Title

Multiperson Tracking by Online Learned Grouping Model With Nonlinear Motion Context

Journal

IEEE Transactions on Circuits and Systems for Video Technology, 26(12)

ISSN

1051-8215

Authors

Chen, Xiaojing

Qin, Zhen

An, Le

et al.

Publication Date

2016-12-01

DOI

10.1109/tcsvt.2015.2511480

Peer reviewed

Multiperson Tracking by Online Learned Grouping Model With Nonlinear Motion Context

Xiaojing Chen, Zhen Qin, Le An, and Bir Bhanu, *Fellow, IEEE*

***Abstract*— An online approach to learn elementary groups containing only two targets, i.e., pedestrians, for inferring high-level context is introduced to improve multiperson tracking. In most existing data association-based tracking approaches, only low-level information (e.g., time, appearance, and motion) is used to build the affinity model, and each target is considered as an independent agent. Unlike those previous methods, in this paper, an online learned social grouping behavior model is used to provide more robust tracklet affinities. A disjoint grouping graph is used to encode social grouping behavior of pairwise targets, where each node represents an elementary group of two targets, and two nodes are connected if they share a common target. Probabilities of the uncertain target in two connected nodes being the same person are inferred from each edge of the grouping graph. Relationships between elementary groups are discovered by group tracking, and a nonlinear motion map is used for explaining nonlinear motion pattern between elementary groups. The proposed method is efficient, able to handle group split and merge, and can be easily integrated into any basic affinity model. The approach is evaluated on four data sets, and it shows significant improvements compared with state-of-the-art methods.**

***Index Terms*— Data association, elementary grouping model, multitarget tracking, social grouping behavior.**

I. Introduction

AUTOMATIC tracking of multiple targets simultaneously in real-world scenes has been an active research topic in computer vision for many years, as it is crucial for many industrial applications and high-level analysis, such as visual surveillance, human–computer interaction, and anomaly detection. The goal of multitarget tracking is to recover trajectories of all targets while maintaining consistent identity labels. There are many challenges for this problem, such as illumination and appearance variation, occlusion, and sudden change in motion [1], [2]. As great improvement has been achieved in object detection, data association-based tracking (DAT) has become popular recently [3]–[7]. In the DAT framework, often a prelearned detector is applied on each frame to produce detection responses of all targets, and short-term tracking results (i.e., tracklets) are generated by associating responses from consecutive frames that have high probability to contain the same target. These tracklets are further linked to produce long-term tracking results. An affinity model integrating multiple visual cues (appearance and motion information) is formulated to find the linking probability between tracklets, and the global optimal solution is often obtained by solving the maximum *a posteriori* problem using various optimization algorithms.

Fig. 1. Examples in which grouping information is helpful under the challenging conditions for tracking in a video. The same color indicates the same target. Note that for both targets with bounding boxes, there are significant appearance and motion changes due to occlusions and cluttered background. Images are from CAVIAR data set [8].

Although much progress has been made in building more discriminative appearance and motion models, problems such as identity switch and track fragmentation still exist in current association-based tracking approaches, especially under challenging conditions where appearance or motion of the target changes abruptly and drastically, as shown in Fig. 1. The goal of association optimization is to find the best set of associations with the highest probability for all targets, which makes it not necessarily capable of linking each of the difficult tracklet pairs. In this paper, we explore high-level contextual information, i.e., social grouping behavior, for associating tracklets that are very challenging using only lower level features (time, appearance, and motion).

Manuscript received February 28, 2015; revised July 13, 2015; accepted September 15, 2015. Date of publication December 22, 2015; date of current version December 2, 2016. This work was supported in part by the National Science Foundation under Grant 1330110 and in part by the Office of Naval Research under Grant N00014-12-1-1026. This paper was recommended by Associate Editor J. Zhang. *(Corresponding author: Le An.)*

X. Chen and Z. Qin are with the Department of Computer Science, University of California at Riverside, Riverside, CA 92521 USA (e-mail: xchen010@ucr.edu; zqin001@cs.ucr.edu).

L. An is with National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: lan004@ucr.edu).

B. Bhanu is with the Center for Research in Intelligent Systems, University of California at Riverside, Riverside, CA 92521 USA (e-mail: 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/TCSVT.2015.2511480

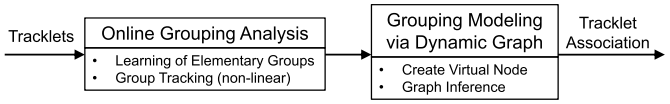


Fig. 2. Overview of the elementary grouping model.

When there are only a few interactions and occlusions among targets, DAT achieves robust performance. Discriminative descriptors of targets are usually generated using appearance and motion information from tracklets. Appearance model often uses global or part-based color histograms to match tracklets, and a linear motion model that assumes all targets maintain constant speed without motion direction change is often adopted to constrain motion smoothness of two tracklets. However, these low-level descriptors generally fail to associate tracklet pairs with long time gap. This is because the appearance of a target might change drastically due to heavy occlusion, and the linear motion model is unreliable for predicting location of a target after a large time interval.

Nevertheless, there is often other useful high-level contextual information in the scene, which can be effectively used to mitigate the aforementioned shortcomings. For instance, sociologists have found that up to 70% of pedestrians tend to walk in groups in a crowd, and people in the same group are more likely to have a similar motion pattern and be spatially close to each other for better group interaction [9]. Moreover, pedestrians in the crowd often either consciously or unconsciously follow other individuals with a similar destination to facilitate navigation [10]. It is also observed in many real-world surveillance videos that if two people are walking together at certain time, then it is very likely that these two people will still walk together after a short time period.

Based on the above observations, we propose an elementary grouping model with nonlinear motion context to compensate for the errors caused using basic appearance model and linear motion model. A grouping graph is constructed based on input tracklets with high confidence, where each node represents a pair of tracklets that form an elementary group (a group of two targets) and each edge indicates that the connected two nodes (two elementary groups) have at least one target in common. The group trajectories of any two linked nodes are used to estimate the probability of the other target in each group being the same person. Neighboring tracklets that have time overlap and a similar motion pattern are possible candidates for elementary groups. Relationships between elementary groups are further discovered with the help of group tracking, in which a nonlinear motion map is used to explain a large time gap between two elementary groups. The elementary grouping model is summarized in Fig. 2.

The size of a group may change dynamically as people join and leave the group, but a group of any size can always be considered as a set of elementary groups. Therefore, focusing on finding elementary groups instead of the complete group makes our approach capable of modeling flexible group evolution [11] in the real world. Note that the social group in this paper refers to a number of individuals with correlated movements and does not indicate a group of people who know each other.

The rest of this paper is organized as follows. Section II discusses related work and contributions of this paper, the proposed elementary grouping model is described in Section III, experiments are presented in Section IV, and Section V concludes this paper.

II. Related Work and Contributions

A. Related Work

Visual tracking has attracted extensive research efforts in recent years, from individual tracker design [12], [13] to multitracker fusion [14]. The method proposed in [14] was the first work that can jointly exploit both the spatial and the temporal correlation from multiple tracking results, leading to the state-of-the-art tracking performance. Specifically, the temporal information helps to identify individual tracker consistency, and the spatial information is used to establish pairwise correlation among multiple trackers.

Traditional filtering-based multitarget tracking methods process videos on a frame-by-frame basis, which are more suitable for time-critical applications [12], [15]. However, such greedy methods tend to get stuck at a local optimum, with the possible solution space growing exponentially in the presence of observation gaps. Recently, the focus of multitarget tracking has shifted to robust DAT schemes, due to their global reasoning ability of the solution space. With a deferred global inference, DAT is more robust against observation gaps resulting from heavy interactions and occlusions [16].

Huang *et al.* [13] first propose to hierarchically associate detection responses for multiperson tracking. Since then, most follow-up works focus on designing features for more reliable association scores or developing effective optimization schemes. In the first regime, affinity scores are generally extracted from appearance information such as color histograms and motion features such as motion smoothness. Global appearance constraints are exploited to prevent identity switches in multitarget tracking [17]. Part-based appearance models have been applied in multitarget tracking to mitigate occlusions [18]. For optimization, bipartite matching via the Hungarian algorithm is among the most popular and simplest algorithms [7], [13]. A lot of other optimization frameworks have been proposed, such as K-shortest path [19], set-cover [20], linear programming [21], and quadratic Boolean programming [22].

Most of the works consider only pairwise similarities, without referring to high-level contextual information. Thus, problems such as possible abrupt motion changes cannot be properly accounted for. Yang and Nevatia [23] use a conditional random field for tracking while modeling motion dependencies among associated tracklet pairs. Butt and Collins [24] carry out a Lagrangian relaxation to make higher order reasoning tractable in the min-cost flow framework. These methods focus on higher order constraints such as constant velocity. However, both of them [23], [24] concentrate on individuals and may fail in real-world scenarios, in which individuals may possess a lot of freedom.

In this paper, we focus on utilizing social grouping information for more natural *high-level contextual constraints*.

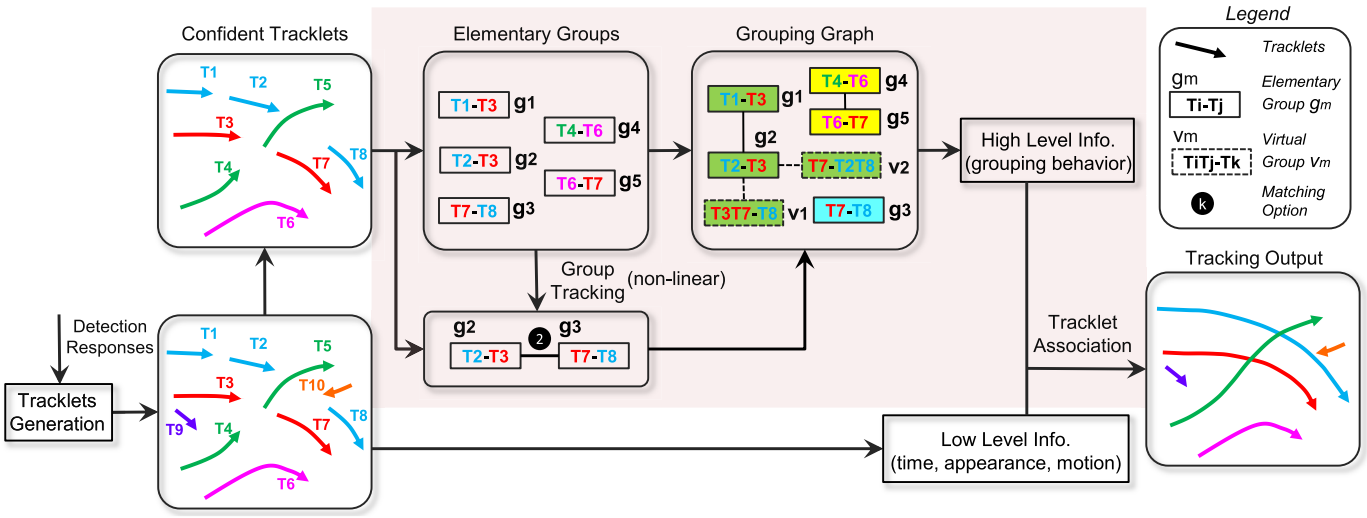


Fig. 3. Block diagram of our tracking system. After initial tracklets are generated by linking detection responses, confident tracklets are selected to form elementary groups. The relationships between elementary groups are identified by group tracking with nonlinear motion context. Then a disjoint grouping graph is constructed, from which high-level information (i.e., grouping behavior) is extracted. Finally, tracklet association is carried out based on the affinity model that combines both high-level and low-level information. Tracklets with the same color contain the same target. For the legends, please see the box on the upper right-hand side. Best viewed in color.

Social factors have attracted a lot of attention in multitarget tracking recently, since they are complementary to unreliable visual features and are motivated by sociology research. Pellegrini *et al.* [25] propose a more effective dynamic model by leveraging nearby people's positions. Brendel *et al.* [26] also consider nearby tracks as contextual constraints. Alahi *et al.* [27] study large-scale crowd destination forecasting with social context. Pellegrini *et al.* [28] improve trajectory prediction accuracy by inferring pedestrian groups. In the DAT context, Qin and Shelton [29] seek the consistency of trajectories in both tracklet association space and tracklet group assignment space based on visual and grouping cues. They use gradient-based optimization and K-means clustering with multiple random initializations. Bazzani *et al.* [30] consider joint individual group tracking, with a decentralized particle filter sampling in both individual and group spaces. Yan *et al.* [31] explicitly consider group structures to improve tracking consistency across time. Compared with these methods, our approach is deterministic with a closed-form solution. Furthermore, the previous work assumes a static group structure or a fixed number of groups, while our grouping scheme is more flexible using elementary groups and allows for more local refinements.

B. Contributions of This Paper

The contributions of this paper include the following.

1) An approach estimating elementary groups online is proposed, which infers grouping information to adjust the affinity model for DAT. This approach is independent of detection methods, affinity models, and optimization algorithms.
2) A motion model that takes advantage of nearby nonlinear motion patterns is integrated into group tracking. It enables the proposed method to explain reasonable nonlinear motions of targets.
3) The proposed approach based on elementary grouping is simple and computationally efficient, while it is effective and robust.
4) Four real-world surveillance data sets are used for evaluation, and extensive experiments are carried out to validate the effectiveness of the proposed method.

A preliminary version of this paper appeared in [32]. In this paper, we have the following major changes and improvements compared with [32].

1) We study the related work more extensively, and more recent advances are discussed.
2) We improve our elementary grouping framework by incorporating a modified motion model for group tracking to handle nonlinear motions of targets.
3) We include more details for better understanding of the technical approach.
4) We conduct more in-depth experiments on more data sets and provide more comparisons with the state-of-the-art methods.

III. Technical Approach

In this section, we introduce how the elementary grouping model is integrated into the basic tracking framework for tracklet association. An overview of the proposed method is presented in Fig. 3.

A. Tracking Framework With Grouping

Given a video sequence, a human detector is first applied to each frame to obtain detection responses. Finding the best set of detection associations with the maximum linking probability is the aim of detection-based tracking. In an ideal association, each disjoint string of detections should correspond to the trajectory of a specific target in the ground truth (GT). However, object detector is prone to errors, such as

false alarms and inaccurate detections. Also, directly linking detections incur a high computational cost. In order to generate a set of reliable tracklets (trajectory fragments), therefore, it is a common practice to prelink detection responses that have high probability to contain the same person. Next, a global optimization method is employed to associate tracklets according to multiple cues. Finally, missed detections are inserted by interpolation between the linked tracklets. Detections that do not belong to any tracklet or tracklets that are too short are considered as false alarms and removed from the final results.

A mathematical formulation of the tracking problem is given as follows. Suppose a set of tracklets $\mathcal{T} = \{T_1, \ldots, T_n\}$ is generated from a video sequence. A tracklet T_i is a consecutive sequence of detection responses or interpolated responses that contain the same target. The goal is to associate tracklets that correspond to the same target, given certain spatial–temporal constraints. Let association a_{ij} define the hypothesis that tracklets T_i and T_j contain the same target, assuming that T_i occurs before T_j. A valid association matrix A is defined as follows:

$$A = \{a_{ij}\}, \quad a_{ij} = \begin{cases} 1, & \text{if } T_i \text{ is associated to } T_j \\ 0, & \text{otherwise} \end{cases}$$
$$\text{s.t.} \sum_{i=1}^{n} a_{ij} = 1 \text{ and } \sum_{j=1}^{n} a_{ij} = 1. \tag{1}$$

The constraints for matrix A indicate that each tracklet should be associated with and associated by only one other tracklet (the initial and the terminating tracklets of a track are discussed in Section IV-A).

We define S_{ij} as the basic cost for linking tracklet T_i and T_j based on low-level information (time, appearance, and motion). It is computed as the negative log likelihood of T_i and T_j being the same target (explained in detail in Section IV-A). Note that $S_{ij} = \infty$ if T_i and T_j have overlap in time.

Let Ω be the set of all possible association matrices, and the multitarget tracking can be formulated as the following optimization problem:

$$A^* = \underset{A \in \Omega}{\operatorname{argmin}} \sum_{ij} a_{ij} S_{ij}. \tag{2}$$

This assignment problem can be optimally solved by the Hungarian algorithm in polynomial time. In order to reduce computational cost, the video is segmented by a predefined time sliding window, which is fixed to be 12-s long. Tracklet association is carried out in each time sliding window. There has to be a 50% overlap between two neighboring time windows. To handle association conflicts in the overlapping part of two windows, we use a method similar to [23]. More specifically, the overlapped part is evenly divided into two parts. In the first half, the tracking results produced by the previous time window are kept, while in the second half, the original input tracklets are used despite the association results from the previous time window.

As low-level information is not sufficient to distinguish targets under challenging situations, we consider integrating high-level information from social grouping behavior into the cost matrix to regularize the solution. However, group configuration is often not known *a priori*. Also, it is not fixed for the entire video, as people might change groups. Therefore, we propose elementary groups that are learned and updated online, during the tracking process to provide useful social grouping information while maintaining the flexibility of the group structure. Two tracklets T_i and T_j are likely to correspond to the same target if they satisfy the following constraints.

1) Each of them forms an elementary group with the same tracklet, namely, the same target.
2) The trajectory obtained by linking T_i and T_j has a small distance to the group mean trajectory.

The first constraint is based on the observation that if two people are walking together for a certain time, then there is high probability that they will still walk together after a short time period. The second constraint prevents us from linking a wrong pair of tracklets. Let P_{ij} be the inferred high-level information for T_i and T_j, and the tracklet association problem can be refined as

$$A^* = \underset{A \in \Omega}{\operatorname{argmin}} \sum_{ij} a_{ij} (S_{ij} - \alpha P_{ij}) \tag{3}$$

where α is a weighting parameter. It is selected by coarse binary search in only one time window and kept fixed for all the others.

In the following, we introduce an online method for group analysis and obtain P_{ij} by making inferences from the grouping graph.

B. Learning of Elementary Groups

In this section, we explain how the nodes (elementary groups) of the grouping graph are created. A set of tracklets is generated after low-level association, but only confident tracklets are considered for grouping analysis, as there might be false alarms, which may lead to incorrect associations in the input tracklets. Based on the observation that inaccurate tracklets are often the short ones, we define a tracklet as confident if it is long enough (e.g., it exists for at least ten frames).

Two tracklets T_i and T_j form an elementary group if they have the following properties: 1) T_i and T_j have overlap in time for more than l frames (l is set to 5 in our experiments); 2) they are spatially close to each other; and 3) they have similar velocities. Mathematically, we use G_{ij} to denote the probability of T_i and T_j forming an elementary group

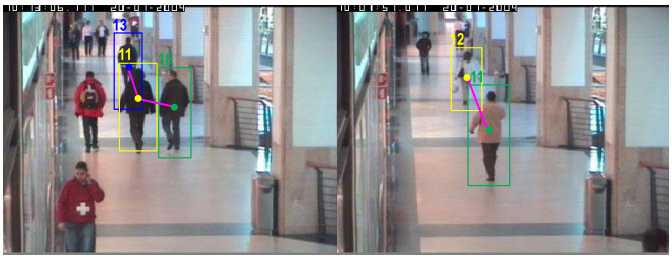
$$G_{ij} = P_t(T_i, T_j) \cdot P_d(T_i, T_j) \cdot P_v(T_i, T_j) \tag{4}$$

where $P_t(\cdot)$, $P_d(\cdot)$, and $P_v(\cdot)$ are the grouping probabilities based on overlap in time, distance, and velocity, respectively. Their definitions are given in

$$P_t(T_i, T_j) = \frac{L_{ij}}{L_{ij} + l} \tag{5}$$

$$P_d(T_i, T_j) = \frac{1}{L_{ij}} \sum_{n=1}^{L_{ij}} \left(1 - \frac{2}{\pi} \arctan(\text{dist}_n)\right) \tag{6}$$

$$P_v(T_i, T_j) = \frac{\cos\theta + 1}{2} \tag{7}$$



(a) (b)

Fig. 4. Two examples (a) and (b) generating incorrect elementary groups if the distances are not normalized.

where L_{ij} is the length of overlapped frames for T_i and T_j, dist_n is the normalized center distance for T_i and T_j on the nth overlapped frame, and θ is the angle between the average velocities of the two tracklets during the overlapped frames. In our experiments, dist_n is set as follows:

$$\text{dist}_n = \text{ratio}_n \cdot d/0.5(\text{width}_i + \text{width}_j) \tag{8}$$

where ratio_n is the size of the larger target over the size of the smaller target, d is the Euclidean distance between the two object centers, and $0.5(\text{width}_i + \text{width}_j)$ is the smallest distance in the image space for two people that walk side by side. The term ratio_n prevents tracklets as shown in Fig. 4 to be considered as a group, where the distance in the image space is small, while the distance in the 3D space is quite large.

We create a node for each pair of tracklets that have nonzero grouping probability G. Thus, each node contains two tracklets/targets and is associated with a probability G; its value indicates the similarity of motion patterns for these two tracklets during their coexistence period.

Note that if two tracklets form an elementary group, their group mean trajectory is obtained by computing the mean position using only their overlapping parts, as the grouping is meaningful only for the overlapped time period. For example, if T_a and T_b are in the same elementary group, this indicates that only T_a and T_b have similar motion patterns for the period that they have time overlap. During the nonoverlapping period, T_a may form elementary groups with other tracklets/targets that are even in a different group than the group of T_b. Such a property makes the elementary group flexible to handle group split and merge.

C. Group Tracking

The relationship between two elementary groups is identified by group tracking. Inspired by association-based multitarget tracking, we define our group tracking as a problem of finding globally optimal associations between elementary groups based on the three most commonly used features: time, appearance, and motion. More specifically, given a set of elementary groups, we compute the linking cost for any two groups and obtain the association results by finding the association set with the minimum total cost.

Let $\{T_1^{g_i}, T_2^{g_i}\}$ denote the two tracklets in an elementary group g_i. Given two elementary groups g_i and g_j, assuming that g_i starts before g_j, their linking cost is $C^g(g_i, g_j) = C_t^g(g_i, g_j) + C_{\text{appr}}^g(g_i, g_j) + C_{mt}^g(g_i, g_j)$, where $C_t^g(\cdot)$, $C_{\text{appr}}^g(\cdot)$, and $C_{mt}^g(\cdot)$ are the linking costs based on time, appearance, and motion, respectively. Similar to (2), let Φ be the set of all possible group association matrices, then the group tracking can be formulated as the following optimization problem:

$$A^{g*} = \underset{A^g \in \Phi}{\text{argmin}} \sum_{ij} a_{ij} C^g(g_i, g_j). \tag{9}$$

The Hungarian algorithm is used to solve this assignment problem.

1) Time Model for Group Tracking: For the linking cost based on time, we defined it as

$$C_t^g(g_i, g_j) = \begin{cases} 0, & g_i \text{ is not overlapped with } g_j \\ \infty, & \text{otherwise} \end{cases} \tag{10}$$

where the nonoverlapping constraint that means any tracklet in g_i has no time overlap with any tracklet in g_j.

If g_i and g_j contain the same two targets, there are only two matching possibilities.

1) $T_1^{g_i}$ and $T_1^{g_j}$ are the same target, and $T_2^{g_i}$ and $T_2^{g_j}$ are the same target.
2) $T_1^{g_i}$ and $T_2^{g_j}$ are the same target, and $T_2^{g_i}$ and $T_1^{g_j}$ are the same target.

We explain in detail matching option 1), note that the computation for matching option 2) is similar. For each matching option, we compute the linking cost based on appearance and motion, and use the one with the smaller sum for $C_{\text{appr}}^g(g_i, g_j) + C_{mt}^g(g_i, g_j)$. Also, the matching option is recorded for each group association.

2) Appearance Model for Group Tracking: Let $S(\cdot)$ be the appearance similarity for two tracklets, and the group linking cost based on appearance is defined as

$$C_{\text{appr}}^g(g_i, g_j) = -\ln\left(0.5\left(S\left(T_1^{g_i}, T_1^{g_j}\right) + S\left(T_2^{g_i}, T_2^{g_j}\right)\right)\right). \tag{11}$$

As there might be appearance variations in a single tracklet due to occlusion and lighting changes, it is hard to generate features that can robustly represent the appearance of a target. In order to more reliably compute the similarity between two tracklets, we adopt the modified Hausdorff metric [33], which is able to compute the similarity of two sets of images. Given a tracklet T_i that has length m_i, let $T_i = \{d_1^i, d_2^i, \ldots, d_{m_i}^i\}$, where d_x^i is the xth estimation of T_i, then $S(\cdot)$ is defined as

$$S(T_i, T_j) = \min\left(\frac{1}{m_i}\sum_{d_x^i \in T_i} s\left(d_x^i, T_j\right), \frac{1}{m_j}\sum_{d_y^j \in T_j} s\left(d_y^j, T_i\right)\right) \tag{12}$$

where $s(d, T) = \max_{d' \in T}(s_{\cos}(d, d'))$ is the Hausdorff similarity between an estimation and a tracklet. A modified cosine similarity measure [34] $s_{\cos}(\cdot)$ is used to compute the similarity between two estimations, which is defined as

$$s_{\cos}(u, v) = \frac{|u^T \cdot v|}{\|u\|\|v\|(\|u - v\|_p + \epsilon)} \tag{13}$$

where u and v are the feature descriptors from two images, $\|\cdot\|_p$ is the l_p norm (we set $p = 2$), and ϵ is a small positive

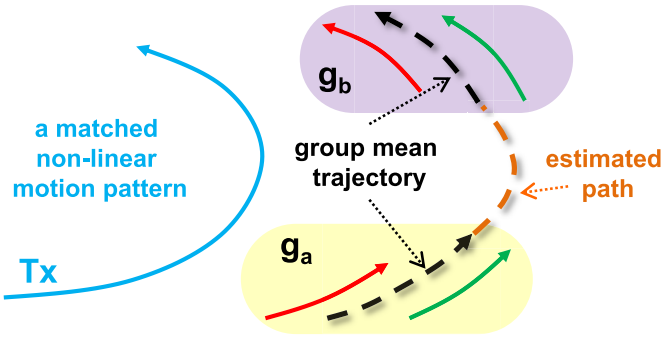


Fig. 5. Example of estimating motion affinity using the nonlinear motion map.

number to avoid dividing by zero. In our experiments, we use the concatenation of HSV color histogram and HOG features as the feature descriptors.

3) Motion Model for Group Tracking: We measure the motion affinity of two elementary groups by the motion smoothness between the group mean trajectories of the two corresponding elementary groups. The motion cost for linking two group mean trajectories is defined as the negative logarithm of the motion affinity

$$C_{mt}^{g}(g_i, g_j) = -\ln\left(G\left(f_{\text{predict}}(g_i, +\Delta t) - p_{\text{head}}^{g_j}, \Sigma_p\right) \cdot G\left(f_{\text{predict}}(g_j, -\Delta t) - p_{\text{tail}}^{g_i}, \Sigma_p\right)\right) \quad (14)$$

where $G(\cdot)$ is a zero-mean Gaussian distribution, Δt is the time gap between g_i and g_j, $f_{\text{predict}}(g_i, \pm\Delta t)$ gives the location prediction for the group mean trajectory of g_i after $(+)$ or before $(-)$ Δt, and p_{head} and p_{tail} are the head and tail locations for a group mean trajectory.

In most previous tracking frameworks [2], [29], [35], targets are commonly assumed to maintain a linear motion pattern. Thus, $f_{\text{predict}}(g_i, +\Delta t) = p_{\text{tail}}^{g_i} + v_{\text{tail}}^{g_i}\Delta t$ and $f_{\text{predict}}(g_j, -\Delta t) = p_{\text{head}}^{g_j} - v_{\text{head}}^{g_j}\Delta t$. However, in real-world scenarios, it is common to observe several nonlinear motion patterns in the scene. In order to produce more robust motion affinity for two elementary groups, we use the nonlinear motion map [5] to explain large nonlinear time gaps between group mean trajectories. Note that in [5], the nonlinear motion map is directly used to estimate the motion affinity of two tracklets, whereas we use it for explaining nonlinear gap between two elementary groups.

The nonlinear motion map M is a set of all existing nonlinear tracklets in current time sliding window, and the tracklets are selected only from the confident ones. An example of estimating motion affinity between g_a and g_b using a nonlinear motion pattern T_x in the motion map is illustrated in Fig. 5. The tracklet $T_x \in M$ is a nonlinear motion pattern that has coexisted in time with both g_a and g_b, and is a *matched tracklet* for the group mean trajectories of g_a and g_b. T_x is a matched tracklet that indicates that it is spatially close to the elementary group and has a similar motion direction as the elementary group. Then a quadratic curve that best fits positions at the tail part of g_a and the head part of g_b is estimated to fill the path between g_a and g_b. Therefore, each group association has a specific quadratic function for its nonlinear motion estimation. The estimated path is valid only if T_x is a matched tracklet for it. The motion cost for linking g_a and g_b based on nonlinear prediction of locations can be computed according to (14).

For each pair of elementary groups, both linear and nonlinear motion models are used, and the score with a lower cost is selected. Note that when only a linear motion model is used, any trajectory not following the pattern is penalized. With the nonlinear motion model, we are able to explain nonlinear motion in the scene without producing extra penalties for individuals who do not follow a linear motion pattern.

D. Creation of Virtual Nodes

Our goal is to encode grouping structure of the tracklets by the elementary grouping graph. With elementary groups as nodes of the graph, we define an edge between two nodes indicating the existence of at least one common target in the corresponding two elementary groups. For simple cases where two nodes have one tracklet in common, we link these two nodes directly, such as nodes g_1 and g_2 and g_4 and g_5 shown in Fig. 3. For difficult cases where there are four different tracklets in two nodes, we use the results of group tracking to find their relationship.

Suppose that g_i and g_j are associated by group tracking, namely, these two elementary groups contain the same two targets. We create two virtual nodes v_p and v_q, set their grouping probability G to be the same as that of node g_j, and build edges between g_i and the virtual nodes. Note that the virtual nodes can also be added in the other way (i.e., set G to be the same as g_i and link the virtual nodes to g_j), but these two options are exclusive to each other. Each virtual node also contains two tracklets: 1) a virtual tracklet generated by linking a pair of matched tracklets in g_i and g_j and 2) the tracklet left in g_j. An example of virtual node creation is presented in Fig. 3. Based on the association of g_2 and g_3, two virtual nodes v_1 and v_2 are created and connected to g_2. Two virtual nodes are used since there are two pairs of tracklets that need inference (edge for g_2 and v_1 indicates inference for T_2 and T_8; edge for g_2 and v_2 indicates inference for T_3 and T_7). In the following, we show that using the virtual node inference can be easily done.

E. Inference From Grouping Graph

In the grouping graph, each node is an elementary group and each edge indicates that the two connected elementary groups have one target in common. According to the observation that two people walking together at certain time are likely to walk together after a short period, given two directly connected groups, we can infer the probability of the uncertain target in each group to be the same.

Suppose that there is an edge between nodes g_i and g_j in the grouping graph; assuming $T_1^i = T_1^j = T_k$, $T_2^i = T_l$, and $T_2^j = T_m$ without loss of generality, the probability of T_2^i and T_2^j containing the same target is defined as follows:

$$p_{lm} = 0.5(G_{kl} + G_{km}) \times TSimi(T_{\{l,m\}}, G_{\{k,l,m\}}) \quad (15)$$

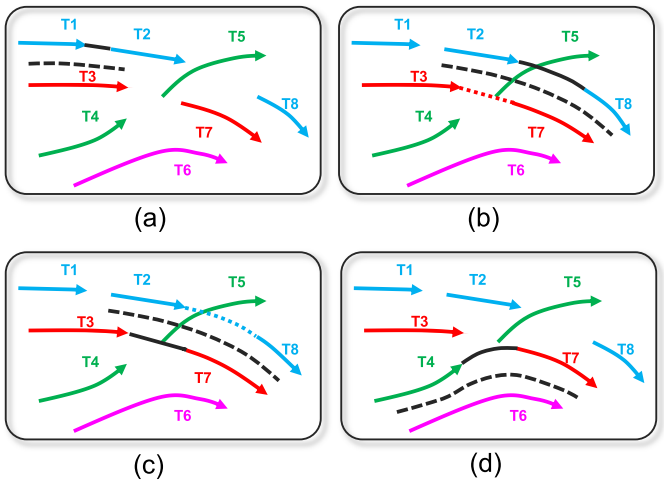


Fig. 6. Inference for each edge in the grouping graph in Fig. 3. (a) Edge between g_1 and g_2. (b) Edge between g_2 and v_1. (c) Edge between g_2 and v_2. (d) Edge between g_4 and g_5 (see Fig. 3 for group annotations). Black solid line: interpolation between the two tracklets that need inference. Black dashed line: group mean trajectory. Colored dotted line: virtual tracklet. Best viewed in color.

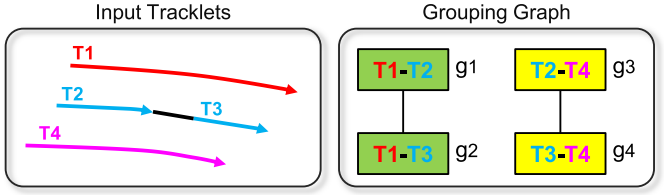


Fig. 7. Example of multiple inferences related to the same two tracklets. According to the proposed elementary grouping model, a grouping graph (shown on the right) is created based on the input tracklets (shown on the left). Thus, inferences based on the edge between nodes g_1 and g_2 and the edge between nodes g_3 and g_4 are all related to tracklets T_2 and T_3.

where $TSimi(T_{\{l,m\}}, G_{\{k,l,m\}})$ is the trajectory similarity between trajectory $T_{\{l,m\}}$ (created by linking T_l and T_m) and the group mean trajectory $G_{\{k,l,m\}}$ (created by computing the mean position of T_k and $T_{\{l,m\}}$). We define the trajectory similarity as follows:

$$TSimi(T, G) = 1 - \frac{2}{\pi}\arctan(\text{Dist}) \tag{16}$$

where Dist is the average Euclidean distance of trajectory T and group mean trajectory G.

For edges connecting two normal nodes and edges connecting to one virtual node, the same inference function can be used. The only difference is that the latter uses one virtual tracklet and two normal tracklets as input. Examples of making inference for a grouping graph are shown in Fig. 6. Note that there might be multiple inferences related to the same two tracklets, as the same tracklet may be contained in multiple elementary groups, as shown in Fig. 7. Therefore, P_{ij} in (3) is the sum of all inferences that relate to T_i and T_j

$$P_{ij} = \sum p_{ij}. \tag{17}$$

A summary of the proposed elementary grouping model is shown in Algorithm 1.

Algorithm 1 Learning Algorithm for Elementary Grouping Model

Input: Tracklet set $T = \{T_1, .., T_n\}$
Output: Inference matrix P, where P_{ij} is the inference for T_i and T_j

$P \leftarrow$ *empty set*, $Nodes \leftarrow \emptyset$, $Edges \leftarrow \emptyset$
for $i = 1, \ldots, n$ **do**
  for $j = i + 1, \ldots, n$ **do**
    if T_i and T_j are confident tracklets **then**
      $G_{ij} = P_t(T_i, T_j)P_d(T_i, T_j)P_v(T_i, T_j)$
      if $G_{ij} > 0$ **then**
        Create node $g = \{T_i, T_j\}$
        $Nodes = Nodes \cup \{g\}$
for $i = 1, \ldots, size(Nodes)$ **do**
  for $j = i + 1, \ldots, size(Nodes)$ **do**
    if $\exists T \in g_i$, $T = T_1^{g_j}$ or $T = T_2^{g_j}$ **then**
      Create an edge $e_{\{g_i,g_j\}}$ for g_i and g_j
      $Edges = Edges \cup \{e_{\{g_i,g_j\}}\}$
Update *Nodes* and *Edges* according to group tracking
for all $e \in Edges$ **do**
  Compute p_{xy} for the corresponding tracklet pair using Eq. (15)
  Update P: $P_{xy} = P_{xy} + p_{xy}$

IV. Experiments

We evaluate our approach on four data sets: the CAVIAR data set [8], the TownCentre data set [35], the PETS2009 data set [36], and the UNIV data set [37]. The popular evaluation metrics defined in [38] and the CLEAR MOT metrics defined in [39] are used for performance comparison.

1) *GT:* the number of trajectories in the GT.
2) *MT:* the ratio of mostly tracked trajectories, which are successfully tracked for more than 80% of the time.
3) *ML:* the ratio of mostly lost trajectories, which are successfully tracked for less than 20% of the time.
4) *Frag Fragments:* the number of times that a GT trajectory is interrupted.
5) *IDS—ID Switches:* the number of times that a tracked trajectory changes its matched id.
6) *FP—False Positive:* the number of tracker hypotheses for which no real object exists.
7) *FN—False Negative:* the number of times that targets have no matched hypothesis.
8) *MOTA—Multiple Object Tracking Accuracy:* a combined measure that takes into account false positives, false negatives, and identity switches.
9) *MOTP—Multiple Object Tracking Precision:* measures the alignment of the tracking results with respect to GT.

The following tracking approaches are tested.

1) *Our Model (Nonlinear):* the proposed elementary grouping model with nonlinear motion context for group tracking.
2) *Our Model (Linear):* the proposed elementary grouping model with only linear motion model for group tracking.
3) *Baseline Model 1:* the basic affinity model.

TABLE I
COMPARISON OF TRACKING RESULTS ON CAVIAR DATA SET. NUMBER OF TRAJECTORIES IN GT IS 75

Method	MT	ML	Frag	IDS	FP	FN	MOTA	MOTP	Time
Baseline Model 1	74.7%	6.7%	11	12	1459	10827	79.2%	78.8%	1.5s
Baseline Model 2	78.7%	6.7%	10	8	1535	9134	82.0%	81.7%	4.2s
SGB Model [29]	89.3%	2.7%	7	5	1597	8497	83.0%	82.1%	50s
Our Model (linear)	90.7%	2.7%	6	5	1668	8081	83.5%	82.0%	4.6s
Our Model (non-linear)	90.7%	2.7%	6	5	1668	8081	83.5%	82.0%	6.1s

4) *Baseline Model 2:* the proposed elementary grouping model without group tracking.
5) *SGB:* the social grouping behavior model [29].

For a fair comparison, the same input tracklet set, GT, and basic affinity model are used for all the methods. All the results for the SGB model are kindly provided in [29]. Both quantitative comparisons with the state-of-the-art methods and the visual results of our approach are presented.

A. Implementation Details

1) Tracklets' Generation: Two different ways of generating tracklets are employed to demonstrate that the proposed grouping model can be easily integrated into any DAT-based tracking system, regardless of the method used to extract the initial tracklets. In the first method, targets on each frame are detected using the discriminatively trained deformable part-based models [40]. We apply a nearest neighbor detection association method similar to [7] to generate the initial tracklets. For each unassociated detection, a Kalman filter-based tracker is initialized with position and velocity states. A detection A is associated with a detection B in the next frame if B has the minimum distance to the predicted location and overlaps at least 50% [measured as $size(A \cap B)/size(A \cup B)$] in size with detection A. Then the corresponding Kalman filter is updated with the newly associated detection. The tracker terminates if no association is found for more than two consecutive frames, or a detection is associated by multiple trackers.

In the second method, the popular HOG-based human detector [41] is used. Tracklets are generated by connecting detections in consecutive frames that have high similarity in appearance and have large overlap in size. A simple two-threshold strategy [13] is used to generate reliable tracklets. In our experiments, two detections are connected if and only if: 1) their affinity is higher than 90% and 2) their affinity is at least 20% larger than the affinities of any other alternatives.

2) Basic Affinity Model: In order to produce reasonable basic affinity for a pair of tracklets, three commonly used features are adopted: time, appearance, and motion. The basic affinity P_{basic} for two tracklets T_i and T_j is defined as

$$P_{\text{basic}}(T_i, T_j) = f_t(T_i, T_j) \cdot f_{\text{appr}}(T_i, T_j) \cdot f_{mt}(T_i, T_j). \quad (18)$$

The time affinity model f_t assigns zero affinity to tracklet pairs whose time gap is greater than a predefined threshold GAP, and it is defined as

$$f_t(T_i, T_j) = \begin{cases} 0, & \text{if } \text{Gap}_{ij} > \text{GAP} \\ 1, & \text{otherwise.} \end{cases} \quad (19)$$

The appearance affinity model f_{appr} is based on the Bhattacharyya coefficient of two average HSV color histograms. For the motion affinity model f_{mt}, the same method as shown in (14) with linear motion for f_{predict} is used to measure the motion smoothness of two tracklets in both forward and backward directions. Given $P_{\text{basic}}(T_i, T_j)$, the basic cost S_{ij} in (2) is computed as $S_{ij} = -\ln(P_{\text{basic}}(T_i, T_j))$.

3) Cost Matrix S: Due to the constraints in (1), the traditional pairwise assignment algorithm is not able to find the initial and the terminating tracklets. Therefore, instead of using the cost matrix S $(n \times n)$ directly, we use the augmented matrix $(2n \times 2n)$ proposed in [29] as the input for the Hungarian algorithm. This enables us to set a threshold for association, and a pair of tracklets can only be associated when their cost is lower than the threshold. In our experiments, the threshold is set to $-\ln 0.5$ for all data sets.

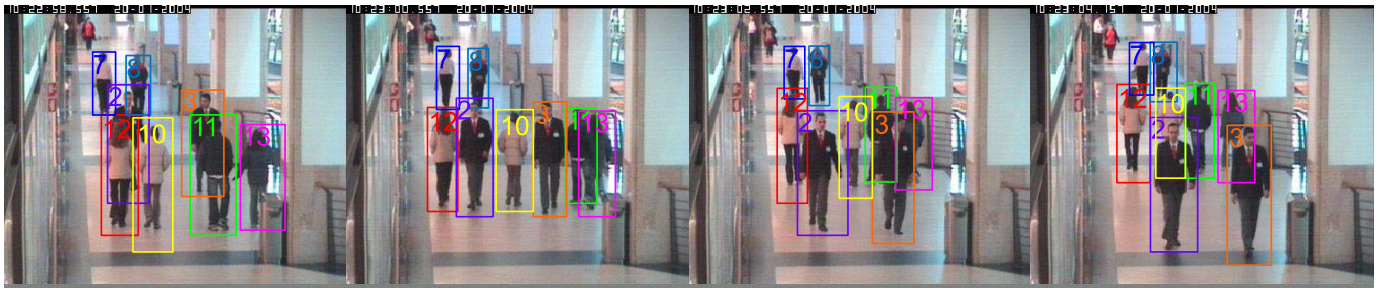
B. Results on CAVIAR Data Set

The videos in the CAVIAR data set are acquired in a shopping center where frequent interactions and occlusions occur and people are more likely to walk in groups. We select the same set of test videos as in [29], which are the relatively challenging ones in the data set. We generate input tracklets using the first method described in Section IV-A. The comparative results are shown in Table I. Our proposed models (both linear and nonlinear) achieve the best overall tracking accuracy (MOTA) with the high tracking precision (MOTP) compared with the other alternatives. It is observed that the basic affinity model (baseline model 1) can produce reasonable tracking results, and the performance is further improved by integrating high-level grouping information [baseline model 2, our model (linear), and our model (nonlinear)]. Both linear and nonlinear versions of our model have comparable or better performances in most metrics compared with the SGB model (e.g., better results in MT and Frag, and the same results in ML and IDS), but with much less computational time. The comparisons between baseline model 2 and our model (both linear and nonlinear) demonstrate the importance of group tracking, as they reveal more grouping information. Since most pedestrians in the videos are walking linearly along a corridor in this data set, there is barely any nonlinear context in the scene. Therefore, the linear and nonlinear versions of our model have the same performance (except computational time) on this data set. The sample tracking results are shown in Fig. 8.

C. Results on TownCentre Data Set

The TownCentre data set has one high-resolution video that captures the scene of a busy street. There are 220 people in

(a)



(b)

(c)

Fig. 8. Examples of tracking results of our approach on CAVIAR data set. The same color indicates the same target. Best viewed in color. (a) Track targets (4, 5, 6) when appearances vary a lot due to occlusions. (b) Successfully tracking targets (11, 13) with long time gap. (c) Track targets (4, 5, 6) when sudden motion change and occlusion happen.

TABLE II

COMPARISON OF TRACKING RESULTS ON TOWNCENTRE DATA SET. NUMBER OF TRAJECTORIES IN GT IS 220

Method	MT	ML	Frag	IDS	FP	FN	MOTA	MOTP	Time
Baseline Model 1	76.8%	7.7%	37	60	2746	28493	56.1%	68.8%	350s
Baseline Model 2	78.6%	6.8%	34	46	3155	22236	64.3%	71.3%	457s
SGB Model [29]	83.2%	5.9%	28	39	4387	15871	81.8%	69.7%	4861s
Our Model (linear)	85.5%	5.9%	26	36	4105	14804	73.4%	69.2%	465s
Our Model (non-linear)	86.4%	5.9%	25	36	4938	13910	73.5%	69.2%	505s

total, with an average of 16 people visible per frame. We test all models using the first 3 min of the video and generate input tracklets using the second method described in Section IV-A. The comparative results are shown in Table II. Similar to the observations from Table I, Table II suggests that the performance of our method is consistent on both data sets. As there are some nonlinear motions in this data set, the tracking performance is slightly improved by the incorporation of nonlinear context. The sample tracking results are shown in Fig. 9.

D. Results on PETS2009 Data Set

We select sequence *S2L2* in the PETS2009 data set to evaluate the performance of the proposed method. This sequence

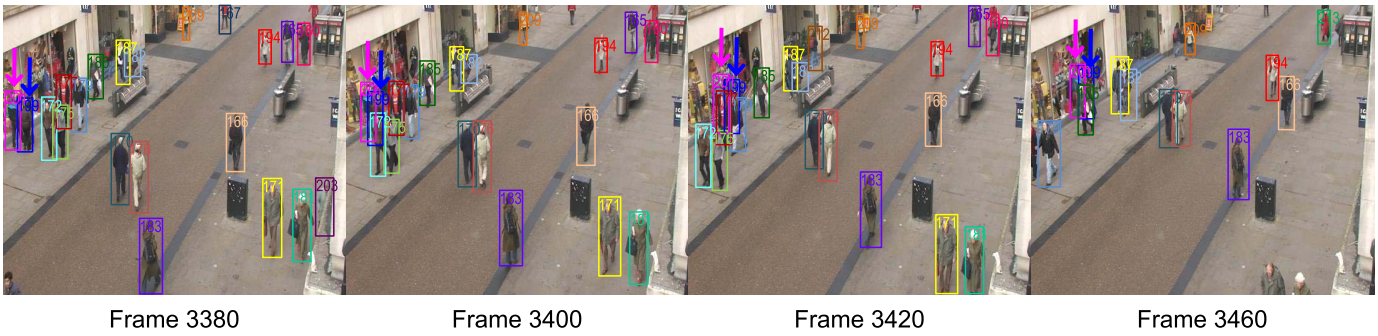


Fig. 9. Examples of tracking results of our approach on TownCentre data set. With grouping information, targets (199 and 201) pointed by arrows are correctly tracked under frequent occlusions. The same color indicates the same target. Best viewed in color.

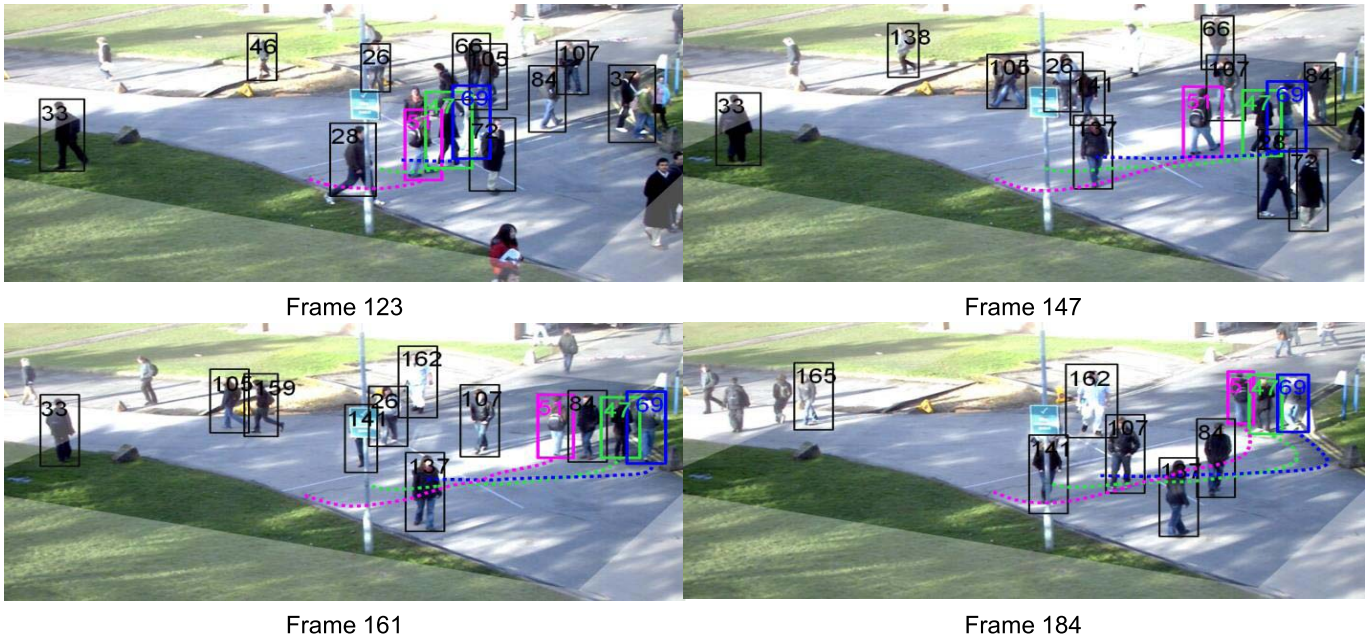


Fig. 10. Examples of tracking results of our approach on PETS2009 data set. Track targets (47, 51, 69) with nonlinear motion successfully. Best viewed in color.

TABLE III

COMPARISON OF TRACKING RESULTS ON PETS2009 DATA SET. NUMBER OF TRAJECTORIES IN GT IS 74

Method	MT	ML	Frag	IDS	FP	FN	MOTA	MOTP	Time
Baseline Model 1	14.9%	64.9%	120	88	271	5414	32.4%	60.5%	297s
Baseline Model 2	21.6%	50%	104	102	436	4773	37.8%	59.7%	381s
SGB Model [29]	23%	41.9%	95	91	691	3828	46.0%	59.9%	4962s
Our Model (linear)	28.4%	44.6%	93	97	683	3987	44.1%	58.8%	477s
Our Model (non-linear)	33.8%	35.1%	79	89	729	3081	54.3%	60.1%	612s

TABLE IV

COMPARISON OF TRACKING RESULTS ON UNIV DATA SET. NUMBER OF TRAJECTORIES IN GT IS 40

Method	MT	ML	Frag	IDS	FP	FN	MOTA	MOTP	Time
SGB Model [29]	75%	5%	38	7	213	443	96.7%	82.9%	47s
Our Model (linear)	87.5%	5%	26	5	224	287	97.4%	83.1%	3.9s
Our Model (non-linear)	87.5%	5%	26	5	224	287	97.4%	83.1%	4.2s

captures the outdoor scene of a campus from an elevated viewpoint. Unlike the widely used sequence *S2L1*, sequence *S2L2* is more challenging as it has higher crowd density (up to 33 targets per frame) and includes many nonlinear motion patterns. A rectangular area is defined in the world coordinates and used as the boundary of the tracking area

Fig. 11. Examples of tracking results of our approach on UNIV data set. Using only the grouping model, we correctly tracked targets (1, 2, 3, 4) in situations where the group split and merge occurs. The same color indicates the same target. Best viewed in color.

(as shown in Fig. 10), and trajectories outside the area are excluded from our solution. The first method described in Section IV-A is used to generate input tracklets. The comparative results are shown in Table III. We can see that when many nonlinear walking patterns are present in the data set, significant improvements are achieved by integrating nonlinear motion context into the tracking system. Our model with nonlinear context gives the best MOTA and has a higher MT (33.8%) and a lower ML (35.1%) compared with the SGB method (MT: 23% and ML: 41.9%) and our model (linear) (MT: 28.4% and ML: 44.6%) that consider only linear motion during grouping. Also the numbers of fragments and ID switches are greatly reduced when social grouping and nonlinear context are employed in the tracking system. The sample tracking results of the proposed method with nonlinear motion context are shown in Fig. 10. In the first row of Fig. 12, we present tracking examples of our method with linear motion model on the same sample sequence as shown in Fig. 10.

E. Results on UNIV Data Set

To further evaluate the effectiveness of the proposed method in handling dynamics of social groups (e.g., group merge and split), four video sequences are collected from an elevated viewpoint that allows the capture of rich group evolving scenarios. Each video is about 30-s long with an average of nine pedestrians visible in each frame, and some sample frames are shown in Fig. 11. The input tracklets for this data set are produced using the second method described in Section IV-A. Multitarget tracking is carried out using only the grouping information, namely, the linking costs for tracklet pairs are based only on P_{ij} in (3). The comparative results are shown in Table IV. Our models with both linear and nonlinear motion have the same performance, as this data set contains little nonlinear motion pattern. Compared with the SGB model that assumes a fixed number of groups in the scene, our grouping model improves MT by 12.5%, reduces the fragments by 31.5%, and also achieves higher MOTA and MOTP. The results imply that our grouping model is better at handling group dynamics in the scene, as it focuses on analyzing elementary groups instead of the complete groups. The sample tracking results of the proposed method are shown in Fig. 11. In the second row of Fig. 12, we show tracking examples of SGB model on the same sample sequence as shown in Fig. 11.

F. Computational Time

The computational time is greatly affected by the number of targets in a video and the length of the video. All methods are implemented in MATLAB without code optimization or parallelization and tested on a PC with 3.0-GHz CPU and 8-GB memory. The average computational times for all the data sets are shown in the last columns in Tables I–IV. Note that the computational times for object detection, tracklet generation, and appearance and motion feature extraction are not included in the above estimates of computational time. It is clear that our models (both linear and nonlinear) improve the computational efficiency by an order of magnitude compared with the SGB model that also uses social grouping information in tracking. For the relatively short videos (30–66 s) in CAVIAR and UNIV data sets, our approach takes 292 frames/s for the linear version and 235 frames/s for the nonlinear version on average. For the video in TownCentre (3 min), the computational time is 10 frames/s for the linear version and 9 frames/s for the nonlinear version. When our approach is applied on the high crowd density video in PETS2009, the computational time is 0.9 frames/s for the linear version and 0.7 frames/s for the nonlinear version. It is observed that integrating nonlinear context into the motion model increases the computational cost, but still our model is significantly more efficient than the SGB model and produces better tracking results.

From a theoretical perspective, the optimization of SGB is a gradient-based iterative method. To compute the gradient, an alternative approach involving the Hungarian algorithm and K-means clustering is applied. K-means clustering needs multiple initial starts to reach a reasonable local optimum, which leads to high computational cost. Our solver, on the

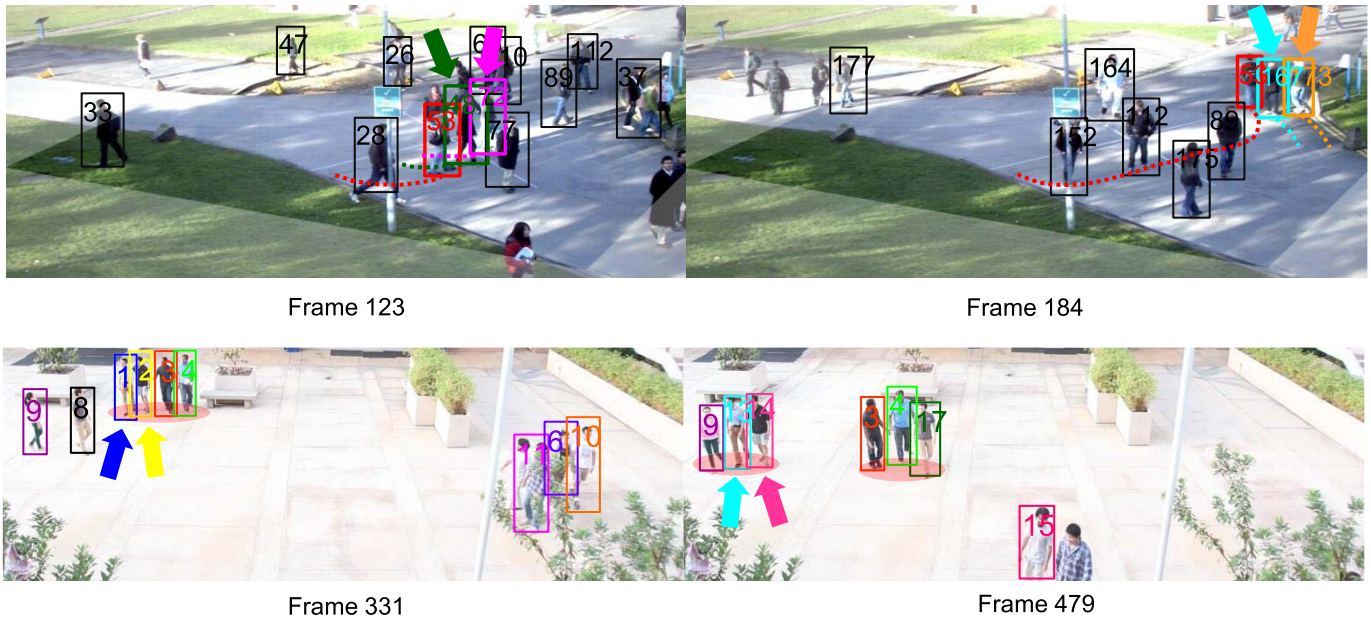


Fig. 12. Examples of tracking results from referenced models. First row: our model (linear) on PETS2009 data set. Targets (48, 72) cannot be correctly tracked, as tracklet associations generating nonlinear motion pattern are penalized when only linear motion model is used. Second row: SGB model on UNIV data set. Trajectories of targets (1, 2) cannot be fully recovered, because SGB model is not able to link tracklets that are not assigned to the same group. Best viewed in color.

other hand, has a closed-form solution based only on the deterministic Hungarian algorithm, and thus can be computed much more efficiently.

V. Conclusion

In this paper, we have presented an online approach that integrates high-level grouping information into the basic affinity model for multitarget tracking. The grouping behavior is modeled by a novel elementary grouping graph, which not only encodes the grouping structure of tracklets but is also flexible to cope with the evolution of a group (i.e., group split and merge). We have used nonlinear motion context explicitly for discovering relationships between elementary groups. The experimental results on four challenging data sets demonstrated the superior tracking performance by integrating elementary grouping information. Compared with the state-of-the-art social grouping model, our approach provides better performance in a more computationally efficient manner. However, if there is not much grouping or all the targets follow a linear motion pattern in the input video, the integration of the elementary grouping model will have limited improvement on the tracking performance. Possible future work would be extending the elementary grouping model to multiperson tracking in multiple cameras.

References

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

[2] 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. (CVPR)*, Jun. 2009, pp. 1200–1207.

[3] J. F. Henriques, R. Caseiro, and J. Batista, "Globally optimal solution to multi-object tracking with merged measurements," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2470–2477.

[4] Z. Qin, C. R. Shelton, and L. Chai, "Social grouping for target handover in multi-view video," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

[5] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1918–1925.

[6] 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. (CVPR)*, Jun. 2011, pp. 1201–1208.

[7] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2006, pp. 666–673.

[8] *Caviar Dataset*. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[9] 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, Apr. 2010.

[10] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, pp. 4282–4286, May 1995.

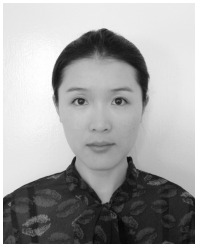
[11] N. Ghosh and B. Bhanu, "Evolving Bayesian graph for three-dimensional vehicle model building from video," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 563–578, Apr. 2014.

[12] 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. 12th IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 1515–1522.

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

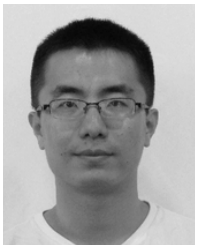
[14] Y. Gao, R. Ji, L. Zhang, and A. Hauptmann, "Symbiotic tracker ensemble toward a unified tracking framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 7, pp. 1122–1131, Jul. 2014.

[15] Z. Jin and B. Bhanu, "Analysis-by-synthesis: Pedestrian tracking with crowd simulation models in a multi-camera video network," *Comput. Vis. Image Understand.*, vol. 134, pp. 48–63, May 2015.
[16] W. Luo, J. Xing, X. Zhang, X. Zhao, and T.-K. Kim, "Multiple object tracking: A review," *CoRR*, vol. abs/1409.7618, 2014.
[17] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua, "Tracking multiple people under global appearance constraints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 137–144.
[18] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1815–1821.
[19] J. Berclaz, F. Fleuret, E. Türetken, 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.
[20] Z. Wu, T. H. Kunz, and M. Betke, "Efficient track linking methods for track graphs using network-flow and set-cover techniques," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1185–1192.
[21] H. Jiang, S. Fels, and J. J. Little, "A linear programming approach for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
[22] B. Leibe, K. Schindler, and L. Van Gool, "Coupled detection and trajectory estimation for multi-object tracking," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[23] B. Yang and R. Nevatia, "Multi-target tracking by online learning a CRF model of appearance and motion patterns," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 203–217, 2014.
[24] A. A. Butt and R. T. Collins, "Multi-target tracking by Lagrangian relaxation to min-cost network flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1846–1853.
[25] 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 12th Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 261–268.
[26] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1273–1280.
[27] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2211–2218.
[28] S. Pellegrini, A. Ess, and L. Van Gool, "Improving data association by joint modeling of pedestrian trajectories and groupings," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 452–465.
[29] Z. Qin and C. R. Shelton, "Improving multi-target tracking via social grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1972–1978.
[30] 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.
[31] X. Yan, A. Cheriyadat, and S. K. Shah, "Hierarchical group structures in multi-person tracking," in *Proc. IEEE 22nd Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 2221–2226.
[32] X. Chen, Z. Qin, L. An, and B. Bhanu, "An online learned elementary grouping model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1242–1249.
[33] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. 12th IAPR Int. Conf. Pattern Recognit. (ICPR)*, vol. 1. Oct. 1994, pp. 566–568.
[34] C. Liu, "Discriminant analysis and similarity measure," *Pattern Recognit.*, vol. 47, no. 1, pp. 359–367, 2014.
[35] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3457–3464.
[36] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *Proc. 12th IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS-Winter)*, Dec. 2009, pp. 1–6.
[37] L. Feng and B. Bhanu, "Understanding dynamic social grouping behaviors of pedestrians," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 317–329, Mar. 2015.
[38] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2953–2960.
[39] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *J. Image Video Process.*, vol. 2008, pp. 246309-1–246309-10, May 2008.
[40] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 886–893.



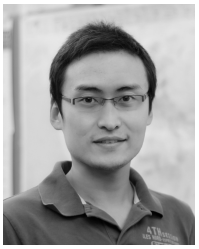
Xiaojing Chen received the B.S. degree in information management and information systems from Beijing Language and Culture University, Beijing, China, in 2007, the M.S. (Hons.) degree in computer science from Leiden University, Leiden, The Netherlands, in 2009, and the Ph.D. degree from the University of California, Riverside, CA, USA, in 2015.

Her recent research has been concerned with multitarget tracking in surveillance cameras. Her current research interests include computer vision, pattern recognition, and machine learning.



Zhen Qin received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the Ph.D. degree from the University of California at Riverside, Riverside, CA, USA, in 2015.

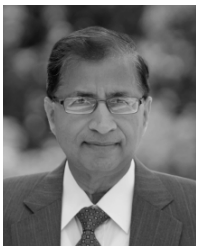
His current research interests include computer vision and machine learning.



Le An received the B.Eng. degree in telecommunications engineering from Zhejiang University, Hangzhou, China, in 2006, the M.Sc. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2008, and the Ph.D. degree in electrical engineering from the University of California at Riverside, Riverside, CA, USA, in 2014.

His current research interests include image processing, computer vision, pattern recognition, and machine learning.

Dr. An received the Best Paper Award from the 2013 IEEE International Conference on Advanced Video and Signal-Based Surveillance.



Bir Bhanu (F'95) received the B.S. (Hons.) degree in electronics engineering from IIT Varanasi, Varanasi, India, the M.E. (Hons.) degree in electronics engineering from the Birla Institute of Technology and Science, Pilani, India, the S.M. and E.E. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, and the M.B.A. degree from the University of California at Irvine, Irvine, CA, USA.

He was a Senior Honeywell Fellow with Honeywell Inc., Minneapolis, MN, USA. He was the Founding Professor of Electrical Engineering with the University of California at Riverside (UCR), Riverside, CA, USA, where he served as the First Chair from 1991 to 1994. He has been the Cooperative Professor of Computer Science and Engineering since 1991, Bioengineering since 2006, and Mechanical Engineering since 2008, and the Director of the Visualization and Intelligent Systems Laboratory since 1991. He has been a Faculty Member with the Computer Science Department, University of Utah, Salt Lake City, UT, USA. He has been with Ford Aerospace and Communications Corporation, Newport Beach, CA, USA; the French Institute for Research in Computer Science and Automation, Paris, France; and the IBM San Jose Research Laboratory, San Jose, CA, USA. He has been the Principal Investigator of various programs for the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), NASA, the Air Force Office of Scientific Research, the Office of Naval Research, the Army Research Office, and other agencies and industries in the areas of video networks, video understanding, video bioinformatics, learning and vision, image understanding, pattern recognition, target recognition, biometrics, autonomous navigation, image databases, and machine-vision applications. He is currently the Distinguished Professor of Electrical Engineering and serves as the Founding Director of the Interdisciplinary Center for Research in Intelligent Systems with UCR. In addition, he serves as the Director of the NSF Interdisciplinary Graduate Education Research and Training Program on Video Bioinformatics and the Interim Chair of the Department of Bioengineering with UCR. He has co-authored seven books and edited four books. He has authored 500 reviewed technical publications, including over 140 journal papers, 44 book chapters, and 15 patents.

Dr. Bhanu is a fellow of the American Association for the Advancement of Science, the International Association of Pattern Recognition, the American Institute for Medical and Biological Engineering, and the International Society for Optics and Photonics. He has been a recipient of many best conference paper and outstanding journal paper awards, industrial and university awards for research excellence, outstanding contributions, and team efforts, and the Doctoral/Dissertation Advisor/Mentor Award. He has been on the Editorial Board of many journals and has edited, as the Lead Editor, the special issues of over 15 journals, including some of the top IEEE publications. He was the General Chair of the IEEE Conference on Computer Vision and Pattern Recognition, the Conference on Advanced Video and Signal Based Surveillance, the International Conference on Distributed Smart Cameras, the IEEE Winter Conference on Applications of Computer Vision, and the DARPA Image Understanding Workshop. Recently, he served on the IEEE Fellow Committee from 2010 to 2012.