GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs

Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah

UCF Computer Vision Lab, Orlando, FL 32816, USA

Abstract. Data association is an essential component of any human tracking system. The majority of current methods, such as bipartite matching, incorporate a limited-temporal-locality of the sequence into the data association problem, which makes them inherently prone to ID-switches and difficulties caused by long-term occlusion, cluttered background, and crowded scenes. We propose an approach to data association which incorporates both motion and appearance in a global manner. Unlike limited-temporal-locality methods which incorporate a few frames into the data association problem, we incorporate the whole temporal span and solve the data association problem for one object at a time, while implicitly incorporating the rest of the objects. In order to achieve this, we utilize Generalized Minimum Clique Graphs to solve the optimization problem of our data association method. Our proposed method yields a better formulated approach to data association which is supported by our superior results. Experiments show the proposed method makes significant improvements in tracking in the diverse sequences of Town Center [1], TUD-crossing [2], TUD-Stadtmitte [2], PETS2009 [3], and a new sequence called Parking Lot compared to the state of the art methods.

Keywords: Data Association, Human Tracking, Generalized Graphs, GMCP, Generalized Minimum Clique Problem.

1 Introduction

In the context of tracking, data association is the problem of finding the detections corresponding to one particular object in different frames of a video. The input to the data association problem in a sequence can ideally be represented by a graph in which all the detections in each frame are connected to all the other detections in the other frames, regardless of their closeness in time. Similarly, the output can be ideally represented by several subgraphs of the input in which the detections belonging to common entities are connected.

Finding the ideal subgraphs which represent the exact solution to data association requires solving an optimization problem which remains unsolved to date due to its extreme complexity. Therefore, approximation methods are employed in order to simplify the conditions and find an acceptable solution. The natural, and probably the simplest, approximation is considering a limited-temporal-locality, e.g. two or few frames of the input graph, and solving the optimization

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 343–356, 2012.

problem for the smaller, less complex subgraph. This natural approximation forms one main category of data association methods which has been investigated thoroughly in literature. Best known examples of such methods are bipartite matching (considering only two frames) and its extensions [4,5], which use an extended, yet limited, number of frames. In [4], an approach based on k-partite matching in an extended temporal window is proposed. However, the complexity of their method is proportional to the size of the temporal window (the number of frames), which makes their method impractical for a window size larger than 5-10 frames. The approach in [6] first generates low level tracklets and then employs heuristics in order to merge tracklets into trajectories. The authors of [7,8,9] use a similar approach; however the motion models are improved by incorporating social interaction using a social force model. Despite good complexity performance of the aforementioned methods, their approximation assumption makes them inherently prone to ID-switches and difficulties caused by long-term occlusions, complex motion scenarios, inaccurate bounding boxes, background domination, etc.

Several methods, generally termed *global*, have recently been proposed that make an effort to reduce or remove the limited-temporal-locality assumption. The common point among all of these methods is that they consider the whole temporal span of the sequence rather than limiting it to a number of frames. Zhang et al. [10] define a graph similar to [4] and solve the data association problem by optimizing the cost flow network in order to find the globally optimal trajectories. Brendel et al. [11] use Maximum Weight Independent Set to merge short tracks of size 2 into full trajectories. Berclaz et al. [12] use a grid of potential spatial locations and formulate their data association problem using the K-Shortest Path algorithm. However, their method does not actively incorporate appearance features into the data association process, which makes their approach prone to ID-switches in complicated scenarios. In contrast, Horesh et al. [13] address this limitation by exploiting the global appearance constraints in a similar framework.

In this paper, we propose a method for data association which incorporates both appearance and motion in a global way. In the proposed framework, we incorporate the whole temporal span of the sequence into the data association problem, but we focus on one object at time rather than addressing all of them simultaneously. This is to avoid dealing with an extremely complex optimization problem. Although we focus on solving the data association problem for one object at time, we also incorporate all the other objects implicity. Therefore, our approximation in the object-domain is significantly less restrictive than those used by other approximate methods, such as limited-temporal-locality. This is because the limited-temporal-locality methods are literally *blind* to the information outside of the temporal neighborhood they are focused on, while the proposed method incorporates the whole approximation domain, i.e. all objects, implicitly. We argue that this fundamental difference, along with our proposed framework, yields a better formulation of the data association problem which is supported by our superior results.

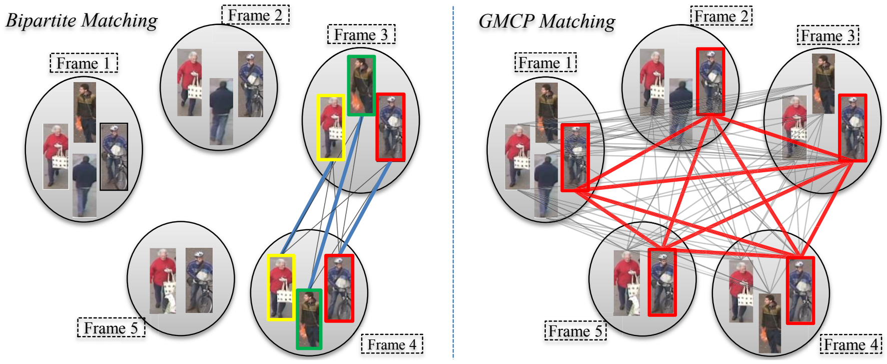

Fig. 1. Bi-partite vs. GMCP matching. Gray and colored edges represent the input graph and optimized subgraph, respectively. Bi-partite matches all objects in a limited temporal window. On the other hand, the proposed method matches one object at a time across full temporal span, while incorporating the rest of the objects implicitly.

The proposed method and bi-partite matching are both shown schematically in Fig. 1 for a sequence of 6 frames. The gray edges show the input graph for the data assosiation for one iteration of the algorithms. The colored thick edges show the optimized subgraph representing the results of that iteration. As shown in the right, the resultant subgraph of our method determines the detections of one person over the whole temporal span, while keeping a notion of other pedestrians in the optimization process. In contrast, the limited-locality-methods do not carry the information out of their current locality (i.e. no edges to or from the frames other than 3 and 4).

The proposed framework is composed of detecting humans in frames[14], calculating tracklets in video segments (subsec. 2.1) and merging them into trajectories(subsec. 2.2). We utilize Generalized Minimum Clique Graphs to solve our data association problem (subsec. 2.1). We tested the proposed method on the diverse sequences of Town Center [1], TUD-corssing [2], TUD-Stadtmitte [2], PETS2009 [3], and a new sequence called Parking Lot, all with promising results.

2 GMCP-Tracker

The block diagram of the proposed global data association algorithm is shown in Fig. 2. The first step is to detect the humans in each frame. We used Felzenszwalb et al.'s [14] part-based human detector. However, any other detector could be used. Next, we divide the input video into a number of segments and find the tracklet of pedestrians within each segment using the proposed global method for tracklet generation utilizing GMCP. In the last step, we merge the tracklets found in all of the segments to form the trajectory of each person over the course of the whole video.

Despite the appearance of the pedestrians remaining rather consistent throughout a video, the pattern of motion tends to differ significantly in short and long

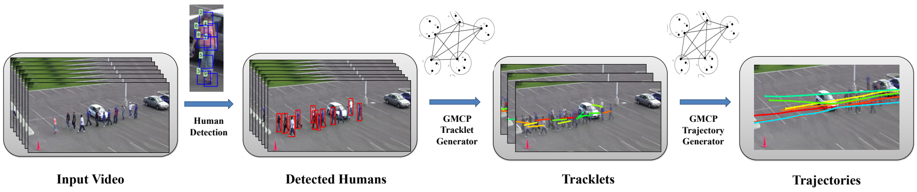

Fig. 2. The block diagram of the proposed human tracking method

term. In principle, it's difficult to model the motion of one person for a long duration without having the knowledge of the destination, structure of the scene, interactions between people, etc. However, the motion can be modeled sufficiently using constant velocity or acceleration models over a short period of time. Therefore, the way motion is incorporated into the global data association process should be different in short and long terms. This motivated us to employ the hierarchical approach, i.e. finding tracklets first and then merging them into full trajectories.

The rest of this section is organized as follows: 2.1 explains the proposed method for finding tracklets along with an overview of Generalized Minimum Clique Problem, our global motion-cost model and occlusion handling method. Merging the tracklets to form global trajectories is explained in 2.2.

2.1 Finding Tracklets Using GMCP

We divide a video into s segments of f frames each. We propose a data association method for finding tracklets which are globally consistent in terms of motion and appearance over the course of a segment. The input to our data association problem for finding tracklets is a graph $G = (\boldsymbol{V}, E, w)$, where $\boldsymbol{V}$, E and w denote the set of nodes, set of edges and weights of edges, respectively. $\boldsymbol{V}$ is divided into f disjoint clusters. Each represents one frame, and the nodes therein represent the human detections in that particular frame. Let C_i, where $i \in \mathbb{Z} : 1 \leq i \leq f$, denote the frame (≡cluster) i, and v_m^i denote the mth detection (≡node) in the ith frame. Therefore $\boldsymbol{C}_i = \{v_1^i, v_2^i, v_3^i, ...\}$. The edges of the graph are defined as $E = \{(v_m^i, v_n^j) | i \neq j\}$ which signifies that all the nodes in G are connected as long as they do not belong to the same cluster. A node, v_m^i, is associated with a location feature, $\boldsymbol{x}_m^i$, which is the 2-dimensional spatial coordinates of the center of the corresponding detection and appearance features, $\underset{l}{\boldsymbol{\phi}}_m^i$, which represents the color histogram of the lth body part of the detection v_m^i. The weight of an edge between two nodes, $w : E \rightarrow \mathbb{R}^+$, represents the similarity between the two corresponding detections:

$$w(v_m^i, v_n^j) = \sum_{l=1}^{8} k(\underset{l}{\boldsymbol{\phi}}_m^i, \underset{l}{\boldsymbol{\phi}}_n^j), \tag{1}$$

where k represents histogram intersection kernel.

The task of finding the tracklet of one particular person in a segment requires identifying the detections of that person in each frame. Therefore, a feasible solution to this problem can be represented by a subgraph of G in which one node ($\equiv$detection) is selected from each cluster($\equiv$frame). We call this subgraph which represents a feasible solution $G_s = (\boldsymbol{V}_s, E_s, w_s)$. Therefore, G_s contains a set of nodes which includes the general form $\boldsymbol{V}_s = \{v_a^1, v_b^2, v_c^3, ...\}$ which means the ath node from 1st cluster, bth one from 2nd cluster, and so on are selected to be in $\boldsymbol{V}_s$. By definition, $E_s = \{E(p,q)|p \in \boldsymbol{V}_s, q \in \boldsymbol{V}_s\}$ and $w_s = \{w(p,q)|p \in \boldsymbol{V}_s, q \in \boldsymbol{V}_s\}$. Bear in mind that the feasible solution G_s represents the tracklet of *one person* and not all the people visible in the segment. Fig 3 (a) shows the human detections in a small segment of 6 frames along with the graph G they form in (b). (d) shows a feasible solution G_s with the tracklet it forms in (c). Since the set of nodes $\boldsymbol{V}_s$ is enough to form G_s, we use $\boldsymbol{V}_s$ to denote a feasible solution hereafter.

We define the appearance cost of the feasible solution V_s as:

$$\gamma_{appearance}(\boldsymbol{V}_s) = \frac{1}{2}(\sum_{i=1}^{f} \sum_{j=1, j\neq i}^{f} w(\boldsymbol{V}_s(i), \boldsymbol{V}_s(j))), \tag{2}$$

which is the cost of the complete graph induced by the nodes in $\boldsymbol{V}_s$. Eq. 2 is a global cost function since it is based on comparing all pairs of detection in a feasible solution no matter how close they are temporally. This is based on the assumption that the appearance of people does not change drastically in a segment. Overlapping bounding boxes, occlusion, noisy descriptors, background domination, etc in part of a trajectory can potentially cause an ID-switch in the majority of current methods, in particular the limited-temporal-locality ones. The formulation defined in eq. 2 minimizes the chance of such cases of ID-switches as all possible pairs of detections are compared regardless of their temporal order.

By finding the feasible solution with the minimum appearance cost, i.e. $\underset{\boldsymbol{V}_s}{\text{argmin}}(\gamma_{appearance}(\boldsymbol{V}_s))$, the tracklet of the person with the most stable color histogram features in the segment will be found. In the following subsection, we explain that Generalized Minimum Clique Graph is a perfect fit for our problem, and can be used for solving the aforementioned optimization task for finding tracklets.

Generalized Minimum Clique Problem (GMCP). Generalized Graph Problems, more formally known as Generalized Network Design Problems [15], are a class of problems which are commonly built on generalizing the standard subgraph problems. The generalization is commonly done by expanding the definition of a node to a cluster of nodes. For instance, the objective in the standard Traveling Salesman Problem (TSP) is to find the minimal Hamiltonian cycle which visits all the nodes of the input graph exactly once. In the *Generalized* Traveling Salesman Problem, the nodes of the input graph are grouped into disjoint clusters, and the objective is to find the minimal Hamiltonian cycle which visits all the clusters of the input graph exactly once. From each cluster, exactly one node should be visited.

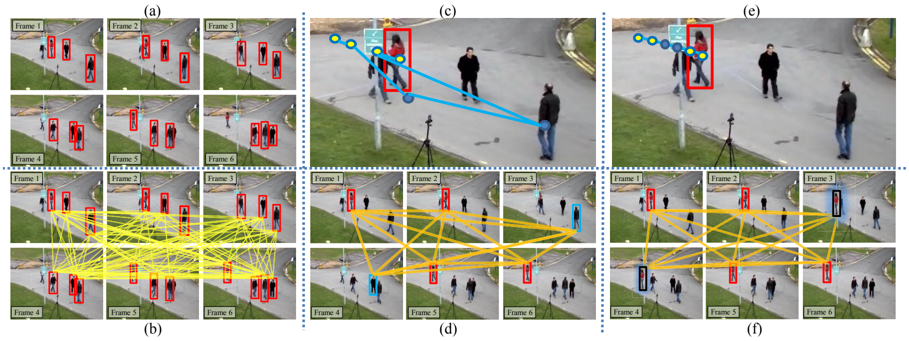

Fig. 3. Finding a tracklet for a small segment of 6 frames. The left column shows the detections in each frame along with graph G they induce. The middle column shows the feasible solution with minimal cost along with the tracklet it forms, *without* adding hypothetical nodes. The right column shows the feasible solution with minimal cost *with* hypothetical nodes added for handling occlusion, along with the tracklet it forms.

Similarly, in the Generalized Minimum Clique Problem (GMCP) the nodes of the input graph are grouped into disjoint clusters. The objective is to find a subset of the nodes that include exactly one node from each cluster while requiring the minimum cost for the complete graph they produce [15]. Recently, GMCP has been used in the fields of biology and telecommunications [15]. However, its potential applications in Computer Vision have not been studied to date.

In order to have a more formal definition for GMCP, assume a graph $G = (\boldsymbol{V}, E, w)$ exists, where G is undirected and weighted, $\boldsymbol{V}$ is the set of all nodes, E is the set of edges, and $w : E \to \mathbb{R}^+$ is the weight of a given edge. The set of nodes V is divided into f clusters $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_f$ such that all of the clusters are completely disjoint: $\boldsymbol{C}_1 \cup \boldsymbol{C}_2 \cup \ldots \cup \boldsymbol{C}_f = V$ and $\boldsymbol{C}_1 \cap \boldsymbol{C}_j = \emptyset$ $(1 \leq i \neq j \leq k)$. A feasible solution of the GMCP instance is a subgraph $G_s = (\boldsymbol{V}_s, E_s, w_s)$, where $\boldsymbol{V}_s$ is a subset of $\boldsymbol{V}$ which encompasses only one node from a given cluster. E_s is a subset of E which includes the nodes V_s induces, and w_s is their corresponding weights from w. The goal of the GMCP is to find the feasible solution with the minimal cost, where the cost is defined to be the sum of all the weights along the solution subgraph.

In this formulation, there exists an edge in E for all possible pairs of nodes of $\boldsymbol{V}$, as long as they do not belong to the same cluster. Therefore, the subgraph G_s is essentially complete, which makes any feasible solution of GMCP a clique.

As can be seen from the formulation of our data association problem explained in 2.1, GMCP essentially solves the same optimization problem we need to solve for finding a tracklet if the input graph G is formed as explained in 2.1. Therefore, by solving GMCP for the graph G, the optimal solution which corresponds to the feasible solution with the most consistency in appearance features over the course of the segment, i.e. $\underset{\boldsymbol{V}_s}{\text{argmin}}(\gamma_{appearance}(\boldsymbol{V}_s))$, is found. More details about solving GMCP will be discussed in section 3.

In order to incorporate motion, as well as appearance, into the data association problem, we add one more term to the cost function and define our global data association task as the following optimization problem:

$$\hat{\boldsymbol{V}}_s = \underset{\boldsymbol{V}_s}{\operatorname{argmin}}(\gamma_{appearance}(\boldsymbol{V}_s) + \alpha.\gamma_{motion}(\boldsymbol{V}_s)), \tag{3}$$

where $\hat{\boldsymbol{V}}_s$ is the optimal solution to determine the data association for *one* tracklet, and α is the mixture constant which balances the contribution of appearance and motion.

Finding $\hat{\boldsymbol{V}}_s$ by solving eq. 3 yields the tracklet of *one* person in the segment. Therefore, in order to find the tracklets of all the pedestrians in the segment, the optimization problem of eq. 3 has to be solved several times. The first time eq. 3 is solved, the algorithm finds the tracklet which has the lowest total cost, i.e. the most stable appearance features and most consistent motion with the model. Then, the vertices selected in $\hat{\boldsymbol{V}}_s$ are excluded from G and the above optimization process is repeated to find the tracklet for the next person, and so on. This process is repeated until zero or few nodes remains in G.

Since the algorithm finds the tracklets in order of how stable and consistent their appearance and motion features are, the tracklets which are less likely to be confused are calculated and excluded from G first. Therefore, our method does not lower the chance of successful extraction for the tracklets found at the last iterations. Our global motion-cost model, which defines the term $\gamma_{motion}(\boldsymbol{V}_s)$, is explained in the next subsection.

Tracklet-global Motion Cost Model. In order to incorporate motion into the optimization process of eq. 3, we need to calculate a cost for the feasible solution $\boldsymbol{V}_s$ based on motion. The spatial velocity vector for the feasible solution $\boldsymbol{V}_s$ is defined as: $\dot{\boldsymbol{X}}_{\boldsymbol{s}}(i) = \boldsymbol{X}(i+1) - \boldsymbol{X}(i)$, where $1 \leq i \leq (f-1)$. One common approach to computing the motion cost is to calculate the deviation from a presumed model, such as constant velocity. This can be done by using each velocity vector to predict the spatial location of the detection immediately after it, and summing up the errors between the predicted locations and corresponding locations in the feasible solution. This piecewise approach is mainly used in bipartite matching and similar approaches [4,5]. However, in our global framework, one feasible solution is meant to represent one tracklet over the course of the whole segment; therefore we can calculate the motion cost in a more effective way, which assures both piecewise and global consistency within the model. We assume the constant velocity model for the motion of pedestrians in one segment and calculate the motion cost as:

$$\gamma_{motion}(\boldsymbol{V}_s) = \sum_{i=1}^{s} \sum_{j=1}^{s-1} \overbrace{|\boldsymbol{X}_s(i) - \underbrace{[\boldsymbol{X}_s(j) + \dot{\boldsymbol{X}}_s(j).(i-j)]}_{prediction}|}^{deviation}, \tag{4}$$

where the term in brackets in eq. 4 is the predicted location for the node $\boldsymbol{V}_s(i)$ using $\dot{\boldsymbol{X}}_s(j)$. In eq. 4, we assumed a person moves at a constant velocity manner in one segment, and each element of $\dot{\boldsymbol{X}}_{\boldsymbol{s}}$ vector is used to predict the location of all other nodes in the feasible solution $\boldsymbol{V}_s$.

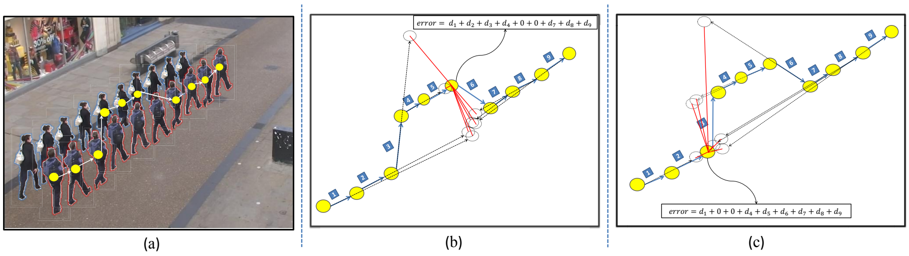
error = $d_1 + d_2 + d_3 + d_4 + 0 + 0 + d_7 + d_8 + d_9$
error = $d_1 + 0 + 0 + d_4 + d_5 + d_6 + d_7 + d_8 + d_9$
1
2
3
4
5
6
7
8
9
(a)
(b)
(c)

Fig. 4. Tracklet-Global motion cost. (a) shows the tracklet of a feasible solution with three outliers. (b) and (c) show the cost for an outlier and inlier, respectively.

Fig. 4 shows this in more detail. Fig. 4 (a) shows a feasible solution which is being generated for the person with the red boundary. However, three detections of another person are mistakenly selected in the feasible solution. Therefore, we expect the three wrong selections to add a large value to the motion cost, while the rest of the selected nodes, which are consistent, to add low cost values. The value of eq. 4 is shown for two nodes of $i = 6$ and $i = 3$ in parts (b) and (c), respectively. The black circles show the predicted locations for the node i. The red lines depict the distance between the predicted locations and $\boldsymbol{X}_s(i)$, which shows the deviation from the model. The value node i, which adds to the motion cost, is the sum of these distances. As can be seen, the node $i = 6$, which is not consistent with the majority of the tracklet, adds a large value to the cost, while $i = 3$ adds a lower value.

Therefore, in contrast to the piecewise motion models, the motion cost in eq. 4 is calculated by measuring the deviation from the constant velocity model in a tracklet in a global manner, because all nodes are contributing to the cost of the other nodes. Although we use the constant velocity model in eq. 4, the extension to the constant acceleration and higher order models for more complicated scenarios is straightforward.

Handling Occlusion Using Hypothetical Nodes. In some cases, a given frame may not include a detection for a particular person due to occlusion, missed detection, etc. In order to cope with this issue, we add a Hypothetical Node to each cluster, thus if one frame does not include an appropriate detection, the hypothetical node is selected.

We need to define appearance and motion features for the hypothetical nodes, as each node in G has these two features. Solving the optimization problem of eq. 3 is an iterative process. In each iteration these two features for the Hypothetical Nodes are re-estimated using the method explained in the rest of this subsection and the hypothetical nodes are updated.

$\boldsymbol{V}_s$ is expected to include the detections for one person (inliers), along with other detections (outliers) for the frames, which do not include a detection for that particular person. Due to the short temporal span of one segment, we assume a person moves at a constant velocity in a segment and use this assumption in order to identify inliers and outliers in $\boldsymbol{V}_s$. According to 2-dimensional constant velocity motion, the spatial location of the detections can be modeled using $\boldsymbol{X}_s(i) = \boldsymbol{a_1} i + \boldsymbol{a_o}$, where $\boldsymbol{a_1}$ and $\boldsymbol{a_0}$ are 2-dimensional constant vectors. Therefore, we identify the inliers and outliers to the constant velocity model determined by $\boldsymbol{a_1}$ and $\boldsymbol{a_0}$ using:

$$\boldsymbol{V}_s^{inliers} = \{\boldsymbol{V}_s(i) : |\boldsymbol{a_1} i + \boldsymbol{a_0} - \boldsymbol{X}_s(i)| < \delta\}, \tag{5}$$

where δ is the fitting tolerance. The best parameters of the tracklet's constant velocity model are the ones which maximize the number of inliers:

$$\hat{\boldsymbol{a_1}}, \hat{\boldsymbol{a_0}} = \underset{\boldsymbol{a_1}, \boldsymbol{a_0}}{argmax}(\#\{\boldsymbol{V}_s^{inliers}\}), \tag{6}$$

where # represents the cardinality of the set. Since V_s is composed of inliers and outliers, we employ *RANSAC*, which is a robust estimation method, to compute $\boldsymbol{a_1}$ and $\boldsymbol{a_0}$ in eq. 6 with the error criterion of eq. 5

The inlier nodes in V_s are those which survive the criterion of eq. 5 using $\hat{\boldsymbol{a_1}}$ and $\hat{\boldsymbol{a_0}}$. The spatial coordinates of the hypothetical nodes are computed using the estimated model $\boldsymbol{x}_H^i = \hat{\boldsymbol{a_1}} i + \hat{\boldsymbol{a_0}}$, where $\boldsymbol{x}_H^i$ denotes the spatial location of the hypothetical node of cluster i. The appearance feature of the hypothetical nodes is the average appearance of the inlier nodes' appearances (mean of their color histograms). However, a constant penalty is added to the weights of the edges connected to the hypothetical nodes in G, in order to avoid selecting the hypothetical node if the frame includes a proper detection.

As mentioned earlier, the hypothetical nodes are updated at the end of each iteration when solving the optimization problem of eq. 3. In the first few iterations, hypothetical nodes are not likely to be selected as the algorithm is still selecting the existing detections. However, as the optimization process progresses, the clusters which include correct detections are exhausted and the hypothetical nodes will start to contribute until the algorithm converges to the final solution $\hat{V_s}$. Fig. 3 (f) shows $\hat{\boldsymbol{V}}_s$ two hypothetical nodes selected for the frames with occlusion. The trajectory $\hat{\boldsymbol{V}}_s$ forms is shown in (e).

2.2 Merging Tracklets into Trajectories Using GMCP

As explained earlier, we divide the video into s segments and find the tracklets of all the pedestrians in each segment using the method described in subsection 2.1. In order to generate a trajectory of a person over the course of the full video, we need to merge the tracklets belonging to each person. This is a data association problem for which we can use any available data association method, such as bipartite matching[5,4]. However, in order to have a fully global framework, we use the same GMCP-based data association method we used for finding tracklets to merge them. Therefore, the clusters and nodes in G' now represent segments and tracklets, respectively (vs. representing frames and human detections in 2.1)[1]. The appearance feature of a node, which represents one tracklet, is defined as the average appearance of the human detections in the tracklet (the mean of their color histograms), and its spatial location, $\boldsymbol{x'}^{i}_{m}$, is defined as the spatial location of the middle point of the tracklet. Fig. 5 (a) shows six consecutive segments with their tracklets, along with the complete graph their representative nodes induce in (b). Only four tracklets out of fifteen are shown to avoid cluttering of the plots.

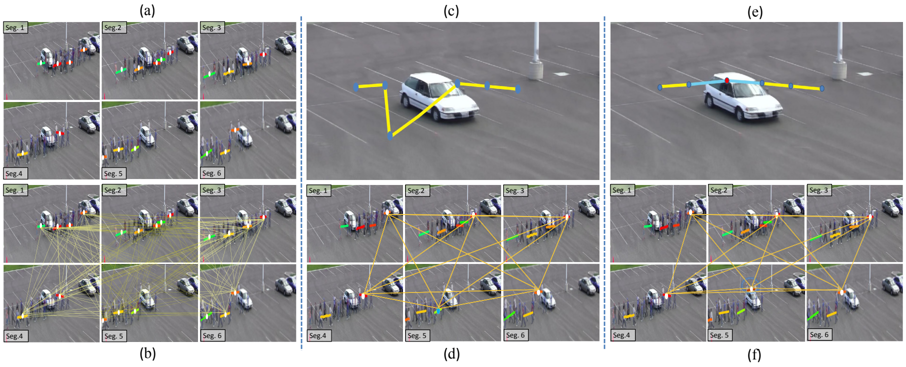

Fig. 5. Merging tracklets into trajectories. The left column shows six consecutive segments with four tracklets in each, along with $\boldsymbol{G'}$. The middle column shows a feasible solution without adding the hypothetical nodes to handle tracklet-occlusion. The right column shows the converged solution, $\hat{\boldsymbol{V}}'_s$, along with the generated full trajectory.

Note that the data association at the track level is fundamentally different from the one used for finding tracklets; there we assumed a pedestrian moves at a constant velocity within one segment, but modeling the human motion over long periods of time in a track becomes extremely difficult. Generally, it's difficult to model the motion of pedestrians for a long duration without the knowledge of scene structure, intentions, destination, social interaction, etc. This major difference prevents us from using the global motion cost model explained in subsection 2.1 and the hypothetical node feature estimation described in subsection 2.1. However, the full trajectory is expected to be temporally smooth. Therefore, at this level, it is essential to change the method for computing the motion cost and estimating hypothetical nodes to piecewise, rather than global, as follows.

Motion-Cost: In order to compute the motion cost for a feasible solution $\boldsymbol{V}'_s$, we use the piecewise extension of the velocity vector immediately preceding each node:

$$\gamma'_{motion}(\boldsymbol{V}'_s) = \sum_{i=3}^{s} |\boldsymbol{X}'_s(i) - [\boldsymbol{X}'_s(i-2) + 2\dot{\boldsymbol{X}}'_s(i-2)]|. \qquad (7)$$

[1] We show the notations related to tracklet merging level with prime (′) to preserve their correspondence to the ones in tracklet generation level, yet not confuse them.

Compared to the global approach of eq. 4, the second summation in eq. 4 is omitted, so that only the preceding piece of the trajectory contributes to the motion cost, because of this eq. 7 represents a piecewise extension.

Handling occluded tracklets by hypothetical nodes: Short term occlusions and missed detections are already handled at the tracklet level by introducing hypothetical nodes. However, if a pedestrian remains occluded for a period longer than the temporal span of a segment, e.g. over 50 frames, then he/she will not have a tracklet in the corresponding segment, which leads to the *tracklet occlusion.* In order to handle such cases, we add hypothetical nodes to the clusters of the tracklet association input graph G' with piecewise prediction for computing their spatial location:

$$\boldsymbol{x'}^{i}_{Hf} = |\boldsymbol{X}'_s(i-2) + 2\dot{\boldsymbol{X}}'_s(i-2)|. \tag{8}$$

In eq. 8, only the preceding piece of the trajectory is used to compute the spatial location of the hypothetical node, rather than using the global method of subsection 2.1 which incorporated all the nodes.

At the tracklet association level, we add two hypothetical nodes to each cluster rather than one: The *Forward* hypothetical node, denoted by Hf, in which the preceding piece of trajectory is used for prediction, and the *backward* hypothetical node, denoted by Hb, for which the following piece of the trajectory is leveraged for prediction. The backward hypothetical node is necessary for the cases where the tracklets at the beginning of the trajectory are occluded. That way, since there does not exist any tracklet before the beginning of the trajectory to be used to predict the location of the hypothetical nodes of the missing tracklets, the backward hypothetical nodes are added. They use the tail of the trajectory to predict the location of the hypothetical nodes for the missing tracklets at the beginning of the trajectory. Therefore, the spatial location of the backward hypothetical node is computed using $\boldsymbol{x'}^{i}_{Hb} = |\boldsymbol{X'}_s(i+2)+2\dot{\boldsymbol{X}}'_s(i+1)|$.

The appearance feature of the forward and backward hypothetical nodes are the same as the nodes, which were used to predict the spatial location, i.e. $\boldsymbol{V'}_s(i-2)$ for the forward and $\boldsymbol{V'}_s(i+2)$ for the backward ones. The hypothetical nodes added at the level of tracklet association are useful for solving the problem of entry/ exit as well. For instance, if one pedestrian exits the field of

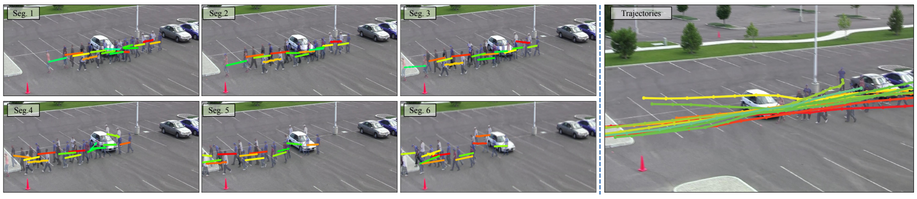

Fig. 6. Left: The tracklets of six sample consecutive segments from Parking Lot sequence. Right: The trajectories resulting from associating tracklets of all the segments.

view before the end of the video, or enters the view later than the beginning of the video, there will be some segments in which they won't have a tracklet. The optimization process will select hypothetical nodes for those segments. However, the computed spatial location of such hypothetical nodes will be out of the view of the frame as they correspond to the time before the pedestrian enters the view or after exiting.

3 Experimental Results and Discussion

We evaluated our method on four publicly available sequences, which provide a wide range of significant challenges: TUD-Crossing [16], TUD-Stadtmitte [17], Town Center [1], and sequence S2L1 from VS-PET2009 benchmark [3]. We also present experimental results on a new data set called Parking Lot. We set δ to 5 in eq. 5 for all the test sequences. We normalize appearance and motion cost values in eq. 3 in order to make them comparable. A typical value for α in eq. 3 is one which assigns equal weights to appearance and motion in the overall cost. However, based on our experiments, the appearance features are more informative than motion in our formulation. In fact, in many cases using appearance features only is sufficient for finding the appropriate tracklets.

Standard CLEAR MOT [18] are used as evaluation metrics. False positives, false negatives and ID-Switches are measured by MOTA. MOTP is defined as the average distance between the ground truth and estimated targets locations. MOTP shows the ability of the tracker in estimating the precise location of the object, regardless of its accuracy at recognizing object configurations, keeping consistent trajectories, and so forth. Therefore, MOTA has been widely accepted in the literature as the main gauge of performance of tracking methods.

Town Center [1]: The sequence consists of 4500 frames. The size of each segment is 50 frames in this experiment. The quantitative results of the competitive methods for this sequence are presented in [8] and shown in Table 1. Our precision and recall values for this sequence are 92.65% and 81.64% respectively.

Table 1. Tracking results on Town Center sequence

	MOTA	MOTP	MODP	MODA
Benfold et al. [1]	64.9	80.4	80.5	64.8
Zhang et al. [10]	65.7	71.5	71.5	66.1
Pellegrini et al. [9]	63.4	70.7	70.8	64.1
Yamaguchi et al. [7]	63.3	70.9	71.1	64.0
Leal-Taixe et al. [8]	67.3	71.5	71.6	67.6
Ours/GMCP	**75.59**	**71.93**	**72.01**	**75.71**

Table 3. Tracking results on Parking Lot sequence

	MOTA	MOTP	Prec.	Rec.
Shu et al. [5]	74.1	79.3	91.3	81.7
Ours/GMCP	**90.43**	**74.1**	**98.2**	**85.3**

Table 2. Tracking results on TUD and PETS 09 sequences

Dataset	MOTA	MOTP	Prec.	Rec.	IDsw
TUD-Crossing. [20]	84.3	71.0	85.1	98.6	2
TUD-Crossing. [11]	85.9	73.0	89.2	98.8	2
TUD-Crossing-Ours	**91.63**	**75.6**	**98.6**	**92.83**	**0**
TUD-Stadtmitte. [2]	60.5	65.8	-	-	7
TUD-Stadtmitte-Ours	**77.7**	**63.4**	**95.6**	**81.4**	**0**
PET2009-View1. [12]	80.00	58.00	81.00	60.00	28
PET2009-View1. [13]	81.46	58.38	90.66	90.81	19
PET2009-View1. [2]	81.84	73.93	96.28	85.13	15
PET2009-View1. [19]	84.77	68.742	92.40	94.03	10
PET2009-View1-Ours	**90.3**	**69.02**	**93.64**	**96.45**	**8**

PET2009-S2L1-View One [3]: The sequence consists of 800 frames. In sec. 2.1, we assumed the motion of pedestrians in one segment is near constant velocity in order to identify the outliers and estimate the location of hypothetical nodes. Therefore, segment size should be set in a way that this assumption is not severely violated. Otherwise, the location of hypothetical nodes will be inaccurate which results in a misplaced tracklet. Typically, the segment size is determined with regard to the frame rate of the video. Therefore, regarding the low frame rate of PET2009-S2L1, we chose a smaller segment size of 15 frames. The quantitative comparison is provided in Table 2. All reported results are calculated using original tracking outputs provided by the authors [12,13,2,19] with the same overlap threshold for CLEAR MOT metrics.

TUD Data Set [2]: TUD-Crossing and TUD-Stadtmitte are two sequences in this data set with low camera angle and frequent occlusions. Crossing and Stadtmitte include 201 and 179 frames respectively. Due to the short length of these sequences, we divided each one into three segments. Quantitative results are provided in Table 2 [11,20].

Parking Lot: This sequence consists of 1,000 frames of a relatively crowded scene with up to 14 pedestrians walking in parallel. It includes frequent occlusions, missed detections, and parallel motion with similar appearances. Quantitative results are shown in Table 3. No ID-switch was observed in our results. Six sample segments and the merging results are shown in fig. 6.

As can be see in Tables 1, 2 and 3, the proposed method constantly outperforms the state of the art on all the standard sequences.

Several application-oriented methods for solving GMCP, such as branch-and-cut algorithm and multi-greedy heuristics [15], have been proposed to date. Inspired by the solutions proposed for the other generalized graph problems, such as GMST [15], we employ Tabu-search to solve our optimization problem of eq. 3. The size of the search neighborhood is changed at each iteration in order to make the optimization process faster and avoid becoming stuck in suboptimal regions. The main contributing factors to the complexity of GMCP are the number of clusters and number of nodes within each cluster. Regarding the small size of GMCP instances we need to solve in our tracking problem, one instance of the optimization problem of eq. 3 is typically solved in the negligible time of a fraction of a second. For a segment of 50 frames with approximately 15 pedestrians, processing a frame using the full proposed framework, excluding human detection, takes an average time of 4.4 seconds on a 4core 2.4 GHz machine running non-optimized Matlab code. Using an optimized parallel implementation in C, the algorithm is likely to work in real time.

4 Conclusion

We propose a global framework for data association. We utilized Generalized Minimum Clique Problem to solve the formulated optimization problem. Our method is based on shifting the approximation from the temporal domain to the

[1] This work was partially supported by Pacific Northwest National Laboratory, WA.

object domain while keeping a notion of the full object domain implicitly. We argue that this framework yields a better developed approach to data association. A two level method, i.e. finding tracklets first and forming full trajectories later, is employed to cope with different characteristics of human motion in the short and long term. Also, we utilize a global approach to compute the motion cost at the tracklet level. Experiments show superior results over the state of the art.

References

1. Benfold, B., Reid, I.: Stable multi-target tracking in real time surveillance video. In: CVPR (2011)
2. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: CVPR (2011)
3. Ferryman, J., Shahrokni, A.: Pets2009: Dataset and challenge. In: International Workshop on Performance Evaluation of Tracking and Surveillance (2009)
4. Shafique, K., Shah, M.: A noniterative greedy algorithm formultiframe point correspondence. In: PAMI (2005)
5. Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M.: Part-based multiple-person tracking with partial occlusion handling. In: CVPR (2012)
6. Leibe, B., Schindler, K., Gool, L.V.: Coupled detection and trajectory estimation for multi-object tracking. In: ICCV (2007)
7. Yamaguchi, K., Berg, A., Ortiz, L., Berg, T.: Who are you with and where are you going? In: CVPR (2011)
8. Leal-Taixe, L., et al.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: ICCV 2011 Workshops (2011)
9. Pellegrini, S., Ess, A., Van Gool, L.: Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 452–465. Springer, Heidelberg (2010)
10. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008)
11. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximumweight independent set. In: CVPR (2011)
12. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. In: PAMI (2011)
13. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV (2011)
14. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. In: PAMI (2010)
15. Feremans, C., Labbe, M., Laporte, G.: Generalized network design problems. In: EJOR (2003)
16. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: CVPR (2008)
17. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR (2010)
18. Kasturi, R., et al.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. In: PAMI (2009)
19. Henriques, J.F., Caseiro, R., Batista, J.: Globally optimal solution to multi-object tracking with merged measurements. In: ICCV (2011)
20. Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009)