Tracklet Siamese Network with Constrained Clustering for Multiple Object Tracking

Jinlong Peng [1], Fan Qiu [1], John See [2], Qi Guo [3], Shaoshuai Huang [3], Ling-Yu Duan [4], Weiyao Lin [#1]
[1] *Department of Electronic Engineering, Shanghai Jiao Tong University*
[2] *Faculty of Computing and Informatics, Multimedia University, Malaysia*
[3] *Research & Advanced Technology Division, SAIC Motor Corporation Limited*
[4] *The National Engineering Lab for Video Technology, Peking University*
[#] Corresponding author:wylin@sjtu.edu.cn

***Abstract*—Multiple object tracking (MOT) is an important yet challenging task in video understanding and analysis. Basically, MOT aims to associate detected objects into trajectories based on their temporal relationships. The occlusion among moving objects poses a major challenge towards robust modeling of these relationships. In this paper, we propose a novel Tracklet Siamese Network (TSN) for learning similarities between tracklets characterized by appearance information, achieving superior performance on two MOTChallenge benchmark datasets. Our framework constructs short tracklets from highly-related object detections by excluding inaccurate object detections. We also adopt a constrained clustering technique to piece tracklets together into long trajectories, thus recovering many missing detections caused by original detector or the detection removing in the previous step. Comparisons against state-of-the-art methods were reported while ablation studies further substantiate the viability of components in our approach.**

***Index Terms*—Multiple object tracking, tracklet, tracklet siamese network, local temporal pooling, constrained clustering**

I. Introduction

Multiple object tracking (MOT) is an important field in computer vision, which is useful for applications such as behavior analysis, traffic security and robotics [1], [2]. The purpose of MOT is to associate detected objects from different frames to generate complete object trajectories [3], [4]. In many high-density scenes, object occlusion is a major problem as shown in Fig. 1. As such, misdetections or tracking drift often occur, resulting in a deterioration in performance.

Since the key task in MOT involves associating or linking detected objects, most research works focus on developing effective object association strategies [5], [6], [7]. For example, Tang et al. [5] modeled their technique as a Minimum Cost Subgraph Multicut Problem, which clusters targets jointly across space and time. Yang et al. [6] proposed a min-cost multi-commodity network flow within a hybrid framework to fuse global optimization and local optimization in data association. Gao et al. [7] designed a social-topology model which combines intra- and inter-group structures and learns typical topology patterns to improve the accuracy of target association. Since these methods are inattentive towards the quality of detection during association, their results are often riddled by noisy detections caused by object occlusions.

Moreover, some recent research used deep learning methods to develop robust object-wise similarity measures for better detection association [8], [9], [10], [11]. Schulter et al. [8] proposed a deep network flow method while Tang et al. [9] advocated a DeepMatching algorithm; both of which were introduced to learn features for object similarity by minimizing pairwise association costs. In addition, Son et al. [10] designed a Quadruplet Convolutional Neural Networks to achieve end-to-end tracking. Wang et al. [11] proposed a Siamese Network with a metric learning based loss function that learns directly from pairs of warped target images. Although these methods show improvements over their predecessors, they do not make full use of the appearance information of the *tracklets*, or object track fragments, which may provide less noisy input and better representation than images.

Fig. 1. An example of occlusion during tracking. Different colors denote different tracklets. The dashed boxes indicate noisy detections that could lead to inaccurate trajectories.

To address these problems, we propose a novel approach to MOT based on high-confidence tracklet generation and constrained clustering. We first construct short tracklets from temporally highly-related object detections, which can exclude most imprecise detections. Then, we propose a tracklet siamese network (TSN) to learn the similarity between tracklets by fully utilizing the appearance information of tracklets. To associate tracklets that are similar, the short tracklets are clustered to generate the long trajectories, with constraints in place to handle temporal overlaps. This process ensures that missing detections are complemented by the interpolation between two neighboring tracklets representing the same object.

In summary, the main contributions of our approach are two-fold: (1) We design a new tracklet siamese network to improve the accuracy and robustness of learning tracklet similarities; (2) We extend a constrained clustering method to associate short tracklets into long trajectories.

The rest of the paper is organized as follows. The details of our proposed method are introduced in Sec. II. Sec. III shows the experimental results, with ablation studies and comparisons against other methods. Finally, Sec. IV concludes the paper.

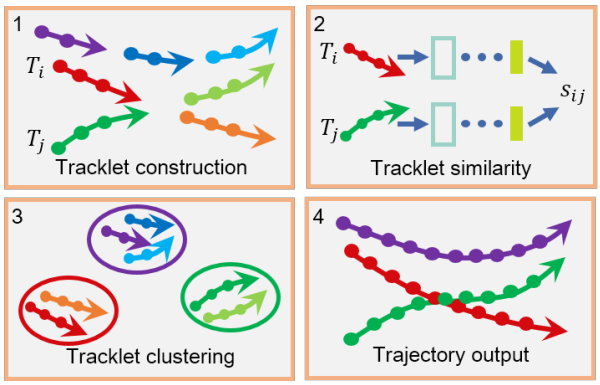


Fig. 2. Framework of the proposed approach. First, the detections (the dots) are linked to form different tracklets (the arrows). Then, the tracklet similarities are learned by the tracklet siamese network. Finally, the tracklets are clustered and associated to generate complete trajectories.

II. Method

As shown in Fig. 2, our proposed method aims to link detections to form high-confidence tracklets and further associate them into long trajectories by clustering. We describe our method in three steps: tracklet construction, tracklet similarity calculation and constrained tracklet clustering.

A. Tracklet Construction

The tracklet construction process aims to merge highly-related objects into high-confidence tracklets, such that they can be used as basic units for the clustering process later.

We first apply a min-max normalization on the detection confidence scores of each detected object, and exclude low confidence ones. Then, the classic Kuhn-Munkres (KM) algorithm is used to dynamically associate temporally related objects into tracklets [12]. During object association, we model the similarity between an object and a tracklet by appearance similarity and motion similarity [12]. To ensure the reliability of the tracklet, we enforce a strict merging bound; an object D is allowed to join a tracklet T only when both of its appearance and motion similarity to T is larger than 0.8.

B. Tracklet Similarity Calculation

To improve the accuracy of tracklet association, it is important to design an effective tracklet similarity calculation method. For simplicity, some recent methods model tracklet-wise similarity by the similarity between the features of tracklets' terminal objects [5], [6]. However, since the terminal object of the tracklet is not representative of the entire tracklet, and there is likelihood of noise in the terminal object detection caused by occlusion, it is inaccurate to directly use features from the terminal object alone. Therefore, we design a tracklet siamese network (TSN) to learn tracklet similarity by using all the appearance information of the tracklets.

Architecture. The architecture of TSN is showed in Fig. 3. The input of the TSN is two tracklets of any length while the output is the similarity score between the two tracklets. All images of the input tracklets will be input to a backbone CNN to obtain a series of feature vectors for each input tracklet. The

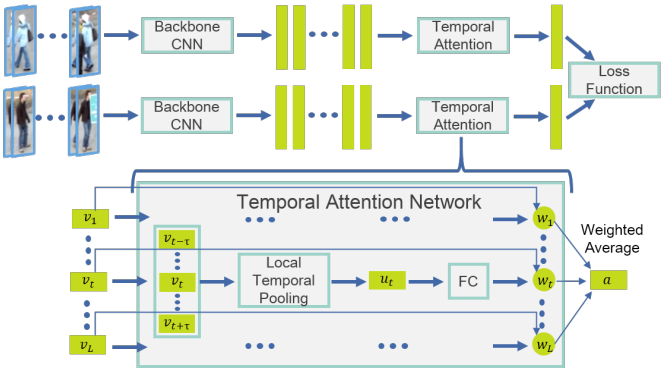


Fig. 3. The Tracklet Siamese Network (TSN). Its framework is displayed on top while the details of the temporal attention network is displayed at the bottom. The input to TSN is two tracklets and the output is a similarity score.

temporal attention network then takes these features to fuse the appearance information of all detected objects in the tracklet.

In the temporal attention network, every feature vector is concatenated with its neighboring 2τ vectors before feeding it to a *local* temporal pooling layer, which only considers a localized part of the tracklet information. The reason behind this is to accommodate potential occlusions of the object, which usually occur in a short few continuous frames. The local temporal pooling layer is defined as:

$$u_t^k = \max(v_{t-\tau}^k, \cdots, v_{t-1}^k, v_t^k, v_{t+1}^k, \cdots, v_{t+\tau}^k) \quad (1)$$

where v_t^k is the kth element of the input feature vector v_t and u_t is the corresponding pooled vector of v_t. The value of the constant τ is discussed in detail in Sec. III.

Intuitively, this mechanism generates a pooled vector which includes the local temporal information. Further to this, all pooled vectors are input to a fully connected layer, which in turn produces a set of weights. The final feature vector is represented by the weighted average of the input feature vectors:

$$a = \sum_{t=1}^{L} w_t v_t \quad (2)$$

where a is the final tracklet feature vector. w_t is the corresponding weight of v_t and L is the length of the input tracklet.

Learning and loss function. The TSN consists of two parallel networks that share the same parameters, and is trained by using the contrastive loss:

$$\mathcal{L} = y||a_i - a_j||_2^2 + (1-y)max(m - ||a_i - a_j||_2, 0)^2 \quad (3)$$

where a_i and a_j are feature vectors of two input tracklets. The margin m is set to 0.2, $y = 1$ denotes the two tracklets are of the same objects while $y = 0$ denotes different objects. During test phase, the similarity between two tracklets is calculated by the cosine similarity between their feature vectors.

C. Constrained Tracklet Clustering

After obtaining tracklets and their respective similarities, we need to associate them to form long trajectories. To make full use of the global information of all tracklets, we extend the clustering method of Rodriguez and Laio [13] to perform the association. The basic idea of this clustering method is

that cluster centers are characterized by a higher local density than their neighbors and by a relatively large distance from sample points with higher local densities [13]. In our method, tracklet is the basic clustering unit and the local density ρ_i of tracklet T_i is measured by similarity (instead of distance):

$$\rho_i = \sum_{j:O(T_i,T_j)=0} \chi(s_{ij} - s_c) \tag{4}$$

where $\chi(x) = 1$ if $x > 0$ and $\chi(x) = 0$ otherwise. The similarity threshold s_c is set at 0.5 while $O(T_i, T_j) = 0$ constrains that tracklets T_i and T_j do not overlap in the temporal domain, which ensures that two tracklets with temporal overlapping could not be considered as the same object.

The maximum similarity δ_i between tracklet T_i and any other tracklet T_j with higher density is defined as:

$$\delta_i = \max_{j:\rho_j>\rho_i, O(T_i,T_j)=0} (s_{ij}) \tag{5}$$

In our method, we recognize tracklets that fulfill $\delta_i < s_c$ as cluster centers so that the similarity between any two cluster centers is always lower than s_c.

After finding the cluster centers, we can assign the remaining tracklet to the same cluster as its most similar tracklet of higher density. Note that two tracklets with temporal overlapping may not belong to the same object. Thus, if a new tracklet is in conflict with any tracklet in the cluster, it should be assigned to the next most similar tracklet of higher density; this step is repeated until the conflict is resolved. If the tracklet is in conflict with all the clusters, it is then deleted.

When the clustering process is finished, the tracklets belonging to the same cluster will be sorted according to the frame index. The neighboring tracklets in the same cluster are then associated to generate a long trajectory while the missing detections are added by linear interpolation.

III. Experiments

This section reports the experiments that were conducted. We first introduce the datasets used and experimental details. Then, we report some ablation studies on our method and make comparisons against recent state-of-the-art methods.

A. Datasets and Implementation Details

We perform experiments on two benchmark datasets: **MOT15** [14] and **MOT16** [15]. The MOT15 dataset contains 11 training sequences and 11 test sequences. The MOT16 dataset contains 7 training sequences and 7 test sequences. All sequences in MOT15 and MOT16 are captured by static or moving cameras, with different scenes and resolutions. These datasets are challenging as the sequences contain a large amount of objects and various types of object occlusions.

We evaluate our method by the CLEAR MOT metrics [16], including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), False Negatives (FN), False Positives (FP), Identity Switches (IDs), Mostly Tracked Trajectories (MT) and Mostly Lost Trajectories (ML). Specifically, MOTA is the primary evaluation metric involving FN, FP, IDs which typically depicts the overall performance.

TABLE I
Ablation study on MOT16 validation dataset.

Method	MOTA↑	MOTP↑	MT↑	ML↓
KM	38.1	77.2	13.7%	50.3%
Greedy+Terminal	39.2	77.3	15.2%	48.1%
Clustering+Terminal	40.6	77.3	16.4%	44.6%
Clustering+Siamese	**42.4**	**77.5**	**18.2%**	**40.8%**

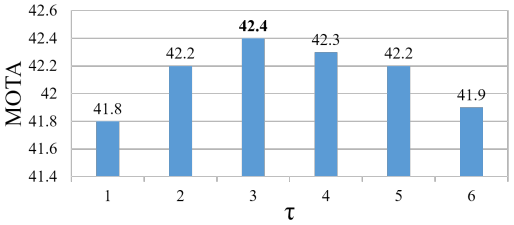


Fig. 4. The tracking performance on MOT16 validation dataset for different values of τ.

For fair comparison with other MOT methods, our experiments utilize the publicly available detection results provided by the MOT15 and MOT16 datasets [14], [15]. The tracklet construction and constrained tracklet clustering steps do not require any training. TSN is trained by the sequences with ground-truth from MOT15 and MOT16 while the testing process is performed by using all their respective test sequences. A VGG-16 pre-trained model is adopted as the backbone of the TSN, with τ in the temporal attention network set to 3. In training, the batch size is set to 8 and the learning rate is set to 0.0001 initially. We train for 96000 iterations, with the learning rate repeatedly halved every 16000 iterations.

B. Ablation Studies

We also conduct ablation studies to demonstrate how our approach would fare without certain essential components. For ease, we ran these experiments using image sequences (with publicly available detection results) from the MOT16 validation set to make comparisons:

(1) *KM*. Directly applying the classic Kuhn-Munkres matching approach on the detected objects to generate trajectories. This method uses less stricter merge bound of 0.5.

(2) *Greedy+Terminal (G+T)*. Greedy algorithm is used to associate the tracklets, which merges tracklets with the highest similarity in turn. Similarity between tracklets is calculated by features of the tracklets' terminal objects.

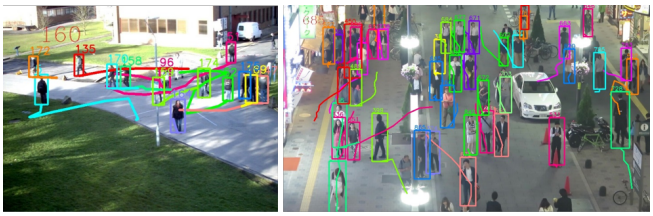
(3) *Clustering+Terminal (C+T)*. Our proposed constrained clustering algorithm is used to associate tracklets. Similarity between tracklets is calculated by terminal objects' features.

(4) *Clustering+Siamese (C+S)*. Our proposed constrained clustering algorithm is used to associate tracklets. Similarity between tracklets is calculated by using our proposed TSN.

Table I compares the tracking results on MOT16 validation dataset. We can see that *G+T* performs better than *KM*. This indicates that using a stricter merge bound with tracklet association process can reduce the noisy effects by excluding some inaccurate detections. Comparing *C+T* with *G+T* demonstrates that the constraint clustering algorithm utilizes the global tracklet information more effectively than the greedy algorithm, which leads to better results. Meanwhile,

TABLE II
TRACKING PERFORMANCE OF TSN VS. STATE-OF-THE-ART METHODS ON MOT15 AND MOT16 TEST DATASETS

Dataset	MOT15							MOT16						
Method	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓
CNNTCM [11]	29.6	71.8	11.2%	44.0%	7786	34733	712	–	–	–	–	–	–	–
CDA-DDAL [1]	32.8	70.7	9.7%	42.2%	**4983**	35690	614	43.9	74.7	10.7%	44.4%	6450	95175	676
NOMT [2]	33.7	71.9	12.2%	44.0%	7762	32547	442	46.4	76.6	18.3%	41.4%	9753	87565	**359**
MTEV [7]	33.8	71.1	12.1%	34.8%	9232	**31743**	722	–	–	–	–	–	–	–
Quad-CNN [10]	33.8	**73.4**	12.9%	**36.9%**	7898	32061	703	44.1	76.4	14.6%	44.9%	6388	94775	745
STAM [3]	34.3	70.5	11.4%	43.4%	5154	34848	**348**	46.0	74.9	14.6%	43.6%	6895	91117	473
JMC [9]	–	–	–	–	–	–	–	46.3	75.7	15.5%	39.7%	6373	90914	657
NLLMPa [4]	–	–	–	–	–	–	–	47.6	**78.5**	17.0%	40.4%	**5844**	89093	629
TSN	**35.5**	71.5	**14.4%**	43.6%	5682	33515	454	**48.2**	75.0	**19.9%**	**38.9%**	8447	**85315**	665



PETS09-S2L2 #160 (MOT15) MOT16-03 #685 (MOT16)

Fig. 5. Qualitative results of TSN on MOT15 and MOT16 test datasets.

C+S has obviously better performance than *C+T*, which shows the strength of the proposed TSN in making full use of the tracklets' appearance information and further eliminating unwanted effects caused by object occlusion.

In our proposed TSN, the parameter τ is set to 3. The tracking performance on MOT16 validation dataset for different values of τ is shown in Fig. 4. We can see that the difference in performance is not obvious if τ is around 3. The lowest MOTA (41.8) with $\tau = 1$ is larger than the next best performing method in Table I, which further proves the robustness of our method. We consider that some useful temporal information may be lost if τ is too small while there may be redundancy in information if τ is too large.

C. Compare with State-of-the-art Methods

Table II compares our approach against state-of-the-art MOT methods on the test sequences of the MOT15 and MOT16 datasets. For a fair comparison, all the methods are evaluated based on the detection results provided by the datasets. Examples of our tracking results are shown in Fig. 5. From Table 2, we make the following observations:

(1) Our approach outperforms existing MOT methods in terms of MOTA (the primary evaluation metric), which demonstrates the effectiveness of our approach.

(2) Our approach produces the highest MT score on both datasets and the lowest FN score on MOT16. This shows that our approach associates the tracklets more accurately and is able to properly cater for the missing detections.

(3) On MOT16, our approach produces the lowest ML score among all methods. This demonstrates TSN's advantage in effectively handling the easily confusable objects.

IV. CONCLUSION

This paper introduces a new approach to MOT that competently addresses the object occlusion problem. Our approach consists of two key ingredients: A tracklet siamese network that learns the feature vectors with full appearance information of tracklets to calculate tracklet similarity, and a constrained clustering method that accurately associates the short tracklets into long trajectories. Experiments on two MOT benchmark datasets demonstrate the effectiveness of our approach, particularly in the MOTA and MT metrics.

ACKNOWLEDGMENT

This work is supported in part by NSFC(61471235), in part by Shanghai 'The Belt and Road' Young Scholar Exchange Grant(17510740100), and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation.

REFERENCES

[1] S.-H. Bae and K.-J. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *TPAMI*, 2018.
[2] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *ICCV*, 2015.
[3] Q. Chu, W. Ouyang, *et al.*, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *ICCV*, 2017.
[4] E. Levinkov, J. Uhrig, *et al.*, "Joint graph decomposition & node labeling: Problem, algorithms, applications," in *CVPR*, 2017.
[5] S. Tang, B. Andres, *et al.*, "Subgraph decomposition for multi-target tracking," in *CVPR*, 2015.
[6] M. Yang, Y. Wu, and Y. Jia, "A hybrid data association framework for robust online multi-object tracking," *TIP*, 2017.
[7] S. Gao, X. Chen, *et al.*, "Beyond group: Multiple person tracking via minimal topology-energy-variation," *TIP*, 2017.
[8] S. Schulter, P. Vernaza, *et al.*, "Deep network flow for multi-object tracking," *CVPR*, 2017.
[9] S. Tang, B. Andres, *et al.*, "Multi-person tracking by multicut and deep matching," in *ECCVW*, 2016.
[10] J. Son, M. Baek, *et al.*, "Multi-object tracking with quadruplet convolutional neural networks," in *CVPR*, 2017.
[11] B. Wang, L. Wang, *et al.*, "Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association," in *CVPRW*, 2016.
[12] F. Yu, W. Li, *et al.*, "Poi: multiple object tracking with high performance detection and appearance feature," in *ECCVW*, 2016.
[13] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, 2014.
[14] L. Leal-Taixé, A. Milan, *et al.*, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv*, 2015.
[15] A. Milan, L. Leal-Taixé, *et al.*, "Mot16: A benchmark for multi-object tracking," *arXiv*, 2016.
[16] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP JIVP*, 2008.