This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

On Detection, Data Association and Segmentation for Multi-target Tracking

Yicong Tian, *Member, IEEE,* Afshin Dehghan, *Member, IEEE,* and Mubarak Shah, *Fellow, IEEE*

Abstract—In this work, we propose a tracker that differs from most existing multi-target trackers in two major ways. Firstly, our tracker does not rely on a pre-trained object detector to get the initial object hypotheses. Secondly, our tracker's final output is the fine contours of the targets rather than traditional bounding boxes. Therefore, our tracker simultaneously solves three main problems: detection, data association and segmentation. This is especially important because the output of each of those three problems are highly correlated and the solution of one can greatly help improve the others. The proposed algorithm consists of two main components: structured learning and Lagrange dual decomposition. Our structured learning based tracker learns a model for each target and infers the best locations of all targets simultaneously in a video clip. The inference of our structured learning is achieved through a new *Target Identity-aware Network Flow (TINF)*, where each node in the network encodes the probability of each target identity belonging to that node. The probabilities are obtained by training target specific models using a global structured learning technique. This is followed by proposed Lagrangian relaxation optimization to find the high quality solution to the network. This forms the first component of our tracker. The second component is Lagrange dual decomposition, which combines the structured learning tracker with a segmentation algorithm. For segmentation, multi-label Conditional Random Field (CRF) is applied to a superpixel based spatio-temporal graph in a segment of video, in order to assign background or target labels to every superpixel. We show how the multi-label CRF is combined with the structured learning tracker through our dual decomposition formulation. This leads to more accurate segmentation results and also *helps better resolve typical difficulties in multiple target tracking, such as occlusion handling, ID-switch and track drifting*. The experiments on diverse and challenging sequences show that our method achieves superior results compared to competitive approaches for detection, multiple target tracking as well as segmentation.

Index Terms—Multiple target tracking, Object segmentation, Network flow, Lagrangian relaxation, Dual decomposition.

1 Introduction

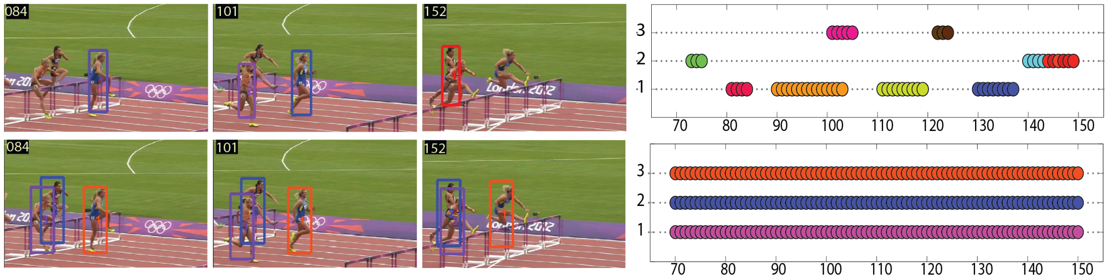
Multi-target tracking is, undoubtedly, one of the fundamental problems in computer vision, with a variety of applications ranging from surveillance to sport analysis and medical image analysis. The goal of tracking is to detect targets and associate them across sequence of frames. Traditionally, the output of the detection stage is a set of bounding boxes corresponding to targets present in the video, where each bounding box is assigned a target label. However, the ultimate way of detecting a target is to localize it and provide pixel wise segmentation, so that the fine contour of a target can be achieved. This way each pixel is assigned a target label.

Formulating tracking where each target pixel is assigned a label, requires solving three major problems: detection, data association and segmentation. Each of these problems has their own line of research, which have been active for decades in the computer vision community. Most existing tracking methods limit themselves to bounding box level target representation and mainly focus on improving either the detection or the data association component of the tracker. Though convenient, bounding boxes are coarse approximations of targets. Moreover, since bounding boxes usually include non-target pixels, the features extracted from them could be contaminated by background pixels. When these features are used as target representation in tracking, they may cause drift, ID-switches and inaccurate target localization. Therefore, the ultimate goal of tracking should be to determine the pixel-wise localization of targets instead of just coarse bounding boxes.

The focus of most previous multi-target tracking algorithm is to improve the data association component of the tracker. A majority of these algorithms assume the existence of pre-trained object detector. Some of these methods heavily rely on the results of the pre-trained detector [1], [2], while some have more tolerance [3], [4], [5] toward miss-detections and false detections that commonly happen when using a pre-trained detector. One solution to address this issue is to design trackers that internally train a detector for each target, eliminating the need for a pre-trained detector. However, there is only a handful of trackers that focus on solving both detection and tracking in the context of multi-target tracking [6], [7], [8], [9]. Due to the lack of a good pre-trained object detector in several scenarios, e.g those in which objects undergo heavy articulation and occlusion, and also due to heavy correlation between performance of a data association method and object detector, solving detection and data association simultaneously is very natural. An example to further motivate this approach is shown in Fig. 1, where we show a scenario in which poor detection propagates into data association and results in poor tracking performance.

Another problem that is highly correlated with tracking is video segmentation. Looking back at the literature, these two problems are almost always considered as separate problems. However, we argue that tracking and segmentation are actually closely related and solving them should help each other (See Fig. 2). On one hand, the object track, which is a set of bounding boxes with one bounding box in every frame, would provide strong high-level guidance for the target/background segmentation task. Pixels within

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) (b)

Fig. 1. Correlation between detection and tracking. (a) shows the tracking results of our proposed tracker (bottom row) and the method from [10] (top row). A pre-trained object detector fails when objects go under heavy articulation. This error is propagated to the data association step, which consequently cause failure in tracking. Differently, our proposed tracker is based on online discriminative learning and solves detection and global data association simultaneously, thus handles articulated targets well. The same observation can be made from (b). Each row represents one of the three identities in the scene. Each circle shows a corresponding match in a frame and the color represents the ID that is assigned to that detection. As can be seen, the top row has a lot of fragmented tracks while the bottom row only contains three tracks corresponding to the three identities.

a target box are highly likely to be labelled as the target. Conversely, the chance that pixels far away from the box belonging to the target is quite low. On the other hand, the object segmentation would separate object from other objects and background, which will be useful for determining track locations in every frame. This will help in resolving common issues in tracking. For example, during occlusion, the bounding box based appearance score of the occluded target is typically low, posing difficulty in tracking. However, the pixel labels in the visible part of the target would guide tracker to find the correct location of the target. In addition, labels of pixels in target/background contain information about target identities and locations, thus will help in avoiding track drifting and ID-switches.

In this paper, we propose to combine detection, data association and segmentation in one framework. The key idea to couple these three tasks is the high correlation between them. As discussed above, poor detection results in poor data association, therefore pixel level segmentation can help further improve tracking. At the heart of our tracker lies a Lagrange dual decomposition that combines an online discriminative tracker with segmentation. Our tracker is a new online discriminate learning tracker that solves detection and data association simultaneously. This online tracker is later combined with a segmentation method through Lagrange dual decomposition. In each iteration, the two subproblems of online tracking and segmentation are solved independently with the Lagrange variables serving as a connection between them.

For the tracking subproblem in dual decomposition, we propose an algorithm based on online discriminative learning, which solves detection and global data association simultaneously by integrating a new global data association technique into the inference of a structured learning tracker. Despite other online trackers which are temporally local, our tracker provides the tracks across a segment of a video. The input to our tracker, in every frame, is densely sampled candidate windows instead of sparse detections. This allows our tracker to impose temporal consistency between the frames and correct poor detections (mostly caused by occlusion or severe pose change), thus avoiding error propagation. We propose to perform the inference through a new Target Identity-aware Network Flow graph (TINF), which is a variant of multi-commodity flow graph [11].

For the segmentation subproblem, a foreground Gaussian Mixture Model (GMM) is constructed for each target along with one universal background GMM. These GMMs are used to compute target/background confidence maps. For a segment of video (a few frames), a superpixel based spatio-temporal graph is built and multi-label CRF is applied to the graph to obtain target/background labeling.

The tracking and segmentation subproblems are coupled through dual decomposition. We introduce a new coupling energy term, which penalizes background labels inside target bounding boxes as well as foreground labels outside target bounding boxes. Iterative optimization is applied to solve the problem. In each iteration, Lagrange variables are updated based on the inconsistency between tracking and segmentation results. The algorithm converges when tracking and segmentation results are consistent.

In the earlier version of this paper [12], we introduced the online discriminative learning component of Lagrange dual decomposition. This submission further extends [12] by combining segmentation and online discriminative learning tracking through dual decomposition. To summarize, this paper makes following important contributions. First, we propose a novel framework which combines multiple target tracking and segmentation in one energy function. The two tasks benefit from each other, thus leading to both better tracking and better segmentation results (See Fig. 2). The unified energy function is optimized effectively using dual decomposition. Second, to solve the tracking subproblem, we present a new multiple object tracking method which combines discriminative learning and global data association. We introduce a new Target Identity-aware Network Flow (TINF) and efficiently optimize it through Lagrangian relaxation. Our soft-spatial constraint replaces the ad-hoc non-maximum suppression step of object detection methods and further improves the results. Finally, the proposed approach is able to track multiple targets in terms of finer segments (regions) supported by corresponding target pixels rather than coarse bounding boxes, and achieve better or comparable results for both tracking and segmentation than state-of-art on challenging sequences.

The reminder of this paper is organized as follows. Section 2 describes the work most related to this paper. Section

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

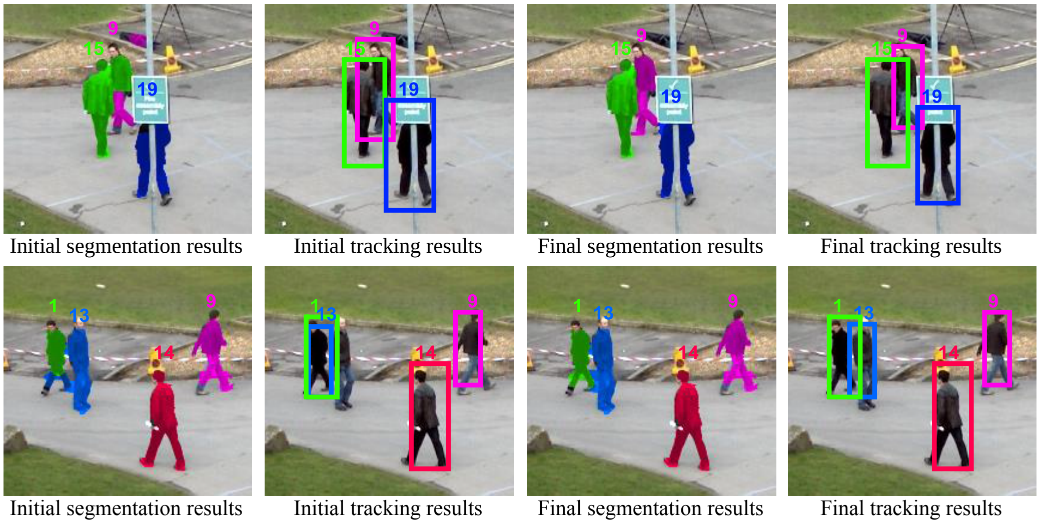


Fig. 2. Two examples of the tracking and segmentation tasks benefiting from each other (zoomed in views are shown). **First row**: By applying pure segmentation, the upper body of target 9 is mislabelled as target 15 due to similar color. But the tracking part is able to track target 9 correctly. After dual decomposition, the whole body of target 9 is labelled correctly and more accurate box is obtained for target 9. **Second row**: Without incorporating segmentation, the track for target 13 drifts to target 1. However, the segmentation results for target 13 are correct using pure segmentation. After dual decomposition, target 13 is tracked successfully and the segmentation results for target 1 are also improved. Combining the two subproblems lead to both better tracking and better segmentation results.

3.1 introduces the proposed multiple target tracking method based on a new target identity-aware network. Section 3.2 presents the approach used for target segmentation. In Section 3.3, the proposed novel framework, which combines multiple target tracking (Section 3.1) and segmentation (Section 3.2) through dual decomposition, is presented. The experimental results on diverse and challenging datasets are shown in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

2.1 Multiple target tracking

Most approaches for multiple target tracking (MOT) follow tracking-by-detection framework. First, a pre-trained object detector is applied to find a set of candidate locations for targets. Then these candidates are fed into a data association mechanism to form tracks. A majority of previous work on MOT focuses on designing data association techniques which can be divided into two groups of local and global techniques. Local data association methods [1], [13], [14], [15] are temporally local, which means they consider only a few frames while solving the association problem. Whilst most techniques in this class of methods are computationally inexpensive, their use of two frames makes them prone to ID-switches and other difficulties in tracking such as occlusions, pose changes and camera motion.

To better deal with above problems, global data association techniques have recently received a lot of attentions. In global association methods, the number of frames during data association is increased [3], [16], [17]. Recent approaches have formulated the data association as a network flow problem and its variations [2], [4], [18], [19], [20]. Despite popularity of these methods, their performance heavily depend on object detector outputs, which are usually poor when dealing with occlusion and articulated objects. Recent approaches have focused on improving the performance of the generic object detector [15] or designing a better data association techniques [2], [21] to improve tracking. Shu *et al.* in [15] proposed an extension to deformable part-based human detector [22], which can handle occlusion up to a scale. Additionally some recent work have directly used the dense detection output, before the non-maximum suppression, as the input to their tracking algorithm [23], [24], [25]. This is mostly to overcome the limitations of pre-trained detectors and non-maximum suppression algorithms when targets are occluding each other or are too close to each other. An alternative method to overcome the drawbacks of object detector when dealing with articulated objects or arbitrary objects (when a good pre-trained detector does not exist) is online learning of the object classifier [26], [27], [28]. Online discriminative learning approaches allow training target specific classifiers for a given sequence using different features including video specific features like color histogram. Moreover, these classifiers can adapt themselves as the appearance of targets change, which is not the case in pre-trained object detector.

Online discriminative learning methods have been used extensively for tracking deformable objects in the context of *single object tracking*. However, its extension to multiple objects remains relatively unexplored and is limited to only few works. The work of Zhang and Maaten [29] is probably the first attempt to apply online discriminative learning in tracking multiple objects. In [29], the spatial constraint among the targets is modeled during tracking. It is shown that the tracker performs well when the structure among the objects remains the same (or changes very slowly). However, this is only applicable to very limited scenarios and it will perform poorly in others, specially when the targets are moving independently.

Multi-commodity network flows have been used recently for multi-target tracking [20], [30], [31], [32]. In the earlier version of this work [12], we show that multi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

commodity network flows can be used in an inner loop of structured learning. Additionally the network design in our work is different from [20], [30], where our network includes the target identities by considering more than one node per candidate location and each node encodes the probability of assigning one of the target identities to that candidate location. Moreover, the network consists of multiple source and sink nodes, where each pair accounts for entry and exit of one target. Also, we show that a high-quality solution to the network can be found through Lagrange relaxation of some of the hard constraints, which is more efficient than Integer Program (IP) or Linear Program (LP) solutions. Thus we do not need to prune the graph as in [20], [30].

2.2 Object Segmentation in Video

Video object segmentation [33], [34], [35] aims to segment foreground pixels belonging to the object from the background in every frame. Video Object segmentation has been used in combination with single object tracking in [36], [37], [38]. However, the videos which are typically used in this work contain only one or two main moving objects. Different from these approaches, we solve video object segmentation along with multiple target tracking. The goal is to segment multiple interacting targets and preserve targets' identities at the same time. Authors in [39], [40], [41] track contours of targets using a level-set framework. Chen *et al.* [42] formulate tracking as constrained sequential labeling of supervoxels and obtain tracking results with object segmentation. In contrast, we propose an energy function coupling the tracking and segmentation subproblems, which is solved using dual decomposition by taking advantage of synergies between them. Milan *et al.* [43] propose a CRF model to jointly optimize over tracking and segmentation. First, a large number of trajectory hypotheses are generated by two trackers ([2] and [44]) using human detection results. Then they assign detections and superpixels to trajectory hypotheses. However, our approach does not rely on human detection or other trackers.

2.3 Dual Decomposition

Dual decomposition is a general and powerful technique widely used in optimization. It solves a problem by decomposing the original problem into multiple subproblems, solving the subproblems separately and then merging the solutions to solve the overall problem. Wu *et al.* [45] propose to incorporate both object detection and data association in a single objective function to avoid error propagation. The objective function is optimized by dual decomposition. Wang and Koller [46] construct a unified model over human poses as well as pixel-wise foreground/background segmentation and optimize the energy function using dual decomposition. To the best of our knowledge, we are the first ones to utilize dual decomposition to solve the multiple target tracking and segmentation problems.

3 Proposed Approach

Two main components of our framework are an online discriminative learning based tracker and a GMM-based spatial temporal video segmentation algorithm. These two components collaborate through a Lagrange dual decomposition to help improve performance of each task of detection, data association or segmentation. In the next subsection, we first explain the online tracker and then give details for each of its components. In the following subsection, we present the spatial temporal segmentation algorithm used in our approach. Finally, in the last subsection we show how dual decomposition is used to combine the above two modules.

3.1 Online Discriminative Learning Tracker

Our online discriminative learning tracker starts by training a model for each of the objects through structured learning (section 3.1.1). The tracker is provided the initial bounding boxes for the objects entering the scene in the first few frames (from annotation or using an object detector). During learning, the most violated constraints are found by searching for a set of tracks that minimize the cost function of our target identity-aware network flow. Later, the same network is used to find the best tracks in the next temporal span (segment) of a sequence (section 3.1.2). The new tracks are later used to update the model through passive aggressive algorithm [47].

3.1.1 Target-specific Model

Given a set of τ training images, $X = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^\tau\} \subset \mathcal{X}$, along with labels (target identities) $Y = \{\mathbf{y}_1^1, \mathbf{y}_2^1, ..., \mathbf{y}_K^1, ..., \mathbf{y}_{K-1}^\tau, \mathbf{y}_K^\tau\} \subset \mathcal{Y}$, where $\mathbf{y}_k^t$, defines the bounding box location of object k in frame t, the target models are obtained through structured learning [48]. The aim of learning is to find a prediction function $f : \mathcal{X} \mapsto \mathcal{Y}$, which directly predicts the locations of all the objects in a set of frames. The task of structured learning is to learn a prediction function of the form

$$f_{\mathbf{w}}(X) = \arg\max_{Y \in \mathcal{Y}} \sum_{t=1}^{\tau} \sum_{k=1}^{K} \mathbf{w}_k^T \phi(\mathbf{x}^t, \mathbf{y}_k^t), \tag{1}$$

where $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K\}$ is the concatenation of the models for all the K objects. $\phi(\mathbf{x}^t, \mathbf{y}_k^t)$ is the joint feature map which represents the feature extracted at location $\mathbf{y}_k^t$ in frame t. The optimal parameter vector $\mathbf{w}^*$ is obtained by solving the following optimization problem:

$$\begin{gathered}\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \qquad s.t. \quad \xi \geq 0 \\ \sum_{t=1}^{\tau} \sum_{k=1}^{K} \mathbf{w}_k^T \left(\phi(\mathbf{x}^t, \mathbf{y}_k^t) - \phi(\mathbf{x}^t, \bar{\mathbf{y}}_k^t)\right) \geq \Delta(Y, \bar{Y}) - \xi \\ \forall \bar{Y} \in \mathcal{Y} \setminus Y.\end{gathered} \tag{2}$$

The loss function is defined based on the intersection-over-union overlap between ground truth Y and prediction $\bar{Y}$:

$$\Delta(Y, \bar{Y}) = \frac{1}{\tau} \sum_{t=1}^{\tau} \sum_{k=1}^{K} (1 - (\mathbf{y}_k^t \cap \bar{\mathbf{y}}_k^t)). \tag{3}$$

Due to exponential number of possible combinations of bounding boxes in $\mathcal{Y}$, exhaustive verification of constraint in 2 is not feasible. However [48], [49] showed that high quality solution can be obtained in a polynomial time by using only the *most-violated constraints*, i.e a set of bounding boxes that maximize the sum of scores and loss functions. Once

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

the model parameters are learned ($\mathbf{w}$), we use the same inference that we use for finding the *most-violated constraints* to find the best set of tracks for all the K objects in next segment of the video.

3.1.2 Finding Tracks

Given the model parameters, $\mathbf{w}$, and *densely overlapping bounding boxes* in each frame, the goal is to find a sequence of candidate windows, called a track, for each object which maximizes the score in Eq. 1. This maximization requires searching over exponentially many configurations. We propose to formulate the inference as a global data association, which helps reducing the search space by enforcing some temporal consistency across the candidates in consecutive frames. Recently, such global data association has been formulated using network flow [2], [18], for which there exists an exact solution. In order to be able to use such networks as inference of our structured learning, the solution to the network needs to maximize the score function in Eq. 1. This requires the nodes in the graph to encode the probability of the target identity assignment using the learned parameters $\mathbf{w}_k$. This is not possible through traditional network flow methods.

We propose a new network called Identity-Aware network, which is shown in Fig. 3. The black circles represent all possible candidate locations in each frame (densely sampled across the entire frame). Each candidate location is represented with a pair of nodes that are linked through K *observation edges*; one *observation edge* for each identity. This is different from traditional network flow, in which only one *observation edge* connects a pair of nodes. Another major difference between our network from traditional network flow is that, our network has K sources and K sinks, each belonging to one object. The rest of the network is similar to that of traditional network flow. Transition edges that connect nodes from different frames, represent a potential move of an object from one location to the other and there is a transition cost associated with that. There is an edge between the start/sink node and every other node in the graph, which takes care of persons entering/leaving the scene. (For simplicity we only show some of the entry/exit edges).

The flow is a binary indicator which is 1, when a node is part of a track and 0 otherwise. A unit of flow is pushed through each source and the tracks for all the objects are found by minimizing the cost assigned to the flows. In addition, we show later that by setting the upper bound of flows passing through *observation edges* of one bounding box, we ensure that at most one track will claim one candidate location. In the following we will first present formulation of the problem as a Lagrangian relaxation optimization and later we will introduce our spatial constraint, which replaces the greedy non-maximum suppression in object detectors.

3.1.3 Target Identity-aware Network Flow

First we need to build our graph $G(V,E)$. For every candidate window in frame t, we consider a pair of nodes which are linked through K different *observation edges*, each belonging to one identity. For every node v_p, in frame t and v_q in frame $t+1$, there has to be a transition edge between

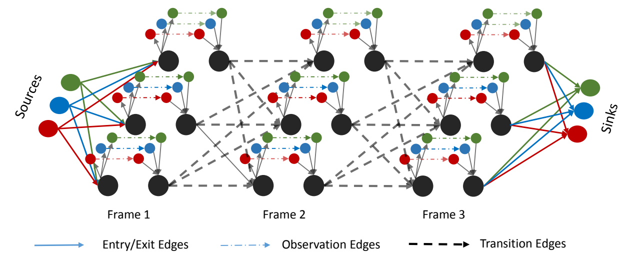


Fig. 3. Shows the network used in our inference for three identities. Each identity is shown with a unique color. The flow entering each node can take only one of the three observation edges depending on which source (identity) it belong to. The constraint in Eq. 8 ensures that one candidate can belong to only one track, so the tracks will not overlap.

the two if v_q belongs to the neighborhood, $N_\sigma(v_p^t)$ of v_p. Neighborhood of the node v_p is defined as

$$v_q^{t+1} \in N_\sigma(v_p^t) \Leftrightarrow \left\| v_p^t - v_q^{t+1} \right\|_2 \leq \sigma,$$

we consider a neighboring area within σ distance of node v_p that connects two candidate windows in two consecutive frames. In addition, we have source/sink edges which connect all the candidate windows to the source and sink nodes.

Different edges in our graph are assigned costs that take into account different characteristics of objects during tracking. Each pair of nodes which represents a candidate window will be assigned K different costs defined by the K target-specific models. Considering $\mathbf{w}_k$ to be the linear weights learned for the k^{th} object, the cost assigned to k^{th} *observation edge* representing the candidate location $\mathbf{y}_p^t$ in frame t is computed as follow:

$$c_{mn}^k = -\mathbf{w}_k^T \phi(\mathbf{x}^t, \mathbf{y}_p^t).$$

Transition edges which connect the nodes in consecutive frames are assigned costs, which incorporate both appearance and motion direction. The cost of a transition edge (c_{mn}^k) which connects two candidate windows $\mathbf{y}_p^t$ and $\mathbf{y}_q^{t+1}$ in two consecutive frames is computed as:

$$c_{mn}^k = -\alpha_1 H(\phi_c(\mathbf{x}^t, \mathbf{y}_p^t), \phi_c(\mathbf{x}^{t+1}, \mathbf{y}_q^{t+1})) - \alpha_2 \frac{V_{pq} V_{ref}^k}{\|V_{pq}\| \left\| V_{ref}^k \right\|}, \tag{4}$$

where $H(\phi_c(\mathbf{x}^t, \mathbf{y}_p^t), \phi_c(\mathbf{x}^{t+1}, \mathbf{y}_q^{t+1}))$ is the histogram intersection between the color histograms extracted from the location $\mathbf{y}_p^t$ and $\mathbf{y}_q^{t+1}$. $\frac{V_{pq} V_{ref}^k}{\|V_{pq}\| \left\| V_{ref}^k \right\|}$ is the cosine similarity between the reference velocity vector V_{ref}^k for the k^{th} object[1] and the velocity vector between the two candidate windows V_{pq}. Once the graph $G(V,E)$ is constructed, our aim is to find a set of K flows (tracks) by pushing a unit of flow through each source node. The flow $f_{m,n'}^k$ is found by minimizing the following cost function:

$$E_{track}(F) = \sum_{k=1}^{K} \sum_{(m,n)\in E} c_{mn}^k f_{mn}^k. \tag{5}$$

The flow passing through these edges need to satisfy some constraints to ensure that it can actually represent a

1. Average velocity vector for the k^{th} identity in previous batch.

track in a real world. The set of constraints that we define in our graph are as follow:

$$\sum_n f^k_{mn} - \sum_n f^k_{nm} = \begin{cases} 1 & \text{if } m = s_k \\ -1 & \text{if } m = t_k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$f^k_{mn} \in \{0,1\} \qquad \forall (m,n) \in E \ \ and \ \ 1 \le k \le K \tag{7}$$

$$\sum_{k=1}^{K} f^k_{mn} \le 1 \tag{8}$$

The constraint in Eq. 6 is the supply/demand constraint, enforcing the sum of flows arriving at one node to be equal to the sum of flows leaving that node. Constraint in Eq. 8 is the bundle constraint, ensuring that the tracks of different identities will not share a node by setting the upper bound of sum of flows passing through each node to be one.

One can formulate Eq. 5 as an Integer Program (IP). Since IP is NP-Complete, in practice, the problem can be relaxed to Linear Program (LP) in which the solution can be found in polynomial time. However, our experiments show that without pruning steps like the one in [20], [30], which reduces the number of candidate windows, it is intractable to find a solution for a large number of people in a long temporal span (one should note that the input to our tracker is dense candidate windows sampled from the entire frame). Instead, we propose a Lagrange relaxation solution to this problem. We show that after relaxing the hard constraints, the problem in each iteration, reduces to finding the best track for each target separately. The global solution to this can be found in linear time through dynamic programming. Moreover, our iterative optimization allows us to incorporate spatial constraint which further improves the tracking results.

3.1.4 Lagrange Relaxation Solution to TINF

The key idea of Lagrange relaxation is relaxing the hard constraints and moving them into the objective function, in order to generate a simpler approximation. We start by relaxing the bundle constraints in Eq. 8, where we introduce the non-negative Lagrange multiplier λ_{mn}, a vector of Lagrange multipliers that has the same dimension as the number of edges in the graph. After relaxing the bundle constraint the new objective function becomes:

$$E_{track}(F) = \sum_{k=1}^{K} \sum_{(m,n)\in E} c^k_{mn} f^k_{mn} + \sum_{(m,n)\in E} \lambda_{mn} (\sum_{k=1}^{K} f^k_{mn} - 1), \tag{9}$$

Subject to constraints in Eq. 6 and Eq. 7.

The second term in Eq. 9 is a constant for any given choice of Lagrange multipliers, therefore we can ignore it. The new objective function has a cost of $c^k_{mn} + \lambda_{mn}$, associated with every flow variable f^k_{mn}. Since none of the constraints in this problem contains the flow variables for more than one of the identities, we can decompose the problem into separate **minimum cost flow** problem for each identity. Since only one unit of flow is pushed through each source, the solution to a minimum cost flow can be found optimally through dynamic programming in $O(N)$. Thus the complexity of our optimization in each iteration is $O(KN)$, where K is the number of targets and N is the number of frames in the temporal span. Consequently, to apply the sub-gradient optimization to this problem, we alternate between the following two steps:

- For a fixed value of Lagrange multipliers we solve the minimum cost flow for each identity separately considering the cost coefficients $c^k_{mn} + \lambda_{mn}$.
- Update the Lagrange multipliers according to Eq. 10.

$$\lambda^{q+1}_{mn} = \left[\lambda^q_{mn} + \theta^q (\sum_{k=1}^{K} f^k_{mn} - 1) \right]^+ , \tag{10}$$

where λ^q is the Lagrange multipliers at iteration q, θ^q is the step size defining how far we would like to move from current solution and $[\alpha]^+ = max(0, \alpha)$.

3.1.5 Spatial Constraint

One major difference between our tracker and other data association based trackers is that, the input to our tracker is dense candidate windows instead of human detection output. When pedestrians with similar appearance and motion are walking next to each other, it is very likely to have ID-Switches in tracking results. Also when a pedestrian becomes partially occluded, the track for that person tend to pick candidates that highly overlap with other nearby pedestrians. This issue is addressed by non-maximum suppression in human detection [22] or by using other techniques like the one in [29], where the objects are forced to maintain the spatial configurations between consecutive frames. Instead, we introduce a soft-spatial constraint which penalizes the tracks that highly overlap. Our spatial constraint can be easily integrated into our iterative optimization. Similar to our Lagrange multipliers, we introduce a new set of variables that penalizes the cost of *observation edges* that highly overlap. Now the cost associated to each *observation edge* becomes $c^k_{mn} + \lambda_{mn} + \rho_{mn}$. ρ is a vector which has the same dimension as the number of *observation edges* in the graph. It is initialized with a zero vector in the first iteration and is updated according to Eq. 11.

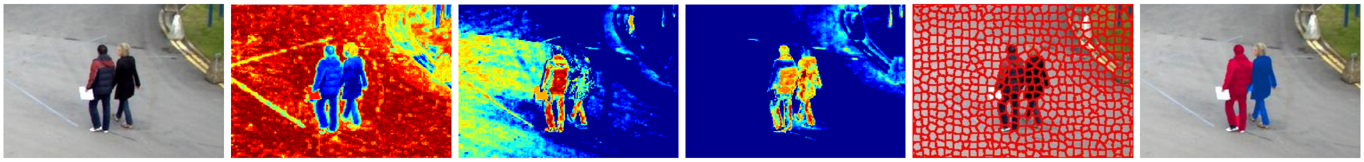
$$\rho^{q+1}_{mn} = \left[\rho^q_{mn} + \theta^q [(y^t_m \cap y^t_n) - 0.5]^+ \exp^{((y^t_m \cap y^t_n) - 0.5)/2} \right]^+ , \tag{11}$$

where $y^t_m \cap y^t_n$ is the overlap between neighboring bounding boxes in the same frame. ρ_{mn} penalizes the observation node which is associated with the cost c_{mn}. One should note that the spatial constraint only penalizes the bounding boxes that overlap more than 50% and the penalty increases exponentially as the overlap increases. After adding the spatial constraint the cost of the nodes are updated at each iteration according to the following:

$$c^{q+1,k}_{mn} = c^k_{mn} + \lambda^{q+1}_{mn} + \rho^{q+1}_{mn} . \tag{12}$$

We observe that penalizing both nodes that highly overlap, sometimes lead to inaccurate bounding boxes for one of the tracks. Therefore, we only penalize the observation nodes of the track that have lower score according to the score function in Eq. 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) (b) (c) (d) (e) (f)

Fig. 4. An illustration of target/background confidence maps and segmentation results. (a) A new frame (part of the frame is shown for clarity). (b) Background confidence map. Red represents higher confidence value while blue represents lower value. (c) and (d) show confidence maps for the target on the left and the target on the right respectively. (e) Superpixels in the part of the frame. (f) The final segmentation results after applying CRF to the superpixel based spatio-temporal graph. Red and blue masks represent foreground pixels for the two targets respectively.

3.2 Spatiotemporal Segmentation

Segmentation aims to find foreground pixels corresponding to each target, so that precise object contour can be determined, instead of typical bounding box representation. In this section, we describe the procedure to get foreground/background segmentation for all targets in a segment of video.

We determine the segmentation mask of target k in its first frame automatically from its initial box $\bar{\mathbf{y}}_k$. GrabCut algorithm [50] is applied to target k's small surrounding region, by initializing pixels within box $\bar{\mathbf{y}}_k$ as foreground while pixels outside box $\bar{\mathbf{y}}_k$ as background. GrabCut starts from this initial segmentation and iteratively refines foreground/background boundary. Then based on the foreground pixels obtained by GrabCut, we build a pixel-level foreground GMM model $\mathbf{w}_{fg(k)}$ for target k. In addition, a background image, obtained by averaging frames in the video, is used to build a universal background GMM model $\mathbf{w}_{bg}$. $CIELAB$ color space is used. A foreground confidence map $S_{fg(k)}$ for target k and a background confidence map S_{bg} are computed by applying $\mathbf{w}_{fg(k)}$ and $\mathbf{w}_{bg}$ to every pixel in a new frame respectively. An example is shown in Fig. 4.

Given K targets in the scene, the goal of segmentation is to assign one of $K+1$ labels (K targets or background) to every pixel. Inspired by [43], the segmentation problem in upcoming frames is solved by multi-label CRF. Since superpixels naturally preserve the boundary of objects and are computationally efficient for processing, we build a superpixel based spatio-temporal graph. Simple Linear Iterative Clustering (SLIC) [51] is employed to generate N superpixels in every frame. There are two types of edges in the graph: spatial edges, ε_S, and temporal edges, ε_T. Spatial edges connect all neighboring superpixels in a frame. Two superpixels s_m and s_n are considered as spatial neighbors if they share an edge in image space. Temporal edges connect all neighboring superpixels in two consecutive frames. Superpixels s_m and s_n are considered as temporal neighbors if at least $1/3$ of the pixels in s_m move to s_n in the next frame as predicted by optical flow. Temporal edges help preserve segmentation consistency across frames.

With the spatio-temporal graph, the multi-label Conditional Random Field (CRF) energy function is defined as

$$\begin{aligned} E_{seg}(Z) = &\sum_{s_m} Q(s_m, z_{s_m}) + \beta_1 \sum_{(s_m,s_n)\in\varepsilon_S} D(s_m, s_n) \\ &+ \beta_2 \sum_{(s_m,s_n)\in\varepsilon_T} D(s_m, s_n), \end{aligned} \tag{13}$$

where Z denotes the target/background labeling of all superpixels in a segment of video. z_{s_m} is the labeling of superpixel s_m. $z_{s_m} = k$ if s_m is labelled as target k and $z_{s_m} = 0$ if s_m is labelled as background. The energy function is optimized using graph cuts with α-expansion [52].

The unary term $Q(s_m, z_{s_m})$ in Eq. 13 is the cost of labeling superpixel s_m:

$$Q(s_m, z_{s_m}) = \begin{cases} -log(S_{fg(k)}(s_m)), & \text{if } z_{s_m} = k \\ -log(S_{bg}(s_m)), & \text{if } z_{s_m} = 0 \end{cases} \tag{14}$$

Here $S_{fg(k)}(s_m)$ represents the probability that superpixel s_m belonging to target k. It is computed as the average confidence value of $S_{fg(k)}$ over all pixels in s_m. $S_{bg}(s_m)$ denotes the probability that superpixel s_m belongs to the background.

The pairwise terms in Eq. 13 incorporate pairwise constraints by combining color similarity and the mean flow direction similarity between two neighboring superpixels. The pairwise potential $D(s_m, s_n)$ between two spatial/temporal neighboring superpixels s_m and s_n is defined as

$$\begin{aligned} D(s_m, s_n) =& \mathbf{1}(z_{s_m} \neq z_{s_n}) \cdot D_c(s_m, s_n) \cdot D_f(s_m, s_n), \\ D_c(s_m, s_n) =& \frac{1}{1 + \|LAB(s_m) - LAB(s_n)\|}, \\ D_f(s_m, s_n) =& \frac{V_{s_m} V_{s_n}}{\|V_{s_m}\| \, \|V_{s_n}\|}, \end{aligned} \tag{15}$$

where $\mathbf{1}(\cdot)$ is the one-zero indicator function. $LAB(s_m)$ is the average LAB color of superpixel s_m and $D_c(s_m, s_n)$ defines the color similarity between superpixels s_m and s_n. V_{s_m} denotes the mean optical flow of superpixel s_m and $D_f(s_m, s_n)$ is the direction similarity between the mean flows of superpixels s_m and s_n.

3.3 Dual Decomposition

As discussed previously, the two tasks: online discriminative tracker (Sec. 3.1) and spatial temporal target segmentation (Sec. 3.2) are highly correlated. To take advantage of synergies between them, dual decomposition is employed to couple these two tasks. We aim at minimizing the following energy function:

$$\min_{F,Z} E(F, Z) = \min_{F,Z} (E_{track}(F) + E_{couple}(F, Z) + E_{seg}(Z)), \tag{16}$$

where $E_{track}(F)$ and $E_{seg}(Z)$ are defined as in Eq. 5 and Eq. 13 respectively. F denotes the set of bounding boxes found by the tracking procedure in Sec. 3.1 and Z denotes

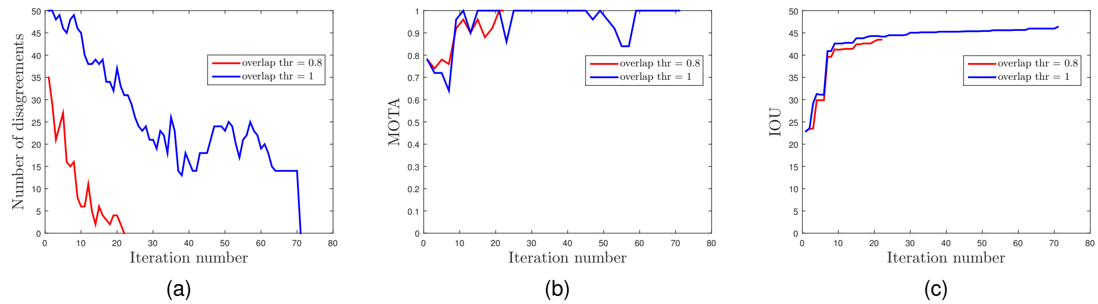


Fig. 5. Number of Disagreements, MOTA and IOU as function of number of iterations. The curves are generated based on a 10-frame segment in TUD-Crossing with 5 persons in the scene. (a) The number of disagreements between tracking and segmentation solutions drops over iterations. The algorithm converges when the two solutions are consistent. (b) The MOTA increases over iterations and reaches the best value at convergence. (c) The IOU (metric detailed in Sec. 4.2.2) increases over iterations. Since the segmentation annotations are available in every 10 frame, IOU is evaluated on the one frame in the 10-frame segment which has segmentation annotations.

target/background segmentation obtained in Sec. 3.2. The coupling term contains both bounding boxes and segmentation information:

$$\begin{aligned} E_{couple}(F,Z) = \sum_{k,m} (&\mathbf{1}(m \in f_k, z_m \neq k)\theta_{f_k}^m \\ &+ \mathbf{1}(m \notin f_k, z_m = k)\varphi_{f_k}^m). \end{aligned} \tag{17}$$

This energy introduces penalties for background labels inside target bounding boxes as well as foreground labels outside target bounding boxes. k denotes a target and m denotes a pixel. The first term penalizes pixels that are not labelled as target k, but are in target k's tracking boxes. f_k denotes the bounding boxes for target k, and $\mathbf{1}(m \in f_k, z_m \neq k)$ represents pixels in f_k which are not labelled as target k. Since a target's bounding box is highly likely to include some non-target pixels near the border of box, but not at the center of box, the resulting penalty is weighted by a human shape prior θ_{f_k}. Thus, background pixels at the center of box induce higher penalty while those close to the border of box result in lower penalty. The same human shape prior θ is used as in [43]. The second term penalizes pixels that are labelled as target k but are outside target k's boxes. $\mathbf{1}(m \notin f_k, z_m = k)$ represents pixels outside f_k which are labelled as target k. The corresponding penalty is weighted by φ_{f_k}, which has a zero weight within f_k and uniform non-zero weight outside f_k.

By introducing an equality constraint, Eq. 16 can be rewritten as

$$\begin{aligned} \min_{F^0,F^1,Z} E(F^0,F^1,Z) = \min_{F^0,F^1,Z} &(E_{track}(F^0) + E_{couple}(F^1,Z) \\ &+ E_{seg}(Z)) \\ &s.t. \quad F^0 = F^1. \end{aligned} \tag{18}$$

Now, the energy function is separable. We form the Lagrangian dual form of the above problem by introducing Lagrange multipliers λ'

$$\begin{aligned} L(\lambda') = &\min_{F^0,F^1,Z} (E_{track}(F^0) + E_{couple}(F^1,Z) + E_{seg}(Z) \\ &+ \lambda'(F^0 - F^1)), \\ = &\min_{F^0}(E_{track}(F^0) + \lambda' F^0) + \min_{F^1,Z}(E_{couple}(F^1,Z) \\ &+ E_{seg}(Z) - \lambda' F^1). \end{aligned} \tag{19}$$

Here λ' has the same dimension as F^0 and F^1.

Eq. 19 can be further decomposed into two independent subproblems:

$$g(\lambda') = \min_{F^0}(E_{track}(F^0) + \lambda' F^0), \tag{20}$$

$$h(\lambda') = \min_{F^1,Z}(E_{couple}(F^1,Z) + E_{seg}(Z) - \lambda' F^1). \tag{21}$$

The first subproblem (Eq. 20) is equivalent to a set of network flow problems, thus $g(\lambda')$ can be solved efficiently using dynamic programming. The second subproblem (Eq. 21) involves both tracking boxes and segmentation. When F^1 is fixed, $E_{couple}(F^1,Z)$ becomes a unary term on Z, thus $h(\lambda')$ can be solved by graph-cut. When Z is fixed, $h(\lambda')$ can be optimized by evaluating all candidate boxes. So a two-step procedure is employed to optimize $h(\lambda')$.

We use a sub-gradient method to optimize the Lagrangian dual problem. The algorithm works by repeating the following steps:

1) Get F^0 by solving the tracking subproblem $g(\lambda')$ (Eq. 20).
2) Get F^1 by solving the segmentation subproblem $h(\lambda')$ (Eq. 21).
3) Stop if $F^0 = F^1$.
4) Otherwise, update dual variable λ' by $\lambda' \leftarrow \lambda' + \alpha_p(F^0 - F^1)$, where α_p is the step size in iteration p and is computed as $\alpha_p = 1/(10+p)$.

In each iteration, we check the consistency between solutions of the two subproblems. The dual variable λ' changes, based on the inconsistent parts among F^0 and F^1, thus adjusting F^0 and F^1 accordingly to make them to be more and more consistent. Suppose in some iteration, boxes f_k are selected for target k by the tracking subproblem, but the segmentation subproblem selects another set of boxes. Then the corresponding element in λ' will increase, such that the penalty of selection of f_k by the tracking subproblem would increase and the penalty of selection of f_k by the segmentation subproblem would decrease. When F^0 and F^1 achieve agreement, λ' will not change and the optimal solution is found.

The spatial constraint described in Section 3.1.5 replaces the non-maximum suppression step in object detection methods and penalizes tracks that highly overlap. When two bounding boxes are highly overlapping, it adds cost to both observation nodes that are involved or the one with lower detection score. In some cases, this scheme leads to inaccurate tracks since it would push both tracks away, no matter if any of the tracks are actually correct or the detection score may not be very accurate. However, we can now utilize the segmentation results to make better decision on the spatial constraint. Assume that from the tracking results in iteration $p-1$, a box $\mathbf{y}_k$ is selected for target k. If the overlap between $\mathbf{y}_k$ and any box in other tracks is larger than 50%, and no pixel in $\mathbf{y}_k$ is labelled as target k from the segmentation results, there is a large chance that box $\mathbf{y}_k$ does not correspond to target k. Thus the cost of the observation node corresponding to $\mathbf{y}_k$ is updated as in Eq. 12. In this way, box $\mathbf{y}_k$ will introduce larger penalty and be less likely selected in tracking in iteration p. However, if there are pixels in $\mathbf{y}_k$ labelled as target k, then the observation node corresponding to $\mathbf{y}_k$ will not be penalized, no matter if it has a large overlap with other boxes. Note that the segmentation results are considered along with the tracking results, so the spatial constraint introduces penalty only if a box is too close to another box and is not supported by the segmentation results. This happens when two targets are close to each other, and the track of one target incorrectly jumps to the other target. On the contrary, when one target is occluded by another target, even though their boxes are close, they both have supporting pixels from the segmentation results, therefore spatial constraint is not applicable.

Due to the dense and overlapping candidate boxes used in our approach, we observe it is not necessary to have F^0 and F^1 to be exactly the same for convergence. In most cases, the results in early iterations are already good enough, though some boxes found by the two subproblems may shift a little. In our experiments, boxes returned by the two subproblems are considered consistent if their overlap is larger than 0.8 and the corresponding element in λ' would not be updated. This greatly reduces the number of iterations to solve the Lagrangian dual problem, with almost no performance loss. As shown in Fig. 5a, when overlap threshold of 0.8 is used, the number of disagreements drops more quickly compared to that case when overlap threshold is 1. The number of iterations to solve the Lagrangian dual problem is reduced by more than three times. Meanwhile, the performance remains almost the same as illustrated in Fig. 5b and 5c. Coupling tracking and segmentation lead to both better tracking and better segmentation results as demonstrated in experiments. It can also be observed in Fig. 5 that both MOTA and IOU are increasing over iterations. On one hand, the object tracks provide strong high-level guidance for target/background segmentation. On the other hand, segmentation helps resolve typical difficulties encountered in multiple target tracking in a couple of ways. First, in traditional tracking-by-detection approach, the tracking results highly depend on the detection performance. Miss-detections are common especially when there is occlusion. So special scheme, such as dummy nodes in network flow, needs to be designed in order to handle them. However, our approach does not rely on pre-trained object detector. We assume densely sampled candidate boxes instead of sparse detection boxes, so the tracker is able to infer temporal consistency between frames naturally. In addition, when target gets occluded, its visible part is segmented correctly even though its overall appearance score may be low. The segmentation results guide tracker to find correct box for the target. Second, the segmentation result provides more information about target location and target identity. Therefore, it helps tracker avoid drifting and ID-switch.

4 EXPERIMENTS

In our evaluation, we focus on tracking humans, due to its importance. We evaluate our proposed TINF tracker and the approach that couples multiple target tracking and segmentation on a set of standard multiple target tracking sequences. Along with tracking, we also provide both segmentation and detection results on a few sequences.

4.1 Experimental Setup

To initialize the target, similar to [29], [53], we use manual annotation. We annotate four initial bounding boxes for each object entering the scene. We also report results where targets are initialized automatically using a pre-trained object detector. For manual annotation, the target is initialized only once and there is no re-initialization of targets. We use histogram-of-oriented gradient [54] and color histogram [55] as our features. We found the combination of both features to be important. HOG captures the edge information of target and is helpful in detecting target from the background, while color histogram is a video specific feature and helps in distinguishing different targets from each other. The sequence is divided into segments of 20 frames each. At the end of each temporal span we check if a track is valid or not by comparing its appearance score from structural SVM ($\mathbf{w}_k^T\phi(\mathbf{x}^t, \mathbf{y}_k^t)$) with a pre-defined threshold. If the track is valid then it is used to update the model. When a target is close to the scene border and its velocity is towards outside of the scene, that target is treated as exiting the scene and the algorithm stops tracking that target. In this way, our approach is able to handle a variable number of targets.

4.2 Experimental Results

In this section, we conduct three sets of experiments. First we compare our approach with the state-of-the-art methods on publicly available sequences. For those sequences, where the object detection performs well, excellent results are already reported. However, we show that, using our approach, one can further improve the performance. Second, we evaluate our approach on two new sequences of [12] where targets experience heavy articulation and we show that we can significantly improve the performance of data-association based trackers as well as online trackers. Third, we test our approach on the popular and complex MOT16 Benchmark. *Parking Lot 1* [15], *Parking Lot 2* [61], *TUD Crossing* [10], *TUD-Stadtmitte* [62] and PET [63] are the five publicly available sequences used in our experiments and the two new sequences are called *Running* and *Dancing*.

4.2.1 Tracking

To quantitatively evaluate the tracking performance of our approach, both popular CLEAR MOT metrics [64] and Trajectory Based Metrics (TBM) [65] are used. CLEAR metrics

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Quantitative tracking results comparison of our methods ("TINF" and "TINF + Seg") with competitive approaches of LPD [56], LDA [57], DLP [58], H2T [5], GMCP [16], PF [59], SegTrack [43], CET [10], DCT [60], STRUCK [27] and SPOT [29] using tracking metrics.

Dataset	Method	MOTA	MOTP	MT	ML	IDS
Running	CET	46.3	50.8	0.67	0	0
	DCT	37.6	50.4	0	0	0
	SPOT	66.1	66.2	0.67	0	0
	STRUCK	79.9	64.3	1	0	0
	TINF	98.7	66.5	1	0	0
	TINF + Seg	**99.1**	**68.3**	**1**	**0**	**0**
Dancing	CET	36.6	62	0.57	0	64
	DCT	36.3	63.6	0	0.14	81
	SPOT	55.4	65.9	0.43	0	16
	STRUCK	69.1	67.1	0.71	0.14	9
	TINF	89.9	**65.9**	0.86	0	1
	TINF + Seg	**91.2**	65.7	**0.86**	**0**	**0**
Parking Lot 1	LPD	89.3	77.7	-	-	-
	GMCP	90.4	74.1	-	-	-
	H2T	88.4	81.9	0.78	0	21
	TINF	90.7	**69.3**	0.86	0	3
	TINF + Seg	**91.5**	67.4	**0.86**	**0**	**0**
Parking Lot 2	CET	71.7	55.8	0.6	0	59
	DCT	73.6	56.5	0.8	0	48
	TINF	89.3	66.3	1	0	0
	TINF + Seg	**90.5**	**68.7**	**1**	**0**	**0**
TUD Crossing	SegTrack	59.2	73.1	0.67	0	8
	PF	84.3	71	-	-	2
	GMCP	91.6	75.6	-	-	0
	TINF	92.9	**69.2**	1	0	0
	TINF + Seg	**93**	68.2	**1**	**0**	**0**
TUD Stadtmitte	SegTrack	68	55.9	0.6	0	3
	GMCP	77.7	63.4	-	-	0
	TINF	81.6	75.4	0.8	0	0
	TINF + Seg	**83.8**	**78.7**	**0.8**	**0**	**0**
PETS	SegTrack	85.3	77.5	1	0	9
	LDA	90	75	0.89	-	6
	DLP	91	70	-	-	5
	GMCP	90.3	69	-	-	8
	TINF	90.4	63.1	0.95	0	3
	TINF + Seg	**92.5**	**68.2**	**0.95**	**0**	**0**

TABLE 2
Performance of TINF tracker with automatic and manual initialization of the targets. For automatic initialization of targets a pre-trained human detector is used [22].

Method	MOTA	MOTP	MT	ML	IDS
Parking Lot 1 - Auto	90.5	65.2	0.86	0	5
Parking Lot 1 - Manual	90.7	69.3	0.86	0	3
Parking Lot 2 - Auto	83.4	63.2	0.7	0	5
Parking Lot 2 - Manual	89.3	66.3	1	0	0
TUD Crossing - Auto	90.8	68.8	0.92	0.08	0
TUD Crossing - Manual	92.9	69.2	1	0	0

TABLE 3
Performance of TINF tracker with and without spatial constraint.

Method	MOTA	MOTP	MT	ML	IDS
Running	97.2	68.1	1	0	0
Running - SP	98.7	66.5	1	0	0
Dancing	88	64.9	0.86	0	2
Dancing - SP	89.9	65.9	0.86	0	1
Parking Lot 1	88	62.9	0.79	0	4
Parking Lot 1 - SP	90.7	69.3	0.86	0	3
Parking Lot 2	82.2	65.6	0.9	0	2
Parking Lot 2 - SP	89.3	66.3	1	0	0
TUD Crossing	86.6	69.8	0.92	0	1
TUD Crossing - SP	92.9	69.2	1	0	0

consider the entire video as a whole, while TBM consider the behavior of each track separately. Each of these metrics captures different characteristics of a tracker and it is important to look at both of them to better capture strength and weakness of a tracker.

First, we evaluate and compare the proposed TINF tracker with two main sets of trackers: data-association based trackers and online trackers. On sequences for which no other tracking results are reported, we compare our method with three data-association based trackers for which we have access to their code, CET [10] and DCT [60]. We use Deformable Part based model [22] as the human detector. For online discriminative learning-based trackers, we selected STRUCK [27] as well as structure preserve multi-object tracking (SPOT) approach [29]. For details about the parameter selection of each method please refer to [12]. The results comparison is shown in Table 1.

Initialization. For initialization, besides manual annotation, we use human detection to automatically initialize the targets. During each segment, a new track is initialized if there are at least four confident detections in consecutive frames that highly overlap and are not associated to any other tracks. We test automatic initialization of targets on publicly available sequences, where human detection performs reasonably well. As can be seen in Table 2, the performance of our method doesn't change much, when using automatic initialization. The main difference is that some of the tracks in some sequences start late compared to manual annotation which causes a small drop in MOTA due to the added false negatives.

Effect of Spatial Constraint. In order to clearly see the effect of our spatial constraint, we run our method on different sequences with and without the spatial constraint. As can be seen in Table. 3, when spatial constraint is added, the performance increases, specially for sequences involving interaction between objects.

The results of our proposed approach that couples multiple target tracking and segmentation are shown in Table 1, denoted as "TINF + Seg". The coupled approach achieves better tracking results compared to TINF. In particular, the number of ID-switches is substantially reduced compared to other methods and TINF.

MOT16 Benchmark. We also test our approach on the popular MOT16 Benchmark. It contains 7 test sequences, including sequences of crowded scenes, sequences captured by moving and static cameras. Due to the large number of targets, human detection is used to automatically initialize targets. The results comparison is shown in Table 4. We compare our results with other published online trackers that use non-standard detections [66], [67], [68] as well as a top performer that uses the standard detections [69]. All better results reported on MOT16 use deep learning based human detection or deep learning based data association. Considering that our approach does not need training and does not involve deep learning features, the results are quite competitive. In addition, our approach achieves low number of ID-switches compared to most state-of-the-art methods. In particular it is interesting to mention that our approach, using simple hand-crafted features, outperforms

TABLE 4
Tracking performance comparison on MOT16 Benchmark.

Method	MOTA	MOTP	FP	FN	MT	ML	IDS
[66]	52.5	78.8	4407	81223	0.19	0.35	910
[67]	59.8	79.6	8698	63245	0.25	0.23	1423
[68]	66.1	79.5	5061	55914	0.34	0.21	805
[69]	47.2	75.8	2,681	92,856	0.14	0.42	774
Ours	57.6	77.9	12121	64401	0.3	0.22	733

the top performer which uses standard publicly available detections along with powerful deep pipeline. Finally the large number of FPs in our approach is mainly due to the way we sample dense candidates. This sometimes leads to inaccurate bounding boxes. (The number of FPs will significantly reduce if we lower the overlap threshold for computing the metrics.)

4.2.2 Segmentation

Besides improving the tracking performance, our proposed dual decomposition based approach is able to track multiple targets with pixel-level target/background labeling. In order to evaluate the segmentation performance, we use the segmentation annotations for TUD-Crossing from [40] and manually annotate pixel-level target masks every 10 frames in the other sequences. The segmentation annotations will be released to facilitate future research in this area.

For evaluation, the segments are optimally assigned to ground truth masks and multiple segments can be assigned to the same ground truth mask (pixel-wise labeled segmentation). Identity-based IOU is the average intersection-over-union overlap with target identity information incorporated. Traditional IOU used in video segmentation evaluation [37] computes the mean IOU of foreground regions over all frames. However, it has no notion of target identities. Therefore, in order to better evaluate the segmentation performance for multiple targets, we extend the traditional foreground IOU to **identity-based IOU**. Identity-based IOU computes the intersection-over-union overlap between ground truth mask and segments assigned to it for every target in every frame and then takes the average over all of them. Overall error is the percentage of wrongly labelled pixels, while average error computes the percentage of mis-classified pixels per ground truth mask. Over-segmentation counts the number of segments merged to cover the ground truth masks.

We compare the above four metrics with [43][2] and [40][3] in Table 5. The proposed approach achieves much higher identity-based IOU and much lower overall error as well as average error compared to previous methods. “TINF + Seg” outperforms “Seg Only”, by a large margin, demonstrating that incorporating tracking leads to more accurate segmentation results. Some qualitative results are shown in Fig. 7.

2. We test the code available on the author’s website with default parameters on TUD-Crossing. The results on the other two sequences are obtained from the author.

3. Note that the identity-based IOU of Horbert et al.’s [40] results is computed using the segmentation results provided by the author, while the IOU reported in [40] is the traditional foreground IOU without notion of target identities.

TABLE 5
A quantitative comparison of segmentation results of our method with competitive approaches in Milan et al. [43] and Horbert et al. [40].

Dataset	Method	Identity -based IOU	Overall err.	Avg. err.	Over -seg.
PETS	[43]	54.82	0.78	40.08	1.65
	Seg Only	19.51	1.68	66.86	1
	TINF + Seg	**73.51**	**0.43**	**17.79**	**1**
TUD Crossing	[43]	25.35	6.68	63.87	2.23
	[40]	46.50	4.13	35.88	3.23
	Seg Only	15.64	7.96	71.85	1
	TINF + Seg	**55.36**	**3.88**	**26.93**	**1**
TUD Stadtmitte	[43]	27.33	6.10	48.59	1.09
	Seg Only	18.87	6.85	56.48	1
	TINF + Seg	**41.62**	**3.35**	**23.65**	**1**
Parking Lot 1	**Seg Only**	20.97	5.12	49.38	1
	TINF + Seg	**68.27**	**1.36**	**20.57**	**1**
Parking Lot 2	**Seg Only**	14.79	8.91	59.54	1
	TINF + Seg	**58.66**	**4.94**	**26.09**	**1**
Running	**Seg Only**	24.67	5.35	58.56	1
	TINF + Seg	**67.18**	**2.31**	19.57	1
Dancing	**Seg Only**	15.5	12.93	67.8	1
	TINF + Seg	**58.17**	**7.88**	**17.33**	**1**

TABLE 6
This table shows the detection performance comparison between our detector and DPM [22] in terms of average precision.

Seq	PL1	PL2	PETS	TUD Crossing	TUD Stadmitte
DPM	87.61	74.61	68.54	85.95	77.62
Ours	88.12	81.77	84.23	83.78	79.99

Note that targets are segmented and tracked correctly even when being occluded or when they are close to other targets.

Moreover, we show number of extracted objects with varying threshold α on ratio of correctly labelled pixels per ground truth mask in Fig. 6. An object is extracted if more than α of its ground truth mask is correctly covered. Our approach (“TINF + Seg”) is able to extract more objects for all different thresholds compared to previous methods and “Seg Only”.

Since MOT16 Benchmark is designed for evaluating multiple object tracking performance, there are no segmentation annotations available. Qualitative results on one sequence are shown in Fig. 7.

4.2.3 Detection

We also present the detection performance of our proposed approach. The comparison with DPM [22] is shown in Table 6. Our detector is much simpler compared to DPM, while coupled with segmentation and data association, it can achieve better performance on almost all the sequences.

In addition, we evaluate our approach on MOT17DET Benchmark. The detection performance comparison is summarized in Table 7. With the help from tracking and segmentation, our approach outperforms DPM [22] and Faster R-CNN [70] on the videos of complex scenes.

4.3 Run Time and Convergence

In order to compare the complexity of the proposed Lagrangian relaxation method with integer program (IP) and linear program (LP), we implemented the IP and LP version

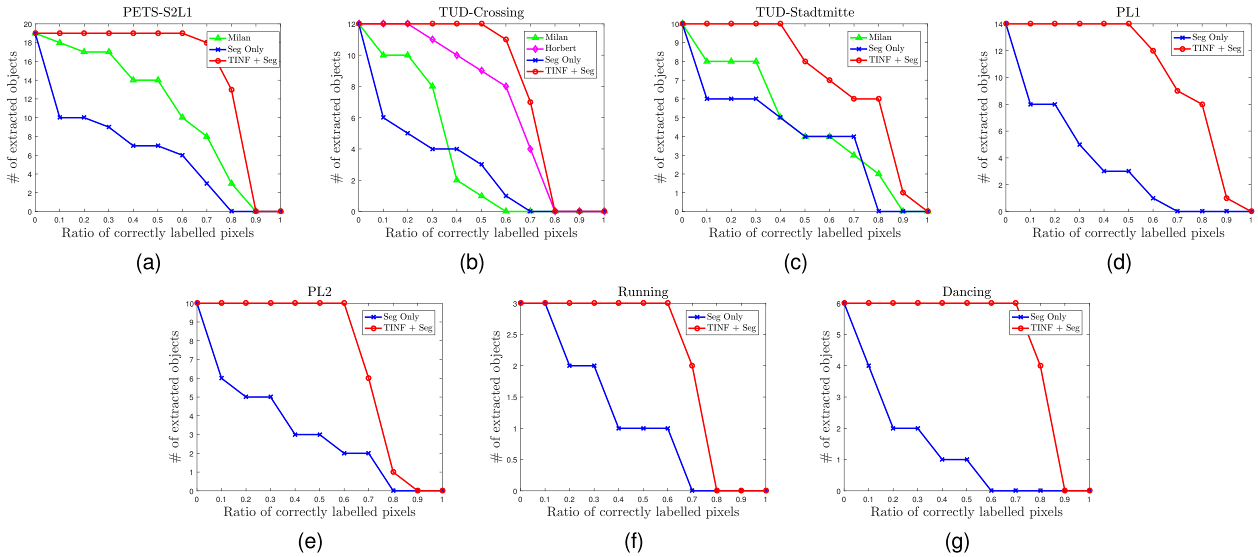


Fig. 6. The curves show the number of extracted objects as a function of correctly labeled pixels per ground truth mask.

Fig. 7. Examples of segmentation and tracking results on PETS-S2L1, TUD-Crossing and MOT16. Each target is shown by a unique color.

TABLE 7
Detection performance comparison with DPM [22] and Faster R-CNN [70] on MOT17DET Benchmark.

Method	AP	Prec.	Rec.	TP	FP	FN
DPM	0.61	64.8	68.1	78007	42308	36557
Faster R-CNN	0.72	89.8	77.3	88601	10081	25963
Ours	0.74	89.3	83.4	95506	11435	19058

of TINF as well. We employ CPLEX [71] as the optimization toolbox. The performance of IP and LP is within $1-2\%$ performance of our Lagrange relaxation formulation when no spatial constraint is used. The runtime for a selected segment of PL2 sequence with different number of targets is shown in Fig. 8. Note that the curves are shown with logarithmic coordinates. As can be observed, the proposed Lagrangian relaxation optimization is a lot more efficient compared to the IP and LP solutions.

In Fig. 9, we further demonstrate the convergence of the proposed TINF and TINF + Seg trackers on PL2 sequence. The number of iterations taken for convergence varies depending on the complexity of the segments. For example, for TINF + Seg, it takes only a few iterations to converge for segments near the beginning or the end of PL2 sequence, since the scene is simpler and it is easy to reach agreement between tracking and segmentation results. While some segments in middle of the PL2 sequence take 40 to 75 iterations to converge. That is because the scene is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

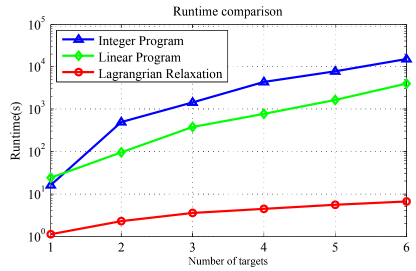


Fig. 8. TINF runtime comparison of the proposed Lagrangian relaxation solution vs IP and LP.

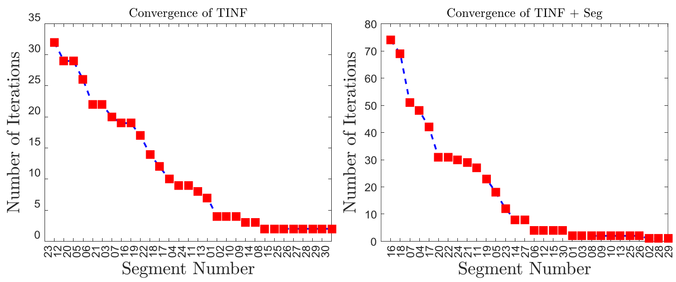


Fig. 9. Convergence of TINF and TINF + Seg on PL2 sequence.

more complex, there are more interacting targets and a lot of occlusions.

4.4 Limitation and Future Work

There are mainly two reasons that the proposed approach does not achieve state-of-the-art tracking results on MOT16 Benchmark. First, the tracking and segmentation are purely based on non deep learning based features. It is not robust enough to handle complex and dynamic scenes well. Second, it lacks an effective mechanism to terminate and re-initialize a track when the track drifts. This would lead to both false positive and false negative at the same time for that track.

One line of research for future work is to explore automatic ways to terminate and re-initialize tracks to avoid drift. This will allow one to utilize the proposed algorithm in scenarios where the camera angle is low and frequent long-term intra object occlusions occur. These sequence are not common in surveillance scenarios. However, recent multi object tracking dataset contains these types of sequences. We also plan to explore the use of more powerful discriminative features, such as deep learning based features, to further improve the performance. In addition, a regressor can be added on the top to improve bounding box accuracy, which is our future work.

5 Conclusion

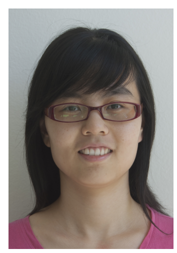
We present a novel framework that combines two main components of most existing trackers, detection and data association, along with segmentation in a single framework. The three tasks are closely related, and solving one helps improve the others. Detection and data association are combined through a structured learning framework, using a novel network flow graph. Additionally, the online discriminative tracking algorithm and segmentation are jointly optimized using dual decomposition, which leads to more accurate segmentation results and also helps resolve typical difficulties in tracking, such as occlusion handling, ID-switch and track drifting. Moreover, more detailed representation of targets - pixel-level target foreground labeling, is obtained rather than coarse bounding boxes.

References

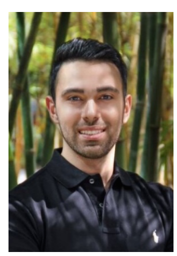
[1] Y. Bar-Shalom, T. Fortmann, and M. Scheffe, “Joint probabilistic data association for multiple targets in clutter,” in *Information Sciences and Systems*, 1980.
[2] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” in *CVPR*, 2011.
[3] A. Dehghan, S. Modiri, and M. Shah, “GMMCP-Tracker:Globally Optimal Generalized Maximum Multi Clique Problem for Multiple Object Tracking,” in *CVPR*, 2015.
[4] A. A. Butt and R. T. Collins, “Multi-target tracking by lagrangian relaxation to min-cost network flow,” in *CVPR*, 2013.
[5] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li, “Multiple target tracking based on undirected hierarchical relation hypergraph,” in *CVPR*, 2014.
[6] K. Cheng-Hao, C. Huang, and R. Nevatia, “Multi-target tracking by on-line learned discriminative appearance models.” in *CVPR*, 2010.
[7] H. Grabner, M. Grabner, and H. Bischof., “Real-time tracking via on-line boosting.” in *BMVC*, 2006.
[8] M. D. B. et al, “Robust tracking-by-detection using a detector confidence particle filter.” in *ICCV*, 2009.
[9] S.-I. Y. et al, “The Solution Path Algorithm for Identity-Aware Multi-Object Tracking.” in *CVPR*, 2016.
[10] A. Andriyenko and K. Schindler, “Multi-target Tracking by Continuous Energy Minimization,” in *CVPR*, 2011.
[11] G. Karakostas, “Faster approximation schemes for fractional multicommodity flow problems,” in *ACM-SIAM*, 2002.
[12] A. Dehghan, Y. Tian, P. H. Torr, and M. Shah, “Target identity-aware network flow for online multiple target tracking,” in *CVPR*, 2015.
[13] D. Reid, “An algorithm for tracking multiple targets,” *IEEE Transactions on Automated Control*, 1996.
[14] K. Shafique and M. Shah, “A noniterative greedy algorithm formultiframe point correspondence,” in *PAMI*, 2005.
[15] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, “Part-based Multiple-Person Tracking with Partial Occlusion Handling,” in *CVPR*, 2012.
[16] A. R. Zamir, A. Dehghan, and M. Shah, “Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs,” in *ECCV*, 2012.
[17] B. Wu and R. Nevatia, “Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors,” *IJCV*, 2007.
[18] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in *CVPR*, 2008.
[19] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, “Multiple Object Tracking Using K-Shortest Paths Optimization,” in *PAMI*, 2011.
[20] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, “Multi-commodity network flow for tracking multiple people,” in *PAMI*, 2013.
[21] S.-H. Bae and K.-J. Yoon, “Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning,” in *CVPR*, 2014.
[22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *PAMI*, 2010.
[23] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool., “Coupled object detection and tracking from static cameras and moving vehicles.” in *PAMI*, 2008.
[24] S. Tang, B. Andres, M. Andriluka, and B. Schiele, “Subgraph decomposition for multi-target tracking,” in *CVPR*, 2015.
[25] S. Tang, Andriluka, B. Andres, and B. Schiele, “Multiple people tracking by lifted multicut and person re-identification,” in *CVPR*, 2017.
[26] Z. K. andKrystian Mikolajczyk and J. Matas, “Tracking-Learning-Detection,” in *PAMI*, 2010.
[27] S. Hare, A. Saffari, and P. H. S. Torr, “Struck: Structured output tracking with kernels,” in *ICCV*, 2011.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849374, IEEE Transactions on Pattern Analysis and Machine Intelligence

[28] S. Wang, H. Lu, F. Yang, and M. Yang, "Superpixel tracking," in *CVPR*, 2011.
[29] L. Zhang and L. Maaten, "Structure preserving object tracking," in *CVPR*, 2013.
[30] X. Wang, E. Turetken, F. Fleuret, and P. Fua, "Tracking interacting objects optimally using integer programming," in *ECCV*, 2014.
[31] D. M. W. Z. Yu, Shoou-I. and A. Hauptmann, "The solution path algorithm for identity-aware multi-object tracking." in *CVPR*, 2016.
[32] Y. Y. X. L. Yu, Shoou-I. and A. G. Hauptmann., "Long-term identity-aware multi-person tracking for surveillance video summarization." in *arXiv:1604.07468*, 2016.
[33] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *ECCV*, 2010.
[34] T. Ma and L. J. Latecki, "Maximum weight cliques with mutex constraints for video object segmentation," in *CVPR*, 2012.
[35] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *CVPR*, 2013.
[36] Z. Yin and R. T. Collins, "Shape constrained figure-ground segmentation and tracking," in *CVPR*, 2009.
[37] F. Li, T. Kim, A. Humayun, D. Tsai, and J. Rehg, "Video segmentation by tracking many figure-ground segments," in *ICCV*, 2013.
[38] L. Wen, D. Du, Z. Lei, S. Z. Li, and M.-H. Yang, "Jots: Joint online tracking and segmentation," in *CVPR*, 2015.
[39] C. Bibby and I. Reid, "Real-time tracking of multiple occluding objects using level sets," in *CVPR*, 2010.
[40] E. Horbert, K. Rematas, and B. Leibe, "Level-set person segmentation and tracking with multi-region appearance models and top-down shape information," in *ICCV*, 2011.
[41] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation," in *ECCV*, 2010.
[42] S. Chen, A. Fern, and S. Todorovic, "Multi-object tracking via constrained sequential labeling," in *CVPR*, 2014.
[43] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *CVPR*, 2015.
[44] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *ECCV*, 2012.
[45] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, "Coupling detection and data association for multiple object tracking," in *CVPR*, 2012.
[46] H. Wang and D. Koller, "Multi-level inference by relaxed dual decomposition for human pose segmentation," in *CVPR*, 2011.
[47] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, , and Y. Singer, "Online passive-aggressive algorithms," in *JMLR*, 2006.
[48] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *JMLR*, 2005.
[49] C. nam Yu and T. Joachims, "Learning structural svms with latent variables," in *ICML*, 2009.
[50] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, 2004.
[51] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *PAMI*, 2012.
[52] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *PAMI*, 2001.
[53] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *PAMI*, 2011.
[54] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.
[55] J. Domke and Y. Aloimonos, "Deformation and viewpoint invariant color histograms," in *BMVC*, 2006.
[56] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, "Learning people detectors for tracking in crowded scenes," in *ICCV*, 2013.
[57] A. V. Segal and I. Reid, "Latent Data Association: Bayesian Model Selection for Multi-target Tracking," in *ICCV*, 2013.
[58] A. K. K.C. and C. D. Vleeschouwer, "Discriminative Label Propagation for Multi-Object Tracking with Sporadic Appearance Features," in *ICCV*, 2013.
[59] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool., "Robust tracking-by-detection using a detector confidence particle Filter." in *ICCV*, 2009.
[60] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-Continuous Optimization for Multi-Target Tracking," in *CVPR*, 2012.
[61] G. Shu, A. Dehghan, and M. Shah, "Improving an Object Detector and Extracting Regions using Superpixels," in *CVPR*, 2013.
[62] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *CVPR*, 2010.
[63] J. Ferryman and A. Shahrokni, "Dataset and challenge." *International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.
[64] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *Journal on Image and Video Processing*, 2008.
[65] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *CVPR*, 2006.
[66] R.Sanchez-Matilla, F.Poiesi, and A.Cavallaro, "Online multi-target tracking with strong and weak detections," in *ECCV*, 2016.
[67] A. Bewley, G. Zongyuan, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *ICIP*, 2016.
[68] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: Multiple object tracking with high performance detection and appearance feature," in *ECCV*, 2016.
[69] A. A. A. Sadeghian and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies." in *CVPR*, 2017.
[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[71] "Ibm ilog cplex optimizer, www.ibm.com/software/integration/optimization/cplex-optimizer."



Yicong Tian received her BSc from Beijing University of Posts and Telecommunications in 2011. She joined Center for Research in Computer Vision at University of Central Florida in 2011, where she is currently working toward the PhD degree in Computer Science. She has published several papers in top computer vision conferences such as CVPR. Her research interests include action detection in video, multiple target tracking and segmentation, and image geo-localization.



Afshin Dehghan received his PhD degree in computer science from University of Central Florida in 2016. He has published several papers in conferences and journals such as CVPR, ECCV, ACM Multimedia, and IEEE Transactions on Pattern Analysis and Machine Intelligence. He was a program committee member of the ACM MM. His research interests include multi-target tracking, deep learning, face/facial attribute recognition, complex event recognition, mathematical optimization, and graph theory. He received UCFs computer science doctoral student of the year award in 2014-2015.

Mubarak Shah , the trustee chair professor of computer science, is the founding director of the Center for Research in Computer Vision at University of Central Florida. He is an editor of an international book series on video computing, was editor-in-chief of Machine Vision and Applications journal, and an associate editor of ACM Computing Surveys journal. His research interests include video surveillance, visual tracking, human activity recognition, visual analysis of crowded scenes, video registration, UAV video analysis, and so on. He is a fellow of the IEEE, AAAS, IAPR, and SPIE.