### Title

Understanding Dynamic Social Grouping Behaviors of Pedestrians




### Journal

IEEE Journal of Selected Topics in Signal Processing, 9(2)

### ISSN

1932-4553



### Authors

Feng, Linan

Bhanu, Bir



### Publication Date

2015-03-01

### DOI

10.1109/jstsp.2014.2365765

Peer reviewed

# Understanding Dynamic Social Grouping Behaviors of Pedestrians

Linan Feng, *Student Member, IEEE*, and Bir Bhanu, *Fellow, IEEE*

***Abstract*—There have been many studies in the literature on social group recognition of crowds of pedestrians. However, most of these studies have approached the problem from a static point of view. A study on the dynamic property of social groups among people over time can provide significant insight into human behaviors and events. Inspired by sociological models of human collective behavior, in this work, we present a framework for characterizing hierarchical social groups based on evolving tracklet interaction network (ETIN) where the tracklets of pedestrians are represented as nodes and the their grouping behaviors are captured by the edges with associated weights. We use non-overlapping snapshots of the interaction network and develop the framework for a unified dynamic group identification and tracklet association. The approach is evaluated quantitatively and qualitatively on videos of pedestrian scenes where manually labeled ground-truth is given. The results of our approach are consistent to human-perceived dynamic social groups of the crowd. The performance analysis of our method shows that the approach is scalable and it provides situational awareness in a real-world scenarios.**

***Index Terms*—Dynamic social grouping behavior, pedestrian social groups, tracklet interaction network.**

## I. Introduction

CONSIDER a video clip recording a number of pedestrians walking in an outdoor (indoor) environment such as a square (hall). Imagine an algorithm that is able to analyze the video and answer the questions like: Are these people evacuating from an emergent situation? Are they gathering for a special event? By just looking at each individual it could be very hard to train the computers to understand these high-level concepts from the low-level visual representations. In this paper we introduce a new model for analyzing social behaviors among pedestrians: rather than treating each person in isolation, we analyze their social grouping behaviors so as to reinforce the recognition of movements of each individual in a group. Our approach is inspired by recent achievements in computer vision and pattern recognition where the correlations of semantic or geometrical concepts are utilized as extra contextual information for recognizing objects in complex scenes [1]. In our work, pedestrian detection and interactions are enforced by taking the advantage of contextual information that comes from within-group positional, velocity and directional distance consistences. This provides our approach the robustness to pedestrian walking behavior analysis from dynamic cluttered background, occlusions among pedestrians, illumination and viewpoint changes, or the variations of backgrounds caused by mobile cameras such as smart-phones.

It is important to understand the collective social behaviors at a group level in many real-world scenarios. For example, people tend to participate or leave an event with herding behavior [2]. When crowd of people evacuate from an emergent situation, they leave with the members in their original group [3], the direction of the group is usually determined by the fastest member and the speed of the group is limited by the slowest member [4]. Computer vision techniques, such as multi-people tracking in crowded scenes [5], [6], crowd segmentation [7] have made tremendous progress in recent years and they provide the opportunities to solve real-world challenging problems such as recognition of human behaviors at the activity and event level that far exceeds the conventional capabilities of a surveillance system.

In this paper, we attempt to achieve a higher level understanding of crowd behaviors in terms of social groups and interaction patterns that are displayed while they are traveling together. A social group of pedestrians consists of people with shared walking patterns such as change of directions, change of speeds, avoiding obstacles, etc. [8]. In particular, we explicitly explore the dynamic properties of social groups that capture the spatio-temporal changes such as splitting and merging of people. Determining the dynamic group structure of a crowd provides the basis for further high-level analysis of events involving social interactions within and across groups.

We propose to detect the social groups of pedestrians based upon the state-of-the-art pedestrian detector and reliable tracklet generation techniques. Our main contribution, is that we explore the evolving social group property among tracklets in a network structure, which we call "*evolving tracklet interaction network*" (ETIN). Based on the social psychological models of collective behavior, the reliable tracklets generated from detection responses are represented as nodes in ETIN with incident edges indicating the social interactions and grouping behaviors (see Fig. 1). The significance of social grouping behavior between nodes is defined by the edge weights. Tracklets from pedestrians in a potential group will have denser spatio-temporal co-occurrences reflected by larger edge weights in ETIN compared to the tracklets from the pedestrians outside the group. We also propose to address the dynamic changes of social groups in ETIN explicitly which is similar to detecting evolving communities that exist in many common social networks such as Facebook and Twitter.

We validate our framework extensively on multiple video datasets that are collected from indoor/outdoor public scenes

Manuscript received March 17, 2014; revised August 13, 2014; accepted October 20, 2014. Date of publication October 28, 2014; date of current version February 11, 2015. This work was supported in part by the National Science Foundation under Grants 0905671 and 1330110 and the Office of Naval Research under Grant N00014-12-1-1026. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Mohan Kankanhalli.

The authors are with the Center for Research in Intelligent Systems, University of California, Riverside, CA 92521 USA (e-mail: fengl@cs.ucr.edu; bhanu@cris.ucr.edu).

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/JSTSP.2014.2365765

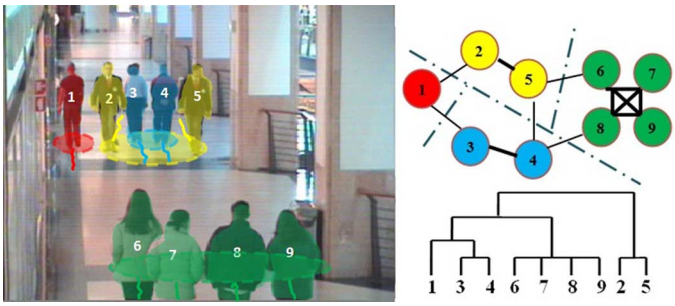


Fig. 1. Left: A real-world video frame (from CAVIAR dataset) shows that people are walking in groups. Individuals and related trajectories are labeled with numbers and the potential social groups among them are marked in different colors. Right: A snapshot restored from evolving tracklet interaction network (ETIN) representation at a given time interval (top) and a hierarchical social group structure discovered by the proposed approach (bottom).

with elevated viewpoints which is the typical setting of surveillance cameras. We compare the results from our group understanding algorithms with manually labeled ground-truth group IDs in a quantitative way. Our work builds upon the recently proposed techniques in the literature on tracking by detection responses and tracklet association [9]–[14]. **Our contributions** are four-fold:

1. We propose a novel evolving tracklet interaction network (ETIN) to depict social grouping behaviors of pedestrians from reliably built tracklets of individuals which embody meaningful spatio-temporal interactions of individuals.
2. We explicitly explore the dynamic property of social groups by providing adaptation schemes for nodes and edges in ETIN representation. Our approach has not only the power of updating the network of tracklets in a very efficient manner, but also has the ability to trace the evolution of the network over time.
3. We introduce a novel modularity optimization based group detection algorithm that detects the *hierarchical* social group structure with a distance metric reflecting the spatio-temporal interactions among the pedestrians. We also provide a unified framework that addresses social group detection refinement and pedestrian tracklet association in an iterative manner.
4. Experimental results and comparison with current techniques using several datasets show that our approach is robust in medium crowd-density scenarios. We find agreement between the predicted social groups and the human-understanding of the group structures.

The rest of this paper is organized as follows. We discuss background and related work in Section II. In Section III we first outline our social grouping behavior understanding framework and then describe in detail each of its major parts. In Section IV, we report test results from our system on the real-life pedestrian videos. Section V concludes the paper.

## II. Background and Related Work

This section explains why analyzing social grouping behaviors of pedestrians is important for understanding high-level activities and events and reviews the related work in both computer vision research for crowd scene analysis and multi-people tracking, and graph partition/clustering techniques for social network analysis.

### A. Recognizing Social Groups in Computer Vision

With the increasing need for surveillance systems monitoring and detecting activities of interests in mass events with their continuing growth in size and frequency, the study of social grouping behavior of pedestrians by using computer vision techniques has become a popular research area [15]–[20]. When people walk, they naturally form groups with smaller distances to the members in the same group and larger distances to the people outside the group [2]. An interesting discovery by MacPhail shows that 89% of people attend events in groups and 94% of them leave with the people they come with [21].

Members in the same group often share same walking behaviorals, known as the *collective behavior* of pedestrians [22], such as change of direction and speed, way of avoiding obstacle, etc., that describes the distinctive and dramatic features of group trajectories and of individual trajectories within groups. In turn, groups can be determined based on the individual spatial location, cardinality and velocity [68]. Recent research efforts [23]–[25] have suggested that social groups that exhibit collective behaviors can be used to improve the understanding of social events in video sequences involving interactions among groups, especially in the cases where the cameras have elevated viewports and monitor crowded environments in which pedestrians are still discernible while partial body occlusions happen frequently. In the context of role understanding of social groups in video sequences, many approaches [26]–[31] have been proposed that combine sociological analysis and computer vision techniques to detect and recognize the behaviors of social groups by using key frames extracted from a video.

There is recent evidence that more efficient algorithms can be developed based on the recognition of high-level social groups detected in a hierarchical structure [32], [33]. The social grouping behavior of people shopping together is captured and evaluated by analyzing the inter-body distances [34]. The velocity similarity has been applied in [35], [36] to group people together for motion prediction and tracking. Ge *et al.* [37] identify small groups of pedestrians based on pre-detected trajectories, however, unlike our approach, they model the social grouping behavior in a pairwise manner, and they overlook the dynamic structural changes of the social groups (merge, split, appear, disappear, etc. [70]).

### B. Multi-Pedestrian Detection and Tracking

We propose to understand the social grouping behavior based on current computer vision techniques for pedestrian detection, multi-people tracking and data association to concatenate short tracks into longer reliable trajectories passing through the scene. State-of-the-art multi-people tracking approaches can be categorized into two classes based on the time sensitivity: real-time tracking and time-delayed tracking. In real-time tracking, the detection responses and the correspondences among them are usually jointly estimated and updated for each frame by using the information acquired from previous frame. Techniques such as particle filter are often adopted [38], [66] to estimate the intermediate states. Many approaches in this category focus on tracking each target separately [39] and they tend to fail when encountered with challenging situations involving from inter-people and scene occlusions, illumination or appearance variations and abrupt motion changes. However, there are also approaches such as [66] that jointly track individuals and groups

and demonstrate that individual tracking can be improved by group tracking and *vice versa*.

For the approaches in time-delayed tracking category, multiple targets are tracked simultaneously [40], [41]. The detection responses produced by pedestrian detectors are formed into tracklets and the final tracks are obtained by associating the tracklets at different granularities [42]. The association of tracklets is addressed by global optimization solutions such as K-shortest path [43], Hungarian algorithm [42], CRF [44] and cost-flow network [45]. The occlusions are modeled as merging and splitting of tracklets and solved by using Markov Chain Monte Carlo (MCMC) [14]. Most of these approaches generally do not use high-level semantics such as social groups to improve data-association for tracking.

Discovering the interactions among pedestrians to improve tracking in crowded scenes has become a new trend of research in the literature. Solmaz *et al.* [67] introduced an approach that identifies individual/group behaviors without any object detection, tracking or training steps. Pelligrini *et al.* [19] proposed a dynamic model for tracking people in complex scenes that exploit the social interactions such as attraction and repulsion. According to recent research by Moussa *et al.* [2], 70% of people in a crowd walk in groups. The grouping property of pedestrians is explicitly analyzed in the computer vision field in [4]. Specifically, groups are used as contextual knowledge for trajectory prediction and refinement [12], [36].

### C. Finding Social Groups in Networks

Network structure has drawn great attention in analyzing social relationships between people. Network structures are proposed in [26], [30] as interaction graph where individuals are indicated by the nodes and the edges between them are weighted by their relatedness in either social or visual sense. A different type of network is presented in [69] with edges express the probability of individuals belonging to a group.

A very common property in many realistic complex networks such as social networks and biological networks is known as the *community structure* [46], [47], i.e., the nodes in the network naturally divide into groups with denser connections inside each group and looser connections among groups. In our tracklet interaction network, the nodes and edges represent pedestrian tracklets and their social grouping behavior, respectively, and the social groups can be viewed as the communities in the network.

Traditional algorithms for detecting groups of nodes in a network can be categorized into partition based methods [48], hierarchical clustering algorithms which can be further classified into agglomerative (e.g., [49]) and divisive (e.g., [50]) algorithms, spectral algorithms [51], modularity-based methods [52], and dynamic algorithms [53]. In most of the work [48], [49], [52] the edges are unweighted in the problem domains, thus, additional computing is required, e.g., in [54] the edge weight is defined by the number of non-independent paths between nodes which can be computed using polynomial-time "max-flow" algorithms, and in [46] it is defined by Freeman's edge betweenness centrality. However, in our network the edge weights are computed directly from the distance metric defined on the spatio-temporal relations between tracklets. As a single node (tracklet of pedestrian) can be present in multiple groups simultaneously (the uncertainty of social groups, for example, a tracklet has equal distances to the other two tracklets), this results in the overlapping of groups, or the sharing of nodes between groups. There are techniques devoted to solve this problem in recent network analysis research [55], [56].

Nodes and edge weights can change over time when the video sequence proceeds. The emergence of new groups as well as the growth, split, merge, and death of old groups can occur over time. As compared to the other algorithms, modularity based approaches have been demonstrated to be the most effective in finding good partitions in an efficient manner in large networks, and they can address weighting, overlapping and evolving problems in a network [57], therefore, we adopt modularity-based approach in our work to find the social groups from tracklet interaction networks.

Our work in this paper provides a novel way for social grouping behavior understanding by representing tracklets of pedestrians and their correlations in a network structure which is original in the field. We also provide a framework that iteratively refines the pedestrian tracklet association and social group detection. Our method differs in three ways from the related work of social group recognition: (1) We detect groups in different sizes. In addition, our detected groups are generated in a hierarchical form where groups are captured at different granularities. (2) Our model explicitly handles dynamic changes of social groups, i.e., merging and splitting, in an effective and efficient manner. (3) Our model is built upon tracking by detection techniques where reliable short-term trajectories, or tracklets are available. The social grouping behaviors are, therefore, captured for a period of time in a consistent manner.

## III. Technical Approach

As illustrated in Fig. 2, the main focus of this work is to understand the dynamic social grouping behavior of pedestrians by using surveillance videos and developing techniques which provide an automated way to quantitatively analyze videos instead of spending hundreds of person hours to watch and manually labeling them. We name our approach the Evolving Tracklet Interaction Network (ETIN) based dynamic social grouping behavior analysis.

The walking behaviors of pedestrians are represented by their trajectories in the frames. However, it is often a non-trivial task to acquire reasonable trajectories in an automated way for pedestrians in a crowded or semi-crowded environment, because of the occlusions among pedestrians. In this regard, it becomes necessary to track people in a given video for a few seconds without occlusions and yield short-term trajectories, called *tracklets*, and hypothesize pedestrian groups based on these reliable tracklets. The next step is to merge and link these tracklets into long-term trajectories using the detected social groups as contextual information. The hypothesis is that pedestrians in the same group should have very similar trajectories. If some of the trajectories are broken because of occlusion, the rest of the trajectories that are complete in the same group can place useful constraints on associating the fragments. This step plays a critical role in accurately detecting long-term groups and their dynamic changes in the future. We, therefore, provide an unified framework that iteratively discovers social groups from reliable tracklets and identify stable and coherent trajectories of pedestrians that benefits from the group contexts.

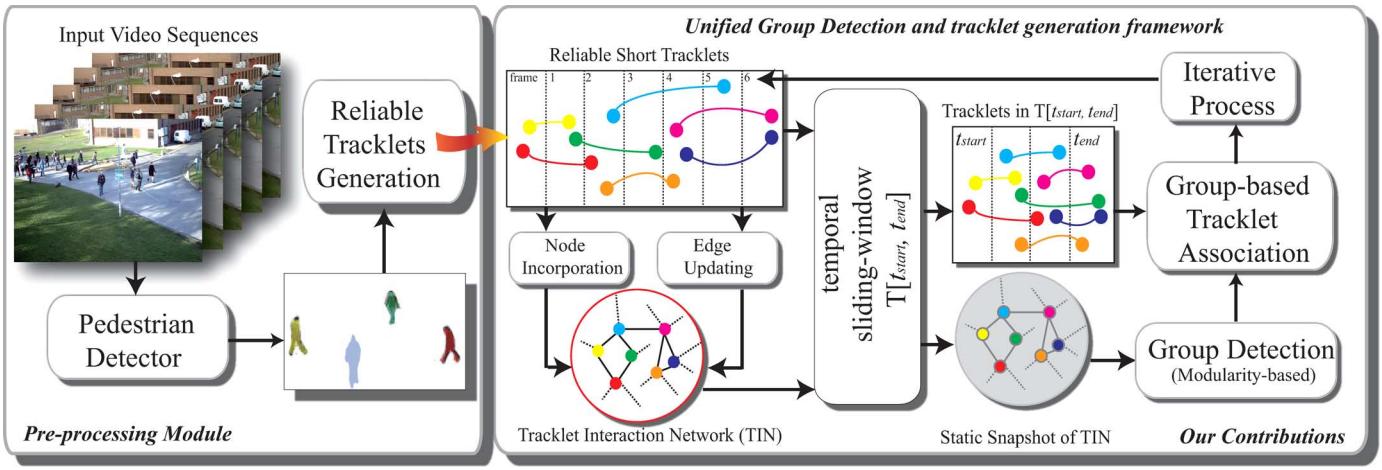


Fig. 2. The block diagram of our evolving tracklet interaction network (ETIN) framework for understanding social grouping behavior of pedestrians.

We represent the interactions among tracklets by using the proposed *evolving tracklet interaction network* (ETIN) and detect social groups using the modularity-based algorithms. Each tracklet is initialized as a node with corresponding information such as the starting and ending frames of the tracklet that is incorporated into ETIN. The relationship between existing nodes and the new node is measured by the edge weights based on the spatio-temporal interactions of the tracklets. Existing edges also need to be updated each time a new node is incorporated because of the transitive property of social grouping behavior, i.e., the social interactions between two existing nodes should be strengthened when a new node is appended with strong connections to both nodes. In order to reduce the time complexity of updating edge weights, we propose an efficient algorithm that takes advantage of the prior social group information and update the edges in an accelerated way.

To study the dynamic property of social groups such as formation, termination, splitting and merging, it is essential to characterize the transitions that go through a network at different time instants along the video. For this purpose, we utilize temporal snapshots to review static versions of the evolving network at different time intervals by applying time sliding windows in the network. In each snapshot, the nodes are kept that have some temporal overlaps with the time sliding window with corresponding edges. The social groups are then detected from the static ETIN for this specific time interval. This is formulated as a community detection problem and solved by modularity optimization that maximizes the within-group connections and minimizes the between-group connections. In the following we describe major components of the system shown in Fig. 2.

### A. Preprocessing Module

We detect pedestrians in each frame using pre-trained deformable part-based detector [58]. In order to lower the percentage of false positives, we explicitly tune the detector to exclude partially occluded people. We also remove detection responses that are of inappropriate sizes as judged by camera calibration. The detections are chained together in a dual-threshold/conflicting pairs data association step to generate short-term tracklets [10]. The output is a set of tracklets that eliminate identity switches.

### B. Evolving Tracklet Interaction Network

For each tracklet $x$ from the output of above procedure, we record the attributes in the format of $x(ID,\ tuple\ set\ \{c_{t_i}, v_{t_i}\}, t_i \in [t_{start}, t_{end}])$, where $ID$ is an unique number used as the index of the tracklet, $t_{start}, t_{end}$ are the corresponding starting and ending frames, tuple $\{c_{t_i}, v_{t_i}\}$ records the centroid of detection $c$ projected onto the ground plane and the estimated velocity vector $v$ at a given time instant (frame) $t_i$. We initialize nodes and incorporate them into TIN for the tracklets in the order of their $t_{start}$ attribute. Each node is also assigned with the corresponding tracklet's attributes. The interactions between individual nodes are modeled as pairwise spatio-temporal co-occurrences and we represent them as edges in the network. Edge weight indicates the significance of a specific interaction. For a given pair of tracklets, we categorize their interaction into two types based on whether they have a temporal overlap: 1) interaction of tracklets with overlap and 2) interaction of tracklets without overlap.

For the *first* type of interaction, we define the temporal overlap as $\Gamma = [t_0, t_1]$ of length $(t_1 - t_0 + 1)$ frames. The interaction between two tracklets is measured by the weighted sum of aggregated positional, velocity and directional distances. Given two tracklets $x_i$ and $x_j$, the distances are defined as:

$$\begin{cases} D^p(x_i, x_j) &= 1 - \exp\left(-\dfrac{\sum_{t=t_0}^{t_1} \|c_i^t - c_j^t\|}{|\Gamma|\rho^p}\right) \\ D^v(x_i, x_j) &= 1 - \exp\left(-\dfrac{\sum_{t=t_0}^{t_1} \|v_i^t - v_j^t\|}{|\Gamma|\rho^v}\right) \\ D^d(x_i, x_j) &= 1 - \exp\left(-\dfrac{c_i^{t_1} - c_i^{t_0}}{\|c_i^{t_1} - c_i^{t_0}\|} \cdot \dfrac{c_j^{t_1} - c_j^{t_0}}{\|c_j^{t_1} - c_j^{t_0}\|}\right) \end{cases} \quad (1)$$

where $\rho^p$ and $\rho^v$ are scaling factors for tuning the aggregated distance. The double vertical bar ($\|\cdot\|$) represents the $L2$ norm of a vector. All the three distance measures are scaled into the range $[0, 1]$ by exponential normalization. Aggregating the distances over time increases the robustness for capturing dynamic social grouping behaviors. Tracklets that are closer to each other and have similar velocities and directions for a longer time will

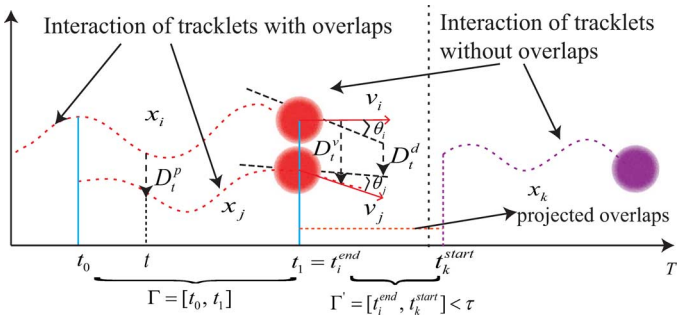


Fig. 3. Two types of tracklet interactions are shown in the left and right side. The two tracklets with overlapped interaction are marked in red and the other tracklet without overlap is marked in purple. The importance of the interaction is either calculated based on their positional, velocity and directional distances based on the temporal overlapping interval or the distances based on the projected overlapping interval.

yield smaller distances. The final pairwise interaction is defined as:

$$e_{ij}^{\Gamma} = \exp(-(\omega_1 \cdot D^p + \omega_2 \cdot D^v + (1-\omega_1-\omega_2)\cdot D^d)) \quad (2)$$

where $\omega_1$ and $\omega_2$ are the weights to adjust the importance of each factor. We use equal weights in our setting to combine the three distance measure into a final tracklet interaction importance measure that is computed over the temporal interval of overlap.

For non-overlapping tracklets $x_i$ and $x_j$, suppose $t_i^{end} < t_j^{start}$ and the time interval $\Gamma = [t_i^{end}, t_j^{start}] < \tau$ where $\tau$ is a threshold, we determine the potential spatio-temporal interaction between them in the projected overlap interval $[t_i^{end}, t_j^{start}]$ based on the motion model. Let $t \in [t_i^{end}, t_j^{start}]$, we estimate the centroids of both tracklets at frame $t$ by (3).

$$\begin{cases} c_i^t &= c_i^{t_i^{end}} + v_i^{t_i^{end}} \cdot (t - t_i^{end}) \\ c_j^t &= c_j^{t_j^{start}} + v_j^{t_j^{start}} \cdot (t_j^{start} - t) \end{cases} \quad (3)$$

The velocities are assumed to be constant in the interval and represented by $v_i^{t_i^{end}}$ and $v_j^{t_j^{start}}$. We compute the interaction importance for the *second* type of interaction by replacing the parameters in (1) with $c_i^t, c_j^t, v_i^{t_i^{end}}, v_j^{t_j^{start}}$ and $\Gamma = [t_i^{end}, t_j^{start}]$, and repeat (2). Finally, the computed values from (2) are used as the edge weights between pairs of nodes representing the tracklets in ETIN. The two types of tracklet interaction and the distances are illustrated in Fig. 3. The interaction importance is used as the edge weights when connecting two nodes representing the tracklets in the ETIN.

For each new node, respective edges are added based on the conditions $t_{new}^{start} < t_{existing}^{end} + \tau$ for a non-negative threshold $\tau$. However, the edge weights between existing nodes also need to be updated because of the social group transitivity. For example, two existing nodes $x_i, x_j$ initially have a small interaction degree. When a new node $x_k$ is added, both $e_{ik}$ and $e_{jk}$ are large which implies a high probability that $x_i$ and $x_k$ are in a social group, so are $x_j$ and $x_k$. And if $1/e_{ij} \geq 1/e_{ik} + 1/e_{jk}$, in this case, $x_i, x_j, x_k$ should be in a same group and $e_{ij}$ also needs to be modified accordingly.

Consider $N$ existing nodes $\{x_1, x_2, \ldots, x_n\}$ and a new node $x_k$, we can calculate $e_{ik}, i \in \{1, \ldots, n\}$ for any pair of nodes $(x_i, x_n)$, and compare if $1/e_{ij} \geq 1/e_{ik} + 1/e_{jk}$, $i, j \in \{1, \ldots, n\}$ & $i \neq j$. However, if the number of nodes in the network is large, the computation will take a lot of time. In order to reduce the computational cost, we propose a group detection based node incorporation and edge updating scheme as illustrated in Fig. 4.

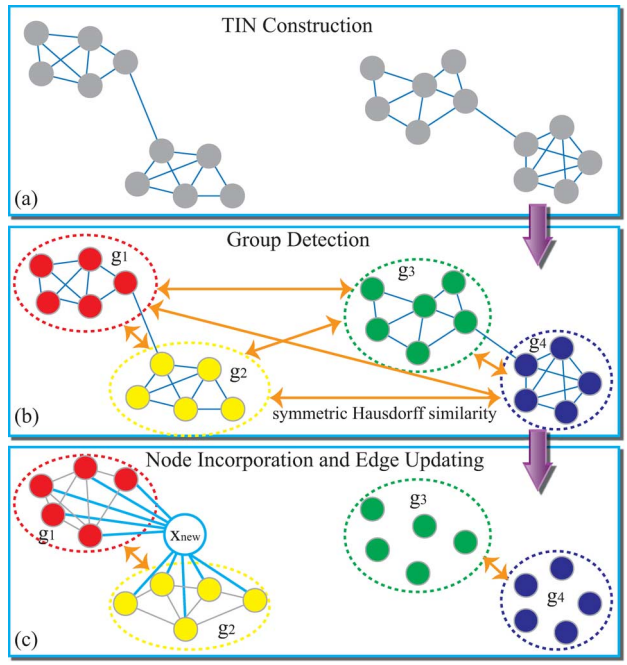


Fig. 4. The new node incorporation and edge updating scheme for the evolving ETIN. (a) The original ETIN. (b) Detection the social groups among nodes based on the modularity optimization. The symmetric Hausdorff similarity is calculated for each pair of groups. (c) When a new node $x_{new}$ is added, the interactions to other nodes are computed only for the nodes in the groups that have distances to $x_{new}$ below a certain threshold.

First, we denote the constructed ETIN at current frame $t$ as $G_t$. We detect the groups of nodes using the approach proposed in Section III-C and the groups are represented as $\{g_1, g_2, \ldots, g_m\}$. Further, we compute the intergroup closeness between any pair of groups by the symmetric Hausdorff similarity measure $H(g_i, g_j) = (h(g_i, g_j) + h(g_j, g_i))/2$ where $h(\cdot, \cdot)$ is defined by (4).

$$h(g_i, g_j) = \frac{\sum_{i=1}^{|g_i|} \cdot \sum_{k=1, g_j}^{\lceil |g_j|/2 \rceil} sort(e_{ij})_k}{|g_i| \times \left\lceil \frac{|g_j|}{2} \right\rceil} \quad (4)$$

where the sort function arrange $e_{ij}$ in descending order and we use the top-$k$, $k$ equals half size of the second group. Hausdorff metric is popular in computing the similarity among nodes in two finite sets. $g_i$ and $g_j$ are considered to be close to each other if every member in $g_i$ has large interaction importance to at least half of the members in $g_j$. The idea is similar to the concept of group expansion introduced in [59].

Using Hausdorff criterion, we set up an appropriate threshold $\epsilon$, and two groups $g_i, g_j$ are considered as neighboring groups if $H(g_i, g_j) > \epsilon$. When a new node comes in, we compare its averaged interaction degree $e_g^{avg}$ to a group $g$ with another properly chosen threshold $\epsilon'$ to see if $e_g^{avg} = (1/|g|) \sum_{i=1}^{|g|} e_{i \cdot new} > \epsilon'$ which indicates the new node is interacting with the group $g$. We chose $\epsilon' \gg \epsilon$ so that any non-neighboring groups of $g$ according to $\epsilon$ are not interacting with the new node. In this way,

**Algorithm 1:** New node incorporation and edge updating

**Input:** Current ETIN, $\{x_1, \ldots x_n\}$, $\{e_{ij}\}, i, j \in \{1, \ldots, n\}, x_{new}$

**Output:** Evolved ETIN with $x_{new}$ incorporated and edges updated

1 **Step one:** Detect social groups $G = \{g_1, \ldots, g_m\}$ in ETIN by the approach proposed in **Section III-C**; /* Group distance calculation. */

2 **foreach** *each* $g_i$ *in* $G$ **do**

3 **foreach** *each* $g_j$ *in* $G$ *and* $j \neq i$ **do**

4 Calculate the inter-group closeness by $H(g_i, g_j)$;

5 **if** $H(g_i, g_j) > \epsilon$ & $g_j \notin g_i^{neighbor}$ **then**

6 Add $g_j$ in $g_i^{neighbor}$;

7 **else if** $g_j \notin g_i^{non-neighbor}$ **then**

8 Add $g_j$ in $g_i^{non-neighbor}$;

9 Copy $G$ to $G'$ as $x_{new}$'s candidate group set;

10 **do**

11 **foreach** $g_i$ *in* $G'$ **do**

12 **if** $e_{g_i}^{avg} = (1/|g_i|)\sum_{j=1}^{|g|} e_{j\cdot new} > \epsilon'$ **then**

13 Updating $e_{j\cdot new}$ where $x_j \in g_i$, delete $g_i$ from $G'$;

14 Updating $e_{k\cdot new}$ where $x_k \in g_i^{neighbor}$;

15 Delete $g_i^{neighbor}, g_i^{non-neighbor}$ from $G'$;

16 **if** $1/e_{jk} \geq 1/e_{j\cdot new} + 1/e_{k\cdot new}$ *where* $x_j, x_k \in g_i \| g_i^{neighbor}$ **then**

17 Update $e_{jk}$ by $\max(e_{j\cdot new}, e_{k\cdot new})$;

18 **while** $G' \neq \emptyset$;

19 **return** Updated ETIN;

we only need to compute the interactions to $g$ and its neighboring groups and update the edge weights in these groups. For the example presented in Fig. 4(c), $e_{g_1}^{avg} > \epsilon'$, we calculate the interactions for $x_{new}$ and nodes in $g_1$ as well as the nodes in the neighboring group $g_2$ and updating the corresponding edges and avoid the calculation for non-neighboring groups $g_3$ and $g_4$. The entire process is summarized in Algorithm 1.

### C. Social Group Detection

We make use of the temporal snapshots to examine static versions of ETIN at different time intervals. We detect the social groups from a restored static ETIN in a given temporal window using *modularity* measure [60].

*Definition:* Let $G = (V, E)$ denote a varying tracklet interaction network where $V$ represents unique tracklets and $E$ the interactions that exist among the tracklets. We define a temporal snapshot $S_i(V_i, E_i)$ of $G$ to be a network representing only tracklets and interactions active in a particular time interval $[t_i^{start}, t_i^{end}]$, called the snapshot interval.

A social group, in our case, is defined as a group of nodes in a specific snapshot that has large internal interaction importance. On the other side, nodes in the group will have weak interactions to the outside nodes. A common way towards detecting communities of people based on the links in a social network is to recursively divide the entities in the complete network into subgroups. We naturally transform the group analysis into finding a method from the social network perspective. In order to quantify the goodness of a network partition, *modularity* has been widely accepted as a measurement of the partition which has been found to be robust and effective in many real world networks [60].

Basically, modularity is the fraction of connections within groups subtracting the expected links of the same quantity of node degrees while the connections are distributed in a random way. Usually, larger value of modularity indicates more significant social grouping phenomenon of nodes. Therefore, our goal is to divide the nodes into groups such that the modularity of the entire network is maximized.

*1) Problem Definition:* Given the evolving $G = (G_0, G_1, \ldots, G_n)$ where $G_0$ is the snapshot at the first snapshot interval, and the rest $Gs$ are the snapshots obtained by $(G_0 + i * \Delta G)$. The problem is to find an adaptive algorithm that efficiently identify the groups at any snapshot interval utilizing the information from the previous interval.

The modularity $Q_{ij}$ of two nodes $x_i, x_j$ measures the difference between their connection strength and expectation of random pair of nodes in the current snapshot of ETIN. Suppose the neighboring node set of node $x_i$ is $N_i$ where each node is connected by an edge to $x_i$, the modularity $Q_{ij}$ is defined as,

$$Q_{ij} = e_{ij} - \frac{\sum\limits_{k\in N_i} e_{ik} \cdot \sum\limits_{k\in N_j} e_{jk}}{\sum\limits_{x_m, x_n \in TIN} e_{mn}} \tag{5}$$

Initially, we assign all the nodes in one group, the modularity $Q$ of the entire network is the summation of the $Q_{ij}$s of any pair of nodes. However, if we divide the nodes into two groups, we use a label vector $s \in \mathbb{R}^n$ to denote the group of each node. If an element $s_i = +1$, the corresponding node is assigned to the first group, and $s_i = -1$ otherwise, and the modularity of the network changes to:

$$Q' = s^T \cdot Q \cdot s = s^T \cdot \sum_{i,j} \left[ e_{ij} - \frac{\sum\limits_{k\in N_i} e_{ik} \cdot \sum\limits_{k\in N_j} e_{jk}}{\sum\limits_{m,n} e_{mn}} \right] \cdot s \tag{6}$$

The element values in the vector $s$ are determined by first representing $Q$ in the matrix format, eigen-decomposing it into eigenvalues and eigenvectors, and then $s_i$ is set to $+1$ if the corresponding eigenvalue is positive and $-1$ otherwise. The strategy for two-subgroup division can be applied to divide the entire network into multiple groups recursively if we change the label vector $s$ into a matrix $S \in \mathbb{R}^{n\times l}$ where $l$ is the number of groups, it starts from 1 and keeps increasing. We record the modularity before and after a new division as $Q_{last}$ and $Q_{new}$, then the modularity gain is measured by $\Delta Q = Q_{new} - Q_{last}$. We stop the recursive division until there is no positive modularity gain, i.e., $\Delta Q \leq 0$. After the top-down division, we assign an unique ID to each of the detected groups based on the path from root to leaf in the hierarchical structure.

Now we address the problem of tracing the dynamic social group changes from one snapshot to the next based on the modularity maximization criteria. As time goes by, new node could be incorporated into the network and old node could also be deleted from the network. Intuitively, adding a new node that

results in the insertion of one or more intra-group edges, or deleting an old node that leads to the removal of one or more inter-group edges in the current snapshot will not weaken the group structure obtained from the previous snapshot. Similarly, removing intra-group edges or inserting inter-group edges will not strengthen the group from the previous snapshot. However, when two groups have less distractions, adding or deleting an edge between them may change the structures of them, leading them either to merge or split further. In this case, we need to determine to which group the new node should join to maximize the modularity gain.

Inspired by an adaptive network analysis approach introduced in [61], we determine that a new node $u$ stays in the original group $C$ or moves to a new group $C'$ by two kinds of forces: $F^{C}_{stay} = e^{C}_{u} - e_u(e_C - e_u)/(2 * \sum_{x_m,x_n \in TIN} e_{mn})$ is the force to keep $u$ stay in $C$ and $F^{C'}_{leave} = e^{C'}_{u} - (e_u * e_{C'})/(2 * \sum_{x_m,x_n \in TIN} e_{mn})$ is the force that $C'$ attract $u$ into it. Based on these two forces, the node $u$ can determine to stay in an old group if $F^{C}_{stay}$ is greater than any of the $F^{C'}_{leave}$, and vice versa. The proof of Theorem 1 in Appendix A demonstrates that joining the group with the largest $F^{C'}_{leave}$ will maximize the modularity gain.

Accordingly, when a node is removed in the current snapshot, it may cause a current group broken into subgroups which may further merge into other groups. To address this problem efficiently and effectively, we utilize the clique percolation method [62]. When a node is removed, a 3-clique is placed to one of its neighbor and the clique percolates until no nodes in the original group are discovered. The subgroups of original group then choose the best groups to merge. The algorithm for detecting the dynamic social groups based on the snapshots is given as Algorithm 2.

**Algorithm 2:** Dynamic Social Group Detection

**Input:** Current snapshot $S_{curr}$, detected social groups from previous snapshot $C_{pre}$: $C_1, C_2, \ldots C_n$, new node set $N(u)$, removal node set $R(v)$

**Output:** New hierarchical group structure for the current snapshot

```
foreach u in N(u) do
    if u has no adjacent edge then
        Create a new group with u as the single member;
        Leave other groups and overall Q intact;
    else if u connects existing groups then
        foreach neighbor group C' do
            Calculate F^{C'}_{leave} = e^{C'}_u - (e_u * e_{C'})/(2 * sum_{x_m,x_n in TIN} e_mn);
            Find the maximum F^{C'}_{leave};
        if F^{C'}_{leave} > F^{C}_{stay} then
            Move u to C';
        else
            Leave all the groups intact;
foreach v in R(v) do
    if v.degree > θ then
        Place a 3-clique to one of v's neighbor group;
        Let the clique percolate until no nodes in C are discovered;
        Let the rest nodes of C merge into other groups based on Q';
    else
        Leave all the groups intact;
return Dynamically updated social groups;
```

### D. Unified Social Group Detection and Tracklet Fragment Association

We introduce a unified social group detection and tracklet association scheme. Sets of short tracklets extracted from two consecutive snapshots are concatenated into longer tracklets by using adapted Hungarian algorithm [63] with contextual social groups. We forward scan the tracklets until the number of non-overlapping tracklet pairs reaches the maximum number of detection responses in the frames. The starting and ending frames of the snapshot are set to the starting frame of the first tracklet and the ending frame of the last tracklet, respectively. We then restore a static version of the ETIN by including all the nodes that have a temporal overlap to the sliding window. We use the approach proposed in Section III-C to detect the social groups of tracklets and obtain the group ID for each tracklet in the time window. Finally, we integrate the group information along with the commonly used appearance and motion models into the affinity matrix $M$ and formulate the linear assignment problem as:

$$\underset{\phi \in \Phi}{argmin} \sum_{i,j} \phi_{ij} M_{ij}, \quad \text{where} \tag{7}$$

$$M_{ij} = \gamma_1[\psi_{ij} \cdot e_{ij}] + \gamma_2 f_{appr}(x_i, x_j) + \gamma_3 f_{motion}(x_i, x_j) \tag{8}$$

where $\psi_{ij} = lcg(x_i, x_j)/L$, $L$ is the total number of levels of the hierarchical grouping and $lcg(\cdot,\cdot)$ is the function for computing the *lowest common group* of two tracklets in the hierarchy. $\Phi$ is the correspondence matrix with an element $\phi_{ij} = 1$ if tracklets $x_i$ and $x_j$ are linked and 0 otherwise. $f_{appr}(\cdot,\cdot)$ and $f_{motion}(\cdot,\cdot)$ denote the appearance and motion models, respectively. $\gamma_i$s are the weighting parameters to determine the importance of each model. After the association process, the newly generated set of longer tracklets is used as input for the next round of social group detection and fragment association until all the tracks for the pedestrians are complete.

## IV. Experiments

We validate our proposed method for understanding the dynamic social grouping behavior of pedestrians on a collection of videos from real-world scenes (shopping mall, University campus, building patio) with different densities of crowds (low and medium), viewports, and sizes of the target in the frames. Sample video frames of each sequence are shown in Figs. 5–7. Each video was recorded using elevated cameras. The videos were converted to sequences of JPEG files using the open source software "Video to Picture Converter" to produce non-interlaced 24-bit color images at a frame rate of 30 frames per second. We apply deformable part-based detectors on all the frames [58]. The detection responses from different frames are connected to form the initial short tracklets. We show and

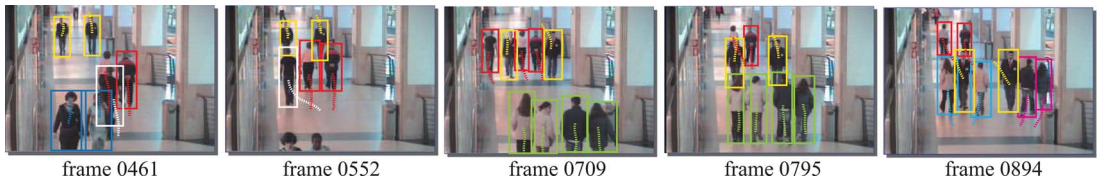


Fig. 5. Group detection results from CAVIAR dataset. The pedestrians that are walking in the same group are marked in the same color. The splitting and merging behaviors are shown in the last four frames.

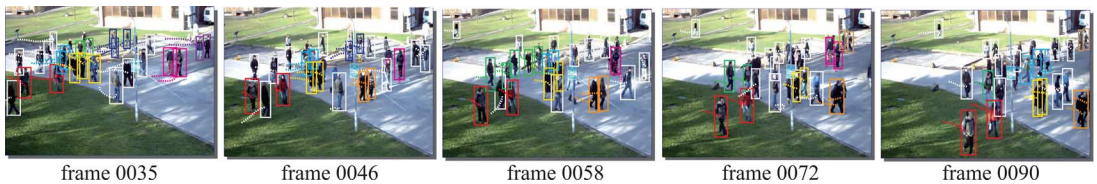


Fig. 6. Group detection results from PETS2009 dataset. The scene is more crowded and complex with a lot of occlusions happen among pedestrians. The dynamic changes of social groups are captured by the color changes of their bounding boxes. The pedestrians with white bounding boxes are walking alone.

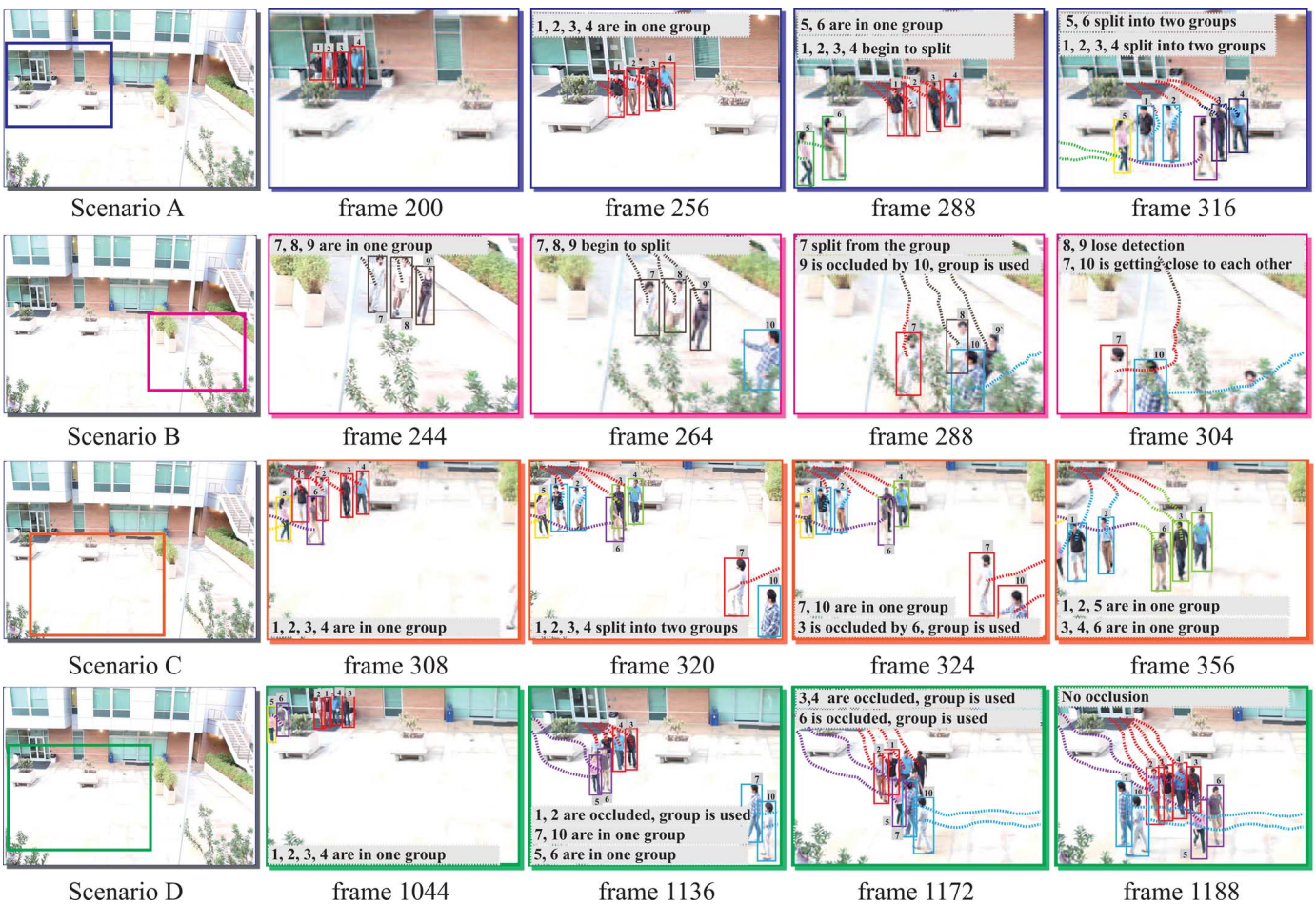


Fig. 7. Group detection results from UNIV dataset. Group splitting and merging behaviors are shown in scenario A, B and C. Scenario D demonstrates the effect of social groups in tracklet association when partial occlusions among group members happen. The social groups are marked in different colors of bounding boxes.

discuss how the detection responses from pedestrian detector will impact the performance of the unified social group detection and tracklet association framework in Section IV-C.

### *A. Data Collection*

The grouping information in the current video datasets is usually unavailable which requires us to manually determine the ground-truth pedestrian groups. We have manually labeled two publicly available datasets that are originally used for multi-people tracking research purpose: CAVIAR dataset [64] (low crowd density), PETS2009 [65] (medium/high crowd density). The ground-truth labeling process is conducted by asking three human judges to identify groups by assigning individuals with group IDs. The judges can rewind and play a video as many times as needed. The final consensus of the ground-truth groups are acquired by using majority voting among the judges. We also introduced a new dataset named "UNIV" which is collected from a University building patio from an elevated camera. The ground-truth for this data are also established by combining decisions made by multiple human judges. There are disagree-

TABLE I
PERCENTAGE DISTRIBUTION OF THE GROUP SIZES

| | size of 1 | 2 | 3 | 4 or more |
|---|---|---|---|---|
| CAVIAR [64] | 21% | 57% | 8% | 14% |
| PETS2009 [65] | 13% | 42% | 21% | 24% |
| UNIV (newly introduced) | 10% | 69% | 5% | 16% |

ments among the judges on some of the groups which indicates baseline ambiguity exists in the video sequences. The average disagreement rate on the number of group over the three datasets is about 5% percent of the total number of groups.

Feedback from the judges indicates that the difficulty of group identification arises when the crowd density increases. This makes PETS2009 the most difficult dataset to label. Also it is easier to identify groups from sequences with a camera viewport direction that is parallel with the walking directions of the pedestrians than the videos from camera with perpendicular views to the direction of the walking people.

CAVIAR dataset captures people walking in an indoor shopping mall environment by an elevated camera. In the dataset, people either walk from near field to far field or vice versa, and a lot of social grouping behavior can be observed during their walking. The merging and splitting of groups also happen frequently over time. There are also partial occlusions between the members of groups. PETS2009 dataset contains video sequence recorded in an outdoor scene from an University campus with a high density of people in each frame (on average 25 people are visible in each frame). Identifying individuals within a group is more challenging due to the frequent occlusions and the abrupt motion changes (direction, velocity, etc.). UNIV dataset is collected at a large camera angle under bright light conditions. The crowd density is larger than CAVIAR but smaller than PETS2009 dataset. However, more grouping behaviors and other social interactions are involved in this dataset.

The percentage distribution of group sizes from the ground-truth labeling for the video sequences are summarized in Table I.

### B. Quantitative Evaluation

We set $\omega$s in (2) and $\gamma$s in (8) to $1/3$. We set the two thresholds $\epsilon = 0.05$ and $\epsilon' = 0.3$ in Algorithm 1. We compare the performance of our proposed group detection with the following baseline approaches:

- **Baseline-I [8]:** A hierarchical agglomerative clustering based group analysis approach that starts with assigning each individual into a separate group and gradually merges the small groups into larger ones. The spatio-temporal dissimilarity between tracklets of individuals is used as a distance measure. However, it does not explicitly address the dynamic changes of social groups.
- **Baseline-II [37]:** It is another bottom-up group detection approach that is built upon algorithms for pedestrian detection and multi-people tracking. The interactions between individuals are measured by pairwise proximity and velocity without using a network representing the interactions and modularity gain as the group measure.

We compare our ETIN approach with two baseline approaches using the evaluation metric as the Percentage of correctly detected Social Groups (**PSG**) of different sizes. We measure the influence of simultaneous groups by using Percentage of correctly detected social groups of Any size as a function of the number of Simultaneous groups (**PAS**). We also compare the performance of different approaches in tracking the dynamic social group changes (splitting and merging) by using Percentage of correctly detected Dynamic group Changes (**PDC**), which is defined as the number of correctly detected group changes by our unified detection and association approach, divided by the total number of ground-truth changes marked manually in the video frames.

TABLE II
QUANTITATIVE EVALUATION ON CAVIAR DATASET

| Metric | Baseline-I [8] | Baseline-II [37] | our ETIN |
|---|---|---|---|
| PSG-1 | 67.4% | 79.2% | 83.5% |
| PSG-2 | 52.2% | 65.7% | 75.4% |
| PSG-3 | 48.5% | 57.1% | 69.4% |
| PSG-4/more | 39.3% | 47.6% | 67.8% |
| PAS-1 | 78.5% | 82.1% | 86.9% |
| PAS-2 | 54.9% | 62.6% | 69.4% |
| PAS-3/more | 47.3% | 51.7% | 61.3% |
| PDC | 34.5% | 31.2% | 79.5% |

*1) Results on CAVIAR Dataset:* We automatically detected pedestrians and generated the tracklets, and carried out the ETIN construction and modularity-based hierarchical group detection to understand the social grouping behaviors. Sample results are shown in Fig. 5. The statistical results are summarized in Table II. From the table we can observe that, all the approaches are able to identify the pedestrians walking alone with a high percentage of correctness. However, when the group size increases, the PSG scores from Baseline-I and Baseline-II degrades more than for our approach, which implies that our approach is more robust in detecting social groups in larger sizes. Further, when the group size is larger than 2, our PSG score is relatively stable which demonstrates the power of our network representation of tracklet interactions is stronger as compared to the pairwise social interaction representation used in other approaches. Baseline-I achieves relatively the same low score of PDC as Baseline-II which indicates that they do not actively address the dynamic group changes. This suggests that our unified framework for social group detection and tracklet association that utilizes temporal snapshots at different time intervals yields better performance in tracking the dynamic changes of social groups. Our approach achieves the best performance when more than one group appear simultaneously measured by the PAS scores. When more than two groups appears at the same time, our approach can still maintain a relatively large score (61.3%) which demonstrates that our approach can effectively handle the influences across groups.

*2) Results on PETS2009 Dataset:* Similar experiments were conducted on the shorter but more challenging PETS2009 dataset. The scores of the evaluation metrics are summarized in Table III. Our approach gives better results (though reduced PSGs and PDC scores) for group detection and dynamic behavior tracking performance as compared to the other two approaches. Some sample detected groups and pedestrian walking behaviors are shown in Fig. 6. Even for this harder problem, our approach still demonstrates a substantial agreement (more than 50% of correctness) with the ground-truth not only on the different group sizes (1, 2, 3, 4 and more than 4), but also on the dynamic changes of the memberships of the groups.

A further investigation on the results shows that the PSG performance of our approach is not as stable as on the CAVIAR dataset. It degrades gradually as the number of members in the groups increase. A potential reason is that the crowd is in

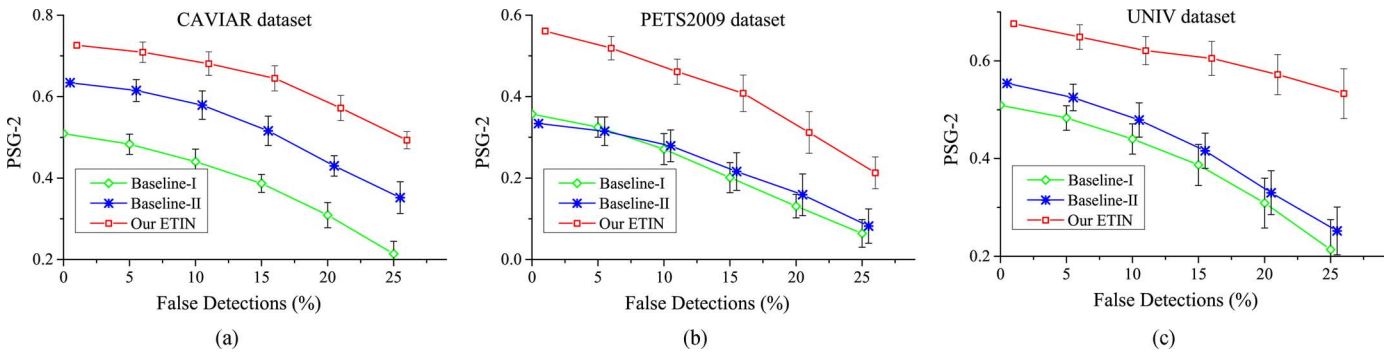


Fig. 8. The PSG measure is compared across the three approaches, as the percentage of false detections varies on (a) CAVIAR, (b) PETS2009 and (c) UNIV.

TABLE III
QUANTITATIVE EVALUATION ON PETS2009 DATASET

| Metric | Baseline-I [8] | Baseline-II [37] | our ETIN |
|---|---|---|---|
| PSG-1 | 49.3% | 53.2% | 74.5% |
| PSG-2 | 39.5% | 35.1% | 67.3% |
| PSG-3 | 29.7% | 31.7% | 59.3% |
| PSG-4/more | 29.1% | 27.5% | 51.6% |
| PAS-1 | 53.6% | 61.7% | 78.7% |
| PAS-2 | 41.9% | 50.4% | 63.2% |
| PAS-3/more | 27.2% | 31.3% | 51.9% |
| PDC | 26.1% | 25.5% | 58.4% |

TABLE IV
QUANTITATIVE EVALUATION ON UNIV DATASET

| Metric | Baseline-I [8] | Baseline-II [37] | our ETIN |
|---|---|---|---|
| PSG-1 | 60.3% | 66.7% | 78.6% |
| PSG-2 | 57.3% | 60.4% | 73.1% |
| PSG-3 | 51.2% | 55.8% | 70.3% |
| PSG-4/more | 49.6% | 51.4% | 69.5% |
| PAS-1 | 54.8% | 62.8% | 74.3% |
| PAS-2 | 44.5% | 49.6% | 61.7% |
| PAS-3/more | 38.6% | 41.1% | 48.9% |
| PDC | 14.3% | 11.9% | 54.2% |

medium/high density and the group members tend to walk in a random pattern to avoid collisions to other pedestrians which results in a weakened social interactions among group members. For a moderate crowd of 25 people per frame, our PDC score is above 0.5, which still indicates a reasonable to good performance of our approach in tracking the dynamic group changes. The PAS scores have decreased to 51.9% compared to the CAVIAR dataset when more than two groups appear simultaneously. This implies that incorrect group information exerts on single person will have negative influence on the tracking performance when the density of the pedestrians is large and occlusion becomes a challenging problem.

*3) Results on UNIV Dataset:* To further evaluate the effectiveness of our approach in understanding the dynamic social groups, we applied the approach on the UNIV dataset where more social grouping behaviors are involved in a natural setting. The inter-group interactions are easier to be distinguished from the intra-group interactions in the first few frames because the groups are coming from different corners in the scene and the walking direction of each group is different. However, it becomes more challenging when the groups begins to merge and re-split in the middle frames of the video. A quantitative comparison is shown in Table IV.

From Table IV we can observe that although the PSG scores drop to some degree compared to the scores from the CAVIAR dataset, the performance of our approach still exceeds the Baselines which demonstrates that the dynamic group analysis model and the unified group detection and tracklet association framework work effectively on this dataset where group information plays a positive role in concatenating tracklets of group members while intense occlusion happens. Overall, our proposed approach achieves the best results over the other approaches in PSG scores in all the group sizes. However, as compared to the CAVIAR dataset, the PDC scores from all the approaches have decreased to some extent as UNIV has much more dynamic social interactions that are interlaced with a large number of occlusions.

There are considerable drops in the PDC scores for the two Baselines compared to our method, particularly for the Baseline-II where the score drops from 31.2% in the CAVIAR dataset to 11.9%. The primary reason is that the other two methods do not handle group changes explicitly by investigating the group member interactions over time. The PAS score shows that our approach can still achieve a relatively good performance when more group dynamics (appear, disappear, merge, split) are involved.

### C. Impact From False Detection

It is to be noted that the underlying detection errors could propagate to the group detection process in all the approaches that are based on pedestrian detection. To show that to what extent these approaches rely on accurate detection responses, we artificially introduce three types of false detections into the correct detection responses. They are: *misdetections* which represent the type of missing data, *false responses* and *inaccurate detections* that represent outliers and noises separately. The first type of false detections is added by randomly erasing correct detections and the rest two types are added by setting detections at random locations that do not cover correct detections. All the three types of false detections are added together at the percentages [0, 5%, 10%, 15%, 20%, 25%] of the total number of detections in the three datasets. The group detection performance measured by PSG-2 as a function of false detection percentages is shown in Fig. 8.

The results from Fig. 8 show that the robustness of our approach given unreliable detection responses. As expected, our approach maintains the best performance when the false detection percentage increases. This indicates that social groups are important contextual cues when the short tracklets are linked to form longer ones; if a group member is occluded by other pedestrians in the scene, the other group members that have close tracklet interactions can contribute to the estimation of the

TABLE V
QUANTITATIVE TRACKING PERFORMANCE ON FM DATASET

| Metric | DEEPER-JIGT [66] | VAR3 [66] | our ETIN |
|---|---|---|---|
| MOTP | 0.80 | 2.80 | 0.31 |
| MOTA | 67.58% | 2.73% | 69.42% |

tracks of the occluded group member. The performance of the other two approaches that do not consider using group information in forming the trajectories drops as more false detections are obtained.

### D. Application: Pedestrian Tracking

The focus of this paper is our novel approach in understanding dynamic social grouping behaviors by clustering trajectories using a social network analysis based method. However, tracking individuals by generating reliable tracks is itself a non-trivial task because of the complexity of the environment. Therefore, for completeness, we utilize our social grouping analysis framework in this section to address the individual tracking problem, which is capable of producing reasonable results that can be compared with other state-of-the-art tracking methods. Tracking individuals in the crowd is formulated as a multi-target tracking problem. We use our modified Hungarian algorithm that is integrated with individual group information to perform multi-target data association between current trajectory hypotheses and the trajectories in the following frames (see Section III-D for more detail). Our modified Hungarian algorithm finds an optimal bipartite marching between tracklets not only based on the physical similarity but also based on the group similarity.

We evaluate our approach using the following dataset:

*1) Friends Meet (FM) [66]:* contains groups of pedestrians that appear, disappear and evolve (split and merge) over time. The dataset is composed of 53 sequences for a total of 16286 frames. We use a subset of 25 sequences that contains sequences in real-life outdoor scenes. The range of the individuals in a single frame is between 3 and 11.

We use the following metrics to evaluate the performance:

*2) MOTP (Multi-Object Tracking Precision) [71]:* which we define as the total error for associated tracklet-hypothesis pairs across all the time sliding windows, averaged by the total number of associations made. The value is the lower the better.

*3) MOTA (Multi-Object Tracking Accuracy) [71]:* which equals one minus the mismatch rate in the data association process. It is similar to metrics widely used in other domains such as the word error rate (WER) used in speech recognition. The value is the larger the better.

We compare with the following approaches as baselines:

- **DEEPER-JIGT** (DEcentralizEd Particle filtER for Joint Individual-Group Tracking) [66]: a joint individual-group tracking framework based on decentralized particle filtering which factorizes the joint individual-group state space in two conditionally dependent subspaces. The approach is specialized in real-time tracking scenario.
- **VAR3** [66]: a variant of DEEPER-JIGT which separates individual from group tracking in two different particle filters thus blocks the contribution of the group clustering.

The results on FM dataset are summarized in Table V. Our approach reaches the best performances in terms of the MOTP and MOTA evaluation. Moreover, the group information has been demonstrated as a crucial source to boosting the individual tracking. By pruning away the group information (VAR3), the performances decrease dramatically compared to other two approaches (DEEPER-JIGT and our ETIN) which build the connection between groups and individuals. In our unified social group detection and tracklet fragment association framework, the individual tracklets consider the influence from the groups in the data association process which shows the effectiveness of injecting group-driven dynamics.

## V. CONCLUSIONS

We proposed a principled method for understanding the social grouping behavior of pedestrians as well as a unified framework for tracking the dynamic social group changes and tracklet association based on the temporal snapshots of the introduced evolving tracklet interaction network (ETIN). Our novel model addressed the social group understanding problem in video sequences from a social network perspective. The novelties included representing tracklets of pedestrians and their interactions in a network which is evolving over time and carrying out modularity to divide the tracklets into hierarchical subgroups. The dynamic changes of social groups are detected using the restored static temporal snapshots of the original network based on the time overlaps. In experiments, we showed that our method is adapted to dynamic grouping behaviors such as merging and splitting and it is robust in detecting social groups of different densities.

## APPENDIX
## PROOF OF THE MAXIMAL MODULARITY GAIN

*Theorem 1: Suppose a new node $u$ with degree $d$ is added into the group that gives the maximum $F^{C'}_{leave}$, then adding $u$ to $C'$ gives the maximal modularity gain.*

*Proof:* Let $C''$ be another group of $G$ and $C'' \neq C'$. We would like to prove that joining $u$ into $C''$ will give less modularity gain than joining $C'$. Let $f_{C'}$ denotes the total degree of nodes in $C'$, and let $M$ denotes half of the summation of the total edge weights in $G$. The overall modularity $Q$ when $u$ joins $C'$ is

$$Q = \frac{e_{C'} + e^u_{C'}}{M+d} - \frac{(f_{C'} + e^u_{C'} + d)^2}{4(M+d)^2} + \frac{e_{C''}}{M+d} - \frac{(f_{C''} + e^u_{C''})^2}{4(M+d)^2} + A \quad (9)$$

where $A$ is the summation of other modularity gains. Similarly, adding $u$ to $C''$ will give

$$Q' = \frac{e_{C'}}{M+d} - \frac{(f_{C'} + e^u_{C'})^2}{4(M+d)^2} + \frac{e_{C''} + e^u_{C''}}{M+d} - \frac{(f_{C''} + e^u_{C''} + d)^2}{4(M+d)^2} + \frac{e_{C''}}{M+d} + A \quad (10)$$

and

$$Q - Q' = \frac{1}{M+d}\left(e^u_{C'} - e^u_{C''} + \frac{d(f_{C''} - f_{C'} + e^u_{C''} - e^u_{C'})}{2(M+d)}\right) \quad (11)$$

since $C'$ is the group that gives the maximum $F^{C'}_{leave}$, we have

$$e^u_{C'} - \frac{d(f_{C'} + e^u_{C'})}{2(M+d)} > e^u_{C''} - \frac{d(f_{C''} + e^u_{C''})}{2(M+d)} \quad (12)$$

which means

$$e^u_{C'} - e^u_{C''} + \frac{d(f_{C''} - f_{C'} + e^u_{C''} - e^u_{C'})}{2(M+d)} > 0 \qquad (13)$$

therefore, $Q - Q' > 0$ and the conclusion is true.

## REFERENCES


[1] L. Feng and B. Bhanu, "Utilizing co-occurrence patterns for semantic concept detection in images," in *Proc. 21st Int. Conf. Pattern Recogn.*, Nov. 2012, pp. 2918–2921.

[2] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behavior of pedestrian social groups and its impact on crowd dynamics," *PLos ONE*, vol. 5, no. 4, p. e10047, Apr. 2010.

[3] X. Pan, C. S. Han, K. Dauber, and K. H. Law, "A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations," *AI Soc.*, pp. 113–132, Nov. 2007.

[4] 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.

[5] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.

[6] R. Eshel and Y. Moese, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2008, pp. 1–8.

[7] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, vol. 5305, pp. 691–704.

[8] M. C. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group level activity recognition in crowded environments across multiple cameras," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.*, 2010, pp. 56–63.

[9] M. Breitenstein, F. Reichlin, B. Leibe, E. K. Meier, and L. V. Gool, "Robust tracking-by-Detection using a detector confidence particle filter," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1515–1522.

[10] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, vol. 5303, pp. 788–801.

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

[12] S. Pellegrini, A. Ess, and L. V. Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, vol. 6311, pp. 452–465.

[13] X. Chen, L. An, Q. Zhen, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2014, pp. 1242–1249.

[14] Q. Yu and G. Medioni, "Multiple-target tracking by spatio-temporal Monte Carlo Markov chain data association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2007, pp. 1–8.

[15] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.

[16] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. Xu, "Crowd analysis: A survey," *J. Mach. Visi. Applicat.*, vol. 19, no. 5, pp. 345–357, Oct. 2008.

[17] A. B. Chan, Z. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 1–7.

[18] P. Kilamba, E. Ribnick, A. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Comput. Vision Image Understand.*, vol. 110, no. 1, pp. 43–59, Apr. 2008.

[19] 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.

[20] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 935–942.

[21] C. McPhail, *Withs Across the Life Course of Temporary Sport Gatherings.* 2003, Univ. of Illinois.

[22] R. W. Brown, "Mass phenomena," in *Handbook of Social Psychology*, G. Lindzey, Ed. Boston, MA, USA: Addison Wesley, 1954, vol. 2, pp. 833–876.

[23] W. Choi and G. Medioni, "Learning context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3273–3280.

[24] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, vol. 7575, pp. 215–230.

[25] Y. Zhang, W. Ge, M. C. Chang, and X. Liu, "Group context learning for event recognition," in *Proc. IEEE Workshop Applicat. Comput. Vis.*, Jan. 2012, pp. 249–255.

[26] L. Ding and A. Yilmaz, "Learning relations among movie characters: A Social network perspective," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, vol. 6314, pp. 410–423.

[27] L. Ding and A. Yilmaz, "Inferring social relations from visual concepts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 699–706.

[28] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2012, pp. 1226–1233.

[29] V. Ramanathan, B. Yao, and F. F. Li, "Social role discovery in human events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2013, pp. 2475–2482.

[30] T. Yu, S. N. Lim, K. Patwardhan, and N. 'Krahnstoever, "Monitoring, recognizing and discovering social networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jan. 2009, pp. 1462–1469.

[31] C. W. Chen and H. Aghajan, "Multiview social behavior analysis in work environments," in *Proc. ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Aug. 2011, pp. 1–6.

[32] M. Ryoo and J. Aggarwal, "Recognition of high-level group activities based on activities of individual members," in *Proc. IEEE Workshop Motion Video Comput.*, Jan. 2008, pp. 1–8.

[33] W. Zhang, F. Chen, W. Xu, and Y. Du, "Hierarchical group process representation in multi-agent activity recognition," *Image Commun.*, vol. 23, pp. 739–739, Jan. 2008.

[34] I. Haritaoglu and M. Flickner, "Detection and tracking of shopping groups in stores," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2001, pp. 431–438.

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

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

[37] W. Ge, R. T. Collins, and B. Ruback, "Automatically detecting the small group structure of a crowd," in *Proc. IEEE Workshop Applicat. Comput. Vis.*, Dec. 2009, pp. 1–8.

[38] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Robust tracking-by-Detection using a detector confidence particle filter," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2009, pp. 1515–1522.

[39] S. Wang, H. Lu, F. Yang, and M. H. Yang, "Superpixel tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1323–1330.

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

[41] B. Song, T. Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury, "A stochastic graph evolution framework for robust multi-target tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, vol. 6311, pp. 605–619.

[42] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, vol. 5303, pp. 788–801.

[43] J. Berclaz, F. Fleuret, E. T. Uretken, 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.

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

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

[46] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *PNAS*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[47] G. Palla, P. Pollner, A. Barabasi, and T. Vicsek, "Social group dynamics in networks," in *Adaptive Networks*. Berlin/Heidelberg, Germany: Springer, 2009, pp. 11–38.

[48] S. Fortunato, "Community detection in graphs," *Phys. Reports*, vol. 486, no. 3–5, pp. 75–174, Feb. 2010.
[49] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech.: Theory Exper.*, no. 10, Oct. 2008.
[50] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E.*, vol. 69, no. 2, p. 026113,, Feb. 2004.
[51] M. Mitrovic and B. Tadic, "Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities," *Phys. Rev. E*, vol. 80, no. 2, p. 026123, Aug. 2009.
[52] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, p. 066133, Jun. 2004.
[53] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di, "Community detection by signaling on complex networks," *Phys. Rev. E.*, vol. 78, no. 1, p. 016115, Jul. 2008.
[54] E. Estrada and N. Hatano, "A vibrational approach to node centrality and vulnerability in complex networks," *Physica A: Statist. Mech. and its Applicat.*, vol. 389, no. 17, pp. 3648–3660, Sep. 2010.
[55] G. Palla 1, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature.*, vol. 435, pp. 814–818, Jun. 2005.
[56] S. Zhang, R. S. Wang, and X. S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy C-means clustering," *Physica A: Statist. Mech. and its Applicat.*, vol. 374, no. 1, pp. 483–490, Jan. 2007.
[57] N. Ghosh and B. Bhanu, "Evolving Bayesian graph for 3D vehicle model building from video," *IEEE Trans. Intell. Transportat. Syst.*, vol. 15, no. 2, pp. 563–578, Apr.. 2014.
[58] P. Felzenszwalb, D. McAllester, and D. Ramaman, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2008, pp. 1–8.
[59] C. McPhail and R. Wohlstein, "Using film to analyze pedestrian behavior," *Sociol. Meth. Res.*, vol. 10, no. 3, pp. 347–375, Feb. 1982.
[60] M. E. Newman and M. Girvan, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, p. 036104, Sep. 2006.
[61] Z. Ye, S. Hu, and J. Yu, "Adaptive clustering algorithm for community detection in complex networks," *Phys. Rev. E*, vol. 78, p. 046115, Oct. 2008.
[62] G. Palla, P. Pollner, A. L. Barabási, and T. Vicsek, "Social group dynamics in networks," *Adaptive Netw.*, pp. 11–38, Sep. 2009.
[63] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, pp. 83–97, 1955.
[64] [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/
[65] [Online]. Available: http://www.pets2009.net/
[66] L. Bazzani, V. Murino, and M. Cristani, "Decentralized particle filter for joint individual-group tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2012, pp. 1886–1893.
[67] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.
[68] G. Gennari and G. Hager, "Probabilistic data association methods in visual tracking of groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2004, vol. 2, pp. 876–881.
[69] M. Chang, N. Krahnstoever, and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 747–754.
[70] Y. D. Wang, J. K. Wu, A. A. Kassim, and W. M. Huang, "Tracking a variable number of human groups in video using probability hypothesis density," in *Proc. Int. Conf. Pattern Recogn.*, 2006, pp. 1127–1130.
[71] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *J. Image Video Process.*, Feb. 2008.

**Linan Feng** received the B.Sc. degree in electrical engineering in 2006 and the M.E. degree in software engineering in 2009 both from Shanghai Jiao Tong University, Shanghai, China. Since 2009, he has been a Ph.D. candidate in computer science, at the University of California, Riverside, CA. His research interests are in computer vision, pattern recognition and machine learning, with emphasis on automated image annotation and concept-based image retrieval. He is the student member of IEEE.

**Bir Bhanu** received the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, the Ph.D. degree in electrical engineering from the Image Processing Institute, University of Southern California and the M.B.A. degree from the University of California, Irvine. He is the Distinguished Professor of Electrical Engineering and Cooperative Professor of Computer Science and Engineering, Mechanical Engineering and Bioengineering, Director of the Center for Research in Intelligent Systems (CRIS), and the Visualization and Intelligent Systems Laboratory (VISLab) at the University of California, Riverside (UCR). He also serves as the Director of NSF IGERT program on Video Bioinformatics. He was the Founding Chair of Electrical Engineering at UCR. Prior to that, he was Senior Honeywell Fellow at Honeywell Inc. His current research interests are computer vision, pattern recognition and data mining, image and video database, graphics/visualization, robotics, biological, medical, military and intelligence applications. He has been the principal investigator of various programs from NSF, DARPA, NASA, AFOSR, ONR, ARO and other agencies and industries. He is Fellow of IEEE, AAAS, IAPR and SPIE.