of assignment $M_{i_a}^{j_b} = \langle O_{i_a}, O_{j_b} \rangle$ is calculated for every possible match to $O_{j_b}$. Then, the accuracy over all the queries is computed for each rank. Area Under the Curve (AUC) for CMC gives a single quantified value over different ranks and an evaluation for overall performance. The advantage of CMC is that it does not require $1-1$ correspondence between matches, and is the optimal choice for evaluating different cost functions or similarity measures.

The CMC curves are meaningful only for linear constraints. Unlike linear constraints which penalize or reward matches (pair of objects), quadratic constraints penalize or reward pairs of matches. Figure 3 illustrates the idea of quantifying both linear and quadratic costs through CMC, since this measure quantifies quality of costs independent of optimization. Given three objects $O_{1_a}, O_{2_a}, O_{3_a}$ and $O_{1_b}, O_{2_b}, O_{3_b}$ in cameras $a$ and $b$, respectively, the black lines in Fig. 3 (a) show linear constraints / matchings. Let us assume we intend to evaluate quadratic constraints for the match between $O_{1_a}$ and $O_{2_b}$. For this, we assume that all other matches are correct (red lines), and proceed with adding relevant quadratic (Fig. 3) and linear costs. For evaluating match between $O_{1_a}$ and $O_{2_b}$, we add linear costs between them, as well as quadratic costs between other matches (shown with red circles in Fig. 3(b)), and pair-wise costs of the match under consideration with all other matches (shown with orange circles). This is repeated for all possible matches. Later, the matches are sorted and evaluated similar to standard CMC. Note that, this approach gives an optimization-independent method of evaluating quadratic constraints. Nonetheless, the explicit use of ground truth during evaluation of quadratic constraints makes them only comparable to other quadratic constraints.

To evaluate $1-1$ correspondence between matches, we use F-score which is defined as $2 \times (\mathsf{precision} \times \mathsf{recall}) / (\mathsf{precision} + \mathsf{recall})$ on the output of optimization. We used Hungarian Algorithm (Munkres) [36] for comparison as it provides a globally optimal solution for linear costs. For the proposed PSE constraints, we use Alg. 1 since we use both linear and quadratic costs.

**Table 2.** This table presents the quantitative results of the proposed approach and other methods on the **PRID Dataset**. We report accuracy (number of correct matches), values of Cumulative Matching Characteristic curves at ranks $1, 5, 10$, $20$ and $50$. As can be seen, the proposed approach outperforms existing methods.

| Method | CMC | | | | |
|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | Rank-20 | Rank-50 |
| KissME [23] + Reranking [26] | 8.00% | 19.00% | 30.00% | 41.00% | 57.00% |
| LMNN [50] + Reranking [26] | 10.00% | 24.00% | 34.00% | 44.00% | 61.00% |
| Mahalanobis [43] + Reranking [26] | 11.00% | 29.00% | 37.00% | 46.00% | 60.00% |
| Non-linear ML [37] | 17.90% | 39.50% | 50.00% | 61.50% | - |
| Desc+Disc [19] | 19.18% | 41.44% | 52.10% | 66.56% | 84.51% |
| **CrowdPSE (w/o topology)** | **21.11%** | **46.65%** | **59.98%** | **76.63%** | **98.81%** |

### 4.3 Results and Comparison

In Table 1, we present the results on Grand Central dataset of our approach using PSE constraints and optimization in Alg. 1 with several baselines. We report accuracy (number of correct matches), values of Cumulative Matching Characteristic curves at ranks $1, 5, 10$, $20$ and $50$, as well as Area Under the Curve (AUC) for CMC between ranks $1$ and $100$. The values of CMC are computed before any optimization. The last column shows the F-Score of $1-1$ assignments post optimization. In Table 1, the first row shows the results of random assignment, whereas next seven rows show results using several re-identification methods. These include LOMO-XQDA [28], SDALF [13], SAM [2], eSDC-knn [56], Manifold Learning [30] - normalized (Ln) and unnormalized (Lu), as well as CNN features [44] which use VGG-19 deep network. Finally, the last two rows show the results of our approach both for the case when camera topology is not known and when it is known. These results show that PSE constraints - both linear and quadratic - significantly improve the performance of human re-identification especially in challenging scenarios such as dense crowds.

Next, we present results on PRID dataset in Table 2. The first three rows show Reranking [26] on KissME [23], LMNN [50], and Mahalanobis distance learning [43] for re-identification. Next two rows show the performance of non-linear Metric Learning [37] and Descriptive & Discriminative features [19]. The last row shows the performance of our method which is better than existing unsupervised approaches for human re-identification. For this dataset, the spatial grouping did not improve the results since the dataset captures a straight sidewalk and does not involve decision makings and different travel times between different gates.

### 4.4 Contribution of Different PSE Constraints

We performed several experiments to gauge the performance of different PSE constraints and components of the proposed approach on Grand Central dataset. The comparison of different constraints using Cumulative Matching Characteristics (CMC) is shown in Figure 4. In this figure, the $x$-axis is the rank, while $y$-axis is accuracy with corresponding rank on $x$-axis. First, we show the results of randomly assigning objects between cameras (blue curve). Then, we use appearance features (Convolutional Neural Network) for re-identification and do not use any personal, social or environmental

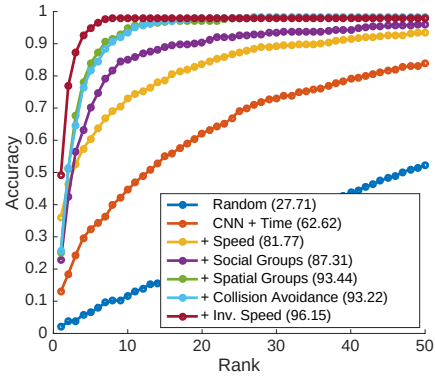



**Fig. 4.** This graph shows the CMC for different PSE constraints proposed in this paper on Grand Central Dataset. The results of random assignment are shown with blue curve, while appearance features with time limits yield the orange curve. Incorporating personal constraint such as preferred speed (amber), and social constraints such as social and spatial grouping (purple and green, respectively) further improve the performance. Given the topology, we can additionally incorporate collision avoidance (light blue) and preferred speed in the invisible region (maroon), which gives the best performance.

constraints (shown with orange curve), which we also use to compute the appearance similarity for our method. The low performance highlights the difficult nature of this problem in crowded scenes. Next, we introduce linear constraint of preferred speed shown with amber curve which gives an improvement of $\sim 19\%$ in terms of Area under the Curve of CMC between ranks $1$ and $50$. Then, we add quadratic constraints of grouping, both of which make an improvement to matching performance, with social grouping contributing about $\sim 6\%$ while spatial grouping adding another $\sim 6\%$. Remember that both these quadratic constraints are antipodal in the sense that former rewards while latter penalizes the loss function. The last two curves show the performance using constraints computable if camera topology is known. Given topology, we employ collision avoidance shown in light blue, whereas the constraint capturing the desire of people to walk with preferred speed between cameras is shown in maroon, which gives the maximum AUC of $96.15\%$ in conjunction with other PSE constraints.

This study shows that except for collision avoidance, all PSE constraints contribute significantly to the performance of human re-identification. We provide real examples of collision avoidance and social grouping in Fig. 5(a) and (b), respectively. In Fig. 5, the bounding boxes are color-coded with time using colormap shown on left. White-to-Yellow indicate earlier time stamps while Red-to-Black indicate later ones. The person under consideration is shown with dashed white line, while the track of two other people in each image are color-coded with costs using colormap on the right. Here, blue indicates low cost whereas red means high cost.

Collision avoidance which has been shown to work for tracking in non-crowded scenes [38] deteriorates the results slightly in crowded scenes. Fig. 5(a) shows a case where collision avoidance constraint assigns a high cost to a pair of correct matches. Due to limitation in space in dense crowds, people do not change their path significantly. Furthermore, any slight change in path between cameras is unlikely to have any effect on matching for re-identification. On the other hand, the grouping constraint yields a strong increase in performance ($\sim 12\%$) as also seen in Fig. 5(b) This is despite the fact that the Grand Central dataset depicts dense crowd of commuters in a busy subway station, many of whom walk alone.

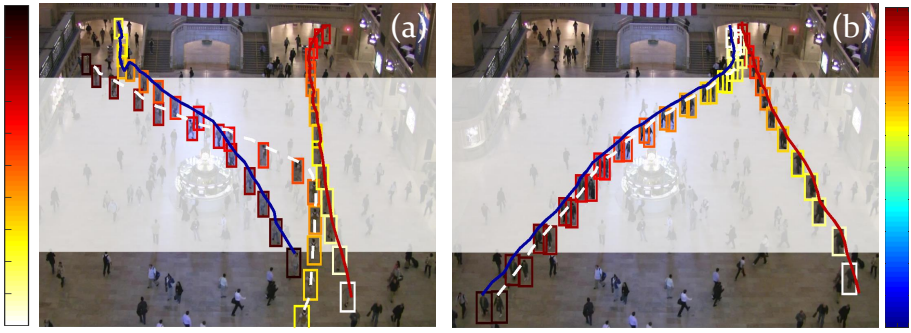



**Fig. 5.** This figure shows two examples of quadratic constraints. The color of bounding boxes indicates time using colorbar on the left, with white signifying start time and black representing end time. The person under consideration is shown with white trajectory, while the other two trajectories have the color of the cost for (a) collision avoidance and (b) grouping, color-coded with bar on the right. That is, blue and red trajectories indicate low and high costs, respectively. In (a), collision avoidance unnecessarily assigns high cost to a correct match, but not to a colliding person. On the other hand, grouping helps in re-identifying people who walk together by assigning a low cost between them.

## 5 Conclusion

This paper addressed the problem of re-identifying people across non-overlapping cameras in crowded scenes. Due to the difficult nature of the problem, the appearance similarity alone gives poor performance. We employed several personal, social and environmental constraints in the form of *preferred speed, destination probability* and *spatial and social grouping*. These constraints do not require knowledge about camera topology, however if available, it can be incorporated into our formulation. Since the problem with PSE constraints is NP-hard, we used a greedy local neighborhood search algorithm that can handle both quadratic and linear constraints. The crowd dataset used in the paper brings to light the difficulty and challenges of re-identifying and associating people across cameras in crowds. For future work, we plan to use discriminative appearance models independently trained on individuals, and inference of topology in an unsupervised manner for crowded scenes.

**Acknowledgment:** This material is based upon work supported in part by, the U.S. Army Research Laboratory, the U.S. Army Research Office under contract/grant number W911NF-14-1-0294.

## References


1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
2. Alahi, A., Ramanathan, V., Fei-Fei, L.: Socially-aware large-scale crowd forecasting. In: CVPR (2014)
3. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: ECCV (2008)
4. Bialkowski, A., Lucey, P., Wei, X., Sridharan, S.: Person re-identification using group information. In: DICTA (2013)
5. Cai, Y., Takala, V., Pietikäinen, M.: Matching groups of people by covariance descriptor. In: ICPR (2010)
6. Chakraborty, A., Das, A., Roy-Chowdhury, A.K.: Network consistent data association. IEEE TPAMI (2014)
7. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: CVPR (2015)
8. Chen, K.W., Lai, C.C., Hung, Y.P., Chen, C.S.: An adaptive learning method for target tracking across multiple cameras. In: CVPR (2008)
9. Chen, X., An, L., Bhanu, B.: Multitarget tracking in nonoverlapping cameras using a reference set. IEEE Sensors Journal 15(5) (2015)
10. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: ECCV (2014)
11. Dick, A.R., Brooks, M.J.: A stochastic approach to tracking objects across multiple cameras. In: AI 2004: Advances in Artificial Intelligence (2005)
12. Ess, A., Leibe, B., Schindler, K., Gool, L.V.: A mobile vision system for robust multi-person tracking. In: CVPR (2008)
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR (2010)
14. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE TPAMI 30(2) (2008)
15. Ge, W., Collins, R.T., Ruback, R.B.: Vision-based analysis of small groups in pedestrian crowds. IEEE TPAMI 34(5) (2012)
16. Gilbert, A., Bowden, R.: Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In: ECCV (2006)
17. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person re-identification 1 (2014)
18. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV (2008)
19. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image Analysis (2011)
20. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. CVIU 109(2) (2008)
21. Javed, O., Shafique, K., Shah, M.: Appearance modeling for tracking in multiple non-overlapping cameras. In: CVPR. vol. 2 (2005)
22. Kelly, H.: After boston: The pros and cons of surveillance cameras. CNN (April 26, 2013), http://www.cnn.com/2013/04/26/tech/innovation/security-cameras-boston-bombings/, [Accessed: Oct 1, 2015]
23. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR (2012)
24. Kuo, C.H., Huang, C., Nevatia, R.: Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In: ECCV (2010)

25. Leal-Taixé, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: ICCV Workshops (2011)
26. Leng, Q., Hu, R., Liang, C., Wang, Y., Chen, J.: Person re-identification with content and context re-ranking. Multimedia Tools and Applications 74(17) (2015)
27. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
28. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
29. Lisanti, G., Masi, I., Bagdanov, A.D., Del Bimbo, A.: Person re-identification by iterative re-weighted sparse ranking. Pattern Analysis and Machine Intelligence, IEEE Transactions on 37(8) (2015)
30. Loy, C.C., Liu, C., Gong, S.: Person re-identification by manifold ranking. In: ICIP (2013)
31. Madden, C., Cheng, E.D., Piccardi, M.: Tracking people across disjoint camera views by an illumination-tolerant appearance representation. MVA 18(3-4) (2007)
32. Makris, D., Ellis, T., Black, J.: Bridging the gaps between cameras. In: CVPR (2004)
33. Mazzon, R., Cavallaro, A.: Multi-camera tracking using a multi-goal social force model. Neurocomputing 100 (2013)
34. Mazzon, R., Tahir, S.F., Cavallaro, A.: Person re-identification in crowd. Pattern Recognition Letters 33(14) (2012)
35. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR (2009)
36. Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics 5(1) (1957)
37. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR (2015)
38. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
39. Porikli, F.: Inter-camera color calibration by correlation model function. In: ICIP. vol. 2 (2003)
40. Prosser, B., Gong, S., Xiang, T.: Multi-camera matching using bi-directional cumulative brightness transfer functions. In: BMVC (2008)
41. Qin, Z., Shelton, C.R., Chai, L.: Social grouping for target handover in multi-view video. In: ICME (2013)
42. Robin, T., Antonini, G., Bierlaire, M., Cruz, J.: Specification, estimation and validation of a pedestrian walking behavior model. Transportation Research Part B: Methodological 43(1) (2009)
43. Roth, P.M., Hirzer, M., Köstinger, M., Beleznai, C., Bischof, H.: Mahalanobis distance learning for person re-identification. In: Person Re-Identification (2014)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
45. Song, B., Roy-Chowdhury, A.K.: Robust tracking in a camera network: A multi-objective optimization framework. IEEE Selected Topics in Signal Processing 2(4) (2008)
46. Stauffer, C.: Learning to track objects through unobserved regions. In: WACV/MOTIONS Volume 1. vol. 2 (2005)
47. Tieu, K., Dalley, G., Grimson, W.E.L.: Inference of non-overlapping camera network topology by measuring statistical dependence. In: ICCV. vol. 2 (2005)
48. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3) (2012)
49. Wang, X.: Intelligent multi-camera video surveillance: A review. Pattern recognition letters 34(1) (2013)

50. Weinberger, K.Q., Saul, L.K.: Fast solvers and efficient implementations for distance metric learning. In: ICML (2008)
51. Wu, Z., Kunz, T.H., Betke, M.: Efficient track linking methods for track graphs using network-flow and set-cover techniques. In: CVPR (2011)
52. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: CVPR (2011)
53. Yi, S., Li, H., Wang, X.: Understanding pedestrian behaviors from stationary crowd groups. In: CVPR (2015)
54. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: Improving person recognition using multiple cues. In: CVPR (2015)
55. Zhang, Z., Chen, Y., Saligrama, V.: Group membership prediction. In: ICCV (2015)
56. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)
57. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
58. Zheng, W.S., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
59. Zhou, B., Tang, X., Wang, X.: Learning collective crowd behaviors with dynamic pedestrian-agents. IJCV 111(1) (2015)
60. Zhou, B., Wang, X., Tang, X.: Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In: CVPR (2012)