An Online Learned Elementary Grouping Model for Multi-target Tracking

Xiaojing Chen, Zhen Qin, Le An, Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside
{xchen010, zqin001}@cs.ucr.edu, lan004@ucr.edu, bhanu@cris.ucr.edu

Abstract

We introduce an online approach to learn possible elementary groups (groups that contain only two targets) for inferring high level context that can be used to improve multi-target tracking in a data-association based framework. Unlike most existing association-based tracking approaches that use only low level information (e.g., time, appearance, and motion) to build the affinity model and consider each target as an independent agent, we online learn social grouping behavior to provide additional information for producing more robust tracklets affinities. Social grouping behavior of pairwise targets is first learned from confident tracklets and encoded in a disjoint grouping graph. The grouping graph is further completed with the help of group tracking. The proposed method is efficient, handles group merge and split, and can be easily integrated into any basic affinity model. We evaluate our approach on two public datasets, and show significant improvements compared with state-of-the-art methods.

1. Introduction

Multi-target tracking in real scenes has been an active research topic in computer vision for many years, due to its promising potential in industrial applications, such as visual surveillance, human-computer interaction, and anomaly detection. The goal of multi-target tracking is to recover trajectories of all targets while maintaining identity labels consistent. There are many challenges for this problem, such as illumination and appearance variation, occlusion, and sudden change in motion [25][27]. As great improvement has been achieved in object detection, data association-based tracking (DAT) has become popular recently [12][22][30]. An affinity model integrating multiple visual cues (appearance and motion information) is formulated to find the linking probability between detection responses or tracklets (trajectory fragments), and the global optimal solution is often obtained by solving the maximum a posteriori problem (MAP) using various optimizaion algorithms.

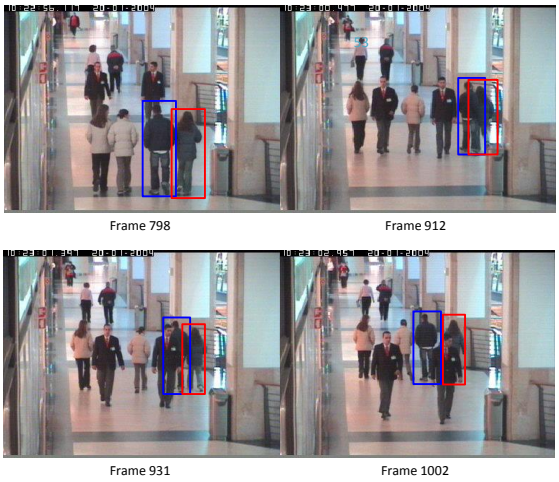

Figure 1. Examples of challenging conditions for tracking. The same color indicates the same target. Note that for both targets with bounding box there are significant appearance and motion changes due to occlusions and cluttered background.

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 Figure 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 difficult tracklet pairs. In this paper, we explore high level contextual information, social grouping behavior, for associating tracklets that are very challenging for lower level features (time, appearance, and motion).

When there are few interactions and occlusions among targets, DAT gives robust performance. Discriminative descriptors of targets are usually generated using appearance and motion information from tracklets. Appearance model often uses global color histograms to match tracklets, and a linear motion model based on velocity and distance is often

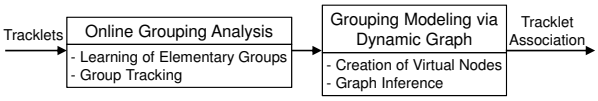

Figure 2. Overview of the elementary grouping model.

assumed to constrain motion smoothness of two tracklets. However, these low level descriptors are likely to fail for tracklets with long time gap. 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.

On the other hand, there is often other useful high level contextual information in the scene which can be used to mitigate such confusions. 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 likely to have similar motion pattern and be spatially close to each other for better group interaction [17]. It is also shown 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 obseravations, we propose an elementary grouping model to construct a grouping graph where each node represents a pair of tracklets that form an elementary group (a group of two targets) and each edge indicates the connected two nodes (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 subject. The elementary grouping model is summarized in Figure 2. The size of a group may change dynamically as people join and leave the group, but a group of any size can be considered as a set of elementary groups. Therefore, focusing on finding elementary groups instead of the complete group makes our approach capable of modelling flexible group evolvement 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 contributions of this paper are:

- We propose an approach that estimates elementary groups online and infers grouping information to adjust the affinity model for association-based tracking. This approach is independent of the detection methods, affinity models, and optimization algorithms.
- Our approach of elementary grouping is simple and computationally efficient, while remaining effective and robust.

2. Related Work

Traditional approaches of multi-target tracking usually use filtering algorithms to enable time-critical applications, where video is processed on frame-by-frame bases [3][31]. Recently, DAT became the major researched area. With the help of the state-of-the-art tracklet extraction methods such as those based on human detectors [13], the focus is shifted to robust tracklet association schemes [27].

To achieve robust association, reliable affinity scores between tracklets are essential. Such scores are generally extracted from appearance information such as color histograms and motion features such as motion smoothness. A discriminative appearance model is learned via combining multiple features in a boosting framework [14]. Part-based appearance models has been applied in multi-target tracking to mitigate occlusions [23].

Given affinity measurements among tracklets, another research focus is on effective and efficient optimization algorithm for association. Bipartite matching via the Hungarian algorithm is among the most popular and simplest algorithms [13][20]. A lot of other optimization frameworks have been proposed, such as K-shortest path [5], set-cover [26], Generalized Minimum Clique Graphs [2].

Most of the work only considers pairwise similarities, without referring to high level information. Thus, problems such as unlikely abrupt motion changes cannot be addressed. In [29] a Conditional Random Field (CRF) is used for tracking while modeling motion dependencies among associated tracklet pairs. In [7] a Lagrangian relaxation is conducted to make higher-order reasoning tractable in the min-cost flow framework. They focus on higher-order constraints such as constant velocity. However, both works concentrate on individuals that may possess a lot of freedom.

We focus on utilizing social grouping information for more natural **high level constraints**. Social factors have attracted a lot of attention in multi-target tracking recently. In [19] a more effective dynamic model leveraging nearby people's positions is proposed. In [18] trajectory prediction accuracy is improved by inferring pedestrian groups. Nearby tracks are also considered as contextual constraints in [6]. In [21] social grouping information is used as a higher-level cue to improve multi-target tracking performance. They seek a balanced explanation of data between K-means clustering for group description and tracklet association. However, their grouping is performed at a pedestrian level and the number of groups is a fixed value which might be too rigid (such as when people in groups split). As a comparison, our grouping scheme is more flexible by using elementary groups. Also their optimization [21] is gradient-based and K-means clustering needs multiple random initializations. The optimization in our approach is deterministic with a closed-form solution.

Some work in computer vision [3][11][24][28] has explored group discovery and group tracking, while our work focuses on using social groups to maintain individual identities in a DAT framework.

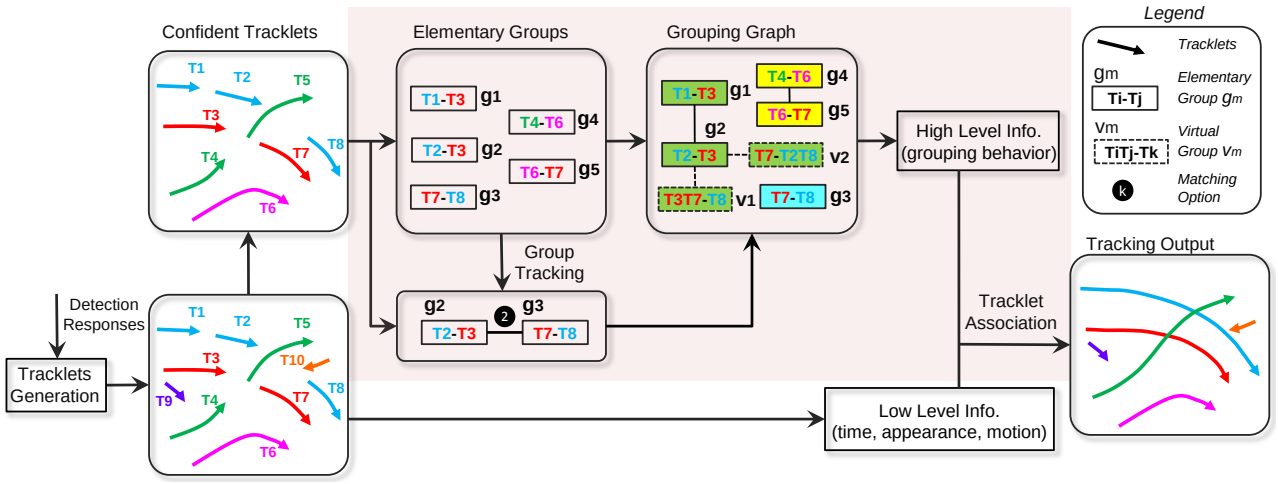

Figure 3. Block diagram of our tracking system. Tracklets with the same color contain the same target. Best viewed in color. For the legends in this figure please see the box in the upper right hand side.

3. Technical Approach

In this section, we introduce how the elementary grouping model is integrated to the basic tracking framework for tracklet association. An overview is presented in Figure 3.

3.1. Tracking Framework with Grouping

Detection-based tracking finds the best set of associations with the maximum linking probability given the detection responses of a video sequence. In an optimal association, each disjoint string of detections should correspond to the trajectory of a specific target in the ground-truth. However, object detector is prone to errors, such as false alarms and inaccurate detections. Also, linking detections directly has a high computational cost. Therefore, it is a common standard to pre-link detection responses with high linking probabilities to generate a set of reliable tracklets (trajectory fragments). Then a global optimization method is employed to associate the tracklets according to multiple cues.

A mathematical formulation of the tracking problem is given as follows. Suppose a set of tracklets $\mathcal{T} = \{T_1, .., T_n\}$ 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 tracklet T_i and T_j contain the same target, assuming T_i occurring before T_j. A valid association matrix A is defined as follows:

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

The constraints for matrix A indicate that each tracklet should be associated to and associated by only one other tracklet (the initial and the terminating tracklets of a track are discussed in Section 4.1).

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 4.1).

Let Ω be the set of all possible association matrices, the multi-target tracking can be formulated as the following optimization problem:

$$A^* = \arg\min_{A \in \Omega} \sum_{ij} a_{ij} S_{ij} \quad (2)$$

This assignment problem can be solved optimally by the Hungarian algorithm in polynomial time.

As low level information is not sufficient to distinguish targets under challenging situations, we consider to integrate high level context into the cost matrix to regularize the solution. The high level context is obtained by analyzing the elementary grouping structure of the tracklets. 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 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 wrong pair of tracklets. Let P_{ij} be the inferred high level information for T_i and T_j, the tracklet association problem can be refined as:

$$A^* = \arg\min_{A \in \Omega} \sum_{ij} a_{ij}(S_{ij} - \alpha P_{ij}) \quad (3)$$

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 learning method for grouping analysis and obtain P_{ij} by making inferences from the grouping graph.

3.2. Online Grouping Analysis

3.2.1 Learning of the 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 and 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 10 frames).

Two tracklets T_i and T_j form an elementary group if they have 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; 3) they have similar velocities. Mathematically, we use G_{ij} to denote the probability of T_i and T_j forming an elementary group:

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

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 Eq. 5, Eq. 6, Eq. 7.

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

$$P_d(T_i, T_j) = \frac{1}{L_{ij}} \sum\nolimits_{n=1}^{L_{ij}} (1 - \frac{2}{\pi} \arctan(dist_n)) \quad (6)$$

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

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 n^{th} overlapped frame, θ is the angle between the average velocities of the two tracklets during the overlapped frames. In our experiments, we set $dist_n = ratio_n \cdot d/0.5(width_i + width_j)$, $ratio_n$ is computed as the size of the larger target over the size of the smaller target, d is the Euclidean distance between the two object centers, and $0.5(width_i + width_j)$ is the largest distance in the image space for two people that walk side by side. The term $ratio_n$ prevents tracklets like in Figure 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.

(a) (b)

Figure 4. Examples of generating incorrect elementary groups if the distances are not normalization.

We create a node for each pair of tracklets that have non-zero 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 co-existing period.

Note that even if two tracklets form an elementary group, the grouping is only meaningful for the overlapped part. For example, if T_a and T_b are in the same elementary group, this only indicates that T_a and T_b have similar motion pattern for the period that they have time overlap. During the non-overlapping period, T_a may form elementary groups with other tracklets/targets that are even in a different group of T_b. Such property makes the elementary group flexible to handle group split and merge easily.

3.2.2 Group Tracking

The relationship between two elementary groups is identified by group tracking. Inspired by association-based multi-target 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 cost for linking any two groups and use Hungarian algorithm to obtain the global optimal solution.

Let $\{T_1^{g_i}, T_2^{g_i}\}$ denotes the two tracklets in an elementary group g_i. Given two elementary groups g_i and g_j, assume g_i starts before g_j, their linking cost based on time, appearance and motion are denoted as $C_t^g(g_i, g_j)$, $C_{appr}^g(g_i, g_j)$ and $C_{mt}^g(g_i, g_j)$, and the summation of these three costs is the final cost for linking g_i and g_j.

The cost for time is defined as:

$$C_t^g(g_i, g_j) = \begin{cases} 0 & g_i \text{ is not overlapped with } g_j \\ \infty & \text{otherwise} \end{cases} \quad (8)$$

where the non-overlapping constraint 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, $T_2^{g_i}$ and $T_2^{g_j}$ are the same target; 2) $T_1^{g_i}$ and $T_2^{g_j}$ are the same target, $T_2^{g_i}$ and $T_1^{g_j}$ are the same target. We explain in detail for matching option 1), the computation for matching option 2) is similar.

Let $S(\cdot)$ be the appearance similarity for two tracklets, the group linking cost based on appearance is defined as $C_{appr}^g(g_i, g_j) = -log(S(T_1^{g_i}, T_1^{g_j}) + S(T_2^{g_i}, T_2^{g_j}))$. As there might be appearance variations in a single tracklet due to occlusion and lighting changes, it is hard to generate features that can well represent the appearance of a target. In order to get more accurate similarity between two tracklets, we adopt the modified Hausdorff metric [9] which is able to compute the similarity of two sets of images. Given a tracklet T_i that has length m_i, let $T_i = \{d_1^i, d_2^i, ..., d_{m_i}^i\}$ where d_x^i is the x^{th} estimation of T_i. Then $S(\cdot)$ is defined as:

$$S(T_i, T_j) = \min(\frac{1}{m_i} \sum_{d_x^i \in T_i} s(d_x^i, T_j), \frac{1}{m_j} \sum_{d_y^j \in T_j} s(d_y^j, T_i)) \quad (9)$$

where $s(\cdot)$ is the Hausdorff similarity between an estimation and a tracklet. We use a modified cosine similarity measure [16] to compute the similarity between two estimations. It is defined as $s_{cos}(u, v) = \frac{|u^T \cdot v|}{\|u\|\|v\|(\|u-v\|_p + \epsilon)}$, where u, 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 regularization number. In our experiments, we use the concatenation of HSV color histogram and HOG features as the feature descriptors.

We define the cost based on motion as follows:

$$C_{mt}^g(g_i, g_j) = C_{mt}^t(T_1^{g_i}, T_1^{g_j}) + C_{mt}^t(T_2^{g_i}, T_2^{g_j}) \quad (10)$$

where $C_{mt}^t(\cdot)$ is the motion model used for estimating the smoothness of two tracklets (explained in Section 4.1).

For each matching option, we compute the linking cost based on appearance and motion, and use the one with the larger sum for $C_{appr}^g(g_i, g_j) + C_{mt}^g(g_i, g_j)$. Also, the matching option is recorded for each group association.

3.3. Grouping Modeling via Dynamic Graph

3.3.1 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, g_4 and g_5 shown in Figure 3. For difficult cases where there are four different tracklets in two nodes, we use the results of group tracking to find their relationship.

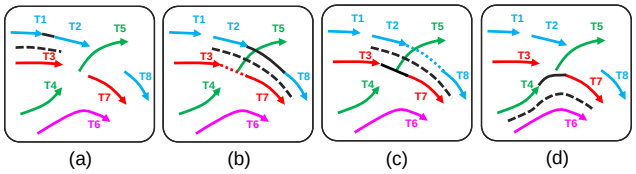

Figure 5. Inference for each edge in the grouping graph in Figure 3: (a) edge between g_1 and g_2, (b) edge between g_2 and v_1, (c) edge between g_2 and v_2, (d) edge between g_4 and g_5. Black solid line represents interpolation between the two tracklets that need inference, black dashed line is the group mean trajectory, and colored dotted line indicates a virtual tracklet.

Suppose 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. Each virtual node also contains two tracklets, one is a virtual tracklet generated by linking a pair of matched tracklets in g_i and g_j, the other is the tracklet left in g_j. An example of virtual node creation is presented in Figure 3. Based on the association of g_2 and g_3, two virtual nodes v_1 and v_2 are created and conneted 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 by using the virtual node we can make inference easily.

3.3.2 Inference from the 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 walk 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 being the same.

Suppose 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 contain the same target is defined as follows:

$$p_{lm} = 0.5(G_{kl} + G_{km}) \times TSimi(T_{\{l,m\}}, G_{\{k,l,m\}}) \quad (11)$$

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:

$$TSimi(T, G) = 1 - \frac{2}{\pi} \arctan(Dist) \quad (12)$$

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 Figure 5. Note that there might be multiple inferences related to the same two tracklets, as the same tracklet may be contained in multiple elementary groups. Therefore, P_{ij} in Eq. 3 is the sum of all inferences that relate to T_i and T_j, as shown below:

$$P_{ij} = \sum p_{ij} \quad (13)$$

4. Experiments

We evaluate our approach on two widely used public pedestrian tracking datasets: the CAVIAR dataset [1] and the TownCentre dataset [4]. The popular evaluation metrics defined in [15] are used for performance comparison: the number of trajectories in ground-truth (GT), the ratio of mostly tracked trajectories (MT), the ratio of mostly lost trajectories (ML), the number of fragments (Frag) and ID switches (IDS).

We compare our approach with the basic affinity model (Baseline Model 1), elementary grouping model without group tracking (Baseline Model 2) and the Social Grouping Behavior model (SGB) in [21]. For a fair comparison, the same input tracklet set, ground-truth, as well as basic affinity model are used for all methods. All the results for the SGB model are provided by courtesy of authors of [21]. Both quantitative comparisons with state-of-the-art methods and visualized results of our approach are presented.

4.1. Implementation Details

Tracklets generation: Two different ways of generating tracklets are used in order to validate that our proposed approach is independent of a specific choice. In the first method, targets on each frame are detected via the discriminatively trained deformable part models [10]. We applied a detection association method similar to [20] to generate conservative tracklets. For each unassociated detection a Kalman filter based tracker is initialized with position and velocity states. A detection is associated to the detection in the next frame that has the minimum distance to the predicted location, and the corresponding Kalman filter is updated. The tracker terminates if no proper association is found, or one detection is associated by multiple trackers.

In the second method, the popular HOG based human detector [8] is used. Tracklets are generated by connecting detections in consecutive frames that have high similarity in position, appearance and size. A simple two-threshold strategy [13] is used to generate reliable tracklets.

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 time model constrict that tracklets can only be linked if their time gap is smaller than a pre-defined threshold. The appearance model is based on the Bhattacharyya distance of two average color histograms [25]. We use a linear motion model [21] to measure the motion smoothness of two tracklets in both forward and backward directions.

The cost matrix S: Due to the constraints in Eq. 1, the traditional pairwise assignment algorithm is not able to find initial and the terminating tracklets. Therefore, instead of using the cost matrix S ($n \times n$) directly, we use the augmented matrix ($2n \times 2n$) in [21] as the input for the Hungarian algorithm. This enables us to set a threshold for association, a pair of tracklets can only be associated when their cost is lower than the threshold.

Method	MT	ML	Frag	IDS	Time
Baseline Model 1	74.7%	6.7%	11	12	1.5s
Baseline Model 2	78.7%	6.7%	10	8	4.2s
SGB Model [21]	89.3%	2.7%	7	5	50s
Our Model	90.7%	2.7%	6	5	4.6s

Table 1. Comparison of tracking results on CAVIAR dataset. The number of trajectories in ground-truth (GT) is 75.

4.2. Results on CAVIAR dataset

The videos in the CAVIAR dataset are obtained in a shopping center where frequent interaction and occlusion occur and people are more likely to walk in groups. We select the same set of test videos as in [21], which are the relatively challenging ones in the dataset. We generate input tracklets using the first method described in Section 4.1. The comparative results are shown in Table 1. Our proposed model achieves the best performance in all aspects. 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 and Our Model). The comparison between Baseline Model 2 and our model demonstrates the importance of group tracking, as it reveals more grouping information. Moreover, our model has better performance compared with the SGB model (better results in MT and Frag, the same results in ML and IDS), but with much less computational time. Sample tracking results are shown in Figure 6.

Method	MT	ML	Frag	IDS	Time
Baseline Model 1	76.8%	7.7%	37	60	350s
Baseline Model 2	78.6%	6.8%	34	46	457s
SGB Model [21]	83.2%	5.9%	28	39	4861s
Our Model	85.5%	5.9%	26	36	465s

Table 2. Comparison of tracking results on TownCentre dataset. The number of trajectories in ground-truth (GT) is 220.

4.3. Results on TownCentre dataset

The TownCentre dataset has one high-resolution video which captures the scene of a busy street. There are 220 people in total, with an average of 16 people visible per frame. We tested all models using the first 3 minutes of the video, and generate input tracklets using the second method described in Section 4.1. The comparative results are shown in Table 2. Results from Table 1 and Table 2 suggest that the performance of our method is consistent on both datasets, which further validate the robustness and efficiency of our model. Sample tracking results are shown in Figure 7.

4.4. Computational Time

The computational time is greatly affected by the number of targets in a video and the length of the video. We implemented our approach in Matlab without code optimization or parallelization and tested it on a PC with 3.0GHz CPU and 8GB memory. For the comparablely short videos in CAVIAR, our approach takes 4.6 seconds on the average. For the video in TownCentre the computational time is 465 seconds. It is observed that our approach is significantly more efficient than the SGB model and produces better tracking results. Note that computational time for object detection, tracklet generation, and appearance and motion feature extraction are not included.

5. Conclusions

In this work we present an online learning approach that integrates high level grouping information into the basic affinity model for multi-target 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 group. Experimental results on challenging datasets demonstrate the superiority of tracking with elementary grouping information. When compared to the state-of-the-art social grouping model, our approach provides better performance and is much more efficient computationally.

Acknowledgements This work was supported in part by NSF grant 1330110 and ONR grants N00014-12-1-1026 and N00014-09-C-0388.

References

[1] Caviar dataset. http://homepages.inf.ed.ac.uk/rbf/caviardata1/.

[2] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, 2008.

[3] L. Bazzani, M. Cristani, and V. Murino. Decentralized particle filter for joint individual-group tracking. In *CVPR*, 2012.

[4] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.

[5] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE TPAMI*, 2011.

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

[7] A. A. Butt and R. T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *CVPR*, 2013.

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

[9] M.-P. Dubuisson and A. Jain. A modified hausdorff distance for object matching. In *ICPR*, 1994.

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

[11] W. Ge, R. Collins, and C. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Trans. PAMI*, 2011.

[12] J. F. Henriques, R. Caseiro, and J. Batista. Globally optimal solution to multi-object tracking with merged measurements. In *ICCV*, 2011.

[13] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[14] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 2010.

[15] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009.

[16] C. Liu. Discriminant analysis and similarity measure. *Pattern Recognition*, 2014.

[17] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 2010.

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

[19] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.

[20] 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 *CVPR*, 2006.

[21] Z. Qin and C. Shelton. Improving multi-target tracking via social grouping. In *CVPR*, 2012.

[22] Z. Qin, C. Shelton, and L. Chai. Social grouping for target handover in multi-view video. In *ICME*, 2013.

(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

Figure 6. Examples of tracking results of our approach on CAVIAR dataset. The same color indicates the same target, best viewed in color.

Figure 7. Examples of tracking results of our approach on TownCentre dataset. With grouping information, targets (199 and 201) pointed by arrow are correctly tracked under frequent occlusions. The same color indicates the same target, best viewed in color.

[23] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012.

[24] J. Sochman and D. Hogg. Who knows who - inverting the social force model for finding groups. In *ICCV Workshops*, 2011.

[25] B. Song, T. Jeng, E. Staudt, and A. K. Roy-Chowdhury. A stochastic graph evolution framework for robust multi-target tracking. In *ECCV*, 2010.

[26] Z. Wu, T. H.Kunz, and M. Betke. Efficient track linking methods for track graphs using network-flow and set-cover techniques. In *CVPR*, 2011.

[27] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, 2009.

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

[29] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *CVPR*, 2011.

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

[31] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 2006.