2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Online Multi-Target Tracking with Tensor-Based High-Order Graph Matching

Zongwei Zhou*†, Junliang Xing*†, Mengdan Zhang*†, Weiming Hu*†
*CAS Center for Excellence in Brain Science and Intelligence Technology,
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
† University of Chinese Academy of Sciences
zhouzongwei2016@ia.ac.cn, {jlxing, mengdan.zhang, wmhu}@nlpr.ia.ac.cn

***Abstract*—In this paper we formulate multi-target tracking (MTT) as a high-order graph matching problem and propose a ℓ_1-norm tensor power iteration solution. Concretely, the search for trajectory-observation correspondences in MTT task is cast as a hypergraph matching problem to maximize a multi-linear objective function over all permutations of the associations. This function is defined by a tensor representing the affinity between association tuples where pair-wise similarities, motion consistency and spatial structural information can be embedded expediently. To solve the matching problem, a dual-direction unit ℓ_1-norm constrained tensor power iteration algorithm is proposed. Additionally, as measuring the appearance affinity with features extracted from the rectangle patch, which is adopted in most methods, has a weak discrimination when bounding boxes overlap each other heavily, we present a deep pair-wise appearance similarity metric based on object mask in this paper where just the features from true target region are utilized. Experimental evaluation shows that our approach achieves an accuracy comparable to state-of-the-art online trackers. The source code of the proposed approach will be released to facilitate further studies on the MTT problem.**

I. Introduction

Multi-target tracking (MTT) [1], and in particular people tracking, is critical for many applications, ranging from vision-based surveillance to human-computer interaction. A popular approach to generate the trajectories of multiple pedestrians is tracking-by-detection, where the targets are detected in a preprocessing step, usually either by background subtraction or using a discriminative classifier. By building on the advances in person detection over the last decade, especially detectors based on deep learning, tracking-by-detection has been very successful [2][4][5][6].

Within this paradigm, some models usually formulate object trajectory generation as a global optimization problem that processes video batches at once. One representative framework is flow network which models potential locations over time and finds trajectories with minimum cost [7][8][11][18][9][10]. The other popular batch solution recently is subgraph multicut formulation that considering data association as a subgraph decomposition problem [14] [19] [12] [13], which implicitly ensures long-term temporal consistency through transitivity constraints while still being based on node to node affinities, keeping the approach computational feasible.

However, due to batch processing, these methods are not applicable in online scenarios where a target identity must be available at each time step. On the contrary, online methods generate trajectories only using information up to the current frame which adopt probabilistic inference[17], [15] or deterministic optimization[16]. Such association based tracking methods are more sensitive to noisy detections since they often relies on matching locally between existing trajectories and new observations which only impose local constraints on the association. On this account, we propose an online multi-target tracking framework with the tensor-based high-order graph matching to introduce more global information for matching.

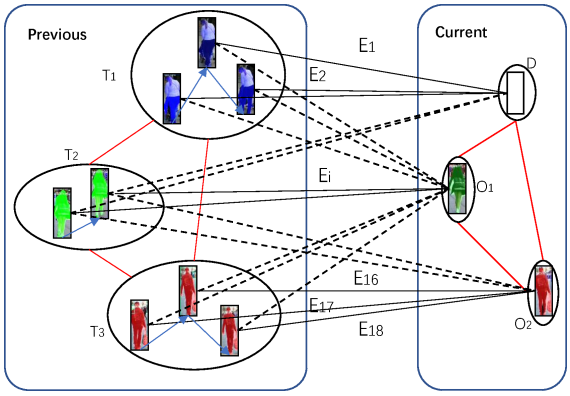

Fig. 1. An example illustrating assigning observations to trajectories with high-order graph matching. Hyper-edges composed by tracked detections sampled from trajectories or observations are represented by ellipses. Candidate assignments are represented by solid and dot lines ($E_1 \sim E_{18}$) and red lines reflect their geometric relations. Our goal is to find the node correspondences which preserve their topology by hypergraph matching and then assign observations to corresponding trajectories according to node correspondences. D represents missing detections and solid lines are the obtained associations finally, then we can assign O_1, O_2 to T_2, T_3 and assume the observation of trajectory T_1 is loss in current time step.

Specifically, searching for correspondences between current observations and some selected tracked detections in history is casted as a hypergraph matching problem with a high-order energy function composed of pair-wise appearance similarity, motion consistency and spatial structural information. In the hypergraph, each target (tracked detection or observation) is denoted by a node, each trajectory or observation is represented by a hyper-edge. Geometric relations between hyper-edges are extracted as spatial structural information. The matching problem is to find hyper-edge correspondences

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

between observations and trajectories which preserve their topology. As shown in Fig. 1, hyper-edges are represented by ellipses, geometric relations are represented by red lines. Weights of associated sides between observations and tracked detections ($E_1 \sim E_{18}$) representing matching probabilities are obtained via hypergraph matching, and then assignments from observations to trajectories can be drawn between them, for example, O_2 is assigned to T_3 considering the high value of $E_{16} \sim E_{18}$. Trajectory T_1 is matched to node D, a dummy node representing missing detection, which is induced relying heavily on the geometric relation while appearance similarity is missing.

To meet the common assumption in MTT task that at most one observation corresponds to one trajectory and one observation at most corresponds to one trajectory within same frame, a dual-direction unit ℓ_1-norm power iteration algorithm is presented to solve the specified hypergraph matching problem, where the overall energy from one trajectory to observations (forward) and that from one observation to trajectories (backward) are all bounded to be unit. The basic idea is to iteratively update the solution by tensor powering followed by ℓ_1 unit normalization on rows and columns of the assignment matrix respectively.

Appearance affinity is often used to characterize pairs' similarity and many works have been proposed to exploit appearance information. Most methods metric appearance affinity with features extracted from the regions in targets' bounding boxes [23][13]. However, it often has a weak discrimination power when bounding boxes are overlapped each other heavily as they share many common features. Thus, we present metric learning network based on object masks in this paper where only the features from true foreground are considered.

Three main contributions of this article are concluded:

- An online MTT framework is proposed by formulating the search of correspondences in MTT as a hyper-order graph matching problem which can integrate high-order information flexibly.
- A dual-direction unit ℓ_1 norm tensor power iteration algorithm is proposed to solve the variant hypergraph matching model considering the prior hypotheses in MTT that at most one observation corresponds to one trajectory and one observation at most corresponds to one trajectory within the same time frame.
- To better characterize pair-wise appearance affinity, a metric learning network based on identity masks is presented.

II. Multi-Target Tracking with High-Order Graph Matching

A. Problem statement

At time t, we assume trajectories $\mathcal{T} = \{T_1, T_2, T_N\}$ are obtained with $N_1 = \sum_n^{|\mathcal{T}|} |T_n|$ tracked detections sampled from them according to some principles (e.g. Fibonacci series) where T_n represents the number of sampled detections from nth trajectory, and N_2 observations are detected. Here, we do not assume that $N_1 = N_2$, i.e., there may be different numbers of trajectories and observations because trajectories may start or finish and false alarms or missing detections may occur at present. Throughout this paper, for s=1,2, all indices i_s, j_s, k_s will be assumed to vary from 1 to N_s. We will also denote by $i = (i_1, i_2), j = (j_1, j_2), k = (k_1, k_2)$ the possible associations of observations and trajectories.

Let $S^1_{i_1}, S^2_{i_2}$ be the tracked detections and observations respectively. The problem of matching observations to tracked detections is equivalent to looking for an $N_1 \times N_2$ assignment matrix X, such that $X_{(i_1,i_2)}$, a.k.a. X_i, is 1 when this assignment $S^1_{i_1}$ to $S^2_{i_2}$ is selected, and 0 otherwise. In MTT task, a common assumption is that a trajectory is matched to at most one observation and vice versa, i.e., the summary of each row and columns is equal to one[1]. Thus we consider the set $\mathcal{X}$ of assignment matrices:

$$\begin{aligned} \mathcal{X} =& \{X \in \{0,1\}^{N_1 \times N_2}\}, \\ \text{s.t.} \quad & \forall i_1, \sum_{i_1} X_{i_1,i_2} = 1; \forall i_2, \sum_{i_2} X_{i_1,i_2} = 1. \end{aligned} \tag{1}$$

Assignment in MTT task is formulated as the maximization of the following score over $\mathcal{X}$ to select $N = \min(N_1, N_2)$ best connections from candidate space considering global information:

$$score(\mathcal{X}) = H_{c_1 c_2 \cdots c_N} X_{c_1} X_{c_2} \cdots X_{c_N} \tag{2}$$

where $c_i \in R^{N_1 N_2}, i = 1, 2, \cdots N$ represents the index of candidate connections and $H_{c_1 c_2 \cdots c_N}$, which is described in Section III, is an integrated potential corresponding to the combination of connections. It is a positive likelihood measure, the higher the value of H is, the more likely the combinations are selected.

However, in practice, it could lead to an exponential growth of the computational complexity with the huge search space to take n-order information into consideration. Thus, just p-order $(p \leq 3)$ information are considered in this paper. $p = 1, 2, 3$ denote one-to-one, pair-to-pair and triple-to-triple comparisons for matching respectively. Of course, higher-order information can be integrated straightforward. For simplify, without loss of generality, our model is illustrated with $p = 3$. Then the matching problem can be formulated as :

$$\begin{aligned} \max \quad & score(\mathcal{X}) = \sum_{i,j,k} H_{i,j,k} X_i X_j X_k \\ \text{s.t.} \quad & \begin{cases} \sum_{i_2} X_{i_1 i_2} = 1, & i_1 = 1, 2, \cdots, N_1 \\ \sum_{i_1} X_{i_1 i_2} = 1, & i_2 = 1, 2, \cdots, N_2 \end{cases} . \end{aligned} \tag{3}$$

The procedure is illustrated in Fig.2. It can be viewed as a variant of rank-1 tensor approximation formulation described in next subsection . And an efficient customized power iteration method, named dual-direction unit ℓ_1-norm power iteration, is presented in II-C to solve the problem above.

[1]Dummy nodes are added to trajectories and observations set respectively to interpret the beginning or ending of trajectories, or missing detections. Unit ℓ_1-norm constraint needs not to be satisfied for dummy node.

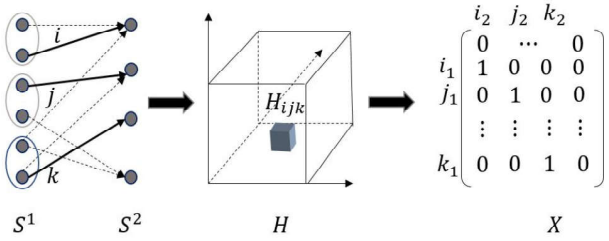

Fig. 2. Illustration of Eq.(3). S^1 denotes 3 trajectories with 6 sampled detections, and S^2 denotes 4 detections in current time-step. H denotes the energy tensor, and H_{ijk}, X corresponding to the energy and solutions respectively when edges i, j, k are selected simultaneously.

Observation O_{i_2} is assigned to corresponding nth trajectory after connections are selected by

$$n = \text{argmax}_n \frac{1}{|T_n|} \sum_{i_1} X_{i_1 i_2}, \quad \forall i_1 \in T_n \tag{4}$$

where $n = 1, 2, \cdots, |\mathcal{T}_n|$ is the index of trajectory. The procedure for online multi-target tracking with tensor-based high-order graph matching is presented in Algorithm 1.

B. Tensor formulation

A tensor is the high dimensional generalization of a matrix. For a K-order tensor $\mathcal{S} \in R^{I_1 \times I_2 \cdots \times I_K}$, each element is denoted as $\mathcal{S}_{i_1 \ldots i_k \ldots i_K}$ and $1 \leq i_k \leq I_k$. In the tensor terminology, each dimension of a tensor is associated with a mode.

A tensor and a vector can be multiplied like matrix-vector as following notation:

$$\mathcal{B} = \mathcal{S} \otimes_n V$$
$$\mathcal{B}_{i_1 \ldots i_{n-1} i_{n+1} \ldots i_K} = \sum_{i_n} \mathcal{S}_{i_1 \ldots i_{n-1} i_n \ldots i_K} V_{i_n}$$

where V is a I_n-dimensional vector. Like a matrix multiplied by a vector produces a vector, a $(K-1)$-order tensor $\mathcal{B} \in R^{I_1 \times \cdots I_{n-1} \times I_{n+1} \cdots \times I_K}$ is obtained in tensor-vector multiplication. $\otimes_n$ represents we multiply on the nth dimension.

According to the definition in [24], Eq.(5) below can be formulated as a tensor rank-1 approximation problem which can be solved efficiently by power iteration algorithm[24].

$$\begin{aligned} &\max \sum \mathcal{S}_{i_1 i_2 \cdots i_K} V_{i_1}^{(1)} V_{i_2}^{(2)} \cdots V_{i_K}^{(K)} \\ &\text{s.t.} \|V^k\|_2^2 = 1, \forall k \in \{1, 2, ..., K\} \end{aligned} \tag{5}$$

C. Dual-direction unit ℓ_1-norm power iteration

Different from the unit l_2-norm constraint in tensor rank-1 approximation of Eq.(5), the matching problem of Eq.(3) in MTT is attempt to obtain an assignment matrix with unit ℓ_1-norm rows (one tracked detection to all observations, forward unit ℓ_1-norm) and columns (one observation to all tracked detections, backward unit ℓ_1-norm). To address the issue raised from these constraints, we advocate a dual-direction unit ℓ_1-norm power iteration algorithm to solve Eq.(3) efficiently with proved convergence. The basic idea is to reformulate the Eq.(3)

Algorithm 1 Online multi-target tracking with tensor-based high-order graph matching

Require: $\eta, \epsilon, w_1, w_2, T = \emptyset, order = 3$

$D \leftarrow$ All detections in sequence.
$F \leftarrow$ The number of frames in sequence.
for $f = 1, 2, \cdots, F$ **do**
  $D_f = \{d_{ij} | i = 1, 2, ...|T|, j = 1, 2, 3...\}$, {Tracked detections sampled from T_i}
  $O_f = \{O_i | i = 1, 2, ...\}$,{Observations on present moment}
  $\mathcal{X} = \{X_i | i = 1, 2, ...\}$,{Candidate connections filtered according to Eq.(7)}
  $H \in R^{c_1 \times c_2 \times c_3}$,{High-order energy tensor calculated according to Eq.(8)}
  $V \in R^{|\mathcal{X}|} \leftarrow$random
  repeat
    $V_0 \leftarrow V$
    $V \leftarrow V \circ H \otimes_1 (V \circ V) \otimes_2 (V \circ V)$
    $\forall i, V(i,:) \leftarrow \frac{1}{\|V(i,:)\|_2} V(i,:)$
    $\forall j, V(:,j) \leftarrow \frac{1}{\|V(:,j)\|_2} V(:,j)$
  until $\|V - V_0\| \leq \epsilon$,{Dual-direction unit ℓ_1-norm power iteration}
  Update T
end for
return T

as follows by introducing $Y_i^2 = X_i, (Y_i \in [0, 1])$ to transfer ℓ_1-norm to l_2-norm firstly.

$$\begin{aligned} \max_Y \quad & \sum_{i,j,k} H_{i,j,k} Y_i^2 Y_j^2 Y_k^2 \\ \text{s.t.} \quad & \begin{cases} \sum_{i_2} Y_{i_1 i_2}^2 = 1, & i_1 = 1, 2, \cdots, N_1 \\ \sum_{i_1} Y_{i_1 i_2}^2 = 1, & i_2 = 1, 2, \cdots, N_2 \end{cases} \end{aligned} \tag{6}$$

And then iteratively update the solution by tensor powering followed by l_2 unit normalization on rows and columns of assignment matrix X respectively. The procedure for customized rank-1 tensor approximation is presented in the inner loop of Algorithm 1, where '$\circ$' indicates the Hadamard product (element-wise product), and '$\otimes_k$' denotes the Kronecker product. The convergence of the iterative process can be proved simply referring to [25]. Though this method is not guaranteed to reach a global optimum, [26] proposes a smart way to initialize it, leading to a quantifiable proximity to the optimal solution.

III. Building Tensor for High-Order Graph Matching

High-order potentials are used in our model to characterize either the geometric invariance of identities, or the motion consistency. Assume trajectory set $\mathcal{T} = \{T_1, T_2, \cdots, T_{N_T}\}$ has been tracked and observation set $\mathcal{O} = \{O_1, O_2, \cdots, O_{N_2}\}$ has been detected on present moment. Tracked detections $\mathcal{D} = \{D_1, D_2, \cdots, D_{N_1}\}$ are sampled from $\mathcal{T}$ in time window $\mathcal{N}$ according to the Fibonacci sequence. D_i, O_j are quadruples

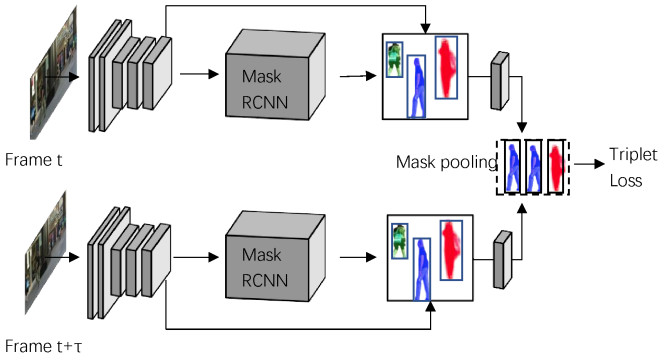

Fig. 3. Architecture of our Siamese network to extract appearance affinity based on identity masks.

with (x, y, w, h) representing center location, width and height of corresponding bounding box. Connections between nodes in $\mathcal{D}$ and $\mathcal{O}$ are filtered by distance threshold η according to Eq.(7) to avoid searching the huge solution space produced by connecting tracked detections to observations one-to-one.

$$\mathcal{C} = \{(i_1, i_2) | D_{i_1} \in T_t, \frac{d(P_t, O_{i_2})}{h_{i_2}} \leq \eta\} \tag{7}$$

where P_t is the predicted location in current time of trajectory T_t with uniform motion assumption, h_{i_2} represents the height of observation O_{i_2} and $d(a, b)$ denotes the Euclidean distance between the centers of a and b.

Then a high-order tensor $H \in R^{K \times K \times K}, K = |\mathcal{C}|$ is constructed to integrate appearance affinity, motion consistency and spatial structural potential as follows:

$$H_{ijk} = \phi_a(i, j, k)\phi_m(i, j, k)\phi_s(i, j, k). \tag{8}$$

where $i, j, k \in \mathcal{C}$ and $\phi_a(i, j, k), \phi_m(i, j, k), \phi_s(i, j, k)$ represents appearance affinity, motion consistency and spatial potential respectively when connections i, j, k are selected simultaneously.

A. Appearance affinity

The distance between appearance features is frequently used for computing the affinity in data association.The affinity value based on the ideal appearance feature should be large for persons of the same identity and be relative small for persons of different identities. In many approaches, appearance affinity is obtained by a siamese network with person patch pair or tuple as import. However, it is hard to distinguish identities that close to each other as the bounding boxes are overlap sharing many common features. Thus, we extract the appearance features just from the mask area using a Siamese network framework based on Mask RCNN with a ResNet-FPN backbone[27] (see Fig. 3). Targets' boxes and masks are produced by Mask RCNN, and the bottom features (final convolutional layer of the 3rd stage of ResNet50) are selected by masks. Then triple samples are selected as input to a shallow siamese network to extract the 128-dimensional appearance features. The triplet loss is defined as follows:

$$\mathcal{L} = \sum_i \max(\cos(pred_i, neg_i) - \cos(pred_i, pos_i) + m, 0) \tag{9}$$

where $pred_i, pos_i, neg_i$ are features of prediction, positive and negative samples respectively, m denotes the margin of separation between correct and incorrect pair and $cos(a, b)$ denotes the cosine distance between a and b . Then the energy produced by appearance affinity when tuple of connections are selected can be calculated by:

$$\phi_a(i, j, k) = a_i a_j a_k \tag{10}$$

where $a_i = a_{(i_1, i_2)} = cos(f_{i_1}, f_{i_2})$, and $f_{i.}$ is the deep appearance feature extracted by network.

B. Motion consistency

In common, the velocity of one target is assumed to be a constant in a short period. Thus, a simple linear model is used to predict the target's state (x, y, w, h) at present. And the motion consistency can be characterized as follows:

$$\phi_m(i, j, k) = \exp\big(-\frac{w_1}{3} \sum_{c=\{i,j,k\}} \|P_{t_c} - O_{l_c}\|_2\big) \tag{11}$$

where w_1 is a weight parameter, t_c is the index of a trajectory which the tracked detections in connection c belong to, and l_c is the index of the observation connected by c.

C. Spatial potential

Linear motion model is too simple to predict desired state in some cases, such as, motion of camera. In these cases, relative structural information of nodes (trajectories or observations) can be utilized as it is an affine-invariant potential. To model this affine-invariant property, we define the spatial potential of connection tuples in H,

$$\phi_s(i, j, k) = \exp\big(-w_2 \sum_{c_1, c_2} \|d(P_{t_{c1}}, P_{t_{c2}}) - d(O_{l_{c1}}, O_{l_{c2}})\|\big) \tag{12}$$

where w_2 is a weight parameter, $c_1, c_2 \in \{i, j, k\}$. Note here, we use absolute difference of distances of nodes rather than relative ratio as the larger the distance, the smaller the possibility of changing the relative position.

IV. Experiments

We evaluate the performance of our tracker on the MOT16 challenge benchmark[6]. The benchmark includes training and test sets composed of seven sequences respectively, including frontal-view scenes with moving camera as well as top-down surveillance setups. Model parameters for the test sequences are provided based on the corresponding training sequences[2]. Evaluation is carried out according to the following metrics:

- Multi-Object Tracking Precision (MOTP): It usually depends on the precision of detectors.

[2] $w_1 = 0.7, w_2 = 0.3$ for all sequences except MOT16-13 and MOT16-14 where $w_1 = 0.4, w_2 = 1.1, \eta = 0.5$

TABLE I

TRACKING RESULTS ON THE MOT16 CHALLENGE. WE COMPARE TO OTHER PUBLISHED METHODS WITH PRIVATE DETECTIONS. THE FULL TABLE OF RESULTS CAN BE FOUND ON THE CHALLENGE WEBSITE.

		MOTA(%) ↑	MOTP (%) ↑	IDF1 (%) ↑	MT (%) ↑	ML (%) ↓	ID ↓	Frag↓	FP ↓	FN ↓	Hz (fps) ↑
HT_SJTUZTE[28]	BATCH	**71.3**	79.3	67.6	46.5	19.5	617	743	9238	**42521**	29.0
LMP_p [19]	BATCH	71.0	**80.2**	**70.1**	**46.9**	21.9	**434**	**587**	7880	44564	0.5
KDNT [29]	BATCH	68.2	79.4	60.0	41.0	**19.0**	933	1093	11479	45605	0.7
MCMOT_HDM[31]	BATCH	62.4	78.3	51.6	31.5	22.2	1394	1318	9855	57257	**34.9**
NOMTwSDP16 [20]	BATCH	62.2	79.6	62.6	32.5	31.1	406	642	**5119**	63352	3.1
POI [29]	ONLINE	**66.1**	79.5	65.1	34.0	20.8	805	3093	**5061**	55914	9.9
DeepSORT_2[32]	ONLINE	61.4	79.1	62.2	32.8	**18.2**	**781**	2008	12852	56668	17.4
SORTwHPD16[34]	ONLINE	59.8	**79.6**	53.8	25.4	22.7	1423	1835	8698	63245	59.5
IOU [33]	ONLINE	57.1	77.1	46.9	23.6	32.9	2167	3028	5702	70278	**3004.6**
HOGM(ours)	ONLINE	64.8	78.6	**73.5**	**40.6**	22.0	794	**1050**	13470	**49927**	18.2

- Multi-Object Tracking Accuracy (MOTA): Combination of three error sources of false positives, missed targets and identity switches.
- ID F1 Score (IDF1): Ratio of correctly identified detections over the average number of ground-truth and computed detections[30].
- Mostly Tracked targets (MT), Mostly lost targets (ML): Ratios of ground-truth trajectories that are covered by a track hypothesis for at least 80%, at most 20% of their respective life span.
- False Positives (FP) and False Negatives (FN).
- Identity switches (IDs): Number of times the reported identity of a ground-truth track changes.
- Fragmentation (Frag): Number of times a trajectory is interrupted during tracking.

The results of our evaluation are shown in Table I. Our tracker is a strong competitor to other online tracking frameworks. In particular, our approach returns the highest identified detection score, MT and fewest fragments of all online methods while maintaining competitive MOTA scores, ML and identity switches. We note that our method returns a higher number of false positives which impairs the reported tracking accuracy. In common sense, applying a larger confidence threshold to the detections can potentially increase the performance. However, visual inspection of the tracking output shows that, similar to DeepSORT, most of the false positives in our model are generated from the sporadic detector responses at static scene geometry. Due to our high-order spatial structural information and larger margin of temporal distance, these are usually associated by trajectories. However, as demonstrated by the score of IDF1, which is more appropriate than MOTA to evaluate the robustness of the tracker, these mismatches do not lead to continually identity switches. Some qualitative results shown in Fig. 6 that targets can be tracked correctly even occlusions are encountered or the scene changes greatly as camera movement. Additionally, our model is even more outstanding than some state-of-the-art methods in batch mode, such as NOMT which is significantly more complex and uses frames in the near future.

To evaluate the effectiveness of proposed high-order affinity containing motion consistency and spatial structural information, change of MOTA is counted under different order and temporal distance $\mathcal{N}$, i.e. the length of history, on the MOT16-10 which is recorded with a moving camera. As illustrated in Fig. 5, the MOTA and IDF1 are all improved along with the increase of order as richer spatial information is extracted. And we observe that more samples from history are benefit to the performance of our tracker within a certain period. However, the performance is not improved or even declined when the temporal distance is too long. The possible reason is that the linear motion assumption is no longer set up and the appearance may change greatly after a long time interval.

We also compare the proposed tracker with the DeepSORT as its code is available and the performance is better than other online trackers demonstrating the effectiveness of our feature extraction network based on the identity mask on training sequences. As shown in Fig. 4, no matter DeepSORT or our model (HOGM), using the features with identity masks is better than that without. And the proposed method in this paper outperform DeepSORT with same features.

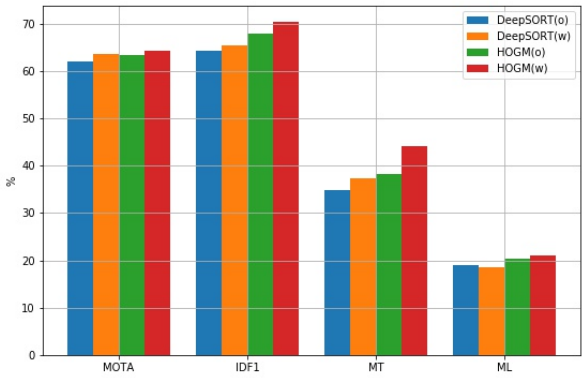

Fig. 4. Tacking results on the training sequences using features with/without (w/o) identity masks.

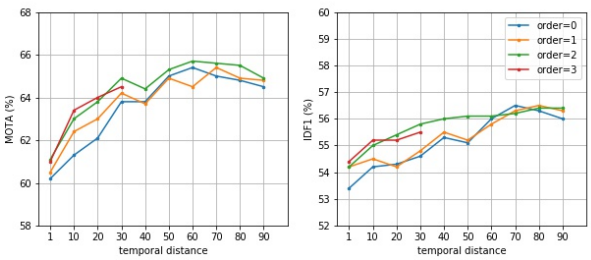

Fig. 5. Performance of our approach on MOT16-10 with different order of energy tensor. We set maximum temporal distance=30 when order=4 considering the huge search space.

Our implementation runs at approximately 18Hz and it can be much faster with parallel operation as the assignment in our model is mainly operated on tensors.

Fig. 6. Qualitative results on some test sequences of the MOT16 benchmark. Same identity labeled by box with same color.

V. Conclusion

In this paper, we propose a multi-target tracker that formulates assignment task as a high-order graph matching problem with an effective solution. With this formulation, high-order information, such as motion consistency and spatial structural invariance, can be embedded expediently. Additionally, with the framework of Siamese network, features based on identity masks are extracted to characterize appearance similarity. Experiments showed that appearance affinity with features extracted from masks area is better than from whole patch. And the proposed framework achieved an accuracy comparable to the state-of-the-art on line trackers. Yet, the algorithm to the formulation remains neat to implement and runs in real time. As clarified before, linear motion model sometimes is not appropriate and generates some false positives impairing the finally accuracy of the tracker, future work will investigate an adaptive model to characterize the motion consistency.

Acknowledgment

This work is supported by the Natural Science Foundation of China (Grant No. 61672519, 61751212, 61472421, 61602478), the NSFC-general technology collaborative Fund for basic research (Grant No. U1636218), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, and the CAS External cooperation key project.

References

[1] W. Luo, J. Xing, A. Milan, X. Zhang, X. Zhao, and T.-K. Kim,*Multiple object tracking: A literature review*. In arXiv:1409.7618, 2014.

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

[3] A. Dehghan, Y. Tian, P. H. Tor, and M. Shah, *Target identity aware network flow for online multiple target tracking.*. In CVPR, 2015.

[4] J. Xing, H. Ai, and S. Lao, *Multiple human tracking based on multi-view upper-body detection and discriminative learning*. In ICPR, 2010.

[5] L. Leal-Taixe, A. Milan, I. Reid, S. Roth, and K. Schindler, *MOTChallenge 2015: Towards a benchmark for multi-target tracking*. In arXiv:1504.01942, 2015.

[6] A. Milan, L. Leal-Taixe, I. D. Reid, S. Roth, and K. Schindler,*MOT16: Abenchmark for multi-object tracking*. In arXiv:1603:00831, 2016.

[7] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, *Multiple object tracking using k-shortest paths optimization*. TPAMI, vol. 33, no. 9, pp. 1806-1819, 2011.

[8] A. Dehghan, Y. Tian, P. H. Tor, and M. Shah, *Target identity aware network flow for online multiple target tracking.*. In CVPR, 2015.

[9] X. Wang, E. Turetken, F. Fleuret, and P. Fua, *Tracking interacting objects using intertwined flows*. TPAMI, vol. 38, no. 11, pp. 2312-2326, 2016.

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

[11] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, *Globally optimal greedy algorithms for tracking a variable number of objects*. In CVPR, 2011.

[12] M. Keuper, S. Tang, Z. Yu, B. andres, T. Brox and B. Schiele,*Hidden Hands: Tracking Hands With an Occlusion Aware Tracker*. In CVPR, 2016.

[13] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, *Learning by tracking: Siamese CNN for robust target association*. In CVPR Workshops, 2016.

[14] S. Tang, B. Andres, M. Andriluka, and B. Schiele, *Subgraph decomposition for multi-target tracking*. In CVPR, 2015.

[15] J. Xing, H. Ai, L. Liu, and S. Lao, *Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling*. TIP, vol. 20, no. 6, pp. 1652-1667, 2011.

[16] S. H. Bae and K. J. Yoon, *Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning*. In CVPR, 2014.

[17] S. Oh, S. Russell, and S. Sastry, *Markov chain monte carlo data association for multi-target tracking*. TAC, vol. 54, no. 3, pp. 481-497, 2009.

[18] S. Gao, Q. Ye, J. Xing, A. Kuijper, Z. Han, J. Jiao, and X. Ji, *Beyond group: multiple person tracking via minimal topology-energy-variation*. TIP, vol. 26, no. 12, pp. 5575-5589, 2017.

[19] S. Tang, M. Andriluka, B. Andres, B. Schiele, *Multiple people tracking by lifted multicut and person re-identification*. In CVPR, 2016.

[20] W. Choi,*Near-onlinr multi-target tracking with aggregated local flow descriptor*. In ICCV, 2015.

[21] 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.

[22] O. Duchenne, F. Bach,I-S. Kweon, and J. Ponce,*A tensor-based algorithm for high-order graph matching*. TPAMI, vol. 33, no. 12, pp. 2383-2395, 2011.

[23] C. Kim, F. Li, A. Ciptadi, and I. M. Rehg, *Multiple hypothesis tracking revisited*. In ICCV, 2015.

[24] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, *On the Best Rank-1 and Rank-(R1 ,R2 ,. . .,RN) Approximation of Higher-Order Tensors*. JMAA, vol. 21, no. 4, pp. 1324-1342, 2000.

[25] X. Shi, H. Ling, J. Xing, W. Hu, *Multi-target tracking by rank-1 tensor approximation*. In CVPR, 2013.

[26] P. A. Regalia and E. Kofidis, *The higher-order power method revisited: convergence proofs and effective initialization*. In ICASSP, 2000.

[27] K. HE, G. Gkioxari, P. Dollar and R. Girshick, *Mask R-CNN*. In ICCV, 2017.

[28] W. Lin, J. Peng, S. Deng, M. Liu, X. Jia and H. Xiong,*Real-time multi-object tracking with hyper-plane matching (v2)*. Tech Report, Shanghai Jiao Tong University & ZTE Corp, 2017.

[29] 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 BMTT, 2016.

[30] E. Ristani, E. Solera, R. Zou, R. Cucchiara, and C. Tomasi, *Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking*. In ECCV Workshops, 2016.

[31] B. Lee, E. Erdenee, S. Jin, M. Nam, Y. Jung and P. Rhee, *Multi-Class Multi-Object Tracking using Changing Point Detection*. In BMTT, 2016.

[32] N. Wojke, A. Bewley and D. Paulus. *Simple Online and Realtime Tracking with a Deep Association Metric*. In arXiv:1703.07402, 2017.

[33] E. Bochinski, V. Eiselein, T. Sikora,*High-Speed Tracking-by-Detection Without Using Image Information*. In AVSS Workshops, 2017.

[34] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, *Simple online and realtime tracking*. In ICIP, 2016.