Detection-Free Multiobject Tracking by Reconfigurable Inference With Bundle Representations

Liang Lin, Yongyi Lu, Chenglong Li, Hui Cheng, and Wangmeng Zuo, *Senior Member, IEEE*

***Abstract*—This paper presents a conceptually simple but effective approach to track multiobject in videos without requiring elaborate supervision (i.e., training object detectors or templates offline). Our framework performs a bi-layer inference of spatio-temporal grouping to exploit rich appearance and motion information in the observed sequence. First, we generate a robust middle-level video representation based on clustered point tracks, namely video bundles. Each bundle encapsulates a chunk of point tracks satisfying both spatial proximity and temporal coherency. Taking the video bundles as vertices, we build a spatio-temporal graph that incorporates both competitive and compatible relations among vertices. The multiobject tracking can be then phrased as a graph partition problem under the Bayesian framework, and we solve it by developing a reconfigurable belief propagation (BP) algorithm. This algorithm improves the traditional BP method by allowing a converged solution to be reconfigured during optimization, so that the inference can be reactivated once it gets stuck in local minima and thus conduct more reliable results. In the experiments, we demonstrate the superior performances of our approach on the challenging benchmarks compared with other state-of-the-art methods.**

***Index Terms*—Graphical inference, object tracking, spatio-temporal analysis, video processing.**

I. Introduction

VISUAL object tracking has long been an active research topic in computer vision, and impressive progresses are made recently. One of the most popular approaches follows the tracking-by-detection framework, where the object tracking can be naturally specified as an online learning and detection task [1]–[4]. The standard procedure of applying these trackers mainly includes the following steps: 1) manually select the desired object at the beginning and usually only one object is specified; 2) train an object classifier by supervised learning; and 3) localize the object in the rest video frames while updating the object classifier. Though very impressive results have been achieved in current research, the problem of long-term robust tracking in unconstrained environments still remains open, particularly with the following scenarios.

1) Frequent partial occlusion and object deformation lower the precision of detector.
2) The detection responses are possibly inconsistent in time, resulting in the risk of tracking drift.
3) For some objects with large intraclass variance (e.g., sports players), the cost of training reliable detectors is expensive.

In this paper, we present a detection-free tracking framework that parses object trajectories in the observed video sequence via spatio-temporal grouping without adopting object detectors. Our framework infers multiple-object tracking with two stages: 1) extract a batch of video bundles by encapsulating dense point tracks to compose object trajectories and 2) associate identities of the bundles for trajectory parsing by a reconfigurable belief propagation (RBP) algorithm. The inference is conducted based on a set of deferred observations (e.g., the entire video or a period of frames). Fig. 1 demonstrates the advantages of our approach: some detection-guided methods may not work as the human detections are unreliable, while the satisfied results are produced by our approach.

The main insight of our bottom-up framework is to oversegment the object trajectories as the intermediate-level representation and then search for the optimal partition during inference. As a result, the spatial partition and the temporal tracking are jointly solved to handle the realistic challenges.

First, we adopted a kind of mid-level features, i.e., dense point tracks, to represent the video sequence. Dense point tracks have longer lifespans than pixel-based features such as superpixels [1] and exploit long term motion difference. Based on the well-known Gestalt principle of "common fate" [5], motion is a salient factor and provides significant information about moving object in videos. Studies with congenitally blind people also show that they learn more easily from moving objects than static images. Based on the dense point tracks, we extracted a set of video bundles, according to the similarity of the tracks.

Manuscript received February 9, 2015; revised July 16, 2015; accepted September 11, 2015. Date of publication September 29, 2015; date of current version October 13, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61271093, in part by the Guangdong Natural Science Foundation under Grant S2013050014548 and Grant 2014A030313201, in part by the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067, and in part by the Science and Technology Program of Guangzhou under Grant 1563000439. This paper was recommended by Associate Editor W. Hu. *(Corresponding author: Wangmeng Zuo.)*

L. Lin, Y. Lu, and H. Cheng are with Sun Yat-Sen University, Guangzhou 510006, China (e-mail: linliang@ieee.org; yylu1989@gmail.com; chengh9@mail.sysu.edu.cn).

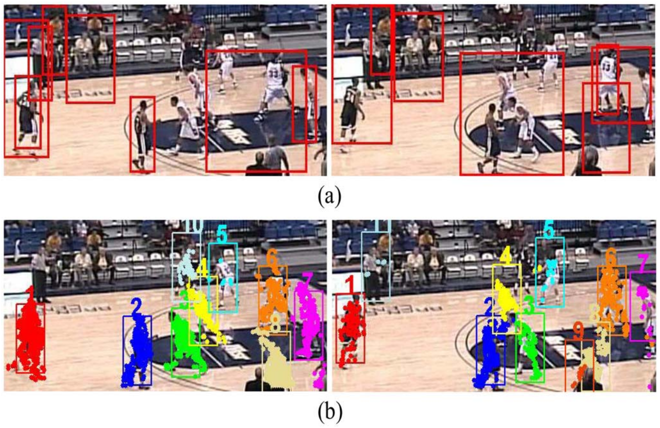
C. Li is with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com).

W. Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: cswmzuo@gmail.com).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2478515


(a)

(b)

Fig. 1. Example of detection-free human tracking on the complex scenario. (a) Two frames from a sport video, where the human detections (denoted by the red boxes) are unreliable due to the appearance variations and occlusions. (b) Tracking results generated by our approach. The object IDs (denoted by the numbers) are retained well during tracking and the human silhouettes are basically preserved.

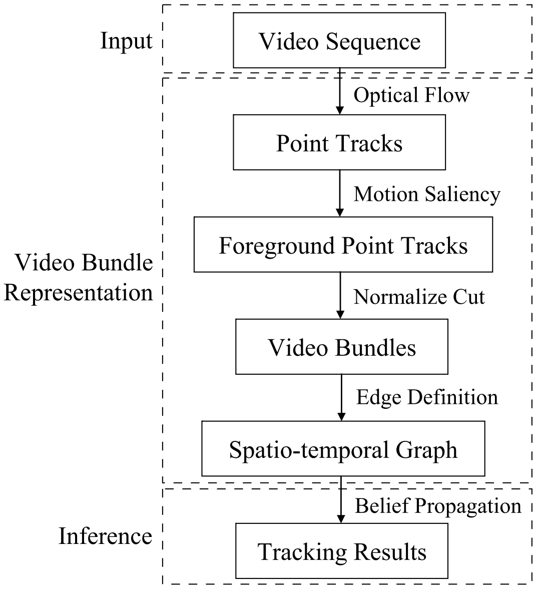


Fig. 2. Flow chart of our framework.

The video bundle can be regarded as a intermediate-level video representation of object trajectory, just like the superpixel in image segmentation. A video bundle comprises of clustered dense point tracks that can be in different lengths over consecutive frames, and one trajectory may include a batch of bundles in the video. The video bundle advances in the following aspects, compared with traditional region-based representations [1], [6]. First, the point tracks are clustered in terms of satisfying both spatial proximity and temporal coherency, so that the bundles are more robust against noises and object conglutinations. Second, the aggregated video bundle provides more reliable and more accurate knowledge about the target location and appearance, compared to each single point track. Also, the bundles, in the form of spatio-temporal volumes, effectively reduce the complexity during inference, i.e., without the need of extracting frame-based correspondences in each step of tracking. Moreover, object silhouettes can be basically preserved by the bundles and we thus obtain fine-grained object trajectories from the video.

Taking video bundles as graph vertices, we link each vertex to its neighbors with an edge in 3-D coordinate to construct an spatio-temporal graph. Building spatio-temporal graph upon middle-level representations has been well explored in existing tracking methods [7], [8]. In this paper, the spatio-temporal graph incorporates trajectory-level relations (the prior model) and global consistency (the likelihood model), while the video bundles capture only low-level cues with point tracks. The edge can be either positive or negative, indicating the two vertices either cooperatively or conflictingly belong to the same trajectory. The negative (competitive) relations serve as important complements to the positive (compatible) ones, both of which should be satisfied with probabilities during inference. We assign an edge to be positive or negative by examining the moving directions of the two video bundles. Specifically, if two video bundles have significantly different moving directions, they are less likely to belong to the same object trajectory. Then we pose the multiobject tracking as a graph partition task under the Bayesian framework.

For the inference of graph partition, we present the RBP algorithm to adapt the real challenges during tracking. In computer vision, belief propagation (BP) algorithm and its extensions are widely employed for graph-based inference [9], [10], providing a general way to assign labels to graph vertices. However, these algorithms sometimes may converge into unsatisfied local minima. This problem might be more serious in tracking, as we tend to simultaneously capture spatial and temporal information of objects and scenes. In this paper, we improve the traditional BP algorithm by allowing the solution to be reconfigured during optimization. We verify the converged solution using the constraint of intrabody variance and three complementary measures, and reactivate the inference (i.e., to jump out from the local minima), by operating on current solution. In brief, our algorithm iteratively performs with following two steps: 1) searching for a solution of graph partition by passing messages and updating beliefs and 2) reconfiguring the graph partition by realizing the merge-and-split operators on graph vertices, once the solution violates the constraint. The flow chart of our framework is presented in Fig. 2.

To the best of our knowledge, this paper is the first to explore the detection-free tracking by exploiting the mid-level representation without relying on off-line trained classifiers.

II. Related Work

Object tracking, in general, is a joint task of object segmentation and temporal correspondence over video sequences. Under the circumstances that reliable object models (e.g., detectors or templates) are invalid or limited, one can solve object tracking as spatio-temporal pixel grouping (or association) in the bottom-up manners [11], [12]. These methods shared some techniques with those for object segmentation [13], [14]. For example, Wang *et al.* [1] utilized structural information captured by superpixels, and proposed to distinguish the targets and background with the mid-level cues. In [2], a trajectory graph and a detection tracklet graph were constructed to encode grouping affinities in space and associations across time, respectively. Basharat *et al.* [15] proposed to construct motion segments based on the spatial and temporal analysis of interest point correspondence. Fragkiadaki and Shi [12] and Brox and Malik [16] utilized dense point tracks to represent object trajectories and demonstrated very good results for capturing motion discontinuities across object boundaries.

The key to success in multiobject tracking is the inference algorithm of data association. The main challenges lie in various ambiguities during optimization caused by abrupt object motion, nonrigid object structures, and changing appearance patterns of both objects and scenes. Previous approaches usually dealt with these problems by exploiting appearance and/or motion cues from different perspectives [14], and the graphical representations were widely adopted [8], [17]–[19]. Exemplar inference algorithms included linear programming [20], dynamic programming [3], joint probabilistic data-association filter [21]. Liu *et al.* [7] performed the stochastic cluster sampling for parsing trajectory in a spatio-temporal graph, but the algorithm is computationally expensive. Our framework is partially motivated by these methods, and advances them in two aspects. First, the representation of video bundles tightly integrates the spatial and temporal information to reduce the ambiguities of multiobject tacking. Second, the proposed RBP algorithm is very robust and fast to conduct reliable results by incorporating both competitive and compatible relations among moving objects.

In the experiments, we compare with both detection-free tracking frameworks [12], [16], [22] and detection-based methods [2], [3] on public benchmarks, and our algorithm performs favorably against all competing methods.

The rest of this paper is organized as follows. We introduce the representations of our approach in Section III, and discuss the formulation and inference procedure in Sections IV and V, respectively. The experiments and comparisons are presented in Section VI, and finally comes the conclusion in Section VII.

III. Video Bundle Representation

We first introduce the video bundle representation and the problem formulation under the spatio-temporal graph.

A. Video Bundle

We first define a point track τ_i to be a sequence of points

$$\tau_i = \left\{p_{i,k} : k \in \left[t_{i,b},\ t_{i,e}\right]\right\} \tag{1}$$

where $p_{i,k}$ indicates the spatial coordinate of τ_i at frame k, $t_{i,b}$ and $t_{i,e}$ the birth and death time of τ_i, respectively.

We obtain point tracks from deferred video sequences using the approach in [23], which tracks points densely with large displacement optical flow. The obtained point tracks are hence dense in space and have various lifespans. Note that [23] produces spatially-denser tracks than conventional sparse point trackers [24], resulting in denser coverage of the moving targets.

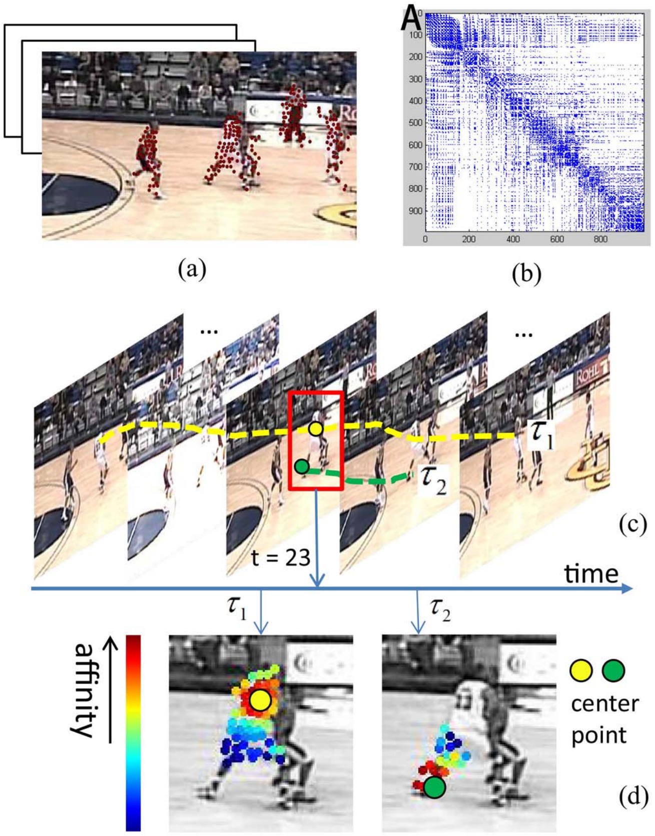


Fig. 3. Illustration of the point tracks and affinity measure. (a) Foreground point tracks in a particular frame, with the (b) point track affinity matrix visualized. (c) Long track τ_1 (yellow dashed line) and a short track τ_2 (green dashed line) are exhibit among video sequence. (d) Corresponding points in frame 23. We can learn that their neighbors cover different parts of the object due to motion differences. Direct clustering to objects may introduce noise.

The obtained track set contains point tracks generated from both foreground and background. Since the point tracks are considerably dense ($\sim 10^4$), taking all of them belonging to foreground and background into consideration results in a high-computational complexity. Moreover, clustering background as well as foreground point tracks simultaneously requires post processing to merge background clusters together. In our framework, however, we only concentrate on the moving targets in foreground. We remove tracks belonging to background using a recently proposed method [12], i.e., motion saliency on point tracks. The nonsalient ones are treated as background tracks and discarded without further consideration in our tracking framework. In motion segmentation scenario, the point tracks assigned to background automatically form the background cluster. Fig. 3(a) shows an example of our extracted foreground point tracks in a single frame.

We further group point tracks based on an affinity matrix A. Each element A_{ij} in the affinity matrix A measures the similarity between two tracks τ_i and τ_j. We define the similarity following two aspects: 1) geometric location and 2) velocity

$$A_{ij} \propto e^{-\mathcal{D}_{\text{tw}}(\tau_i,\tau_j)\cdot\sum_{k\in O_{ij}}\left\|v_{i,k}-v_{j,k}\right\|^2/\left|O_{ij}\right|} \tag{2}$$

where O_{ij} denotes the temporal overlapped frames of τ_i and τ_j, $v_{i,k} = p_{i,k+3} - p_{i,k}$ indicates the velocity for the kth temporally-overlapped point of τ_i aggregated over 3 frames. $\mathcal{D}_{\mathrm{tw}}(\cdot)$ is the dynamic time warping (DTW) distance [25] which measures the aligned geometric distance between two tracks. Given two tracks τ_i and τ_j, DTW seeks the warping path γ with minimum cost to align all points in each track

$$\mathcal{D}_{\mathrm{tw}}(\tau_i, \tau_j) = \min \frac{1}{|\gamma|} \sqrt{\sum_{\gamma_k} \left\| p_{i,k_i} - p_{j,k_j} \right\|^2} \tag{3}$$

where $\gamma_k = (p_{i,k_i}, p_{j,k_j})$ denotes the kth aligned point pair in the warping path, $|\gamma|$ the total number of aligned point pairs. For detailed explanations (see [25]). DTW is more appropriate than traditional distance measure for trajectory clustering as it not only measures position difference, but also shape similarity of the point tracks by finding a time warping, while traditional distance simply performs one-to-one mapping. Note DTW can be used to compare various life-spans sequences which are well-aligned or have shift variance. Nevertheless, it can be replaced by traditional Euclidean distance like [16] does. Fig. 3(b) shows the visualized result of the obtained affinity matrix.

Given the affinity matrix A, we adopt normalized cut, a common and important technique utilized in data clustering and image/video segmentation [16], [26], to group point tracks. Compared with traditional clustering methods such as k-means and the hierarchical clustering algorithm [27], normalized cut is more suitable for our graph partitioning problem. The video bundle is an intermediate-level video representation of object trajectory without directly associating with semantic meaning, just like superpixels in image segmentation. The normalized cut algorithm has good property of robustness against random noisy caused by low-level features. In our implementation, we observe this algorithm can capture well both dissimilarity within the bundles of different objects and similarity within the bundles of the same objects.

However, segmenting object in a bottom-up manner directly using normalized cut often fails to provide meaningful results due to drastic motion differences within a target or tiny motion differences between two targets. By the very nature of an articulated object, some parts are more steady such as the torso and the head. These parts exhibit smooth motion and are usually covered by point tracks which propagate their affinity further and have longer lifespan than others. Other parts like the limbs exhibit drastic motion and are more likely to be covered by the short point tracks. See Fig. 3(c) for reference. Fig. 3(d) visualizes the point track affinity graph with a long track and a short track, respectively. It shows that the long track τ_1 has large number of affinitive neighbors due to its long lifespan, while short track τ_2 is less reliable and has smaller number of neighbors due to large displacement and occlusion of the limb.

Considering the above-mentioned characteristics of the articulated object, clustering all point tracks at once to object via normalized cut would introduce noisy affinity both inside and between objects due to motion difference, leading to over-segmentation and under-segmentation of objects. In the proposed framework, we overcome this unfavorable result by designing a bi-layer clustering strategy, and it involves two steps: 1) over-segmenting foreground point tracks into video bundles via the normalized cut and 2) performing robust inference of the spatio-temporal graph consist of video bundles by the RBP. We first describe the details of the first stage, and then introduce the second stage in Section V.

In the first stage, the point track are over-segmented into a set of clusters via normalized cut. The obtained clusters are served as a robust mid-level representation of video. One analogical task to the proposed method is image segmentation, where we generate superpixels by clustering pixels and further construct region-based graph to conduct inference. By applying normalized cut, the point tracks are embedded into a lower-dimensional subspace. This is done by finding the large eigenvectors of the affinity matrix, then we further discretize the eigenvectors by rotation [28] and obtain K clusters. We treat obtained cluster as a video bundle, denoted by $b_i = (\bar{\tau}_i, \bar{v}_i, \{\tau_j\})$, where $\bar{\tau}_i$ and $\bar{v}_i$ denote the cluster center and the mean velocity of b_i, respectively. $\bar{\tau}_i$ and $\bar{v}_i$ are computed by taking average over all point tracks belonging to b_i

$$\begin{aligned}\bar{\tau}_i &= \left\{ \bar{p}_{i,k} : \bar{p}_{i,k} = \sum_{j=1}^{|b_i|} \frac{p_{j,k}}{|b_i|}, k \in \left[\min_j t_{j,b}, \max_j t_{j,e} \right] \right\} \\ \bar{v}_i &= \left\{ \bar{v}_{i,k} : \bar{v}_{i,k} = \sum_{j=1}^{|b_i|} \frac{v_{j,k}}{|b_i|}, k \in \left[\min_j t_{j,b}, \max_j t_{j,e} - 3 \right] \right\}\end{aligned} \tag{4}$$

where $|b_i|$ denotes the number of tracks within b_i.

The obtained video bundles, as shown in Fig. 4(a), provide robust and compact descriptions for moving objects which respect spatial proximity and temporal coherence. Deferred inference of object can be conducted based on this robust mid-level representation and overcomes the shortage of normalized cut.

B. Spatio-Temporal Graph

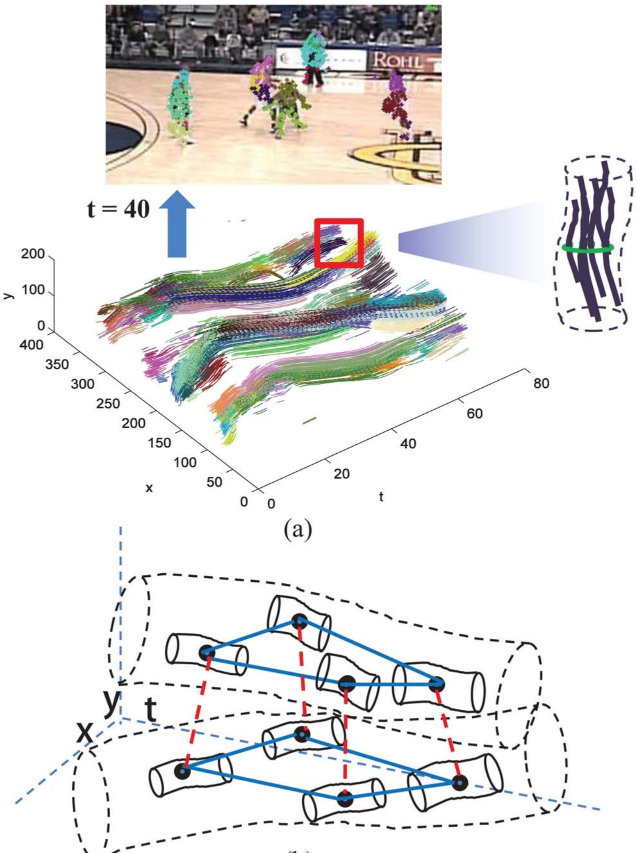
In the previous section, we have presented how the video bundles are defined. This section explains how to use our proposed bundles to construct a graphical model for inference task, i.e., data association for tracking multiple targets among the bundles.

We assume that there are N targets in the video, the objective of multitarget tracking is to identify the trajectory for each object in the video. Given the set of video bundles $B = \{b\}$, we define the solution W as

$$W = \left\{ \Gamma_n = \{b_i\}_{i=1}^{|\Gamma_n|},\ n = 1, 2, \ldots, N,\ b_i \in B \right\} \tag{5}$$

where Γ_n denotes the trajectory for the nth object. We constrain each trajectory encapsulating at least one bundle and each observed bundle belonging to one and only one trajectory. Thus, we can formulate the problem of data association as a graph partition task, i.e., grouping bundles into different object trajectories.

We introduce a spatio-temporal graph $G = <B, E>$ to describe the relations among bundles. Each bundle $b_i \in B$



(a)

(b)

Fig. 4. Illustration of our representations. (a) Video bundle generated by point tracks exhibiting high affinities. (b) Spatio-temporal graph is constructed by taking the bundles as vertices, and the blue and red edges indicate compatible and competitive relations between vertices, respectively. Best view in color image.

is taken as a graph vertex and each edge $e_{ij} = <b_i, b_j> \in E$ describes the relation between two adjacent (neighboring) bundles b_i and b_j. Two bundles b_i and b_j are regarded as neighbors $b_i \in \mathbb{N}(b_j)$ if they have temporal overlap $O_{ij} \neq 0$. We further develop two kinds of edges: 1) negative edges E^- and 2) positive edges E^+ to describe the competitive and compatible relations among them. Two neighboring bundles with significantly different motion directions yield a negative edge and otherwise a positive edge; namely

$$\begin{aligned} E &= E^- \bigcup E^+ \\ &= \left\{e_{ij} : \bar{v}_i \cdot \bar{v}_j < 0\right\} \bigcup \left\{e_{ij} : \bar{v}_i \cdot \bar{v}_j \geq 0\right\}. \end{aligned} \tag{6}$$

For notation simplicity, we drop the notation of the edge index ij in the following discussion.

Negative edges penalize two bundles moving in the opposite direction being coupled together, i.e., these two bundles should belong to two different objects. We define a negative edge probability $\rho^-(b_i, b_j)$ to represent the extent of two bundles repulsing each other

$$\rho_{ij}^- \propto \exp\{\bar{v}_i \cdot \bar{v}_j\}. \tag{7}$$

In other words, two bundles are less likely to belong to the same object if their motions are obviously different.

Positive edges encourage two bundles sharing similar statistics to be assigned with the same label. A positive edge probability ρ_{ij}^+ is defined following two aspects: 1) the geometric distance and 2) the temporal consistency

$$\rho_{ij}^+ \propto \exp\left\{-\frac{\mathcal{D}_{\mathrm{tw}}\left(\bar{\tau}_i, \bar{\tau}_j\right) \cdot \mathcal{D}_{tc}\left(\bar{\tau}_i, \bar{\tau}_j\right)}{g}\right\} \tag{8}$$

where $\bar{\tau}_i$ and $\bar{\tau}_j$ are the cluster centers for b_i and b_j, as defined in (4), and g is a fixed scale. $\mathcal{D}_{\mathrm{tw}}(\cdot)$ is the DTW distance defined in (3). $\mathcal{D}_{tc}(\cdot)$ explores the motion context to provide a complementary cue for identifying the degree of attractiveness of two bundles. For example, in complex scenarios where numerous people move in diverse directions, bundles from different people may not have distinct motion difference. To overcome this problem, we proposed to measure their accumulated temporal consistency. Specifically, we connect one line segment between two bundles and accumulate the derivatives of optical flow along the line, and define

$$\mathcal{D}_{tc}\left(\tau_i, \tau_j\right) = \frac{1}{|O_{ij}|} \sum_{k \in O_{ij}} \sum_{p \in \overline{p_{i,k} p_{j,k}}} \nabla F_k(p) \tag{9}$$

where $\overline{p_{i,k} p_{j,k}}$ denotes the line segment between points $p_{i,k}$ and $p_{j,k}$, and ∇F_k represents the derivative of optical flow $((\partial F_k/\partial x), (\partial F_k/\partial y))$ at the kth frame and is calculated as a size weighted mean of the derivative of separate x and y channel. As $\mathcal{D}_{tc}$ is defined by the sum of gradients of optical flow, which captures the changes in the flow field and depresses the local smooth and constant motions, it penalizes two bundles which exhibit salient motion variance.

An illustration of the spatio-temporal graph representation is shown in Fig. 4(b). Note we only focus on a few bundles in the red bounding box for clear specification.

IV. Bayesian Formulation

We solve W by maximizing a posterior probability under the Bayesian framework

$$W^* = \arg\max_W P(W|B) \propto \arg\max_W P(B|W)P(W). \tag{10}$$

Likelihood $P(B|W)$ measures how well the observed data (video bundle) satisfies a certain object trajectory. Assuming the likelihood of each bundle is calculated independently given the partition, then $P(B|W)$ can be factorized into

$$P(B|W) = \prod_{\Gamma_n \in W} \prod_{b_i \in \Gamma_n} P(b_i|\Gamma_n). \tag{11}$$

Existing related methods [29], [30] usually defined the likelihood model by maintaining a template for each specific object obtained by online learning, or used Kalman or a particle filter [31]–[33] to estimate the locations of targets at each state. Instead of relying on a pretrained detector for measuring the targets, we define the likelihood term using the simple yet effective Gaussian mixture models (GMMs) [34].

For the nth trajectory, we use two GMMs to capture its color and texture information, respectively, $\{\rho_n^c, \rho_n^g\}$. In particular, ρ_n^c represents the hue-saturation-value (HSV) color distribution, and ρ_n^g represents the distribution of orientated gradients. The GMM includes a number of components that are parameterized by the means and covariances. In our implementation, we first extract the color and gradient features for all pixels of the trajectory, and calculate the Gaussian components for ρ_n^c and ρ_n^g, where the Euclidean distance is used for measuring the feature vectors. Therefore, given any video bundle b_i, we can calculate the likelihood $P(b_i|\Gamma_n)$ by matching its features with the GMMs, as

$$\begin{aligned} P(b_i|\Gamma_n) &= P\big(b_i|\rho_n^c, \rho_n^g\big) \\ &\propto P\big(H^c(b_i)|\rho_n^c\big) \cdot P\big(H^g(b_i)|\rho_n^g\big) \end{aligned} \tag{12}$$

where $H^c(b_i)$ and $H^g(b_i)$ are the color and gradient features of b_i, respectively.

Prior $P(W)$ imposes constraints on object trajectories and their interactions. In the proposed framework, this term is defined such that if two bundles are similar then they are supposed to have the same identity. We decompose such constraints into pairwise potentials between video bundles within each trajectory. The pairwise term is defined as

$$\begin{aligned} P(W) &= \prod_{\Gamma_n, \Gamma_m \in W} P(\Gamma_n, \Gamma_m) \\ &= \prod_{b_i, b_j \in \Gamma_n, e \in E^-} \Big(1 - \rho_{ij}^-\Big) \prod_{b_i, b_j \in \Gamma_n, e \in E^+} \rho_{ij}^+ \\ &\quad \prod_{b_i \in \Gamma_n, b_j \in \Gamma_m, e \in E^-} \rho_{ij}^- \prod_{b_i \in \Gamma_n, b_j \in \Gamma_m, e \in E^+} \Big(1 - \rho_{ij}^+\Big) \end{aligned} \tag{13}$$

where ρ_{ij}^- and ρ_{ij}^+ are the negative and positive edge probability defined in (7) and (8).

V. Inference Algorithm

Given the spatio-temporal graph representation, inferring graph partition for W^* is a nonshallow problem, not only because the convexity guaranty of probability distribution $P(W|B)$ does not hold, but also due to the unknown number of targets. We pose the graph partition as the task of assigning labels to graph nodes. Let $\mathcal{L}$ be a set of labels, i.e., $\mathcal{L} = \{l_i = n, n = 1, 2, \ldots, N, b_i \in B\}$. A labeling l_i indicates the bundle $b_i \in B$ belonging to the nth trajectory. This graph-based labeling problem has been extensively discussed in the literature, and a batch of inference methods were proposed. Among them, some stochastic sampling approaches [8], [35] aim to search for the global optimal solution but often limit by relatively low efficiency. Some alternative methods such as BP [9] perform fast and also obtain good if given reliable initializations. They, however, often suffer from getting stuck on unsatisfied local minima, particularly under complex scenarios of multiple object tracking. To alleviate this problem, in this paper, we present an efficient yet effective inference algorithm called RBP that incorporates the splitting and merging operators into the message passing procedure, making the inference reconfigurable to jump from local minima.

A. Initialization

In some traditional BP inferences, the algorithms initialize beliefs for nodes according to their unary likelihoods. In this paper, we improve the initialization by further imposing pairwise similarities between vertices. Specifically, we can find a set of bundles as the representatives $\theta_i = \{\tilde{b}_k\}$ by comparing the similarity measure of bundles defined in (8). This is done by hierarchically merging pairs of bundles that have the least distance. The centers of the obtained groups are served as representatives. We first initialize the beliefs of representative bundles $\tilde{b}_k$ as 1 for their belonging labels and 0 for other labels. For each of the rest bundles b_j, we then compute the mean positive edge probabilities between b_j and the representatives for each label $\tilde{b}_k \in \theta_{l_i}$. Its belief is thus initialized as a distribution proportional to the mean edge probabilities for each labels and we put it into the set with the maximum belief. This process serves as a rough partition on the bundle set.

B. Priority-Based Message Passing

Given the spatio-temporal graph $G = <B, E>$, BP iterates on exchanging messages between nodes and updating node beliefs. In the following discussion, we denote the message passed from b_i to b_j and the belief for a node b_j at the tth iteration as $\Phi_{i\to j}^t(l_i)$ and $\Psi_j^t(l_j)$, respectively.

We adopt the mechanism of priority-based BP (PBP) proposed by [10] to suppress the ambiguous information passed between nodes. The intuition of this mechanism is to disambiguate the labels of nodes in virtue of the strength of their neighbors. The ambiguity of a node b_i at the tth iteration is defined as the entropy of its current belief

$$\zeta^t(b_i) = -\sum_{l_i=1}^{N} \Psi_i^t(l_i) \log\big(\Psi_i^t(l_i)\big). \tag{14}$$

Nodes with less ambiguity are scheduled to transmit their messages with higher priority. Furthermore, to prevent propagating confusing information between nodes, a node only computes the messages passed from its less ambiguous neighbors. At the tth iteration, the message passed from node b_i to b_j is defined as

$$\Phi_{i\to j}^t(l_i) \propto \sum_{l_i=1}^{N} \delta\big(l_i, l_j\big)\delta(b_i|l_i) \prod_{b_k \in \mathbb{N}_<(b_i)} \Phi_{k\to i}^{t-1}(l_i) \tag{15}$$

where $\delta(b_i|l_i)$ and $\delta(l_i, l_j)$ are unary potential and pairwise potential, respectively, which correspond to the likelihood and prior defined in (11) and (13). $\mathbb{N}_<(b_i)$ denotes the less ambiguous neighbors of b_i, i.e., $\mathbb{N}_<(b_i) = \{b_j : \zeta^t(b_j) < \zeta^t(b_i), b_j \in \mathbb{N}(b_i)\}$. Note an implicit requirement for (15) is that b_j is more ambiguous than b_i. After computing the messages passed from its neighbors, the belief of node b_j at the tth iteration is updated by

$$\Psi_j^t(l_j) \propto \delta\big(b_j|l_j\big) \prod_{b_k \in \mathbb{N}(b_j)} \Phi_{k\to j}^t\big(l_j\big). \tag{16}$$

The belief $\Psi_j(l_j)$ at node j represents the posterior probability of bundle b_j having label l_j, and we maximize the posterior probability by searching the maximum belief. After convergence, label l_j is assigned to bundle b_j if it produces the maximum belief at node j. Thus, the label assignment is unique.

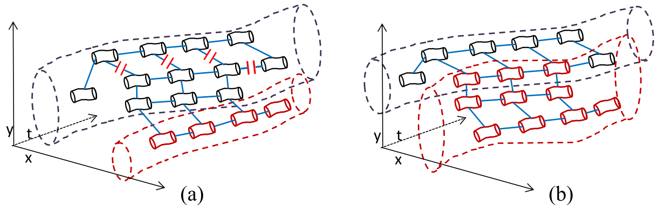


Fig. 5. Illustration of our merge-and-spilt operation to drive the solution reconfiguration. Each small cylinder primitive represents a bundle and the grouped represents a partitioned object trajectory. The red cuts of edges indicate the edges been turned off in the operation.

C. Iterative Belief Reconfiguration

To make the BP more robust, one can impose extra constraints during inference. For example, Kschischang *et al.* [36] adopted the high-order factors into energy potentials, making the inference more robust, but these methods usually lead to expensive computation. In this paper, we provide an alternative way to verify converged solutions and reactivate the inference by reconfiguring the beliefs of nodes.

First, we introduce a global constraint for the divided trajectories based on the intuitive observation in video tracking. For example, tracking targets with very small or very large size probably results from the trajectory being over-segmented or under-segmented, respectively. Specifically, we define the global constraints on an object trajectory Γ_n as bivariate Gaussian distributions on averaged object size in time intervals

$$P(\Gamma_n) = \prod_z P(\Lambda_{n,z}) = \prod_z \mathbf{G}\Big(\big[\bar{w}_{n,z}\ \bar{h}_{n,z}\big]\,\big|\,\mu, \sigma^2\Big) \tag{17}$$

where z denotes the zth time interval of the trajectory, $\Lambda_{n,z}$ the region of nth tracking target in time interval z, $[\bar{w}_{n,z}\ \bar{h}_{n,z}]$ the size of the bounding box of $\Lambda_{n,z}$ averaged over the time interval. The parameters μ and σ are tuned to fit the size of most objects in the dataset. In the experiments, we set the time interval as eight frames and determine an object trajectory violates global constraints if there exists one time interval where the probability is less than 0.01.

For the converged solution, we first identify the most problematic partitioned trajectory, i.e., violating the global constraint. Then we design a corresponding operator, i.e., merge-and-split, to reconfigure the solution by correcting Γ_n over the graph.

1) Merge-and-Split: For the case that we identify the problematic trajectory Γ_n by its very small region, i.e., the size of $\Lambda_{n,z}$ is smaller than a threshold, we scatter every node $b_i \in \Gamma_n$ and merge them with their neighbor nodes according to the affinities specified by the edge connections. The belief vector of b_i is then revised by setting 0 to the nth bin. Note that we need to renormalize the beliefs of all vertices over the graph accordingly. In other case, when one region of the trajectory Γ_n is too large, we split the vertices of Γ_n into two subsets. The vertices in one subset remain unchanged while the rest vertices are merged to their neighbors. The belief of each changed vertex is also need to be revised, i.e., the nth of the belief vector is set as 0. In the implementation, we realize the split operator using the normalized-cuts [37] on positive and negative edges. An illustration is shown in Fig. 5.

Algorithm 1: Sketch of the RBP Algorithm

```
Input: Video bundles $B$
Output: Bundle labels $L$
Initialize: $\Phi_{i\to j}(l_i) = 1$, $\Psi_i(l_i)$ by mechanism proposed
in Section V-A, $l_i = \max_{l_i} \Psi_i(l_i)$ ;
repeat
    Reconfigure beliefs and labels for bundles within the
    problematic object trajectory ;
    for $t = 1$ to $T_2$ do
        repeat
            Prioritize bundle $b_i$ according to its level of
            ambiguity ;
            foreach more ambiguous neighbors of $b_i$ do
                Compute message from bundle $b_i$ to
                bundle $b_j$ by Equ.(15) ;
                Update belief for bundle $b_j$ by Equ.(16) ;
            end
        until all bundles pass messages;
    end
    Assign each bundle $b_i \in B$ the label $l_i$ with the max
    belief ;
until $L$ satisfies variance and three complementary
constraints or algorithm iterates over $T_1$ times;
```

TABLE I
EVALUATION METRICS FOR COMPARISONS OF DETECTION-FREE TRACKING METHODS

Metric	Definition
Clustering error (CE)	The average percentage of pixels with false labels.
Per region CE (PRCE)	The average percentage of pixels with false labels per person mask.
Over-segmentation (OS)	The average number of object trajectories assigned to each mask.
Leakage	The percentage of trajectories exist leaking (overlap with other masks greater than 50% of theirs assigned masks) at one frame.
Recall	The percentage of recalled pixels for each labeled mask by the single best trajectory.
Tracking time (TT)	The percentage of frames whose recall for each mask is above 20% and averages across all masks.

Once a new partition is generated by the operator, the PBP will be reperformed to calculating the beliefs over the graph. These above steps iterate until the target energy converges finally, i.e., the solution will not be changed. The overall procedure is summarized in Algorithm 1.

TABLE II
QUANTITATIVE RESULTS AND COMPARISONS ON MOSEG. EXTRACTED OBJECTS STANDS FOR THE NUMBER OF LABELED MASKS WITH LESS THAT 10% CLUSTERING ERROR

Moseg	Density	CE	PRCE	OS	Extracted Objects
Our Method	**3.61%**	**2.13%**	**18.98%**	**0.25**	**30**
[22]	3.07%	2.29%	20.93%	0.29	29
[12]	3.22%	3.76%	22.06%	1.15	25
[16]	3.32%	3.43%	27.06%	0.4	26

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT DETECTION-FREE TRACKING METHODS. OUR-FULL: OUR METHOD (TRAJECTORY PARSING WITH VIDEO BUNDLES AND RBP ALGORITHM). WE ALSO PRESENT THE RESULTS OF THREE ADDITIONAL APPROACHES BY SIMPLIFYING OUR METHOD

Figment	CE	PRCE	OS	Leakage	Recall	TT
Our-full	**4.24%**	**18.18%**	1.01	**15.84%**	**55.23%**	**86.45%**
Our-1	6.27%	23.60%	**0.92**	19.27%	45.50%	80.99%
Our-2	6.78%	37.79%	1.12	23.89%	43.31%	79.30%
Our-3	9.04%	33.15%	1.62	37.44%	33.27%	76.23%
[22]	7.90%	18.47%	1.5	19.55%	33.28%	82.29%
[12]	4.73%	20.32%	1.57	16.52%	31.07%	75.13%
[16]	20.74%	86.43%	0	81.55%	0.46%	1.03%

VI. EXPERIMENTS

A. Datasets and Settings

Three datasets, motion segmentation (Moseg) [16], figure untanglement (Figment) [12], and TUD [21], [38], are used to evaluate the proposed method. Moseg [16] is a recently released dataset for motion segmentation, which consists of 26 video sequence about objects of various scales and motions and is publicly available. A total of 189 frames are annotated densely in space and sparsely in time, and the software delivered with the dataset is adopted for performance evaluation.

We also evaluate the proposed method for tracking multiple interacting and deforming agents using the Figment dataset that is usually applied for detection-free tracking and video annotation testing [39]. Figment is a dataset of basketball court filmed from a freely moving camera, which consists of 18 challenging basketball clips of with 50–80 frames each.

The detection-based methods usually have unsatisfying performance on abovementioned two datasets, mainly due to their unreliable detection performances under such cluttered scenarios. Moreover, we adopt another benchmark TUD to compare with several detection-based methods, and this dataset addresses pedestrian tracking task, in which the pedestrian detection proposals are relatively reliable. TUD includes three sequences that have hundreds of frames.

All the parameters are fixed in the experiments. The number of video bundles is set to be 200, and the number of trajectories N is between $5 \sim 14$. We use a fixed scale $g = 300$ in positive edge probability. We set the maximum number of reconfiguration $T_1 = 10$ and the maximum number of BP iterations $T_2 = 15$. For intrabody variance, the threshold q for splitting is set to be 100.

B. Comparisons With Detection-Free Methods

We use the six performance metrics listed in Table I for quantitative comparison of detection-free tracking methods. Note the first three metrics are used for evaluating on both Moseg and Figment dataset. For more details about the evaluation metrics please refer to [12].

The proposed method is first compared with three detection-free tracking methods [12], [16], [22] on Moseg. All these competing methods are suggested for automated detection and tracking of objects based on clustering of point tracks. Table II lists the comparison results, where we use the first three metrics together with density and extracted objects for performance evaluation, as [12], [16], and [22] do. Note background clusters are considered here for comparison with other methods. Our proposed method achieve comparable results when compared with the state-of-the-arts. Fig. 6 shows several examples of the segmentation results.

We also compare the detection-free methods on the Figment dataset, which is more challenging in tracking multiple interacting objects. All the six performance metrics listed in Table I are employed for the evaluation of the tracking performance. For fair comparisons, the calculation of metrics follows the standard procedure introduced in [12].

1) All metrics are computed by discarding top and bottom 10% of values and averaging over the remaining ones.
2) Per region clustering error (PRCE) for a mask is set to 100% if it is missed to be assigned a label.
3) Recall for a leaking cluster is set to 0.
4) Recall and tracking time are computed by dilating each trajectory with a radius of eight pixels.

As listed in Table III, our method outperforms all the competing methods in terms of all metrics. Fig. 9 shows some

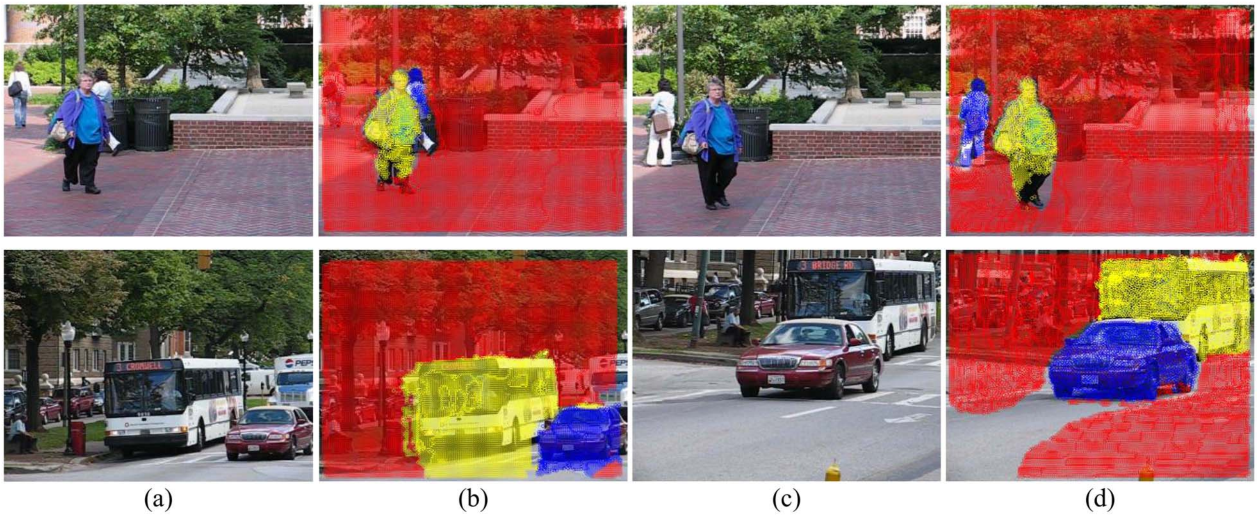


Fig. 6. Segmentation results on *Moseg*. (a) and (c) Sampled frames, with (b) and (d) their corresponding results. Background cluster are considered. Pixels in red are the segmented background, and pixels in blue and yellow are the extracted foreground objects, respectively. Note our method works well for both articulated (row above) and rigid (row below) targets.

TABLE IV
EVALUATION METRICS FOR COMPARISONS WITH DETECTION-BASED TRACKING METHODS

Metric	Definition
Miss detection (MD)	The average percentage of groundtruth objects failed to be detected.
False positive (FP)	The average percentage of detected objects does not belong to ground-truth objects.
ID-switch (ID-sw.)	The average percentage of times a groundtruth object changes its assigned identity.
MOTA	Accuracy $= 1 - \text{MD} - \text{FP} -$ ID-sw.

TABLE V
QUANTITATIVE RESULTS AND COMPARISONS WITH DETECTION-BASED TRACKING METHODS ON FIGMENT AND TUD

Dataset	Method	MD	FP	ID-sw.	MOTA
Figment	Our	26.09%	15.13%	2.07%	**56.71%**
	[2]	50.95%	16.41%	0.43%	32.64%
	[3]	89.19%	0.18%	4.46%	6.17%
TUD	Our	37.23%	12.55%	1.91%	48.30%
	[2]	25.04%	51.70%	1.49%	21.77%
	[3]	46.38%	0.78%	0.57%	**52.27%**

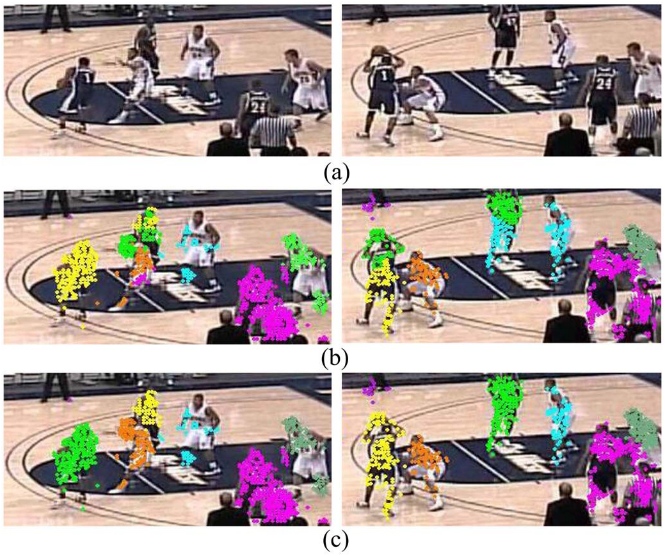


Fig. 7. Tracking results using different inference algorithms. (a) Source frames. (b) and (c) Tracking results by the PBP and the proposed RBP algorithms, respectively. The assigned labels to object IDs are represented by different colors.

examples of the tracking results, and more video results are provided in the supplementary materials.

It worth mentioning that our framework can handle the partial and total occlusion issues well. The video bundles have long-term temporal consistency of visible object pixels, naturally handling partial occlusion in tracking. For total occlusion, our framework can also achieve robust tracking by performing reconfigurable inference on the constructed spatio-temporal graph. The values of Recall and TT in Table III can address the partial and total occlusion issues, and the qualitative illustration is shown in Fig. 6.

C. Comparisons With Detection-Based Methods

We compare our method with two detection-based tracking methods [2], [3], where human detectors [40], [41] are adopted to generate proposals for human tracking. Their publicly available codes are adopted and we tune the parameters to achieve the best performance on Figment and TUD. Note that [3] treats tracking as a network flow problem and solved

TABLE VI
QUANTITATIVE RESULTS OF THE PROPOSED METHOD WITH DIFFERENT VIDEO BUNDLE NUMBERS AND VIDEO BUNDLE FEATURES ON SEVERAL VIDEO SEQUENCES RANDOMLY SELECTED FROM FIGMENT

Bundle Number	CE	PRCE	OS	Leakage	Recall	TT
100	3.18%	38.60%	72.72%	31.82%	15.83%	25.97%
150	7.02%	21.76%	90.91%	22.73%	46.37%	54.55%
200	6.77%	19.77%	90.91%	18.18%	49.05%	59.74%
250	3.45%	23.73%	100%	22.73%	42.86%	55.84%
300	3.97%	27.98%	90.91%	27.27%	40.81%	51.95%
Bundle Feature	CE	PRCE	OS	Leakage	Recall	TT
HSV+Grad	6.77%	19.80%	90.91%	18.18%	48.97%	59.74%
HSV+LBP	6.82%	19.94%	90.91%	18.18%	48.56%	59.74%
RGB+Grad	12.71%	21.13%	90.91%	18.18%	47.97%	59.74%
RGB+LBP	12.60%	19.56%	90.91%	18.18%	47.45%	59.74%

it approximately via dynamic programming, while [2] utilize point tracks as low-level cues. In each frame, the bounding boxes for each target are simply localized as the cluster center of all pixels within the corresponding label. We calculate the one-to-one assignment of object proposals to groundtruth and utilize the widely used CLEAR MOT metrics [42] listed in Table IV for evaluation, where the multiple object tracking accuracy (MOTA) is the combination of the three error ratios. The quantitative results are reported in Table V, and some representative results are shown in Fig. 9. From the results of Figment, one can see that due to the deficiency of detection in such scene, both [2] and [3] suffer from heavy missing detection (MD = 89.19% and 50.95%). Our method recovers substantially more trajectories with much higher MOTA accuracy, i.e., 56.71%, while the competing methods are sensitive to unreliable detections. For TUD, our method can achieve comparable accuracy against [3], and substantially outperforms [2]. This evaluation results demonstrate the effectiveness of the proposed method under the detection-reliable scenarios.

D. Evaluation on Components and Convergence

The proposed framework has two key components, i.e., video bundle representation and RBP. To validate the benefits of the components, we compare the proposed method, i.e., our-full, with three additional methods by replacing one or two components of the framework. Our-1 generates trajectories on video bundles by replacing RBP with PBP [10]. Our-2 generates trajectories by performing spectral clustering on video bundles, i.e., replacing RBP with spectral clustering. Our-3 performs spectral clustering on point tracks directly to generate trajectories, i.e., replacing video bundles with point tracks and replacing RBP with spectral clustering. Table III lists the results of the four variants of the proposed method. Clearly bundle representation and RBP outperform point track representation and conventional PBP, respectively.

We further present the evaluation results of the proposed method with different video bundle numbers and video bundle features on several video sequences selected from Figment. Specifically, we set the number of video bundles as 100, 150, 200, 250, and 300. We use the different color features, i.e., HSV and RGB, and the different texture features, i.e., Grad (oriented gradients) and local binary pattern (LBP). Table VI reports the quantitative results generated under different settings. From the results, we can observe that our algorithm generally achieves satisfying performances.

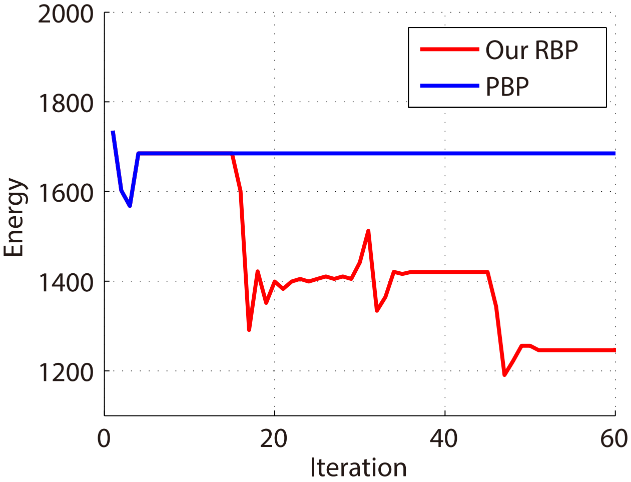


Fig. 8. Energy decreasing of RBP and PBP during inference.

1) Efficiency: Our implementation is coded in MATLAB and all the experiments are conducted on an Intel I5 3.0 GHz PC with 4 GB memory. Given the extracted point tracks, the average runtime for bundle generation is 250 ms per sequence. The runtime of inference is related to the complexity of video content, with the average of 5 $\sim$ 8 min per video sequence.

In addition, we analyze the energy convergence of the RBP algorithm during inference, where the energy is derived based on the posterior probability. Fig. 7 shows the tracking results of RBP and PBP on a clip from figment. Using this clip, Fig. 8 shows the energy values of the proposed RBP and PBP [10] algorithms with increasing iterations. Here, we use the same energy function as in [9]. It is clearly shown that RBP can achieve better convergence while PBP stops after 4 iterations. From Figs. 7 and 8, RBP is effective to avoid being stuck

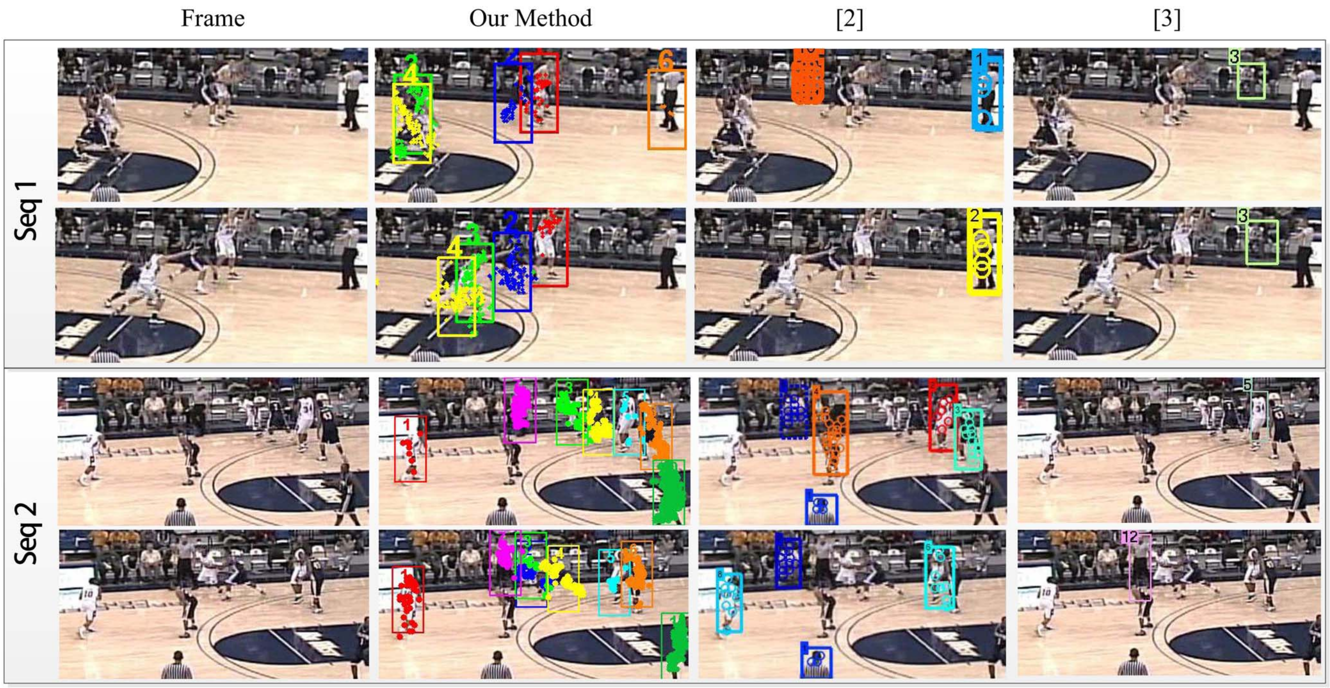


Fig. 9. Segmentation and tracking results on the figment dataset.

in unsatisfied local minima by the reconfiguration operations, and achieves better spatio-temporal assignment results of persons.

VII. Conclusion

This paper proposed a novel video tracking framework in the context of object detectors been limited. The proposed method first constructed a spatial-temporal graph consisting of the video bundles by exploiting multiple cues of motion and appearance, and then generated the trajectories by a novel RBP algorithm. RBP allows the reconfiguration of the clustering results and reactivation of the BP inference to avoid stucking in local minima, and thus can conduct more reliable spatio-temporal association of objects. The experiments and comparisons to the state-of-the-arts demonstrated the effectiveness of our framework on very challenging scenarios.

Acknowledgment

The authors would like to thank Mr. Yuanlu Xu and Mr. Shiyi Hu for their assistance in paper revision and experiments.

References

[1] S. Wang, H. Lu, F. Yang, and M. Yang, "Superpixel tracking," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 1323–1330.

[2] K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi, "Two granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions," in *Proc. ECCV*, Florence, Italy, 2012, pp. 552–565.

[3] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. CVPR*, Providence, RI, USA, 2011, pp. 1201–1208.

[4] L. Lin, Y. Xu, X. Liang, and J. Lai, "Complex background subtraction by pursuing dynamic spatio-temporal models," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3191–3202, Jul. 2014.

[5] K. Koffka, *Principles of Gestalt Psychology*. New York, NY, USA: Hartcourt Brace Jovanovich, 1935.

[6] X. Ren and J. Malik, "Tracking as repeated figure/ground segmentation," in *Proc. CVPR*, Minneapolis, MN, USA, 2007, pp. 1–8.

[7] X. Liu, L. Lin, S.-C. Zhu, and H. Jin, "Trajectory parsing by cluster sampling in spatio-temporal graph," in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 739–746.

[8] Q. Yu and G. G. Medioni, "Multiple-target tracking by spatiotemporal Monte Carlo Markov chain data association," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2196–2210, Dec. 2009.

[9] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, 2006.

[10] K. C. Amit Kumar and C. De Vleeschouwer, "Prioritizing the propagation of identity beliefs for multi-object tracking," in *Proc. BMVC*, Surrey, U.K., pp. 117.1–117.11, 2012.

[11] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.

[12] K. Fragkiadaki and J. Shi, "Detection-free tracking: Exploiting motion and topology for segmenting and tracking under entanglement," in *Proc. CVPR*, Providence, RI, USA, 2011, pp. 2073–2080.

[13] H. Lu, R. Zhang, S. Li, and X. Li, "Spectral segmentation via midlevel cues integrating geodesic and intensity," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2170–2178, Dec. 2013.

[14] Y.-L. Hou and G. K. H. Pang, "Multicue-based crowd segmentation using appearance and motion," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 2, pp. 356–369, Mar. 2013.

[15] A. Basharat, Y. Zhai, and M. Shah, "Content based video matching using spatiotemporal volumes," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 360–377, 2008.

[16] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. ECCV*, Crete, Greece, 2010, pp. 282–295.

[17] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[18] L. Lin, Y. Lu, Y. Pan, and X. Chen, "Integrating graph partitioning and matching for trajectory analysis in video surveillance," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4844–4857, Dec. 2012.

[19] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *Proc. CVPR*, Providence, RI, USA, 2011, pp. 1273–1280.

[20] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "A linear programming approach for multiple object tracking," in *Proc. CVPR*, Minneapolis, MN, USA, 2007, pp. 1–8.

[21] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proc. CVPR*, Anchorage, AK, USA, 2008, pp. 1–8.

[22] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. CVPR*, 2012, pp. 1846–1853.

[23] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by GPU-accelerated large displacement optical flow," in *Proc. ECCV*, Crete, Greece, 2010, pp. 438–451.

[24] J. Shi and C. Tomasi, "Good features to track," in *Proc. ICCV*, Seattle, WA, USA, 1994, pp. 593–600.

[25] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proc. KDD*, Boston, MA, USA, 2000, pp. 285–289.

[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[27] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 686–693.

[28] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. ICCV*, Nice, France, 2003, pp. 313–319.

[29] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, "Level-set person segmentation and tracking with multi-region appearance models and top-down shape information," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 1871–1878.

[30] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[31] D. Wang, H. Lu, and C. Bo, "Visual tracking via weighted local cosine similarity," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1838–1850, Sep. 2014.

[32] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. ECCV*, Prague, Czech Republic, 2004, pp. 28–39.

[33] Q. Wang, F. Chen, and W. Xu, "Tracking by third-order tensor representation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 2, pp. 385–396, Apr. 2011.

[34] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

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

[36] F. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[37] S. X. Yu and J. Shi, "Understanding popout through repulsion," in *Proc. CVPR*, Kauai, HI, USA, 2001, pp. II-752–II-757.

[38] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. CVPR*, San Francisco, CA, USA, 2010, pp. 623–630.

[39] C. Vondrick, D. Ramanan, and D. Patterson, "Efficiently scaling up video annotation with crowdsourced marketplaces," in *Proc. ECCV*, Crete, Greece, 2010, pp. 610–623.

[40] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[41] L. Lin, X. Wang, W. Yang, and J.-H. Lai, "Discriminatively trained and-or graph models for object shape detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 959–972, May 2015.

[42] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *J. Image Video Process.*, vol. 2008, Feb. 2008, Art. ID 1.

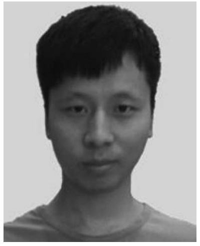
Liang Lin received the B.S. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 1999 and 2008, respectively. From 2006 to 2007, he was a joint Ph.D. Student with the Department of Statistics, University of California, Los Angeles (UCLA), Los Angeles, CA, USA.

He is a Professor with the School of Computer Science, Sun Yat-Sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the Center for Vision, Cognition, Learning, and Art of UCLA. He was supported by several promotive programs or funds for his works such as Guangdong National Science Foundations for Distinguished Young Scholars in 2013. He has published over 80 papers in top tier academic journals and conferences. His current research interests include new models, algorithms and systems for intelligent processing and understanding of visual data such as images and videos. He currently serves as an Associate Editor of *Neurocomputing* and *The Visual Computer*.

Prof. Lin was a recipient of the Best Paper Runners-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, and the Best Student Paper Award in the IEEE ICME 2014.

Yongyi Lu received the B.S. degree in software engineering, and the M.S. degree in computer science from Sun Yat-Sen University, Guangzhou, China, in 2010 and 2013, respectively.

His current research interests include computer vision, machine learning, and intelligent media technology.



Chenglong Li received the B.S. degree in applied mathematics, and the M.S. degree in computer science from Anhui University, Hefei, China, in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree in computer science.

His current research interests include computer vision, machine learning, and intelligent media technology.

Hui Cheng received the B.Eng. degree in electrical engineering from Yan Shan University, Qinhuangdao, China, in 1998, the M.Phil. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2001, and the Ph.D. degree in electrical and electronic engineering from the University of Hong Kong, Hong Kong, in 2005.

She was a Post-Doctoral Fellow with the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong, from 2006 to 2007. She is an Associate Professor with the School of Information Science and Technology, Sun Yat-Sen University, Guangzhou, China. Her current research interests include intelligent robots and networked control.

Wangmeng Zuo (M'09–SM'14) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007.

From 2004 to 2008, he was a Research Assistant with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia, Beijing, China. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His current research interests include image modeling and low-level vision, discriminative learning, and biometrics. He has authored over 50 papers in those areas.

Dr. Zuo is an Associate Editor of the *IET Biometrics*.