Beyond Group: Multiple Person Tracking via Minimal Topology-Energy-Variation

Shan Gao, Qixiang Ye, *Senior Member, IEEE*, Junliang Xing, *Member, IEEE*, Arjan Kuijper, Zhenjun Han, Jianbin Jiao, *Member, IEEE*, and Xiangyang Ji

***Abstract*— Tracking multiple persons is a challenging task when persons move in groups and occlude each other. Existing group-based methods have extensively investigated how to make group division more accurately in a tracking-by-detection framework; however, few of them quantify the group dynamics from the perspective of targets' spatial topology or consider the group in a dynamic view. Inspired by the sociological properties of pedestrians, we propose a novel socio-topology model with a topology-energy function to factor the group dynamics of moving persons and groups. In this model, minimizing the topology-energy-variance in a two-level energy form is expected to produce smooth topology transitions, stable group tracking, and accurate target association. To search for the strong minimum in energy variation, we design the discrete group-tracklet jump moves embedded in the gradient descent method, which ensures that the moves reduce the energy variation of group and trajectory alternately in the varying topology dimension. Experimental results on both RGB and RGB-D data sets show the superiority of our proposed model for multiple person tracking in crowd scenes.**

***Index Terms*— Multiple person tracking, group tracking, RGB-D data, topology.**

I. Introduction

Multiple person Tracking is a fundamental problem in computer vision, contributing to many applications including robotics, video surveillance, and intelligent vehicles [1]. While many researchers consider multiple person tracking in simple scenes a solved problem, in crowded scenes it remains a very challenging problem when considering complex target dynamics and occlusions. Conventional data association methods that link target detections with respect to their appearance, motion, and time gap have been intensively investigated. Modeling complex target dynamics and handling occlusions in crowds; however, are often beyond the scope of their capability.

To model the complex dynamics of moving persons, social behaviour analysis [2], [3] have recently been explored. Sociologists [4] find that up to 70% of the persons in a crowd tend to walk in groups. Persons in the same group are more likely to have similar motion patterns and to be close to each other for a better group interaction. This grouping view treats persons' motion as the result of both their attention and the interactions with the environment. Corresponding tracking models [5]–[9] divide the detections with similar motion patterns into groups and minimize the in-group energy. The recent trend towards group division and group tracking have mitigated the occlusion problem in multiple person tracking to an extent, but the group dynamics remains not being quantified in a principled way. This challenge opens space for new techniques that cope with multiple person tracking, behaviour analysis, or both.

In multiple person tracking, group dynamics refer to person and group events, *i.e.*, person leaving and joining, group merging and splitting, that substantially modify the target spatial distribution and group configuration. From a sociological perspective [10], when an individual leaves a group (split) or joins a group (merge), the social relationships among the remaining persons are revised so that the individuals produce new entities. One extreme example is that persons frequently jump among different groups and introduce frequent group splitting and merging, as well as serious occlusions. Under these circumstances the heterogeneity nature of different groups and many factors, such as individual characteristics, group size, relationships among groups, and influences among group members, need to be investigated.

In this paper, our motivation is to describe topology relation of the persons and groups in a global way. The "topology" is defined in a social context, *i.e.*, to present the relation of persons both in and out of a group. We investigate the spatial and temporal change of the intra- and inter-groups from a socio-topology view, which is higher than the conventional group-level multiple persons tracking methods. The latter ones focus mainly on how to find the similar motion behavior among the tracklets of persons, while the relation among groups is seldom investigated. In our model, the spatial relationships of targets and groups are characterized with intra-group and inter-group structures. The intra-group structure characterizes

Manuscript received August 11, 2016; revised February 19, 2017 and April 5, 2017; accepted May 20, 2017. Date of publication May 26, 2017; date of current version September 1, 2017. This work was supported by Projects of the National Natural Science Foundation of China under Grant 61325003, Grant 61620106005, Grant 61601466, and Grant 61672519, and in part by the China Post-doctoral Science Foundation under Grant 2016M601028. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai. *(Corresponding author: Xiangyang Ji.)*

S. Gao and X. Ji are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: gshan@mail.tsinghua.edu.cn; xyji@tsinghua.edu.cn).

Q. Ye, Z. Han, and J. Jiao are with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qxye@ucas.ac.cn; hanzhj@ucas.ac.cn).

J. Xing is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China. (e-mail: jlxing@nlpr.ia.ac.cn).

A. Kuijper is with the Fraunhofer Institute for Computer Graphics Research, Technology University of Darmstadt, 64283 Darmstadt, Germany (e-mail: arjan.kuijper@mavc.tu-darmstadt.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2708901

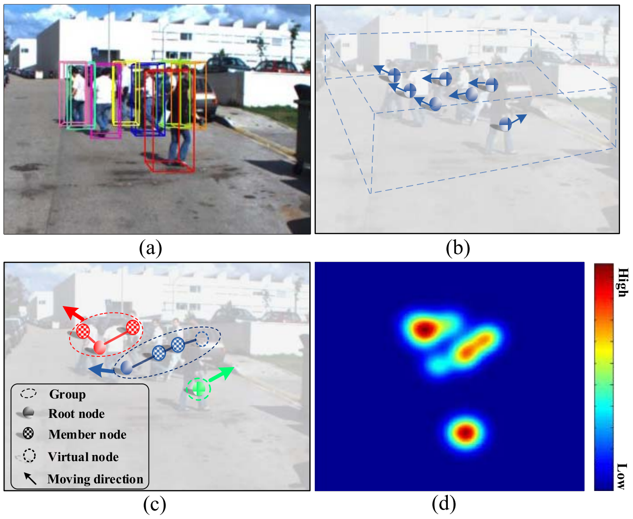


Fig. 1. Illustration of proposed topology-energy model. (a) Multiple person tracking results, where each target is marked with a 3D bounding box. (b) Tracklets of persons in (a). Each node together with an arrow denotes a tracklet with an orientation. (c) Topology configurations corresponding to persons. Each color indicates a group. (d) Energy distribution of the topology.

the connections among the targets inside a group while the inter-group structure models the influence among different groups. To quantify the dynamics, we propose a topology energy function (*cf.* Fig. 1). The typical spatial patterns are off-line learned in different densities on large-scale datasets to accurately characterize such topology-energy distribution. Minimizing the topology-energy-variance via a group-tracklet jump-moves strategy is expected to produce smooth topology transitions, stable group tracking, and accurate target association. Most different with previous group-level models, we minimize the energy variation of the topology rather than the overall energy to infer the grouping dynamics, which is based on the observation that persons moving in two continuous frames usually have slight motion variation and smooth energy variation. The contributions of this paper are summarized as follows:

- We propose a novel social-topology model with a topology-energy function to characterize and quantify the group dynamics of moving persons and groups combining intra- and inter-group structures and learned typical topology patterns.
- We minimize the topology-energy-variance with a group-tracklet jump-moves strategy to solve the multiple person tracking problem, which results in smooth topology transitions, stable group tracking, and accurate target association.
- We explore the proposed model on both RGB and RGB-D datasets, showing the comparable performance in crowd scenes.

The remainder of this paper is organized as follows: the related work is described in Section II. The group-tracklet-level topology energy model is introduced in III and the proposed social topology model with a minimum topology-energy-variation optimization is presented in Section IV and V. Section VI describes the implementation details. Experimental results and conclusions are presented in Sections VI and VII respectively.

II. Related Work

Multiple target tracking approaches in literature [11]–[18] can be coarsely categorized into online tracking and offline tracking.

In online tracking category, target detection responses and their correspondences are jointly estimated and updated for current frame using the information acquired from previous frames [11], [13], [19]–[21]. Methods such as Kalman Filter and Particle Filter [22] are usually adopted to estimate the intermediate states, while sparse represantation [23], online feature learning [19], [20] and Hungarian algorithm [13] are used to calculate target responses. Target occlusions are modeled as merging and splitting of tracklets and solved by using Markov Chain Monte Carlo (MCMC) [24]. Despite of the online advantages of these approaches, they tend to fail when encounter challenges from serious occlusions, target appearance variations and/or complex person motion. Because they can use only short-term target observations, lacking long-term tracklets association or optimization.

In offline tracking category, target detection responses from object detectors are gradually formed into tracklets and the final tracks are obtained by associating the tracklets at different granularities [25]. Existing approaches have widely adopted the data association method which is typically formulated as a graph model, *e.g.,* a cost-flow network [12], [26]–[28], and solved by optimization algorithms including K-shortest path [27], Conditional Random Field (CRF) [29], [30], and metric learning [16]. To guarantee the trajectory smoothness in the graph, for example, Wen *et al.* [31] adopted a tracklets-dense neighborhood searching strategy. Zamir *et al.* [32] and Dehghan *et al.* [33] defined a fully-connected graph to connect all the person detections. To discriminate different person, Yang *et al.* [29], [34] used a trajectory-based CRF function to online learn the affinity and dependency. Yang *et al.* [17] adopted incremental learning to learn temporal dynamic appearance among the person observations. Andriyenko *et al.* [35] and Milan *et al.* [36] presented the energy minimization methods with trajectory-level constraints to distinguish persons. Despite of the global/local optimization advantages, most of these trajectory-level methods ignore the interactions among persons and interactions between persons and the environments, and therefore could be challenged in scenes of target occlusions and complex dynamics.

To perform multiple person tracking in scenes of occlusions and complex dynamics, the socially-aware constraint [37]–[39] has attracted increasing attention. Leal-Taixé *et al.* [40] learned a dictionary of interaction feature on image-level to capture interactions among individual persons, which leaded to a much richer representation for the motion information of persons. Following this clue, Milan *et al.* [41] exploited image-level information and associated super-pixel to a specific target or classified it as background. They then used the segmentation to recover the target information that were missed by the detection response. Choi *et al.* [20] introduced a local flow descriptor that encoded the relative motion pattern between a pair of temporally distant detections using long term interest point trajectories, which provided a robust affinity measure for estimating the likelihood of matching detections. Socially-aware constraint in data association had also been explored. Alahi *et al.* [42] proposed a social affinity map feature to define the motion feature in crowd scenes and utilized it to solve the large-scale pedestrian forecasting problem. Gning *et al.* [43] learned the network structure for the target-group as an evolving graph model, which was propagated combined with a sequential Monte Carlo method.

Social behavior recently has caught more attention in tracking community. Typical social factors include a persons destination, desired speed, and repulsion from other individuals, as well as social grouping behavior. Ge *et al.* [44] automatically discovered small groups of individuals traveling together to infer social groups given a tracking result. Chen *et al.* [5] adopted an online learning strategy to formulate the social behaviour as an elementary grouping model. Leal-Taixé *et al.* [9] proposed a linear programming based group model to track multiple persons, in which the social force [2] was used to model the persons' behaviours.

Most relevant works are from [6]–[8], [45] that leverage behaviour analysis to model person motion and group dynamics. Yamaguchi *et al.* [6] formulated person behavior as an energy minimization model to infer grouping for better trajectory prediction and behavior prediction. The model viewed persons as decision-making agents that considered a plethora of personal, social, and environmental factors to decide where to go next. Using the similar framework, Pellegrini *et al.* [7] modelled the dynamic social behaviour in an energy view, which considered several important aspects of human behavior: future moving destination, environment, and collisions. Qin and Shelton [8] presented a data association approach to exploit social grouping dynamics, where multiple tracklets were clustered together in group entities. They adopted a two-step optimization that provides both tracklet-tracklet associations (to link multiple tracklets of the same person) and tracklet-group associations. Motivation behind their work was that tracklets belonging to the same group should be related to the same individual with higher probability than the tracklets associated to different groups.

Inspired from group-based multiple tergets tracking methods, we propose a topology energy model to quantify person behaviors and group dynamics. Our work is related to [45], constructing a tight relation of mutual support between the modeling of individuals and groups, promoting the idea that groups are better modeled if individuals are considered and vice versa. Our model inherits advantages from above reviewed social grouping models, but differs in two crucial aspects: 1) We propose a topology-level model to describe the group dynamics, expecting higher intra-group but lower inter-group energy distributions rather than minimizing the summation of all the groups' energy. 2) We propose using learned topology patterns and a topology-energy-conservation strategy to track groups and individuals, which is expected to produce smoother topology transitions, more stable group tracking, and more accurate target association.

TABLE I
NOTATIONS IN THIS PAPER

Symbol	Description
l_i^t	coordinates of the tracklet l_i at frame t
A_{ij}	RGB-D affinity between l_i and l_j
X_{ij}	binary indicator between l_i and l_j
T_{ik}	binary indicator between l_i and G_k
G_k	the kth group in $\mathcal{T}$
V_{ik}	velocity affinity between l_i and G_k
Ψ_{ik}	topology affinity between l_i and G_k
ϕ_{ik}	orientation affinity between l_i and G_k
d	threshold of distance between tracklets
E	energy of topology
ΔE	energy variation of topology

III. Topology-Level Multiple Person Tracking

To make the paper self-contained, we review the classic multiple person tracking formulation with data association. For ease of reference, we list all the denotations used in this paper in Table I. Let $L = \{l_1, l_2, \cdots, l_n\}$ denote all person tracklets of a video sequence, supposing N persons in F frames. A tracklet l_i is a consecutive sequence of detection responses that contain the same target. A binary association indicator, X_{ij}, defines the hypothesis that pairwise tracklets l_i and l_j contain the same person (l_i occurs before l_j). A_{ij} denotes the tracklet affinity between tracklets l_i and l_j. The goal of multiple person tracking is to associate tracklets that correspond to the same targets, by minimizing a cost function:

$$\arg\min_X \ \underbrace{\sum_{i,j} A_{ij} X_{ij}}_{\text{tracklet}},$$

$$s.t.\ X_{ij} = \begin{cases} 1 & \text{if } l_j \text{ is associated after } l_i, \\ 0 & \text{otherwise}, \end{cases}$$

$$\sum\nolimits_i X_{ij} \leq 1 \ \text{ and } \sum\nolimits_j X_{ij} \leq 1. \tag{1}$$

Here we would like to obtain the binary label X_{ij} that indicates whether they are the same person (1) or not (0). Constraints $\sum_i X_{ij} \leq 1$ and $\sum_j X_{ij} \leq 1$ imply that each tracklet follows at most one tracklet, except for the first and last ones.

The topology relation in this paper is designed to model individual and group dynamics, which constructs a tight relation of mutual support between the modeling of individuals and groups as a high-level constraint. Based on tracklet-level association function Eq. (1), we add the group association matrix, T_{ik}, which is an indicator matrix similar with the tracklet indicator X_{ij}. If tracklet l_i belongs to G_k, then $T_{ik} = 1$, otherwise, $T_{ik} = 0$. The topology function can be written as

$$\begin{aligned}
&\arg\max_{X,T} \underbrace{\sum_{i,k} \Psi_{ik} T_{ik}}_{\text{group}} + \alpha \underbrace{\sum_{i,j} A_{ij} X_{ij}}_{\text{tracklet}}, \\
&s.t.\ T_{ik} = \begin{cases} 1, & \text{if } l_i \text{ belongs to group } G_k, \\ 0, & \text{otherwise}, \end{cases} \\
&\quad X_{ij} = \begin{cases} 1, & \text{if } l_j \text{ is associated after } l_i, \\ 0, & \text{otherwise}, \end{cases} \\
&\quad \sum\nolimits_k T_{ik} = 1, \quad \sum\nolimits_i T_{ik} \geq 1, \\
&\quad \sum\nolimits_i X_{ij} \leq 1, \quad \sum\nolimits_j X_{ij} \leq 1. \qquad (2)
\end{aligned}$$

Here Ψ_{ik} denotes the affinity between the tracklet l_i and group G_k. The constraints on X_{ij} and T_{ik} satisfy $\sum_i X_{ij} \leq 1$, $\sum_j X_{ij} \leq 1$, $\sum_k T_{ik} = 1$, and $\sum_i T_{ik} \geq 1$. They require that each tracklet belongs to only one group G_k, and one group contains at least one tracklet. The RGB-D affinity A_{ij} that captures appearance and depth consistency of persons is detailed in Sec. VI-D.

The function in Eq. (1) is clearly non-convex, and the function in Eq. (2) is more complex. Solving it directly is computationally expensive and easily getting stuck into a local optimum. In the tracking scenario, however, the energy form of the topology model is still helpful to the tracking results if the model is well initialized and optimized to a better objective. Previous researches [6], [7], [46] design energy minimization methods to assign every possible solution an 'energy' and then find the state with the lowest energy. In this paper, we improve the energy model from tracklet level to a topology level, a joint energy form of group and tracklets in a social topology. To keep the energy of topology as natural as possible, the energy function is defined in consistent with Eq. (2):

$$E = E_{gro} + \alpha E_{tra}, \qquad (3)$$

where the E_{gro} and E_{tra} denote the energy of group and tracklets, respectively. The group term E_{gro} keeps the tracking targets in groups with a high social affinity, the tracklets energy E_{tra} captures the motion and appearance of persons in a tracklet-level. α is a regulation factor.

IV. Topology-Energy Model

The topology energy function, composed of the group energy and the tracklets energy terms, is defined in this section. The definition of the energy function is to quantify the socio-topology relations among groups and tracklets.

A. Group Energy

The topology term in Eq. (3) is inspired by researches on social topology relations [3], [47], which suggest there are three governing sociological factors related to pedestrian movement:

1) Some pedestrians follow the same routes to specific geographical goals.
2) Each pedestrian walks at a maximum speed depending on certain environmental conditions.
3) Pedestrians continuously adjust their positions to facilitate in-group verbal exchange and avoid collisions with out-group pedestrians.

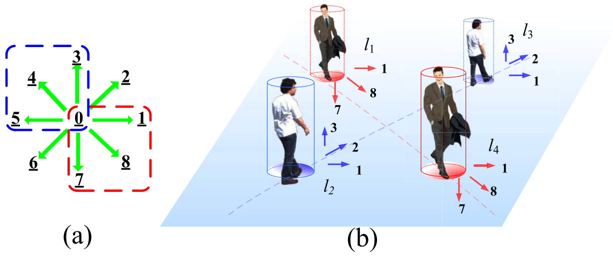


Fig. 2. Illustration of the moving orientations. (a) 1-8 bins denote different orientations, and 0 denotes that a pedestrian keeps still in successive two frames.

The first two factors are modeled as moving orientation and speed terms in the intra-group relation. The third factor, distance term, is modeled as an inter-group relation. The group energy in Eq. (3) is modeled as

$$E_{gro} = E_{intra} - \beta_1 E_{inter}, \qquad (4)$$

where, the parameter β_1 is used to balance the energy between intra- and inter- parts. When performing group-based multiple person tracking, we expect the maximum-intra-group-affinity and minimum-inter-group-affinity.

Intra-Group. The intra-group component is considered from the perspective of each in-group members' motion feature, which defines the affinity between a tracklet l_i and a group G_k. This component contains the speed affinity as

$$V_{ik} = e^{\frac{-\|v_i^t - \bar{v}_k^t\|^2}{2\sigma^2_{\bar{v}_k^t}}}, \qquad (5)$$

where v_i^t denotes the speed of the tracklet l_i at frame t and $\bar{v}_k^t$ denotes the average speed of group G_k at frame t. V_{ik} thus defines the speed affinity between l_i and G_k. In the orientation constraint, a similar Potts model [48] is adopted to define the affinity among different moving orientations:

$$\phi_{ik} = \frac{1 - \cos(\varphi_i^t - \bar{\varphi}_k^t)}{2}, \qquad (6)$$

$$\varphi_n = \frac{2\pi n}{q}, \qquad (7)$$

where ϕ_{ik} defines the moving orientation between tracklet l_i and group G_k. Persons' moving orientations φ are quantized into $q = 9$ bins (*cf*, Fig. 2) with φ_n the bin's label.

The speed and orientation factors together force in-group members have similar motion patterns, at the same time, reduce the identity switching and smooth the in-group trajectory during the poor detections. Upon this motion pattern, intra-group energy is defined as

$$E_{intra} = \sum_{k}^{K} \sum_{i, l_i \in G_k}^{N} V_{ik} \phi_{ik}, \qquad (8)$$

where K is the number of groups at frame t. We omit the superscript t in E_{intra} for simplification. Instead of considering the speed and orientation together as in [5] and [44], we calculate these two factors in two terms, which guarantees the social topology model be applicable in both RGB and RGB-D datasets and enables it effective against the poor detections. Especially, the '0' bin in orientation term is assigned to the stationary object, which makes the stationary pairwise tracklets keeping a stable group energy distribution.

Inter-Group. The inter-group energy in Eq. (4) is designed to make in-group members be close with the group center to facilitate communication and the motion affinity among different groups as large as possible to avoid a collision:

$$E_{inter} = \sum_{k}^{K} \sum_{i,l_i \in G_k}^{N} \frac{d}{\left\| l_i^t - \bar{G}_k^t \right\|^2} + \beta_2 \sum_{k,p,k \neq p}^{K} \frac{\left| V_k^t \phi_k^t - V_p^t \phi_p^t \right|}{\left| \bar{G}_k^t - \bar{G}_p^t \right|}, \tag{9}$$

where $\bar{G}_k$ is the state of the group center. Two tracklets can be divided into the same group G_k only if their distance satisfies $D(l_i, l_j) < d$. The parameter β_2 is used to balance the energy between two parts of distance. In the second component, the average group speed, orientation and distance are jointly measured to discriminate groups. Note that the inter-group term plays a role as clustering, which enforces the trajectory alignment in group. Especially when the in-group members lack of image evidences (without detections), the grouping results are able to provide a soft transition way to smooth the trajectory.

B. Tracklets Energy

The second component of social topology energy model is tracklets energy. For each tracklets, the target should be also matched with detections in the lowest identity switches and fragments. Two factors, motion and appearance constraints, are designed to regulate the tracklets energy form as

$$E_{tra} = E_{mot} + \beta_3 E_{app}. \tag{10}$$

where, E_{mot} and E_{app} denote the motion and appearance energy of tracklets respectively. The parameter β_3 is used to balance two terms.

1) Motion: Tracklets is about track combinations of one person moving in successive frames, which in most cases is smooth. Upon this, the target motion model is measured by minimizing the distance between state vectors:

$$E_{mot} = \sum_{i}^{N} \left\| l_i^{t+1} - l_i^t \right\|^2. \tag{11}$$

The motion energy model poses a tracklet-level constraint to data association. Such a constraint encores short-term smoothness and helps avoid target drifting. In addition, a global tracklet association is used in a relatively long duration to bridge, cut, and grow tracklets in model optimization, detailed in Sec. V.

2) Appearance: Track fragmentation is mainly cased by the missing evidence. To connect two tracklets fragments across the no-evidence area, we add the appearance penalty in the tracklets energy. This is based on hypothesis the appearance feature of a person usually changes smoothly in consecutive frames. To keep the term both robust and smooth, we use a sigmoid function as,

$$E_{app} = \sum_{i}^{N} \frac{1}{1 + \exp(1 - A(l_i^{t+1}, l_i^t))}, \tag{12}$$

where $A(l_i^t, l_i^{t-1})$ is calculated by the widely-used Bhattacharyya coefficient on appearance features. We improve the appearance feature in terms of different datasets to promote it better fit the multiple person tracking problem, as shown in Sec. VI-D. Moreover, the appearance model is designed to fit gradient-based optimization manner (*cf.* Sec. V) according to the property of sigmoid function.

V. Model Optimization

With the defined topology energy model, the conventional multiple person tracking problem is formulated by minimizing the energy in an analytical inference way. Nevertheless, the proposed topology energy model in Eq. (3) is obviously not convex. When adopting the Hungarian algorithm and dual-decomposition method [8] to solve the model, it is computationally expensive and easy getting stuck into a local optimum. We therefore propose a topology-energy-conservation strategy to solve the energy model.

A. Topology Energy Variation

A property of tracking is that persons move slowly and smoothly relative to the frame rate. This consensus means that the motion and appearance of persons change slightly in consecutive frames. In the optimization, we do not primarily minimize the energy function defined in Eq. (3), but the *change* of the topology energy. The change of topology is measured with an energy variation function ΔE between two frames, defined as

$$\Delta E = \sum_{t}^{F} \left| E^{t+1} - E^t \right|, \tag{13}$$

where ΔE measures the energy variation in continuous frames. Fig. 3 visualizes the quantified topology energy in one tracking example. The correct tracking results have lower and stable energy variation, while the false tracking results (identity switch and/or group division erorr) have larger energy variation.

B. Minimizing Topology-Energy-Variation

To obtain a reasonable group and trajectory solution, the energy variation defined in Eq. (13) is minimized as

$$\arg\min_{X,T} \Delta E. \tag{14}$$

The standard conjugate gradient method is adopted to minimize the energy function in each iteration. To speed up the convergence, and get out of the weak local minimum, a two-level jump strategy is proposed for changing the dimension of the current state, which can be regarded as a lite version of reversible jump Markov Chain Monte Carlo (RJ-MCMC)

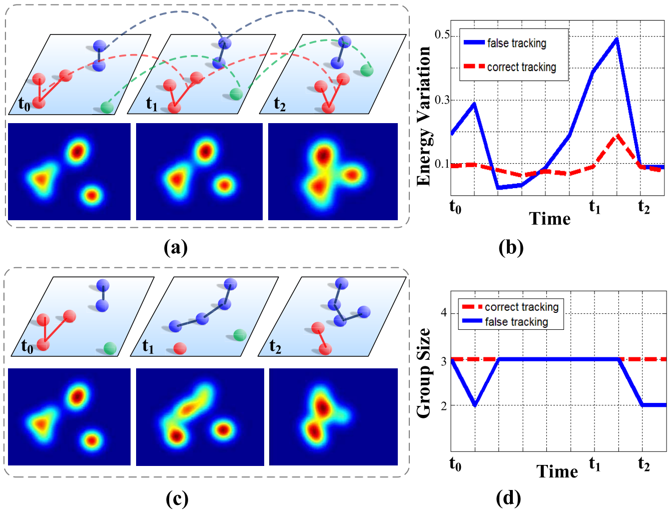


Fig. 3. Correct and false topology tracking results are shown in (a) and (c) respectively. The topology energy in (a) changes slightly, but the topology energy in (c) changes abruptly. (b) shows the energy variation of (a) and (c). (d) shows the change of group's size in (a) and (c). Best viewed in color.

Algorithm 1 Model Optimization

```
Input: Detections in video sequence
Output: Group and trajectory results $T$, $X$
Generate tracklet set $\{l_i\}$ by maximum overlap criterion;
Generate group set $\{G_k\}$ by the threshold $d$; for $k = 1 : K$ do
    while $\neg$ converged do
        for group jump $J_g \in \{bridge, split, merge\}$ do
            if $\Delta E_{new} < \Delta E_{cur}$ then
                execute group jump $J_g$;
            end
            for $i = 1 : N,\ l_i \in G_k$ do
                for tracklet jump $J_t \in \{bridge, grow, cut\}$ do
                    if $\Delta E_{new} < \Delta E_{cur}$ then
                        execute tracklet jump $J_t$;
                    end
                end
                execute conjugate gradient descent;
            end
        end
    end
end
for $i = 1 : N,\ l_i \notin \{G_k\}$ do
    while $\neg$ converged do
        for tracklet jump $J_t \in \{bridge, grow, cut\}$ do
            if $\Delta E_{new} < \Delta E_{cur}$ then
                execute tracklet jump $J_t$;
            end
        end
        execute conjugate gradient descent;
    end
end
```

methods [49]. RJ-MCMC has been improved as different fashions and applied to multiple person tracking problem in [43], [44], and [46]. Most different from the previous researches, we design two-level jump moves embedded in the energy variation framework.

The two-level jump moves, called group-tracklet jump, are designed to fit the group-tracklets energy form. The jump moves are able to change the dimensionality in both of groups and trajectories (X_{ij} and T_{ik}), which give the optimization a high degree of flexibility and a smooth transition among groups and trajectories. With the aid of a topology initialization (*cf.* V-D), the optimization is able to get much closer to a global minimum at limited iterations. Moreover, the initial topology configuration does not require the correct number of targets and groups.

C. Jump Moves

To get rid of weak local energy minima, we introduce two types of jump moves, which change the configuration of the current group and trajectory solution, thereby altering the dimension of the current state $L = \{l_i\}$ and $G = \{G_k\}$ after each iteration. By jumping to different regions of the search space while always lowering the energy variance, the optimization is able to find much stronger local minima.

1) Group Jumps: We design group bridging, merging and splitting jumps. Group bridging jump extends the group length and connects it with another group, which facilitates the tracklets association. Group merging and splitting can effectively improve the group association, which plays a group association role in group transition.

2) Group Bridging: The temporal span of a group could be cut off by missing or false detections, particularly in severe occlusions. We thus extend the group length both spatially and temporally to bridge the gap between groups. Denoting $G_k = G_k^{b_k:e_k}$ the temporal span of a group between frames b_k and e_k, two groups G_i and G_j are connected into one group if the new energy variation $\Delta E_{new} < \Delta E$, as:

$$G_k = (G_i, G_{con}^{e_{i+1}:b_{j-1}}, G_j). \tag{15}$$

Moreover, the group connecting provides a stronger tracklet-association solution for the in-group tracklets. As soon as two groups connected as one group, the trajectory of the members in $G_{con}^{e_{i+1}:b_{j-1}}$ should be connected as well, which refers to tracklet jumps.

3) Group Split and Merge: The group does not always stay stable in the whole tracking. When tracklets stay close enough, they are considered to be merged into a group. While the in-group members show different motion patterns (velocity and/or orientation), those members should be deleted. We check ΔE after each iteration. If ΔE in time t is higher than a certain threshold ε, we give the current group configuration a 'perturbation' to escape the local minimal. The jump goes as a prescribed order (see Alg. 1). For the in-group tracklet l_i^t with the highest energy according to Eq. (4), we delete this member from G_k as an independent target and evaluate the energy variation ΔE_{new}. If $\Delta E_{new} < \Delta E$, l_i^t is deleted from G_k, and $T_{ik} = 0$. For the distance between an independent target l_j^t and group G_k is less than distance threshold d, we add l_j^t in G_k. If $\Delta E_{new} < \Delta E$, merging is executed as $G_k = G_k \cup l_i^t$, and $T_{ik} = 0$.

4) Tracklet Jumps: Tracklet bridging. Let $l_k = l_k^{b_k:e_k}$ denote the state of kth trajectory between frames b_k and e_k. The bridging is kept if the ΔE_{new} is lower than the current energy variation.

$$l_k = (l_i, l_{con}^{e_{i+1}:b_{j-1}}, l_j), \tag{16}$$

5) Tracklet Growing and Cutting: Tracklets can grow in forward and backward as group jumps. If the ΔE_{new} is lower, the growing is $l_i = \left(l_i, l_i^{e_i+1:e_i+t}\right)$ in backward and $l_i = \left(l_i^{b_i-t:b_i-1}, l_i\right)$ in forward. The false detections often make tracklets from different persons connecting to a trajectory. Tracklet cutting is used to eliminate false detections in the long trajectory. Splitting in the frame t with a higher ΔE_{tra} by cutting the trajectory as $l_i = l_k^{b_k:t}$ and $l_j = l_k^{t:e_k}$.

Similar with [46], the jumps among tracklets are able to pick up lost tracks and weed out spurious ones for tracklets. However, there are two types of tracklets in the proposed topology model. One is independent tracklets, $T_{ik} = 0$; the other has the group relation, $T_{ik} = 1$ in the partial or whole time span. For the former, the jump among the tracklets without the group constraint. For the latter, we connect the tracklets in the certain group, which leads a much smaller search space, comparing with the global search. The group connection $G_{con}^{e_{i+1}:b_{j-1}}$ should be smoothed by the tracklet jumps. In the duration $e_i : b_j$, the $E_{G_k}^{e_i:b_j}$ of the group G_k is evaluated by the total energy of its in-group tracklets as

$$E_{G_k}^{e_i:b_j} = \sum_{l_i \in G_k} \left(E_{mot}\left(l_i^{e_i:b_j}\right) + \beta_2 E_{app}\left(l_i^{e_i:b_j}\right)\right), \tag{17}$$

where we minimize the $\Delta E_{G_k}^{e_i:b_j}$ at frames $e_i : b_j$ to make sure each in-group trajectory connecting is correct.

The group-tracklet jumps are executed after each iteration in a fixed order as described in Alg. 1. The group jump is more likely providing a coarse solution as a group solution under socio-constraint. This leads the tracklet jump in a reasonable space (in-group members) rather the jump among trivial tracklets, efficiently improving the optimization to converge. The parameters in jumps: the bridging time step of group and trajectory, the number of frames a trajectory is grown, are executed independently, which conducts the optimization to a lower energy variation.

D. Topology Initialization

Like any non-convex optimization, our algorithm depends on the initial solution from which topology configuration is started. Although the group-tracklet jump strategy is able to conduct the optimization jump out of the local minimal compared with a pure gradient method. If the initial topology solution is far away from the global minimal, it will take more iterations to execute jumps or stuck in a local region with a energy variation.

In initialization, tracklets are generated after low-level association, which is performed in a greedy manner using a maximum overlap criterion. But for grouping, only the tracklets lasting more than f_t frames are considered as confident ones, because most false or ghost tracklets are short ones. The typical group patterns are then adopted to initialize groups via an off-line learning procedure, which is detailed in Sec. VI. If the distance of pairwise tracklets is lower than the threshold d, this grouping result is kept. Actually, this initial solution in greedy fashion is impossible to get to the optimal, but quickly providing a grouping and tracklets association solution for the following group-tracklet jumps. Empirically, different initial solutions converge to similar final group and tracklets results. This is completed by the group-level and tracklet-level jump moves in the following optimization process. The difference is that a better initial solution helps get to the final solution in less jump moves. Thus we investigate the typical group patterns in training datasets, which provides a better initial solution, detailed in Sec. VI.

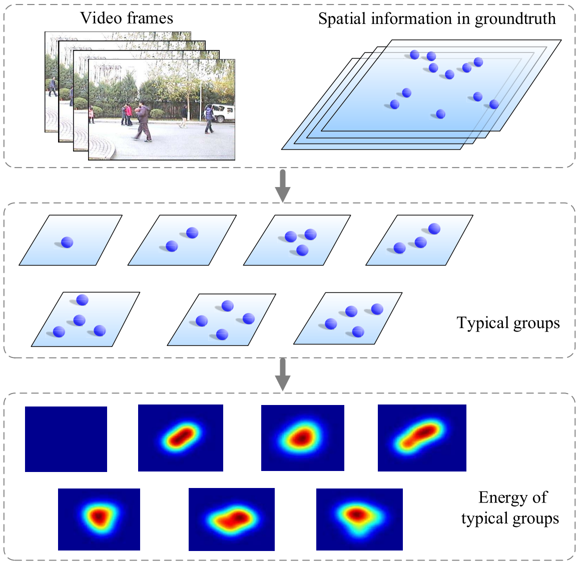


Fig. 4. Visualization of the learned typical group patterns. The size of typical group is ranging from 1 to 4. The energy distribution of the typical group patterns are visualized according to Eq. (4). A brighter color indicates a higher energy distribution. Best viewed in color.

VI. Implementation

In this section, the details about training and optimizing a topology model are described. The details about multiple person tracking implementations are also presented.

A. Typical Group Pattern Training

Typical group patterns are investigated in training datasets to initialize groups by conducting off-line learning. We record typical configuration of groups with a stable energy variation in RGB and RGB-D training dataset, and formulate them as typical group patterns. The topology patterns are learned in RGB and RGB-D training datasets with 13732 frames. In 3D applications with the real-world depth information, d in Eq. (9) denotes the world-coordinate distance (meter) between two persons, while d denotes the distance between the center points of targets' bounding box in images.

In training sequences, given a set of detections and the corresponding ground truth (GT) target and group annotations, the GT targets' IDs are first assigned to each detection as complete trajectories and groups. The parameter in our model

TABLE II
TRAINED PARAMETERS

Parameter	α	β_1	β_2	β_3	d			ε
					low	mid	high	
RGB	0.4	0.3	0.6	0.25	25	20	17	0.05
RGB-D	0.4	0.35	0.6	0.25	2	1.7	1.4	0.05

is not determined manually. We record these parameters in the annotated groups and trajectories: (1) size,[1] (2) average distance d in Eq. (9) among neighbor members. (3) average intra-group and inter-group energy and (4) average tracklets motion and appearance energy. Fig. 4 visualizes several typical group patterns and energy distribution with the size ranging from 2 to 4.

B. Parameter Training

During the typical group pattern training, the intrinsic parameters of topology have been determined, but the weight of all the energy components should be set as well. In training sequences (which have groups of pedestrian), the augmented values of regular parameters α, β_1, β_2 and β_3 are given to evaluate the group and tracklets energy variation. First, the amount of groups and how many times of objects falsely divided into groups are compared with group GT. The optimal α, β_1 and β_2 are selected with the lowest group amount error ε_a and dividing error ε_d as

$$\arg\min_{\alpha,\beta_1,\beta_2} \varepsilon_a + \varepsilon_s. \tag{18}$$

The parameter β_3 is trained in a similar manner, but the reference standard changes as MOTA (*cf.* Sec. VII). We run the tracking algorithm while keeping α, β_1 and β_2 fixed. Table II shows the parameters learned in the training datasets.

C. Crowd Density

According to the density of persons, the group configuration and the energy variation change in three levels. **Low density:** group members tend to move side-by-side in the size of 2, forming a line perpendicular to the moving orientation, thereby occupying a large area. **Middle density:** when the local density level increases, the average distance between group members is, in fact, reduced. Persons in the same group needs to adapt to the reduced availability of space. It is observed that the middle person tends to stand back, while the persons left and right get closer to each other. This is done by configurations of V-like or U-like topology patterns in topology configuration with three and four members. As shown in research [4], these configurations are emergent topology patterns resulting from the tendency of each person to find a comfortable moving position supporting communication with the other intra-group members. **High density:** when the density reaches a high level, the physical constraints would prevail over the social preferences, persons in the same group would start moving one behind another, forming a river-like topology configuration, which corresponds to the aerodynamic features.

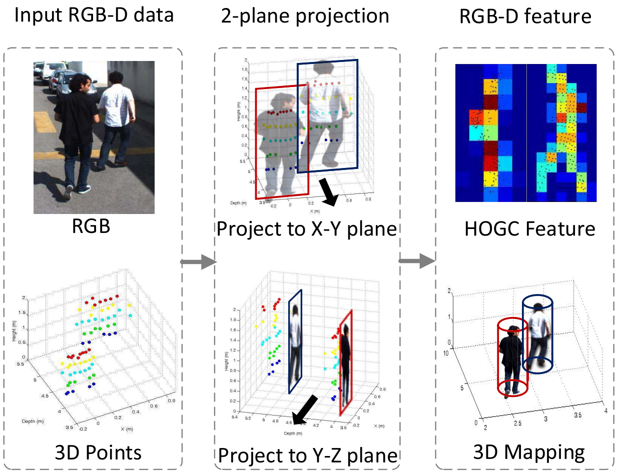


Fig. 5. RGB-D feature extraction. Best viewed in color.

The social distance d in our model changes with crowd density, which is shown in Table II. We divide the crowd density of scenes in three levels according to the number of persons per-frame as well: low ($<$10), middle ([10, 20]) and high ($>$20).

D. RGB-D Affinity

To fit the RGB-D datasets, we fully explore the RGB-D feature for the tracklet appearance. Given a video sequence with depth data (LIDAR or stereo vision), the combined features in ROIs are extracted to describe targets in terms of their appearance and 3D positions.

Assuming that each target is an isolated 3D bounding box, we extract a set of RGB-D based features (Histograms of Oriented Gradient and Color features, and Histogram of Depth (HOD) feature [50], [51]) to discriminate targets from their backgrounds. Note that the one ROI in the image domain can in fact contain more than one targets due to occlusions, which introduces great ambiguity to data association. The RGB-D data is thus projected into two planes, X-Y plane and Y-Z plane, to decrease such ambiguity, as shown in Fig. 5. The Y-Z plane is an auxiliary plane, in which we calculate the average depth value of each target. We define target and background seeds to be the set of pixels inside the bounding box. To compute the target seeds in the projected 2D bounding box, we remove the pixels corresponding to the background seeds as well as a pixel band around the box as [52], which has a larger depth value than the mean depth value in a 2D box. To make this process more robust in the X-Y plane, an online adaptive feature pool, HOGC [53] feature is utilized. In the X-Y plane, 14*7 bins of HOGC features are extracted, and in the X-Z plane, 9*5 bins of variation features [50] on cloud points' locations are extracted. There are 143 bins of RGB-D features in total to represent a target.

[1] Size of a group pattern denotes the number of topology members.

TABLE III

COMPARISONS OF TRACKING RESULTS ON TWO RGB DATASETS IN MULTIPLE OBJECTS CHALLENGE BENCHMARK [54]. FOR THE ITEMS WITH ↑, HIGHER SCORES INDICATE BETTER RESULTS, FOR THOSE WITH ↓, LOWER SCORES INDICATE BETTER RESULTS. RED AND GREEN NUMBERS SHOW THE BEST AND SECOND BEST PERFORMANCE RESPECTIVELY

Dataset	Method	MOTA ↑	MOTP ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDS ↓	Frag. ↓	Hz ↑
MOT Benchmark	DP [28]	14.5	70.8	6.0%	40.8%	13,171	34,814	4,537	3,090	444.8
	SegTrack [40]	22.5	71.7	5.8%	63.9%	7,890	39,020	697	737	0.2
	MotiCon [41]	23.1	70.9	4.7%	52.0%	10,404	35,844	1,018	1,061	1.4
	MDP [19]	30.3	71.3	13.0%	38.4%	9,717	32,422	680	1,500	1.1
	TSMLCDE [16]	34.3	71.7	14.0%	39.4%	7,869	31,908	618	959	6.5
	TDAM [17]	33.0	72.8	13.3%	39.1%	10,064	30,717	464	1,506	5.9
	MHT [18]	32.4	71.8	16.0%	43.8%	9,064	32,060	435	826	0.7
	Ours	33.8	71.1	12.1%	34.8%	9,232	31,743	722	1,257	7.1

VII. EXPERIMENTS

In this section, we evaluate the proposed multiple person tracking algorithm, as well as comparing it with recent state-of-the-art methods. Experimental results clearly show the benefits of utilizing social topology in multiple person tracking.

A. Dataset and Metrics

The proposed methods are tested and evaluated on two kinds of publicly available datasets, RGB and RGB-D datasets, which are summarized in Table V.

The commonly used multiple person tracking metrics defined in [55] is adopted to evaluate the tracking performance following [16]–[19], [28], [40], [41], [46]. The Multi-Object Tracking Accuracy (MOTA) combines three types of errors: false positive (FP), missed targets (FN), and identity switches (IDS), which is normalized such that the score of 100 percent corresponds to no error. All three types of errors are equally weighted. Multi-Object Tracking Precision (MOTP) measures the alignment of the tracker output with respect to the ground truth normalized to the threshold value. Mostly Tracked (MT) and Mostly Lost (ML) scores are computed on the entire trajectories and measure how many Ground Truth trajectories (GT) are successfully tracked (tracked for at least 80%) and lost (tracked for less than 20%). The items Frag. and IDS record how many times the ground truth trajectory is interrupted and switched by a false ID. Hz records the tracker speed in frames per second. In addition, Recall and Precision (Prec.) are two basic metrics. Recall means the rate of correctly matched detections / total detections in ground truth. Precision means the rate of correctly matched detections / total detections in the tracking results.

B. Evaluation on RGB Benchmark

The RGB datasets are from the MOT benchmark [54], which is composed of 11 training and 11 test video sequences, of 11,286 frames (∼16.5 minutes). Some of the videos are recorded using mobile platform and the others are from surveillance videos. As it is composed of videos with various configurations, tracking algorithms that are particularly tuned for a specific scenario would not work well in general.

To keep consistent with previously reported results, we follow the exact same evaluation protocol as other approaches [16]–[19], [28], [40], [41], [46], and use their reported results on MOT website. Unsurprisingly, same detection results are used as inputs to all compared tracking approaches [54]. To verify the robustness, we design three comparing methods based on the proposed model.

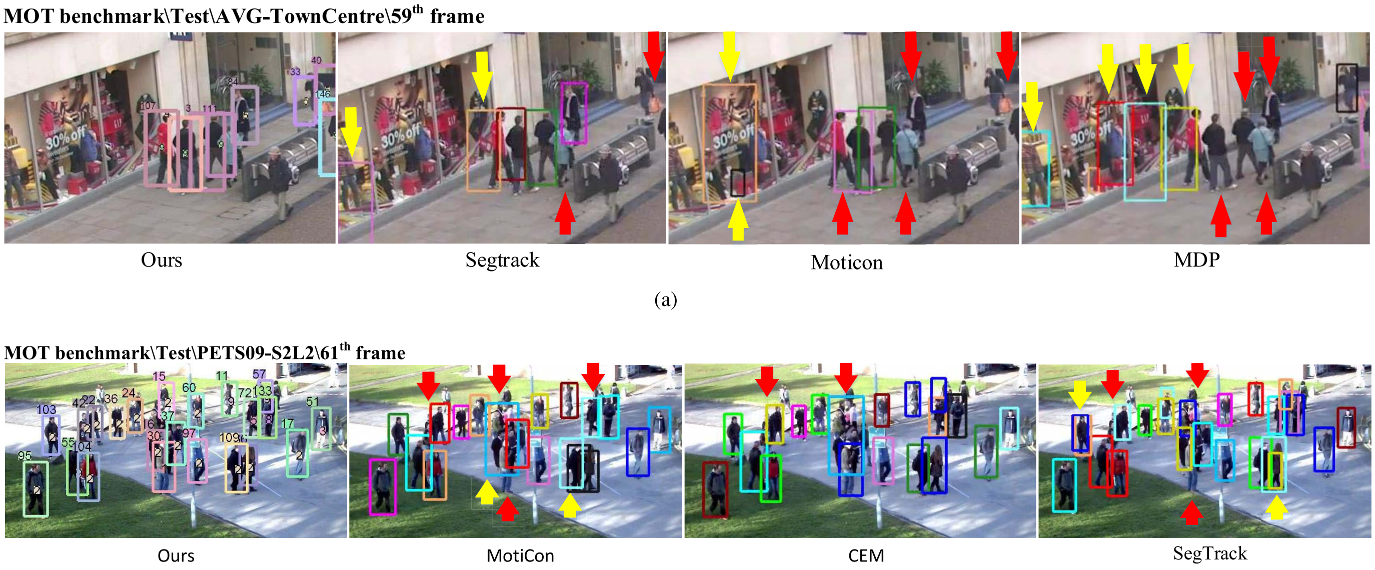
- *Ours* (Our full model, including the inter- and intra-group energy and using the group and tracklet jump moves, the jump threshold of energy variation ε is 0.05);
- *Ours+CGD* (Using purely conjugate gradient descent, which runs until convergence or to maximal number of iterations (here, we set 30), which suffices to get close to a local minimum, this is similar to [46]);
- *Ours+No Inter-G* (Without inter-group term in Eq. (3), only intra-group term and tracklet jump moves);
- *Ours+No Var* (Without topology variance minimization, solving Eq. (3) using the group and tracklet jumps).

Table III summarises the accuracy of the proposed method (GST) and other state-of-the-art methods on the MOT benchmark. It clearly shows that our model achieves the comparable performance. Fig. 6 gives the tracking samples in AVG-TownCentre and PETS09-S2L2 sequences. One frame is randomly selected to show our better performance than results in [19], [40], [41], and [46]. Note that as the AVG-TownCentre sequence is in a high resolution covering the whole street, we zoom in to the street corner at frame 59, and in PETS09-S2L2 sequence we take the result at frame 61. Our approach is able to find more tracklets than the compared methods, particularly, the trajectory in group. Some missed target positions (red arrows in Fig. 6) could be inferred by the group topology, as they keep slight topology variation in successive frames. As the false tracking positions (yellow arrows in Fig. 6) are usually caused by the false detections and usually last short time, our approach that only chooses the confident tracklets in grouping ($f > 5$ frames) shows advantages over compared methods.

Fig. 7 shows the energy minimization with different solution techniques on MOT benchmark. We conclude that the full model "*Ours*" including the complete group-tracklet jumps is able to escape weak local minima, since the purely continuous conjugate gradient descent optimization "*Ours+CGD*" can only search a small local neighborhood of the state space in the case of non-convex energy optimization. The further experiment turning off the inter-group term "*Ours+No Inter-G*" or energy variation minimization "*Ours+No Var*" performs even

TABLE IV
COMPARISONS OF TRACKING RESULTS ON AVG-TOWNCENTRE AND PETS09-S2L2 SEQUENCES IN MOT BENCHMARK

Dataset	Method	MOTA ↑	MOTP ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDS ↓	Frag. ↓
AVG-TownCentre	DP [28]	6.6	69.4	4.4%	35.8%	876	4,482	1,317	562
	SegTrack [40]	3.3	69.3	0.9%	86.3%	235	6,528	151	108
	MotiCon [41]	11.9	70.3	0.9%	69.9%	353	5,872	74	75
	MDP [19]	25.4	69.7	17.7%	33.6%	1,517	3,691	122	264
	TSMLCDE [16]	33.9	68.9	20.4%	31.0%	997	3,604	126	274
	TDAM [17]	25.3	70.3	15.0%	39.4%	1,477	3,794	68	223
	MHT [18]	27.1	70.4	17.3%	44.2%	837	4,300	74	165
	Ours	33.7	70.2	22.1%	30.1%	942	3,756	113	163
PETS09-S2L2	DP [28]	33.8	69.4	7.1%	9.5%	948	4,410	1,029	705
	CEM [46]	44.9	70.2	11.9%	14.3%	657	4,506	150	165
	MotiCon [41]	46.6	67.6	9.5%	14.3%	560	4,354	238	264
	SegTrack [40]	46.1	70.6	26.2%	16.7%	1,213	3,773	211	211
	TSMLCDE [16]	51.5	70.6	14.3%	9.5%	905	3,602	165	198
	TDAM [17]	43.1	69.4	9.5%	11.9%	653	4,673	158	412
	MHT [18]	50.8	70.4	19.0%	7.1%	933	3,667	142	201
	Ours	51.8	70.4	16.7%	11.9%	715	3,812	172	161


MOT benchmark\Test\AVG-TownCentre\59th frame
Ours
Segtrack
Moticon
MDP
MOT benchmark\Test\PETS09-S2L2\61th frame
Ours
MotiCon
CEM
SegTrack

(a)

(b)

Fig. 6. Comparison of our approach with other methods on RGB datasets. The results of comparing methods are downloaded from the MOT website. Red arrows denote the missing targets, while yellow arrows denote the false targets. Best viewed in color. (a) Tracking results on AVG-Towncentre sequence. (b) Tracking results on PETS09-S2L2 sequence.

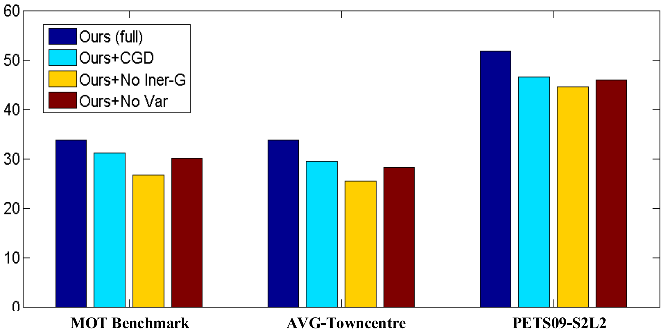


Fig. 7. Tracking results of different energy optimization solutions on MOT benchmark.

worse than the conjugate gradient descent solution. Only by combining the two schemes "Ours" is possible to reach a good optima of the proposed energy-variation-minimization.

TABLE V
DATASETS FOR TRAINING AND TEST

Task	Type	Dataset
Training	RGB RGB-D	MOT benchmark training sequence SDL-Crossing
Test	RGB RGB-D	MOT benchmark test sequence Sync, SDL-Campus, SDL, LIPD

C. Evaluation on RGB-D Datasets

The two evaluations demonstrate that the proposed model is generally applicable to RGB and RGB-D datasets in any application scenario.

Experiments are carried out on four public datasets: the Sync dataset [56], the SDL dataset [57], the SDL-Campus dataset [57], and the LIPD dataset [58]. Each video sequence

TABLE VI

COMPARISONS OF TRACKING RESULTS ON FOUR RGB-D DATASETS. IN THE "BASELINE" ITEM, G=GROUP, T=TOPOLOGY

Dataset	Baseline	Method	Recall ↑	Prec.↑	GT	MT ↑	ML ↓	IDS↓	Frag.↓
Sync [56]	*RGB*	Berclaz *et al.* [27]	69.6%	72.8%	66	9.0%	24.2%	345	323
		Milan *et al.* [46]	73.4%	78.3%	66	19.6%	21.2%	89	115
		Yang *et al.* [29]	75.6%	80.2%	66	21.2%	19.6%	117	146
	RGB+G	Chen *et al.* [5]	76.6%	80.3%	66	25.8%	15.1%	121	133
	RGBD+G	DSA [57]	85.0%	89.7%	66	28.8%	**13.7%**	90	108
	RGB+T	Ours*	80.8%	83.4%	66	27.2%	15.1%	97	124
	RGBD+T	Ours	**87.5%**	**92.2%**	66	**37.8%**	16.7%	**77**	**109**
SDL [57]	*RGB*	Berclaz *et al.* [27]	68.9%	70.5%	92	9.8%	25.0%	168	189
		Milan *et al.* [46]	70.4%	76.4%	92	21.7%	19.6%	89	69
		Yang *et al.* [29]	73.7%	75.1%	92	17.3%	25.0%	106	79
	RGB+G	Chen *et al.* [5]	74.6%	77.5%	92	18.4%	19.6%	103	98
	RGBD+G	DSA [57]	82.4%	87.3%	92	25.0%	15.2%	60	**68**
	RGB+T	Ours*	79.5%	82.1%	92	25.0%	17.3%	84	77
	RGBD+T	Ours	**87.4%**	**89.0%**	92	**30.4%**	**10.9%**	**55**	83
SDL-Campus [57]	*RGB*	Berclaz *et al.* [27]	66.4%	69.8%	74	12.2%	33.8%	130	146
		Milan *et al.* [46]	74.0%	76.5%	74	20.2%	21.6%	78	97
	RGB+G	Chen *et al.* [5]	75.1%	77.3%	74	18.9%	20.2%	80	110
	RGB+T	Ours*	81.9%	83.1%	74	21.6%	18.9%	77	80
	RGBD+T	Ours	**89.2%**	**91.2%**	74	**33.8%**	**13.5%**	**53**	**58**
LIPD [58]	*RGB*	Berclaz *et al.* [27]	72.8%	72.4%	77	10.4%	33.8%	324	219
		Yang *et al.* [29]	76.3%	81.2%	77	19.5%	22.1%	141	164
	RGB+G	Chen *et al.* [5]	77.3%	81.5%	77	13.0%	23.4%	150	172
	RGB+T	Ours*	82.3%	85.6%	77	24.7%	19.5%	102	147
	RGBD+T	Ours	**88.7%**	**90.0%**	77	**37.7%**	**13.0%**	**89**	**92**

in the datasets has corresponding depth information captured with depth sensors (LiDAR or stereo vision). The videos are recorded at 10 FPS, and have a variable number of objects (car, pedestrian, and cyclist). The crowded scenes, moving cameras, and appearance variation of moving objects make the target tracking on them quite challenging. Many conventional assumptions adopted in MOT with a surveillance camera are not applicable in these cases (*e.g.*, fixed entering/exiting location, background modeling, etc).

In RGB-D experiments, we use the codes published by the authors and the parameters (*cf.* Table II) learned in the SDL-Crossing datasets, and compare it with state-of-the-art methods in Table VI:

- *RGB* (RGB-based methods: the network flow method [27], the online learned CRF model [29], the continuous energy minimization [46]);
- *RGB+G* (Grouping method: [5]);
- *RGBD+G* (Grouping method using RGB-D data [57]);
- *RGB+T* (Our model without using RGB-D feature);
- *RGBD+T* (Our full model).

Table VI shows that our approach *"RGBD+T"* significantly outperforms recent state-of-the-art RGB-based methods in significant margins (averagely 12% improvement in both Recall and Precision), and our RGB-based baseline method *Ours** (*"RGB+T"*) outperforms the other trackers averagely 5% in Recall and 4% in Precision. Compared with the depth-based tracker [57], our model improves more than 2% and 5% in Recall and Precision in the Sync and SDL datasets. This validates that the topology-level constraint added in the conventional multiple person tracking framework is a key factor for accurate tracking. With such a topology-energy-variation minimization manner, the in-group target occlusion problem could be better addressed.

The **Sync dataset** is a video sequence with 2147 frames. In this video sequence, long-term and serious occlusion issues are frequent. Cars parking along both sides of the road coincidentally have similar colors with the pedestrians and are close to the pedestrians. Many of the false detections are added to the trajectories by the RGB-based methods, so the MT score in Table VI decreases. It can be seen that our approach outperforms other methods through the topology-energy-variation minimum optimization. In this sequence, frames ranging from 150 to 600 have multiple occlusion scenarios. It is observed that the pedestrian detector outputs numerous false detections. Many of the false detections have been added to the trajectory by the vision-based methods in the first baseline, and the MT item in Table I decreases. That is because the recurring false and missing detections in dynamic traffic backgrounds make the affinity probability of appearance and motion unreliable. Therefore the false detections cannot be easily excluded. The **SDL dataset** is recorded on a straight road and a crossroad, and has fewer occlusions. In Table VI, it is observed that Recall and Precision increase compared with the results in the Sync dataset. Our approach has the fewest Frag. and IDS errors in the SDL dataset. Compared with other datasets, the pedestrians of the **SDL-Campus dataset** is in a relatively low density, and most pedestrians do not walk in groups. In this case, the inter-topology energy makes a key role in the tracking but the intra-topology energy is negligible. Surprisingly, our approach can still have significant performance gain over the state-of-the-art approaches [5], [46]. The **LIPD dataset** is recorded from a sensor acquisition system mounted on

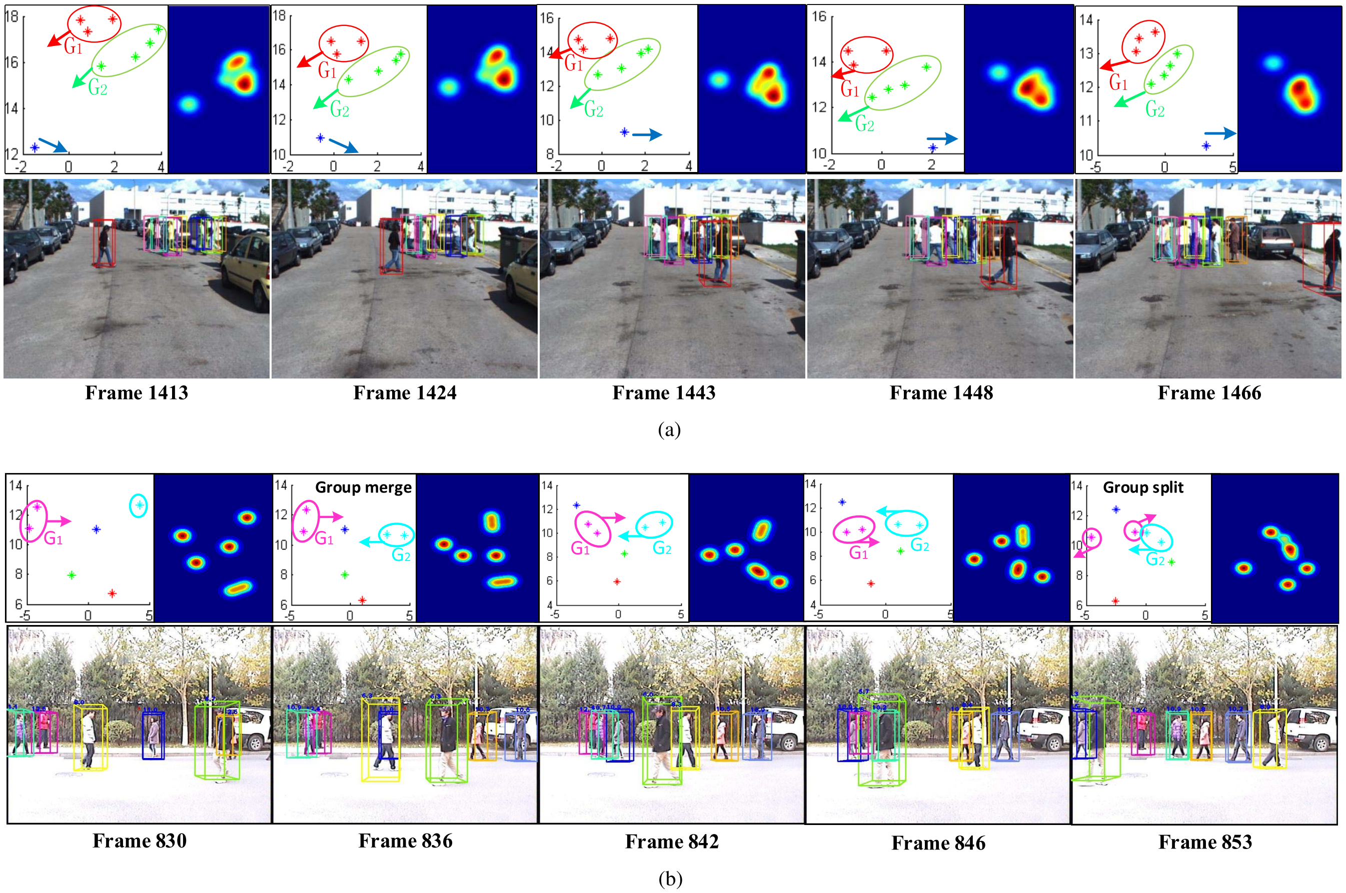


Fig. 8. Tracking results of our approach. In each frame, the tracking outputs include three parts: 3D bounding box (in second row), topology configuration (the same color indicates the same topology), and topology energy distribution. Best viewed in color. (a) Tracking examples on the Sync dataset. (b) Tracking examples on the SDL dataset.

an instrument-equipped Yamaha vehicle, driving in an urban environment. The dataset contains 4823 frames. The dataset is captured at dusk, and the challenge is mainly from lighting variation. The comparative results suggest that our approach is consistent in poor illumination. Some tracking samples in RGB-D datasets are shown in Fig. 8.

D. Group Discovery

Group discovery is provided by the group indicator matrix T in the topology model, which indicates the relationship between individuals and groups. In [44] and [39], the following group discovery evaluation method is adopted: each pedestrian is coded into one of two categories: alone or in a group. Since we do not have all the annotations on the RGB and RGB-D datasets we conduct multiple person tracking experiments on, or any available implementations of related work, we are not able to conduct comparative experiments on all the datasets. In evaluation, we annotate group identity in the PETS09 and AVG-TownCentre sequences. Match rate indicates the percentage of persons that are classified correctly.

Compared with existing methods [37], [38], [43], [44], experiments show that our group discovery component produces more reasonable results. In the test datasets, our approach produces 85% matching rate on more than 300 trajectories. Recall that the same person in different time windows are treated as different persons [44]. In [39], the pedestrians are only divided into along and pair categories. However, in experiments, we keep individual and group identity consistent in frames and achieve substantial agreement with human annotator on this dataset. It is also observed that 36% of the people moving in groups in the AVG-TownCentre dataset, The figure is 65% in the PETS09-S202 sequence.

E. Energy Variation

As described before, the complexly designed energy function in multiple person tracking is always in a non-convex form, so optimization could not guarantee a global minimum in limited time. How to reach a strong local minimum being not far away from the global minimum in limited iterations has been investigated in many previous work [7], [8], [46]. We visualize the energy variation in different optimization strategies applied in our model, shown in Fig. 9. The blue line denotes the purely conjugate gradient descent optimization without any jump move, by which the energy always goes along the way of decreasing greatly. Most trajectory fragmentation caused by false detections and the observations missing evidence can not be modified. The red line denotes the strategy with only tracklet-jump move. Similar to the work [46], parts of the false tracklets solution could be repaired by the tracklet-level jump. However, lacking a soico-topology constraint makes the jumps without the context information. Sometimes, tracklets with different identities are connected

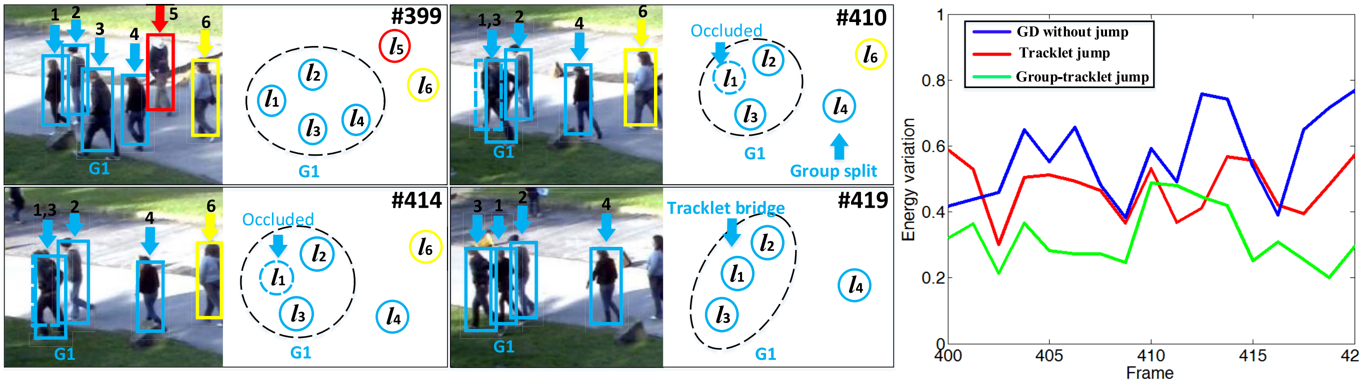


Fig. 9. A tracking example of group-tracklet jump move in PETS09-S2L2 sequence. One group splits into two parts at frame 410. Target l_1 is occluded by other in-group members at frame 414, but found by the tracklet bridging at frame 419. The right figure shows the topology energy variation in this tracking example. Best viewed in color.

to the same target by the 'flexible' jumps. The green line represents the performance of our optimization in group-tracklet jump moves. The energy variation changes slightly compared with the other strategies, which ensures the energy jump out of the weak local minimal but not so far away in a social context. The tracklet-jump could be treated as a kind of tracklet-repair manner in once group jump.

F. Convergence

In the optimization process, the energy variation is minimized until all the group and tracklet jumps run out, which commonly takes less than 20 iterations to reach the minimum under the given threshold $\varepsilon = 0.05$. Another important implementation in solve the energy model is conjugate gradient descent. We adopt Carl Rasmussen's implementation[2] with its default parameters to execute conjugate gradient descent on the energy minimization. In the setting, the Polack-Ribiere formula is used for determining the search directions, while the Wolfe-Powell conditions and the slope ratio method are used for estimating the step. It runs until convergence or reaching a maximal number of iterations without pre-defined jump moves. This severely limits the possible state space changes, and the search largely stays within one region of the solution space in the non-convex model. Finally, the energy is not able to descend much further. This optimization usually leads to a slow convergence and weak local solution.

Another issue is about the move order in the optimization process. We use the ordered moves in both group and tracklet jumps, which significantly speed up the optimization process. Actually, different move orders lead to a similar convergence and tracking solution, which is also verified in [46]. But [46] only uses the tracklet-level jump moves in the model, our two-level jump is able to provide the tracking solution with a social constraint, so that tracklets could be globally associated in a social topology.

G. Time Analysis

The computational time is greatly affected by the number of persons and the length of the video. Experiments are performed on an Intel 3.4GHz PC with 4G memory, and the codes are implemented in Matlab. Without codes optimization, our approach achieves a tracking speed of ~7 Hz when there are averagely 10 persons to be tracked. When our approach is applied on the high crowd density video in AVG-TownCentre, the speed is ~3 HZ, shown in Table VII. It is found that the speed of our full model (Ours) is lower than the conjugate gradient descent solution (Ours+CGD), because the full model gets rid of the region to reach a strong local minimum by the iterative group-tracklet jump moves, when the energy variation is slight.

TABLE VII

TRACKING SPEED OF DIFFERENT OPTIMIZATION SOLUTIONS ON AVG-TOWNCENTRE SEQUENCE

method	Ours(full)	Ours+CGD	Ours+No Inter-G	Ours+No Var
speed	3.4 Hz	4.4 Hz	4.1 Hz	3.6 Hz

VIII. CONCLUSION

We have developed a social topology-energy-variation model and integrated it with the conventional data association method for RGB and RGB-D multiple person tracking. With this model, the dynamics of a collection of moving persons are formulated both in-group and out-group structures in a global manner. To quantify the in- and out-group relations to capture the topology variation, spatial energy distributions are defined. Minimizing the topology-energy-variance in a group-tracklet jump-moves procedure is validated to result in smooth topology transitions, stable group tracking, and accurate target association. Experiments show that the topology constrained data association has the state of the art tracking performance, validating that minimizing the energy variation is the key factor for stable and accurate tracking.

[2] http://www.gatsby.ucl.ac.uk/~edward/code/minimize/minimize.m

REFERENCES

[1] W. Luo, J. Xing, X. Zhang, X. Zhao, and T. K. Kim. (2015). "Multiple object tracking: A literature review." [Online]. Avaliable: https://arxiv.org/abs/1409.7618

[2] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 5, pp. 4282–4286, 1995.

[3] H. Singh, R. Arter, L. Dodd, P. Langston, E. Lester, and J. Drury, "Modelling subgroup behaviour in crowd dynamics DEM simulation," *Appl. Math. Model.*, vol. 33, no. 12, pp. 4408–4423, 2009.

[4] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PLoS ONE*, vol. 5, no. 4, p. e10047, 2012.

[5] X. Chen, Z. Qin, L. An, and B. Bhanu, "Multi-person tracking by online learned grouping model with non-linear motion context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 12, pp. 2226–2239, Dec. 2016.

[6] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1345–1352.

[7] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

[8] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1972–1978.

[9] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multi-people tracker," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 120–127.

[10] A. Kendon, *Conducting Interaction: Patterns Behavior Focused Encounters*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[11] S.-H. Bae and K.-J. Yoon, "Robust online multiobject tracking with data association and track management," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2820–2833, Jul. 2014.

[12] L. Chen, W. Wang, G. Panin, and A. Knoll, "Hierarchical grid-based multi-people tracking-by-detection with global optimization," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4197–4212, Nov. 2015.

[13] H. Jiang, J. Wang, Y. Gong, N. Rong, Z. Chai, and N. Zheng, "Online multi-target tracking with unified handling of complex scenarios," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3464–3477, Nov. 2015.

[14] J. Niño-Castaneda *et al.*, "Scalable semi-automatic annotation for multi-camera person tracking," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2259–2274, May 2016.

[15] X. Shi, H. Ling, J. Xing, and W. Hu, "Multi-target tracking by rank-1 tensor approximation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2387–2394.

[16] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association by online target-specific metric learning and coherent dynamics estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 589–602, Mar. 2017.

[17] M. Yang and Y. Jia, "Temporal dynamic appearance modeling for online multi-person tracking," *Comput. Vis. Image Understand.*, vol. 153, pp. 16–28, Dec. 2016.

[18] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4696–4704.

[19] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4705–4713.

[20] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3029–3037.

[21] S. Gao, Z. Han, C. Li, Q. Ye, and J. Jiao, "Real-time multipedestrian tracking in traffic scenes via an RGB-D-based layered graph model," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2814–2825, Oct. 2015.

[22] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2015, pp. 1515–1522.

[23] W. Hu, W. Li, X. Zhang, and S. Maybank, "Single and multiple object tracking using a multi-feature joint sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 816–833, Apr. 2015.

[24] Z. Khan, T. Balch, and F. Dellaert, "MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1960–1972, Dec. 2006.

[25] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2009, pp. 1200–1207.

[26] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[27] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[28] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1201–1208.

[29] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2034–2041.

[30] A. Heili, A. López-Méndez, and J. M. Odobez, "Exploiting long-term connectivity and visual motion in CRF-based multi-person tracking," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3040–3056, Jul. 2014.

[31] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multiple target tracking based on undirected hierarchical relation hypergraph," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1282–1289.

[32] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 343–356.

[33] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4091–4099.

[34] B. Yang, C. Huang, and R. Nevatia, "Learning affinities and dependencies for multi-target tracking using a CRF model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1233–1240.

[35] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1265–1272.

[36] A. Milan, K. Schindler, and S. Roth, "Detection-and trajectory-level exclusion in multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3682–3689.

[37] I. Chamveha, Y. Sugano, Y. Sato, and A. Sugimoto, "Social group discovery from surveillance videos: A data-driven approach with attention-based cues," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–11.

[38] J. Šochman and D. C. Hogg, "Who knows who—Inverting the social force model for finding groups," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Nov. 2011, pp. 830–837.

[39] Z. Qin and C. R. Shelton, "Social grouping for multi-target tracking and head pose estimation in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2082–2095, Oct. 2016.

[40] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2014, pp. 3542–3549.

[41] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5397–5406.

[42] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2211–2218.

[43] A. Gning, L. Mihaylova, S. Maskell, S. K. Pang, and S. Godsill, "Group object structure and state estimation with evolving networks and monte Carlo methods," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1383–1396, Apr. 2011.

[44] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[45] L. Bazzani, M. Zanotto, M. Cristani, and V. Murino, "Joint individual-group modeling for tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 746–759, Apr. 2014.

[46] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, Jan. 2014.

[47] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1160–1168, 2006.

[48] F.-Y. Wu, "The potts model," *Rev. Mod. Phys.*, vol. 54, no. 1, pp. 235–268, 1982.

[49] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, Dec. 1995.

[50] L. E. Navarro-serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional LADAR data," *Int. J. Robot. Res.*, vol. 29, no. 12, pp. 1516–1528, 2010.

[51] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D data with online boosted target models," in *Proc. Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 3844–3849.

[52] L.-C. Chen, S. Fidler, A. L. Yuille, and R. Urtasun, "Beat the mturkers: Automatic image labeling from weak 3D supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3198–3205.

[53] Z. Han, J. Jiao, B. Zhang, Q. Ye, and J. Liu, "Visual object tracking via sample-based adaptive sparse representation (ADASR)," *Pattern Recognit.*, vol. 44, no. 9, pp. 2170–2183, 2011.

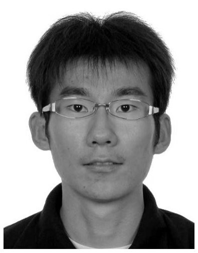
[54] *Multiple Object Tracking Benchmark*, accessed on Jul. 1, 2016. [Online]. Available: http://motchallenge.net

[55] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *Proc. 1st Int. Eval. Workshop Classification Events, Activities Relationships*, 2006, pp. 1–44.

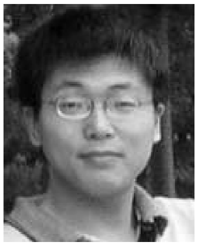
[56] L. Oliveira, U. Nunes, P. Peixoto, M. Silva, and F. Moita, "Semantic fusion of laser and vision in pedestrian detection," *Pattern Recognit.*, vol. 43, no. 10, pp. 3648–3659, 2010.

[57] S. Gao, Z. Han, D. Doermann, and J. Jiao, "Depth structure association for RGB-D multi-target tracking," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4152–4157.

[58] *Laser and Image Pedestrian Detection (LIPD) Dataset in Urban Environment*, accessed on Jul. 1, 2016. [Online]. http://www2.isr.uc.pt/~cpremebida/dataset



Shan Gao received the B.S. degree in communication engineering from Nankai University, Tianjin, China, in 2010, the M.S. and Ph.D. degrees from the University of the Chinese Academy of Sciences, Beijing, in 2013 and 2016, respectively. Since 2016, he has been a Post-Doctoral Researcher with the Automation Department, Tsinghua University, Beijing, China. His current research interests include object detection and tracking, image processing, and multi-sensor fusion.



Qixiang Ye (M'10–SM'15) received the B.S. and M.S. degrees in mechanical and electrical engineering from the Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2006. He was a Visiting Assistant Professor with the Institute of Advanced Computer Studies, University of Maryland at College Park, College Park until 2013. He has been a Professor with the University of Chinese Academy of Sciences since 2009. His research interests include image processing, visual object detection and machine learning. He pioneered the Kernel SVM-based pyrolysis output prediction software which was put into practical application by SINOPEC in 2012. He developed two kinds of piecewise linear SVM methods, which were successfully applied into visual object detection. He has authored over 50 papers in refereed conferences and journals, and received the Sony Outstanding Paper Award.

Junliang Xing (M'09) received the dual B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include computer vision problems related to faces and humans.

Arjan Kuijper received the M.Sc. degree in applied mathematics from Twente University, The Netherlands, and the Ph.D. degree from Utrecht University, The Netherlands, and the Habitation degree from TU Graz, Austria. He holds the Chair in mathematical and applied visual computing, Technology University of Darmstadt. He is a member of the Management of Fraunhofer Institute for Computer Graphics Research, responsible for scientific dissemination. . He was an Assistant Research Professor with the IT University of Copenhagen, Denmark, and a Senior Researcher with RICAM in Linz, Austria. He received He has author of over 250 peer-reviewed publications, and serves as a Reviewer for many journals and conferences, and as a Program Committee Member, and a Organizer of conferences. His research interests cover all aspects of mathematics-based methods for computer vision, graphics, imaging, pattern recognition, interaction, and visualization.

Zhenjun Han received the B.S. degree in software engineering from Tianjin University, Tianjin, in 2006, and the M.S. and Ph. D degrees from the University of the Chinese Academy of Sciences, Beijing, China, in 2009 and 2012, respectively. Since 2013, he has been an Associate Professor with the University of Chinese Academy of Sciences, Beijing. His current research interests include image processing and intelligent surveillance etc.

Jianbin Jiao (M'10) received the B.S., M.S., and Ph.D. degrees in mechanical and electronic engineering from Harbin Institute of Technology (HIT), Harbin, China, in 1989, 1992, and 1995, respectively. From 1997 to 2005, he was an Associate Professor with HIT. Since 2006, he has been a Professor with the School of Electronic, Electrical, and Communication Engineering, University of the Chinese Academy of Sciences, Beijing, China. His current research interests include image processing, pattern recognition, and intelligent surveillance.

Xiangyang Ji received the B.S. degree in materials science and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He joined Tsinghua University, Beijing, China, in 2008, where he is currently a Professor with the Department of Automation, School of Information Science and Technology. He has authored over 100 referred conference and journal papers. His current research interests include signal processing, image/video compression and communication, intelligent imaging.