Multi-target Tracking with Motion Context in Tensor Power Iteration

Xinchu Shi[1,2], Haibin Ling[2], Weiming Hu[1], Chunfeng Yuan[1], Junliang Xing[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]Department of Computer and Information Sciences, Temple University, Philadelphia, USA
{xcshi, wmhu, cfyuan, jlxing}nlpr.ia.ac.cn,   hbling@temple.edu

Abstract

Interactions between moving targets often provide discriminative clues for multiple target tracking (MTT), though many existing approaches ignore such interactions due to difficulty in effectively handling them. In this paper, we model interactions between neighbor targets by pair-wise motion context, and further encode such context into the global association optimization. To solve the resulting global non-convex maximization, we propose an effective and efficient power iteration framework. This solution enjoys two advantages for MTT: First, it allows us to combine the global energy accumulated from individual trajectories and the between-trajectory interaction energy into a united optimization, which can be solved by the proposed power iteration algorithm. Second, the framework is flexible to accommodate various types of pairwise context models and we in fact studied two different context models in this paper. For evaluation, we apply the proposed methods to four public datasets involving different challenging scenarios such as dense aerial borne traffic tracking, dense point set tracking, and semi-crowded pedestrian tracking. In all the experiments, our approaches demonstrate very promising results in comparison with state-of-the-art trackers.

1. Introduction

With great advances in object detection [8, 9], *data association based multi-target tracking* (DAT) has been gaining popularity recently. An effective DAT algorithm needs to address intrinsic association ambiguities due to challenges such as appearance similarity, occlusion and fast motion. A group of DAT algorithms focus on reducing the association ambiguity by collecting multi-frame observations in the time window, and making the association decisions in a batch way. Association across multiple frames is more robust than the recursive tracking counterparts, but meanwhile more difficult to obtain the global solution. Different optimization strategies, such as linear programming [12], network flow [25, 17, 2, 6] and tensor approximation [19], have been proposed to solve the high-dimensional association problem. However, insufficient attention has been devoted to the interactions between target associations, except for the simple constraint that one target belongs to at most one association.

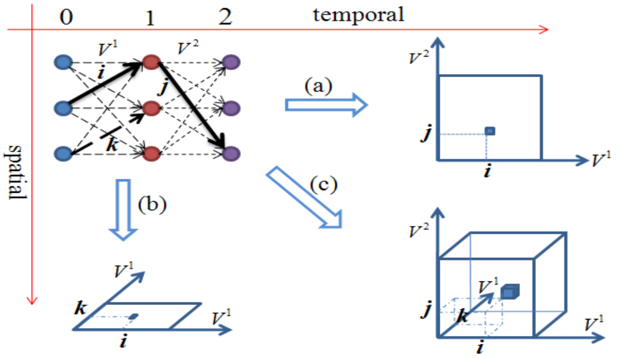

Figure 1. Contextual modeling in a 3-frame association. a) Temporal global trajectory energy. b) Spatial interaction energy (one block). c) Higher-order compound energy (tensor representation).

In this paper, we propose computationally efficient motion contexts to model the interaction between any local associations and integrate seamlessly the contexts into a power iteration association framework. In particular, we unite the pairwise interaction energy and the unary trajectory energy into a single optimization framework. Then, a power iteration solution is proposed for the complex non-convex optimization. Relations among the unary trajectory energy, the pairwise interaction energy and the united energy are illustrated in Fig. 1. The framework has three key ingredients to address challenges in MTT. First, the between-trajectory interaction is treated in a global data association framework; such a combination of context modeling and high-order trajectory information largely alleviates the association ambiguity. Second, the united energy term is encoded in a tensor approximation representation and can be effectively solved via the proposed power iteration solution. Finally, the optimization framework provides the flexibility to use different context information, and we devise two kinds of context representations in this paper.

We applied the proposed method to MTT and tested it on four challenging benchmark datasets involving various scenarios such as wide area traffic scenes, low frame-rate point set sequences and semi-crowded pedestrian videos. In all experiments, our approach produces excellent performances in comparison with several state-of-the-art trackers. The superiority of our approach is especially demonstrated on dense scenes with large association ambiguity between targets.

2. Related work

Most MTT methods can be roughly divided into two groups. Methods in the first group use only observations till the current frame to estimate the current target states, such as recursive filters [13, 4]. The second group contains association-based methods that use information from both previous and future frames to estimate the current states. The association-based approaches become popular recently, since solving the data association jointly across multiple frames are more reliable in general.

By decomposing the global affinity as the product of local pairwise items, the global association can be formulated as a network flow problem [25, 17, 2]. The decomposition on the affinity achieves the efficient global solution at the cost of limited discriminability, since higher-order motion information, which is very useful to ease the association ambiguity, is lost. Addressing this issue, the global affinity is used to enhance the association robustness in some recent methods such as [7, 19]. Our work shares the similar idea with [7, 19] in modeling high-order motion information using global trajectory affinity. In particular, the power iteration solution in our approach is inspired by the tenor approximation solution in [19]. That said, we integrate the context information into the global association, which has not been exploited by previous approaches. We emphasize here that the seamless integration is non-trivial (as shown in the next section), and is very beneficial (as shown in the experiments).

Modeling the interactions among targets is important for crowded scene and traffic analysis, where objects have grouping behavior [10, 16, 21] and follow the similar motion pattern (e.g. velocity) in local temporal-spatial cubes [1]. In the classic *social force model* (SFM) [11], a series of social forces are defined for a pedestrian, to avoid collision and choose a desired direction for the destination. Though powerfully used in pedestrian tracking [15, 18, 14], SFM is complicated and requires pre-training from the similar scenes, as well as the prior knowledge such as the destinations which are not universally available. Further, with the embedding of interaction based motion model, most approaches [1, 15, 14, 21] are limited to the predictive tracking framework, such as recursive filters. However, the local (temporal) association is often troubled by the intrinsic motion ambiguity. In [1], the motion context is a collection of trajectories of objects, and is used to predict and reacquire occluded targets. In [5], the association problem is formulated as finding the maximum weighted independent set, and the interaction between two trajectories is embedded as the soft constraint.

3. Encoding Context in Association

Multi-frame data association is popularly formulated as a multi-dimensional assignment (MDA) problem [7], which is the NP hard in general. In [19], MDA is reformulated as a rank-1 tensor approximation problem and consequently leads to an efficient tensor approximation solution. In the following, we first review the basic optimization formulation, and then show the united optimization framework encoding the motion context. Finally, we give the power iteration solution for the united optimization.

3.1. Problem Formulation

Assume the association is performed on a $K+1$ frame sequence, each frame has N targets[1], and $M=N^2$ denotes the number of possible two-frame associations. Suppose the i_k-th$(1 \le i_k \le N)$ target in the k-th$(0 \le k \le K)$ frame is $o^k_{i_k}$, the global trajectory hypothesis with targets $o^0_{i_0}, o^1_{i_1}, ..., o^K_{i_K}$ is represented as $T_{i_0 i_1 \ldots i_K}$, with trajectory energy $s_{i_0 i_1 \ldots i_K}$.

For targets $o^{k-1}_{i_{k-1}}$ and $o^k_{i_k}$, their association variables are represented interchangably as $x^k_{i_{k-1} i_k}$ (as an element of an assignment matrix) and $v^k_{l_k}$ (as an element of the vectorized assignment vector), such that $l_k = (i_{k-1} - 1) \times N + i_k$.

In the hard decision, $x^k_{i_{k-1} i_k}$ has a binary value as 0 or 1, where 1 means targets $o^{k-1}_{i_{k-1}}$ and $o^k_{i_k}$ are associated, and 0 otherwise. While in the soft decision, $x^k_{i_{k-1} i_k}$ represents the probability of associating $o^{k-1}_{i_{k-1}}$ and $o^k_{i_k}$.

The trajectory affinity is represented using *association index* as

$$a_{l_1 l_2 \ldots l_K} = \begin{cases} s_{\overline{l_1}\,\overline{l_2} \ldots \overline{l_K}\underline{l_K}}, & \text{if } \overline{l_{k+1}} = \underline{l_k}, 1 \le k < K \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\overline{l_k}$ and $\underline{l_k}$ denote the indexes of the two targets connected in association $v^k_{l_k}$ (i.e., i_{k-1} and i_k), and $s_{\overline{l_1}\,\overline{l_2} \ldots \overline{l_K}\underline{l_K}}$ is the affinity for trajectory $\overline{l_1 l_2} \ldots \overline{l_K}\underline{l_K}$. Formally, we have $\overline{l_k} = \lceil \frac{l_k}{N} \rceil$, where $\lceil \cdot \rceil$ is the up rounding operator, and $\underline{l_k} = l_k - (\overline{l_k} - 1) \times N$.

Denote $\mathbb{V} = \{V^k : k = 1, \ldots, K\}$ as the set of association vectors we are seeking, and each vector is defined by $V^k = (v^k_1, v^k_2, \ldots, v^k_M)^\top \in \mathrm{R}^M$. With these notations,

[1] Assuming fixed number of targets is for presentation convenience, variable numbers of targets do not hurt the formulation and derivation.

multi-frame data association can be formulated as the following optimization [19]

$$\max_{\mathbb{V}} \sum_{\mathcal{L}} a_{l_1 l_2 \ldots l_K} v_{l_1}^1 v_{l_2}^2 \cdots v_{l_K}^K, \tag{2}$$

$$\text{s.t.} \begin{cases} \sum\limits_{i_{k-1}} x_{i_{k-1} i_k}^k = 1, k \in \{1, 2, ..., K\} \\ \sum\limits_{i_k} x_{i_{k-1} i_k}^k = 1, k \in \{1, 2, ..., K\} \\ 0 \leq x_{i_{k-1} i_k}^k \leq 1, k \in \{1, 2, ..., K\} \end{cases} \tag{3}$$

For notation conciseness, in the above formulae and hereafter we define $\mathcal{L} = \{l_1, l_2, \cdots, l_K\}$ and use $\sum_{\mathcal{L}}$ to denote the series of summation $\sum_{l_1=1}^{M} \sum_{l_2=1}^{M} \cdots \sum_{l_K=1}^{M}$.

The constrained optimization (2) is challenging due to the very high-dimensional solution space. In work [19], it demonstrates the close relations between the optimization and rank-1 tensor approximation problem, and further presents an efficient iteration approach for the optimization.

3.2. Encoding Context Information

We aim at combining the individual temporal energy and spatial interaction energy of trajectories into a united optimization framework. Modeling the contextual relations of two trajectories over a long term is risky, as the motion patterns of a target are changing over time. We focus on the interaction between two trajectories in a short term. Specially, we consider the pairwise interactions between two associations on neighboring detections or tracklets.

For two-frame association hypothesis $T_{l_k}^k$ and $T_{j_k}^k (1 \leq k \leq K)$, whose association variables are $v_{l_k}^k$ and $v_{j_k}^k$ respectively. We define the interaction energy between $T_{l_k}^k$ and $T_{j_k}^k$ as $c_{l_k j_k}^k$. By embedding the interaction energy, the total energy is represented as the linear combination of two types of energies. In this way, the combinational optimization is formulated as

$$\max_{\mathbb{V}} \sum_{\mathcal{L}} a_{l_1 \ldots l_K} v_{l_1}^1 \ldots v_{l_K}^K + \alpha \sum_{k=1}^{K} \sum_{l_k, j_k} c_{l_k j_k}^k v_{l_k}^k v_{j_k}^k, \tag{4}$$

where α is the weighting parameter, and the optimization has the same constraints as Eq. (3). Intuitively, the second term in (4) models the between-association interaction.

3.3. Power Iteration Solution

The new problem (4) is more difficult than the basic one (2) due to the quadratic context items, where $v_{l_k}^k$ and $v_{j_k}^k$ lie in the same block and couple with each other. In the following, we decouple the interdependency between $v_{l_k}^k$ and $v_{j_k}^k$ to simplify the optimization. If two association hypothesis $T_{l_k}^k$ and $T_{j_k}^k$ share the same target, that is $\overline{l_k} = \overline{j_k}$ or $\underline{l_k} = \underline{j_k}$, we set their interaction energy $c_{l_k j_k}^k$ as 0. This is reasonable as one target can not be in two real associations.

Next, we make some reformulations to make (4) self-consistent. With constraint (3), there are formulations as

$$\begin{cases} \sum\limits_{l_k} {v_{l_k}}^k = N, & k = 1, 2, ..., K \\ \sum\limits_{j_k : \overline{j_k} \neq \overline{l_k}} v_{j_k}^k = N - 1, & \forall l_k, k = 1, 2, ..., K \end{cases} \tag{5}$$

Using the formulation (5), the two components in (4) can be rewritten as

$$\begin{aligned} & \sum_{\mathcal{L}} a_{l_1 \ldots l_K} v_{l_1}^1 \ldots v_{l_K}^K = \\ & \quad \tfrac{1}{N-1} \sum_{\mathcal{L}} \sum_{j_k : \overline{j_k} \neq \overline{l_k}} a_{l_1 \ldots l_K} v_{l_1}^1 \ldots v_{l_k}^k v_{j_k}^k \ldots v_{l_K}^K, \end{aligned} \tag{6}$$

$$\begin{aligned} & \sum_{l_k, j_k} c_{l_k j_k}^k v_{l_k}^k v_{j_k}^k = \\ & \quad \tfrac{1}{N^{K-1}} \sum_{\mathcal{L}} \sum_{j_k : \overline{j_k} \neq \overline{l_k}} c_{l_k j_k}^k v_{l_1}^1 \ldots v_{l_k}^k v_{j_k}^k \ldots v_{l_K}^K. \end{aligned} \tag{7}$$

Merging (6) and (7), optimization (4) is rewritten as

$$\begin{aligned} \max_{\mathbb{V}} & \sum_{\mathcal{L}} \sum_{j_k : \overline{j_k} \neq \overline{l_k}} e_{l_1 \ldots l_k j_k \ldots l_K} v_{l_1}^1 \ldots v_{l_k}^k v_{j_k}^k \ldots v_{l_K}^K \\ & + \alpha (N-1) \sum_{f \neq k} \sum_{l_f, j_f} c_{l_f j_f}^f v_{l_f}^f v_{j_f}^f, \end{aligned} \tag{8}$$

where $e_{l_1 \ldots l_k j_k \ldots l_K}$ is the element of the $(K+1)$-th augmented tensor, which is a combination of the items $a_{l_1 \ldots l_k \ldots l_K}$ and $c_{l_k j_k}^k$, and is computed as

$$e_{l_1 \ldots l_k j_k \ldots l_K} = a_{l_1 \ldots l_k \ldots l_K} + \frac{\alpha (N-1)}{N^{K-1}} c_{l_k j_k}^k. \tag{9}$$

The relation between the new context aware tensor and the original tensor is illustrated in Fig. 1.

We apply the block update strategy [7, 19] to optimize (8) iteratively. When updating block variables in V^k, other block variables $V^f (f \neq k)$ are fixed. In this manner, optimization (8) degenerates into the following formulation

$$\max_{V^k} \sum_{\mathcal{L}} \sum_{j_k : \overline{j_k} \neq \overline{l_k}} e_{l_1 \ldots l_k j_k \ldots l_K} v_{l_1}^1 \ldots v_{l_k}^k v_{j_k}^k \ldots v_{l_K}^K. \tag{10}$$

The optimizations (10) and (2) share the similar form. The former is performed on all block variables V^f, $f = 1, \ldots, K$, while the latter is on the block variable V^k for a certain k. We further reformulate (10) as

$$\begin{aligned} & \max_{V^k} \sum_{\mathcal{L}} \sum_{j_k : \overline{j_k} \neq \overline{l_k}} e_{l_1 \ldots l_k j_k \ldots l_K} v_{l_1}^1 \ldots v_{l_k}^k v_{j_k}^k \ldots v_{l_K}^K \\ & = \sum_{n=1}^{N} \max_{v_{l_k}^k : \overline{l_k} = n} \mathcal{E}_n^k, \end{aligned} \tag{11}$$

where

$$\mathcal{E}_n^k = \sum_{l_1} \ldots \sum_{l_k : \overline{l_k} = n} \sum_{j_k : \overline{j_k} \neq n} \ldots \sum_{l_K} e_{l_1 \ldots l_k j_k \ldots l_K} v_{l_1}^1 \ldots v_{l_k}^k v_{j_k}^k \ldots v_{l_K}^K. \tag{12}$$

This way, (10) is divided into a series of subproblems. In each subproblem, the interdependency between $v_{l_k}^k$ and $v_{j_k}^k$

Algorithm 1 Power iteration with interaction

```
Input: Global energy $\mathcal{A}: a_{l_1\ldots l_k\ldots l_K}$.
       interaction energy $C^k: c^k_{l_k j_k},\ k \in \{1,\ldots,K\}$.
Output: association variables $V^k:\{v^k_1,\ldots,v^k_M\}(1\leq k\leq K)$.
Initialize $V^1,\ldots,V^K$;
repeat
    for $k = 1,\ldots,K$ do
        for $i_{k-1} = 1,\ldots,N$ do
            for $i_k = 1,\ldots,N$ do
                $\varphi_{i_{k-1}i_k} = \sum_{l_f:f\neq k} a_{l_1\ldots l_K} v^1_{l_1}\ldots v^f_{l_f}\ldots v^K_{l_K}$.
                $\phi_{i_{k-1}i_k} = \sum_{j_k:\{\overline{j_k}\neq i_{k-1}\}} c^k_{l_k j_k} v^k_{j_k}$.
            end for
            $\forall i_k,\ x^k_{i_{k-1}i_k} = \frac{x^k_{i_{k-1}i_k}\left(\varphi_{i_{k-1}i_k}+\alpha\phi_{i_{k-1}i_k}\right)}{\sum_{i_k} x^k_{i_{k-1}i_k}\left(\varphi_{i_{k-1}i_k}+\alpha\phi_{i_{k-1}i_k}\right)}$;
        end for
        $\forall i_{k-1},\ x^k_{i_{k-1}i_k} = \frac{x^k_{i_{k-1}i_k}}{\sum_{i_{k-1}} x^k_{i_{k-1}i_k}}$;
    end for
until convergence
```

is decoupled, and the subproblem has the similar formulation with (2). We can then use tensor power iteration [19] for solving each subproblem, and the key iteration is

$$\begin{aligned} v^k_{l_k} &\propto v^k_{l_k} \sum_{\mathcal{L}\backslash\{k\}} \sum_{j_k:\overline{j_k}\neq\overline{l_k}} e_{l_1\ldots l_k j_k\ldots l_K} v^k_{j_k} v^1_{l_1} v^2_{l_2}\ldots v^K_{l_K} \\ &\propto v^k_{l_k}\Big(\sum_{\mathcal{L}\backslash\{k\}} a_{l_1\ldots l_K} v^1_{l_1} v^2_{l_2}\ldots v^K_{l_K} + \alpha \sum_{j_k:\overline{j_k}\neq\overline{l_k}} c^k_{l_k j_k} v^k_{j_k}\Big). \end{aligned} \tag{13}$$

We update the block variables $V^k(1\leq k\leq K)$ in turn to obtain the (local) optimum of (4), the power iteration is presented as Alg. 1.

4. Motion Context

In this section, we define the interaction energy $c^k_{l_k j_k}$. Specifically, we propose two types of motion contexts, low-level context and high-level context, to represent different types of interactions on associations.

4.1. Low-level context

Low-level motion context measures the interaction between two associations on raw detections. First, we give the motion consistency representation for any association pair. Then the specific motion context formulation is presented, by using the non-maximum suppression (NMS) strategy.

Suppose $T^k_{l_k}$ represents the association hypothesis connecting targets $o^{k-1}_{i_{k-1}}$ and $o^k_{i_k}$, $\mathbf{p}^{k-1}_{i_{k-1}}$ ($\mathbf{p}^k_{i_k}$) is the spatial position of the target $o^{k-1}_{i_{k-1}}$ ($o^k_{i_k}$). For another association hypothesis $T^k_{j_k}$, it associates targets $o^{k-1}_{i'_{k-1}}$ and $o^k_{i'_k}$, whose spatial positions are $\mathbf{p}^{k-1}_{i'_{k-1}}$ and $\mathbf{p}^k_{i'_k}$ respectively. Then, the motion consistency between $T^k_{l_k}$ and $T^k_{j_k}$ is defined as

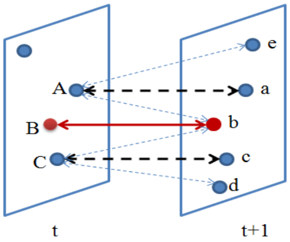

Figure 2. Low-level motion context. For association hypothesis Bb, the neighbor targets are A and C, each has three association candidates. For all association candidates of A, interaction between Ab and Bb is filtered as one-to-one mapping constraint, interaction between Ae and Bb is filtered as non-maximum suppression, only interaction between Aa and Bb is retained. For target C, only energy between Cc and Bb is retained.

$$m^k_{i_{k-1}i_k,i'_{k-1}i'_k} = m^k_{l_k j_k} = \frac{\left|{\mathbf{z}^k_{l_k}}^\top \mathbf{z}^k_{j_k}\right|}{\left\|\mathbf{z}^k_{l_k}\right\|\left\|\mathbf{z}^k_{j_k}\right\|} + \frac{\lambda\left\|\mathbf{z}^k_{l_k}\right\|\left\|\mathbf{z}^k_{j_k}\right\|}{\left\|\mathbf{z}^k_{l_k}\right\|^2+\left\|\mathbf{z}^k_{j_k}\right\|^2}, \tag{14}$$

where $\mathbf{z}^k_{l_k}=\mathbf{p}^k_{i_k}-\mathbf{p}^{k-1}_{i_{k-1}}$ and $\mathbf{z}^k_{j_k}=\mathbf{p}^k_{i'_k}-\mathbf{p}^{k-1}_{i'_{k-1}}$, both of which represent the spatial displacement (velocity) vector; λ is the weighting parameter. Formulation (14) is intuitive, the motion consistency is computed from the orientation similarity and the speed similarity.

Modeling the interaction between any two associations is meaningless, since targets only in the local spatial neighborhood follow the similar motion. Specifically, we define the low-level motion context as a selective representation, the context between $T^k_{l_k}$ and $T^k_{j_k}$ is formulated as

$$c^k_{l_k j_k} = \mathbf{I}_\Omega\left(i_{k-1}, i'_{k-1}, i_k, i'_k, \mathbf{p}^{k-1}_{i_{k-1}}, \mathbf{p}^{k-1}_{i'_{k-1}}, \mathbf{p}^k_{i_k}, \mathbf{p}^k_{i'_k}\right) m^k_{l_k j_k}, \tag{15}$$

where $\mathbf{I}_\Omega(\cdot)$ is the indicator function, it has value 1 when condition set Ω is true, otherwise it is 0. Ω is defined as

$$\begin{aligned} \Omega: \{i_{k-1}\neq i'_{k-1}\}\bigcap\{i_k\neq i'_k\}\bigcap\left\{\left\|\mathbf{p}^{k-1}_{i_{k-1}}-\mathbf{p}^{k-1}_{i'_{k-1}}\right\|<L\right\} \\ \bigcap\left\{\left\|\mathbf{p}^k_{i_k}-\mathbf{p}^k_{i'_k}\right\|<L\right\}\bigcap\left\{i'_k=\max_j m^k_{i_{k-1}i_k,i'_{k-1}j}\right\}, \end{aligned} \tag{16}$$

where L is the distance threshold. Set Ω constitutes of three parts: one-to-one mapping constraint, spatial distance mask and non-maximum suppression.

Low-level context is illustrated in Fig. 2. NMS is very important and effective, because the selection mechanism makes a binding for two association candidates with similar motion patterns and drives them to be true or wrong synchronously. The underlying assumption in this procedure is that real associations around the target follow similar motion patterns, while wrong associations are irregular in motion statistics. Further, it is robust by suppressing the influences from noisy and conflicting association counterparts.

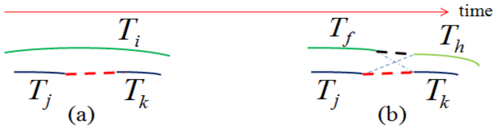

Figure 3. High-level motion contexts. a) Context-A: interaction between association T_{jk} and tracklet T_i; b) Context-B: interaction between any two tracklet associations.

4.2. High-level Context

When frame-between motions are notable and reliable, low-level context is valuable, such as the low-frame rate or fast motion applications. In most pedestrian tracking, bad located object detections (raw zigzag trajectory) along with low-speed motion make raw detection based low-level context unreliable. In this section, we devise two kinds of high-level contexts to model the motion interaction on tracklet associations, which are illustrated in Fig. 3.

Suppose $T_i:\{o_i^{t_s^i}, o_i^{t_s^i+1}, ..., o_i^{t_e^i}\}$ represents the i-th tracklet, where t_s^i and t_e^i denotes the start time and end time of T_i respectively. The spatial displacement from the target o_i^{t-1} to o_i^t is represented as $\mathbf{z}_i^t = \mathbf{p}_i^t - \mathbf{p}_i^{t-1}$ and $\mathbf{p}_i^t(\mathbf{p}_i^{t-1})$ is the spatial position of the target $o_i^t(o_i^{t-1})$. For other tracklets such as T_j and T_k, there are similar notations and definitions.

For two tracklets $T_j:\{o_j^{t_s^j}, ..., o_j^{t_e^j}\}$ and $T_k:\{o_k^{t_s^k}, ..., o_k^{t_e^k}\}$ showed in Fig.3-(a), there exists association hypothesis $T_{jk}:\{o_j^{t_s^j}, ..., o_j^{t_e^j}, o_{jk}^{t_e^j+1}, ..., o_{jk}^{t_s^k-1}, o_k^{t_s^k}, ..., o_k^{t_e^k}\}$, where $o_{jk}^t(t_e^j<t<t_s^k)$ is the interpolated target using T_j and T_k. Then the motion interaction between T_{jk} and T_i in Fig.3-(a) is defined as

$$m_{jk,i} = \frac{1}{t_s^k - t_e^j} \sum_{t=t_e^j+1}^{t_s^k} \frac{\left|{\mathbf{z}_{jk}^t}^\top \mathbf{z}_i^t\right|}{\left\|\mathbf{z}_{jk}^t\right\| \left\|\mathbf{z}_i^t\right\|}, \tag{17}$$

where $\mathbf{z}_{jk}^t$ is the spatial displacement from target o_{jk}^{t-1} to o_{jk}^t.

For context-A of T_{jk}, we consider interactions from all neighbor tracklets around T_{jk}, and give the final formulation. Suppose tracklet set is $\mathcal{T}:\{T_1, ..., T_C\}$, where C is the number of tracklets, then context-A of T_{jk} is computed as

$$sc_{jk} = \frac{\sum_{c=1}^{C} m_{jk,c} \mathbf{I}_\Phi(t_e^j, t_s^k, t_s^c, t_e^c, \mathbf{p}_j^{t_e^j}, \mathbf{p}_k^{t_s^k}, \mathbf{p}_c^{t_e^j}, \mathbf{p}_c^{t_s^k})}{\sum_{c=1}^{C} \mathbf{I}_\Phi(t_e^j, t_s^k, t_s^c, t_e^c, \mathbf{p}_j^{t_e^j}, \mathbf{p}_k^{t_s^k}, \mathbf{p}_c^{t_e^j}, \mathbf{p}_c^{t_s^k})}, \tag{18}$$

where Φ denotes the condition set defined as:

$$\begin{aligned}&\{t_s^c \leq t_e^j\} \bigcap \{t_s^k \leq t_e^c\} \bigcap \\ &\{\|\mathbf{p}_c^{t_e^j} - \mathbf{p}_j^{t_e^j}\| < L\} \bigcap \{\|\mathbf{p}_c^{t_s^k} - \mathbf{p}_k^{t_s^k}\| < L\}.\end{aligned} \tag{19}$$

As shown in Fig.3-(a), Φ selects the spatial neighbor tracklets which are overlapped with T_j and T_k in the time window. Context-A (18) measures the average motion interaction between contextual tracklets and T_{jk}.

Suppose association hypothesis T_{jk} connects tracklet T_j and T_k, association hypothesis T_{fh} connects tracklet T_f and T_h, as is shown in Fig. 3-(b). Motion similarity between T_{jk} and T_{fh} is computed as

$$m_{jk,fh} = \frac{1}{t_s^{kh} - t_e^{jf}} \sum_{t=t_e^{jf}+1}^{t_s^{kh}} \frac{\left|{\mathbf{z}_{jk}^t}^\top \mathbf{z}_{fh}^t\right|}{\left\|\mathbf{z}_{jk}^t\right\| \left\|\mathbf{z}_{fh}^t\right\|}, \tag{20}$$

where $\mathbf{z}_{jk}^t(\mathbf{z}_{fh}^t)$ denotes the spatial displacement from target $o_{jk}^{t-1}(o_{fh}^{t-1})$ to $o_{jk}^t(o_{fh}^t)$. t_s^{kh} and t_e^{jf} are computed as

$$t_s^{kh} = \min\{t_s^h, t_s^k\};\ t_e^{jf} = \max\{t_e^f, t_e^j\}. \tag{21}$$

Eq. (20) computes the temporal average of motion similarities between T_{jk} and T_{fh}. Context-B between T_{jk} and T_{fh} is computed as

$$\begin{aligned}&c_{jk,fh} = \\ &\mathbf{I}_\Psi(t_s^{kh}, t_e^{jf}, j, k, f, h, \mathbf{p}_j^{t_e^{jf}}, \mathbf{p}_f^{t_e^{jf}}, \mathbf{p}_h^{t_s^{kh}}, \mathbf{p}_k^{t_s^{kh}}) m_{jk,fh},\end{aligned} \tag{22}$$

where condition set Ψ is defined as:

$$\begin{aligned}&\{t_e^{jf} < t_s^{kh}\} \bigcap \{j \neq f\} \bigcap \{k \neq h\} \bigcap \{\|\mathbf{p}_j^{t_e^{jf}} - \mathbf{p}_f^{t_e^{jf}}\| < L\} \\ &\bigcap \{\|\mathbf{p}_h^{t_s^{kh}} - \mathbf{p}_k^{t_s^{kh}}\| < L\} \bigcap \{h = \max_g m_{jk,fg}\}.\end{aligned} \tag{23}$$

Condition set Ψ is similar with Ω which is used in low-level context, and Context-B measures the motion interaction between any two tracklet association hypotheses.

Association on tracklets is performed as the extended two-frame association (i.e. $K=1$), thus the global affinity degenerates into $a_{l_1} (1 \leq l_1 \leq C^2)$, and the pairwise interaction element is $c_{l_1 j_1}^1$, i.e., Eq. (22).

5. Experiments

We evaluate the proposed approach on four public datasets, Columbus Large Image Format (CLIF) [22], PSU-data [10], PETS 2009 and TUD-Stadtmitte. The first two datasets are low frame-rate (1∼2 fps) sequences, which are used to test the proposed low-level motion context. The last two pedestrian sequences are used to validate the effectiveness of the high-level motion context.

5.1. Low Frame-rate Sequences

Both CLIF and PSUdata are challenging as the targets have fast motions, along with other challenges. CLIF has extra difficulties such as a large amount of targets, tiny object occupy, similar target appearance and so on. PSUdata are point set sequences, which are challenging as the visual cues relied heavily by many tracking approaches are unavailable.

Three CLIF sequences, seq1, seq2 and seq3, are used in the experiments. There are about 80 targets in each frame for seq1 and seq2, and 200 targets for seq3. Three PSU-data sequences, dense-1fps, dense-2fps and sparse-1fps are tested in the second experiment. The first two contain about

20 targets in each frame, and the last one has 3∼5 objects in each frame.

For CLIF, the global affinity $a_{l_1 l_2 \dots l_K}$ in (4) is defined as

$$a_{l_1 l_2 \dots l_K} = e^1_{l_1} e^2_{l_2} \dots e^K_{l_K} d_{l_1 l_2 \dots l_K} \ , \tag{24}$$

where $e^k_{l_k}$ and $d_{l_1 l_2 \dots l_k}$ are appearance/shape affinity and motion affinity, respectively. Specifically, for an association hypothesis $T^k_{l_k}$, its appearance/shape affinity is defined as

$$e_{l_k} = \frac{2 q^{k-1}_{i_{k-1}} q^k_{i_k}}{(q^{k-1}_{i_{k-1}})^2 + (q^k_{i_k})^2} + \sum\nolimits_b \min\left(h_b^{i_{k-1}}, h_b^{i_k}\right), \tag{25}$$

where $q^{k-1}_{i_{k-1}}$ ($q^k_{i_k}$) denotes the area of the target $o^{k-1}_{i_{k-1}}$ ($o^k_{i_k}$), and $h_b^{i_{k-1}}$ ($h_b^{i_k}$) is b-th bin of the color histogram of the target $o^{k-1}_{i_{k-1}}$ ($o^k_{i_k}$). The motion affinity is defined as

$$\begin{aligned} & d_{l_1 l_2 \dots l_K} \propto \\ & \prod_{k=1}^{K-1} \exp\left(\frac{{\mathbf{z}^k_{l_k}}^\top \mathbf{z}^{k+1}_{l_{k+1}}}{\|\mathbf{z}^k_{l_k}\| \|\mathbf{z}^{k+1}_{l_{k+1}}\|} + \frac{2\|\mathbf{z}^k_{l_k}\| \|\mathbf{z}^{k+1}_{l_{k+1}}\|}{\|\mathbf{z}^k_{l_k}\|^2 + \|\mathbf{z}^{k+1}_{l_{k+1}}\|^2} \right), \end{aligned} \tag{26}$$

where $\mathbf{z}^k_{l_k}$ is the spatial displacement of $T^k_{l_k}$.

The motion affinity (26) has the similar formulation with the motion consistency representation (14), both of which aim at enforcing compatible motion patterns. The difference is that (26) focuses on the motion smoothness of the same target within the temporal window, while the context (14) pays attention to the motion coherence between spatial neighbor targets.

Global affinity $a_{l_1 l_2 \dots l_K}$ used in PSUdata is defined as

$$\begin{aligned} a_{l_1 l_2 \dots l_K} &= E_0 - E_{cont} - E_{curv} \\ &= E_0 - \eta \sum_{k=1}^{K} \left\| \mathbf{z}^k_{l_k} \right\| - \sum_{k=1}^{K-1} \left\| \mathbf{z}^{k+1}_{l_{k+1}} - \mathbf{z}^k_{l_k} \right\|, \end{aligned} \tag{27}$$

where η is the weighting parameter; E_0 is a large constant to make the affinity positive; E_{cont} is used to penalize the large jump in position for any association, and E_{curv} is the constant-velocity model to assure the similar motions for consecutive associations.

The source inputs for the PSUdata are the ground truth data, each point is featured with a spatial coordinate. While the inputs for the CLIF are from vehicle detection [20]. The frame number in a batch is 5 and 6 for CLIF and PSUdata respectively. Some parameters are set as follows: λ in (14) is 0.6 and 2.0 for CLIF and PSUdata respectively; α in (4) is 10 and 5 for two datasets respectively; η in (27) is 0.5. Most parameters are application dependent, such as a smaller λ is used in CLIF than in PSUdata, since the orientation consistency is more important in the CLIF scenarios.

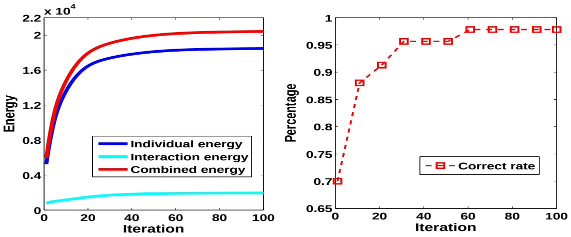

Figure 4. The energy and association performance variations in the iteration process (for one batch in PSUdata). Left: the energy iterated curve; Right: correct match rate curve.

Table 1. Association results of three approaches on the CLIF

	Correct match percentage			Wrong match percentage		
	Seq1	Seq2	Seq3	Seq1	Seq2	Seq3
[17]	65.4	71.6	74.6	34.1	28.1	25.7
[19]	91.1	92.1	91.4	11.9	9.4	9.4
Ours	**94.7**	**96.0**	**95.8**	**6.0**	**4.8**	**4.1**

Table 2. Association results of three approaches on the PSUdata

	Correct match percentage			Wrong match percentage		
	Dense1	Dense2	Sparse	Dense1	Dense2	Sparse
[17]	78.65	98.64	94.57	21.35	1.36	5.43
[6]	**98.54**	99.83	99.59	**1.46**	0.17	0.41
[19]	96.98	99.78	99.45	3.01	0.20	0.50
Ours	98.41	**99.88**	**99.74**	1.58	**0.11**	**0.24**

We compare our work with the tensor method [19], the network flow approach [17] and the min-cost flow [6]. In [19], we employ the same affinity model $a_{l_1 l_2 \dots l_K}$ and parameters. Let $cm(t)$, $wm(t)$ and $g(t)$ represent the numbers of correct associations, wrong associations and ground truth associations at frame t respectively, we use correct match percentage $P_c = 100 \times (\sum_t cm(t) / \sum_t g(t))$ and wrong match percentage $P_w = 100 \times (\sum_t wm(t) / \sum_t g(t))$ to evaluate the association performance.

Quantitative results on the CLIF and PSUdata[2] are presented in Tab. 1 and Tab. 2 respectively. It can be seen that the proposed approach performs better than the tensor method, especially on the CLIF. Both P_c and P_w are improved a lot, and P_w has a remarkable decrease relatively. It demonstrates the proposed solution and the motion context are effective. The motion context is very useful for reducing the association ambiguity, as the decision of the local association is influenced by not only its temporal coherence on the whole trajectory, but also its spatial interaction with other associations. Though the performances of method [19] in PSUdata are close to saturation, yet the embedding of the proposed context improves the results on all sequences remarkably. The min-cost flow [6] has excellent performance on the PSUdata, while our approach is slightly better. The method [17] performs the worst among all algorithms, since the motion information are lost in the network flow formulation. An example to illustrate the variations of the energy and association performance in the power iteration is shown in Fig. 4, both the combined energy and basic energies increase in the iteration, and the association performance is improved gradually too.

Qualitative results are presented in Fig. 5 and Fig. 6.

[2] The results for method [17] is taken from [7] on the PSUdata.

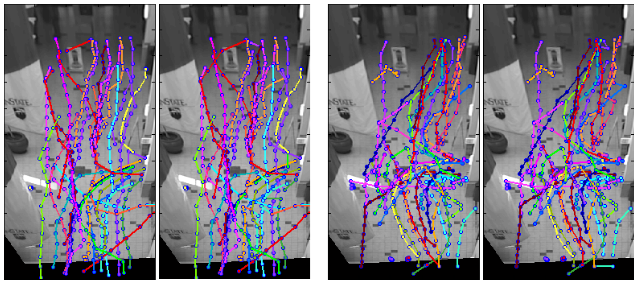

Figure 5. Tracking results on dense-1fps. Left half for the first episode and right half for the second episode. From left to right: our approach (14 mismatches), tensor method [19] (25 mismatches), our approach (12 mismatches), and tensor method (27 mismatches). All trajectories are color-coded with respect to ground truth; edges of good trajectories appear in the same color.

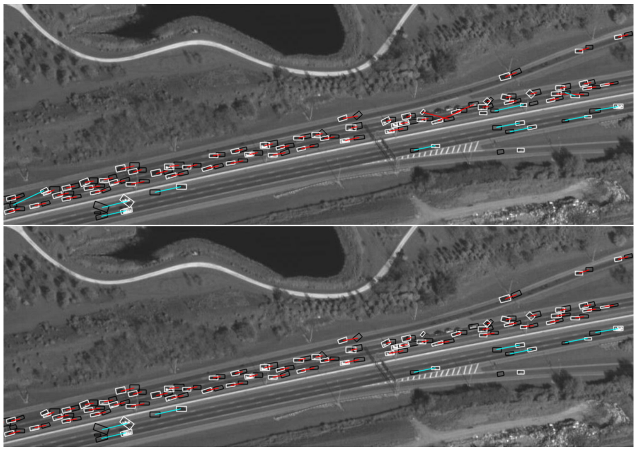

Figure 6. Association results on CLIF (part). Top: tensor method [19] has 6 mismatches; Bottom: our approach has no mismatch; White (black) rectangles: vehicle detections in current (last) frame; Red (green) lines: associations on two orientations.

There are fewer association errors for our approach.

5.2. Pedestrian Datasets

The high-level association is performed on tracklet sets, and the basic tracklets are achieved with the approach [19]. For tracklet $T_j : \{o_j^{t_s^j}, ..., o_j^{t_e^j}\}$ and $T_k : \{o_k^{t_s^k}, ..., o_k^{t_e^k}\}$, the association affinity used in (4) is computed as

$$a_{l_1} = (sa_{jk} + sd_{jk} + sc_{jk})\, st_{jk} \ , \tag{28}$$

where sc_{jk} is the contextual affinity, computed as Eq. (18); sa_{jk}, sd_{jk} and st_{jk} are the appearance, spatial distance and temporal distance affinity respectively, which are defined as

$$sa_{jk} = \sum\nolimits_b \min\left(h_b^j, h_b^k\right), \tag{29}$$

$$st_{jk} = \begin{cases} \exp(-\frac{\Delta t}{TL}), & \text{if } 0 < \Delta t < TL \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

$$sd_{jk} = \frac{1}{2}\exp\left(\frac{\left\|\Delta d - \Delta t \mathbf{z}_j^{t_e^j}\right\|^2}{-2\left\|\mathbf{z}_j^{t_e^j}\right\|^2}\right) + \frac{1}{2}\exp\left(\frac{\left\|\Delta d - \Delta t \mathbf{z}_k^{t_s^k}\right\|^2}{-2\left\|\mathbf{z}_k^{t_s^k}\right\|^2}\right). \tag{31}$$

Table 3. Tracking results on PETS 2009

	Rec	Prec	TA	TP	MT	PT	Frag	IDS
[23]	91.8	**99.0**	-	-	89.5	10.5	9	**0**
[17]	94.0	97.4	88.9	80.9	89.5	10.5	13	10
[19]	96.0	98.2	92.7	81.8	94.7	5.3	11	7
A	97.4	98.5	94.7	81.4	94.7	5.3	8	6
B	96.6	98.8	94.9	81.6	94.7	5.3	8	5
Ours	**97.7**	98.9	**96.1**	**81.8**	**94.7**	**5.3**	**6**	4

Note: 'A' in Tab. 3 and Tab. 4 is the approach with high-level context listed in Fig. 3 (a), and 'B' is the approach with context listed in Fig. 3 (b).

In (29), h_b^j (h_b^k) is the value in the b-th bin of the average color histogram of the tracklet T_j (T_k). In (30), $\Delta t = t_s^k - t_e^j$ is the time gap between T_j and T_k, and TL is the temporal threshold for possible tracklet associations. In (31), $\Delta d = \mathbf{p}_k^{t_s^k} - \mathbf{p}_j^{t_e^j}$ is the spatial displacement from the target $o_j^{t_e^j}$ to $o_k^{t_s^k}$, and $\mathbf{z}_k^{t_s^k}$ ($\mathbf{z}_j^{t_e^j}$) is the velocity of T_k (T_j) at instant t_s^k (t_e^j).

We use pedestrian detection results in [24, 23] as the association inputs. For fair comparison, we also list their tracking results in the experiments. α in (4) is set as 0.4 and 0.2 for PETS 2009 and TUD-Stadtmitte respectively. TL in (29) is set as 25 for both sequences, we do not link two tracklets with a large time gap, since this association may be unreliable. Finally, two kinds of metrics are applied to evaluate the tracking performance. The first is the CLEAR MOT metric [3]. The second metric [24, 23] evaluates the numbers of mostly/partially tracked (MT/PT), mostly lost (ML) trajectories, numbers of fragments and ID switches.

We compare our approach with some state-of-the-art tracking algorithms [19, 17, 23, 24]. Quantitative results are presented in Tab. 3 and Tab. 4, and results of [19] are the performances of further association on tracklets. Our approach is much better than the tensor method, there are less fragments and ID switches, as well as higher TA and TP. The advances on the performances illustrate the effectiveness of the motion context on reducing association errors and merging short tracklets into long tracks. Our approach has more ID switches than the method [23] on PETS 2009, as most errors in our approach are made in the low-level association, where we use the ordinary color histogram. We believe a more powerful appearance model is helpful in reducing ID switches. Generally, our approach has lower fragments and higher MT than methods [23, 24]. For deep analysis on the motion context, we also give the results of the approaches with different high-level contexts. It can be seen both contexts improve the tracking results, and a combination of two kinds of high-level contexts advances the performances significantly. Qualitative illustrations of our approaches on two datasets are presented in Fig. 7.

6. Conclusion

In this paper, we propose a new association-based MTT algorithm by integrating motion context in a power iteration

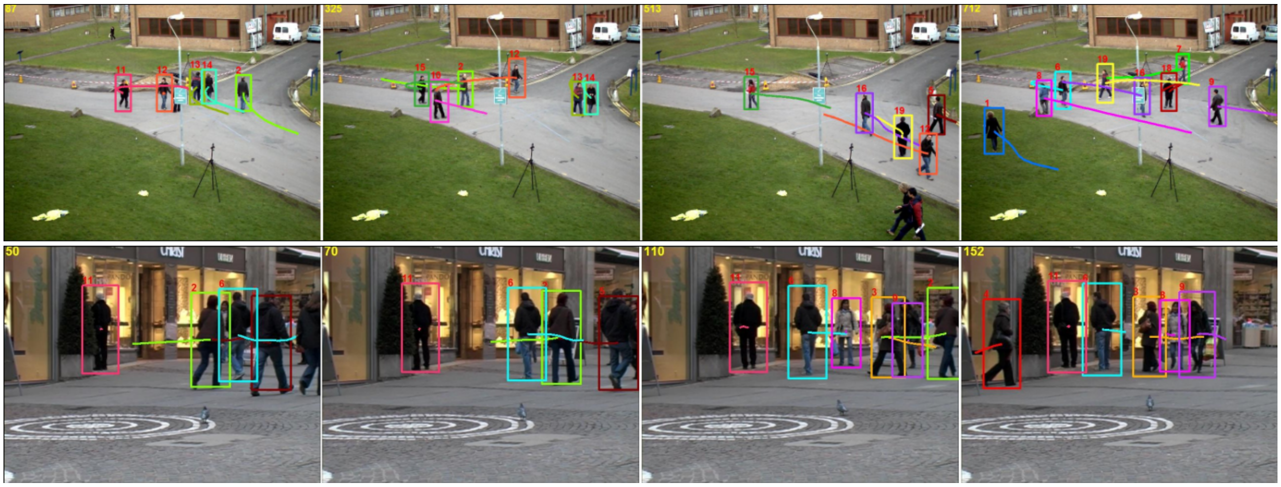

Figure 7. Tracking results of our approach on the pedestrian datasets. Top: PETS 2009 sequence, Bottom: TUD-Stadtmitte sequence.

Table 4. Tracking results on TUD-Stadtmitte

	Rec	Prec	TA	TP	MT	PT	Frag	IDS
[24]	**87.0**	96.7	-	-	70.0	30.0	1	0
[17]	83.8	96.5	75.9	82.6	80.0	20.0	10	8
[19]	83.9	98.8	80.4	87.7	70.0	30.0	5	3
A	85.4	98.6	81.3	87.8	80.0	20.0	2	2
B	83.7	99.7	81.8	88.8	80.0	20.0	2	1
Ours	84.0	**99.9**	**82.5**	**89.3**	**80.0**	**20.0**	**1**	**0**

framework. Our method seamlessly models the interaction energy between target trajectories and the energy of individual trajectories in a united optimization framework. Such integration allows us to use simultaneously contextual cues and high-order motion information to alleviate the association ambiguity. The effectiveness of the proposed approach is demonstrated clearly thorough experiments.

Acknowledgement. This work is partly supported by the 973 basic research program of China (No. 2014CB349303), the National 863 High-Tech R&D Program of China (No. 2012AA012504), the the Natural Science Foundation of China (No. 61303178), the Natural Science Foundation of Beijing (No. 4121003), and the Guangdong Natural Science Foundation (No. S2012020011081). This work is also supported in part by the US NSF Grant IIS-1218156 and the US NSF CAREER Award IIS-1350521.

References

[1] S. Ali, V. Reilly and M. Shah. Motion and appearance contexts for tracking and re-acquiring targets in aerial videos. In *CVPR*, 2007. 2

[2] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *TPAMI*, 33(9):1806–1819, 2011. 1, 2

[3] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP J. on Image & Video Proc.*, 2008. 7

[4] 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. *TPAMI*, 33(9):1820–1833, 2010. 2

[5] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *CVPR*, 2011. 2

[6] A. Butt, and R. Collins Multi-target Tracking by Lagrangian Relaxation to Min-cost Network Flow. In *CVPR*, 2013. 1, 6

[7] R. Collins. Multitarget data association with higher-order motion models. In *CVPR*, 2012. 2, 3, 6

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010. 1

[10] W. Ge, R. Collins, and R. Ruback. Vision-based analysis of small groups in pedestrian crowds. *TPAMI*, 34(5):1003–1016, 2012. 2, 5

[11] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282, 1995. 2

[12] H. Jiang, S. Fels, and J. Little. A linear programming approach for multiple object tracking. In *CVPR*, 2007. 1

[13] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *TPAMI*, 27(11):1805–1819, 2005. 2

[14] M. Luber, J. Stork, G. Tipaldi, and K. Arras. People tracking with human motion predictions from social forces. In *ICRA*, 2010. 2

[15] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. Youll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[16] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*. 2010. 2

[17] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 1, 2, 6, 7, 8

[18] P. Scovanner and M. Tappen. Learning pedestrian dynamics from the real world. In *ICCV*, 2009. 2

[19] X. Shi, H. Ling, J. Xing, and W. Hu. Multiple target tracking by rank-1 tensor approximation. In *CVPR*, 2013. 1, 2, 3, 4, 6, 7, 8

[20] X. Shi, H. Ling, E. Blash, and W. Hu. Context-driven moving vehicle detection in wide area motion imagery. In *ICPR*, 2012. 6

[21] K. Yamaguchi, A. Berg, L. Ortiz, and T. Berg. Who are you with and where are you going? In *CVPR*, 2011. 2

[22] The Columbus Large Image Format CLIF dataset 2006. www.sdms.afrl.af.mil/index.php?collection=clif2006. 5

[23] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *CVPR*, 2012. 7

[24] B. Yang and R. Nevatia. An online learned crf model for multi-target tracking. In *CVPR*, 2012. 7, 8

[25] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 1, 2