Consistent Online Multi-object Tracking with Part-Based Deep Network

Chuanzhi Xu and Yue Zhou(✉)

Institute of Image Processing and Pattern Recognition,
Shanghai Jiao Tong University, Shanghai, China
{ChuanzhiXu,zhouyue}@sjtu.edu.cn

Abstract. Multi-object tracking is still a challenge problem in complex and crowded scenarios. Mismatches will always happen when objects have similar appearance or are occluded with each other. In this paper, we appeal for more attention to the consistency of the trajectories and propose a part-based deep network which employs ROI pooling method to extract full and part-based features for the objects. An occlusion detector is proposed to predict the occlusion degree and guide the procedure of part-based feature fusion and appearance model update. In this way, the feature extraction speed of our tracker is faster, and the objects can be associated correctly even if they are partly occluded. Besides, we train the network based on siamese architecture to learn a dissimilarity metric between pairs of identities. Extensive experiments with multiple evaluation metrics show that our tracker can associate the objects consistently and gain a significant improvement in tracking accuracy.

Keywords: Multi-object tracking · Part-based model
Occlusion detector · Consistent trajectories

1 Introduction

Multi-object tracking (MOT) is an important computer vision task and has a wide application in surveillance, robotics, and human-computer interaction. With recent development of object detectors, MOT has been formulated as tracking by detection framework. Most multi-object tracking benchmarks such as MOT16 [16] provide the tracking video sequences and detection results with public detectors. The key issue of the multi-object tracker is to associate tracklets and corresponding detection responses into long trajectories. Tracklets denote the trajectory set which is established up to current frame.

Recent tracking-by-detection methods could be categorized into batch and online methods. The batch methods process video sequences in a batch mode and take into consideration the frames from the future time steps. These methods always solve the association problem by optimization methods. For example, [17] formulates the MOT problem as minimization of a continuous energy. [5]

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 180–192, 2018.
https://doi.org/10.1007/978-3-030-03335-4_16

models the MOT problem as the min-cost network flow and finds the optimization solution with convex relaxation. Such systems may obtain a nearly global optimal solution but are not suitable for practical application. The online MOT methods only consider the observations up to current frame and associate the tracklets and detection responses frame by frame. The baseline of these online trackers is to build different models to measure the affinities between tracklets and detection responses. Then an online association algorithm is applied to get global optimum. Motion model, appearance model and interaction model are most frequently adopted to build affinity matrix. In [13], integral channel features are adopted to build a robust appearance model. [6] proposes a nonlinear motion model to get reliable motion affinity. [20] establishes an LSTM interaction model to explore the group behavior and compute the matching likelihoods.

In complex and crowded scenarios, many objects are presented with similar appearance and may be occluded with each other. Mismatches always occur in such scenarios. The result is that the tracker can not associate objects consistently. However, the consistency of the trajectories plays an important role in the follow up works such as trajectory prediction and analysis. Spatial constraints and motion model can not handle such problems. To address this problem, a robust appearance model must be established. Appearance model could improve the tracker's ability to associate objects consistently and reduce the mismatch rate. Some online trackers [12] adopt raw pixels or histogram as appearance model. These trackers may get a rapid speed but could not distinguish objects with similar appearance. Recent development on convolutional neural network has drove people to train a deep network to extract deep appearance feature. [1,26] measure appearance similarity with a person re-identification network. However, all these trackers need to crop the objects from images first, then put them into the network in a batch mode. Pre-processing procedure and frequent forward propagations make these trackers time consuming.

The MOTA [2] metric is the widely accepted metric for multi-object tracking evaluation, but it is not capable of evaluating the consistency of the trajectories, and the reasons are explained in Sect. 3.1. In this paper, we adopt ID switch rate and IDF_1 score to evaluate the consistency of the trajectories, which is initially proposed for evaluating the ID consistency for cross camera multi-object tracking.

In this paper, we propose a part-based deep network combined with a confidence-based association metric to address above problems. The main contributions are summarized as below: (i) We propose a part-based deep network which employs ROI pooling method [10] to extract part-based deep appearance feature for all objects by just one forward propagation. The network is trained based on the siamese architecture [7], and this makes our tracker gain the ability to associate correctly even if the objects are partly occluded; (ii) we propose an occlusion detector which could predict the occlusion degree and guide the procedure of part-based similarity fusion and appearance model update; (iii)we appeal for more attention to the consistency of the trajectories and conduct extensive experiments with multiple evaluation metrics introduced in [19] and

[2] on MOT benchmark. The results demonstrate our tracker can associate the objects consistently and gains a significant improvement in tracking accuracy.

2 MOT Framework

The baseline of our tracker is confidence-based association metric. Appearance, motion and shape models are established to measure the affinities between tracklets and object detections. In Sect. 2.1, the structure of fast part-based deep network is described in detail. Section 2.2 introduces the network training procedure. Section 2.3 describes the confidence-based association metric.

2.1 Fast Part-Based Deep Network

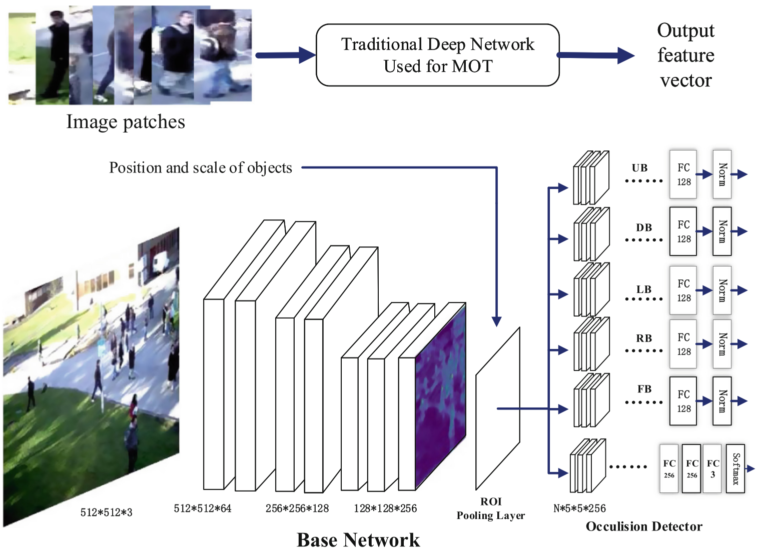


Fig. 1. The feature extraction pipeline of traditional deep network and our part-based deep network

Network Structure. Traditional deep appearance network in MOT field usually takes as input the object regions cropped from the original image in a batch mode. But it is time consuming and needs to do some pre-processing work. The more objects one frame contains, the more times for forward propagation.

The main structure of our part-based deep network is shown in Fig. 1. The network takes as input the entire image and a set of detection responses. The whole image is first processed by several convolutional layers and max pooling layers to generate a shared feature map. Then the ROI pooling method is adopted to generate five feature maps for each detection: the left body (LB), right body (RB), upper body (UB), down body (DB) and full body (FB). Five types of features are fed into the fully-connected layers separately, and the follow up normalization layers normalize the output to obtain the final feature vectors. In this way, our network could extract deep features for all objects by just one forward propagation. Beyond that, an occlusion detector based on the shared feature map is adopted to detect occlusion degree in current detection response, and then guide the procedure of part-based similarity fusion and appearance model update.

The detailed processing steps about ROI pooling are as below: At first, the ROI pooling layer maps the position and scale of the object from original image to the shared feature map, and gets the corresponding ROI window. Then divides the $h*w$ ROI window into an $H*W$ grid of sub-window of approximate size $h/H*w/W$ and maxpools the values in each sub-window into corresponding output grid cell [10]. By adopting ROI pooling layer, the speed for feature extraction gains an improvement compared with other trackers based on deep appearance model.

Part-Based Model. For MOT task, occlusion is still a challenge problem waited to be solved. This can easily cause fragmented trajectories and ID switches especially for online trackers. Mismatches have a great damage to the consistency of the trajectories. We adopt a part-based appearance network combined with a simple occlusion detector to address this problem. It is easy to implement based on the ROI pooling method with almost no speed loss. Persons detected by high position cameras would be easy to be occluded up and down, but they are more likely to be occluded left and right when detected by low position cameras. In this place, we do not design elaborate part detector for the sake of high feature extraction speed and rely more on the representative ability of deep feature. The detected persons are simply divided into UB, DB, LB and RB to overcome multi-view occlusion. During forward propagation, the ROI pooling layer extract features for FB and four divided parts, then a slice layer is added to separate features generated from different parts. So when the object is partly occluded, part-based feature is still reliable for appearance similarity computation. At the same time, the part feature is extracted from the shared convolutional feature map, and there is almost no speed loss for the added part modular.

Occlusion Detector. We propose a novel occlusion detector to detect whether there exist occlusion in current detection and guide the procedure of part-based similarity fusion and appearance model update. At first, the width and height of the detected bounding boxes are enlarged to 1.2 times of original to get more context information. Then the ROI pooling layer is employed to extract corresponding features from the shared feature maps. Follow up classifier takes the features as inputs and outputs the occlusion label, which is composed of three fully-connected layers followed by one softmax layer. The occlusion detector could classify the detections into three types: severe-occluded, part-occluded and non-occluded. For severe-occluded detections, appearance similarity is no more reliable and would not be adopted for final similarity computation. For part-occluded detections, the part-based appearance feature is still reliable would be adopted to measure appearance similarity. For non-occluded detections, FB feature vectors would be employed.

2.2 Network Training

The training procedure is divided into two stages, at first, the part-based deep network is trained based on siamese architecture, then the occlusion detector is trained based on the pretrained base network.

Siamese Architecture Training. To make the deep network gain the ability to distinguish different persons, we select part ALOV300++ sequences [22] which take person as tracking object and MOT training sequences [16] as base training dataset. Then generate positive and negative pairs by randomly sampling same and different identities from video sequences. The part-based deep appearance network is trained based on siamese architecture to learn a dissimilarity metric between pairs of identities. As shown in Fig. 2, we design a siamese network composed of two branches sharing with same structure and filter weights. Each branch has the same architecture with part-based deep network. Two branches are connected with five loss layers for network training. We employ the margin contrastive loss, and the calculation formula is as below:

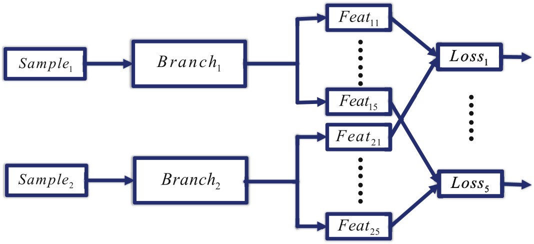


Fig. 2. The structure of siamese training.

$$L\left(x_i, x_j, y_{ij}\right) = \frac{1}{2} * y_{ij} * D + \frac{1}{2}(1 - y_{ij})max(0, \varepsilon - D) \quad (1)$$

Where $D = ||x_i, x_j||^2$ is the Euclidean distance of two normalized feature vector: x_i and x_j, y_{ij} indicates whether the object pairs are same identities, ε is the minimum distance margin that different pairs of objects should satisfy. We set ε to 1 during experiment. The final training loss is the sum of five kinds of losses. After training the siamese architecture network with margin contrastive loss, the part-based deep network could generate good feature representations that are close by enough for positive pairs, whereas they are far away at least by a minimum for negative pairs, and a simple cosine distance metric could measure the appearance similarity.

Occlusion Detector Training. The MOT16 dataset provides the visibility ratio for each annotated bounding box, and we divide these bounding boxes into three types. Bounding boxes with visibility ratio lower than 0.9 and higher than 0.4 is regarded as part-occluded detections, otherwise would be regarded as non-occluded and severe-occluded detections respectively.

After training the part-based network with siamese architecture, the weights of base network are frozen, and the occlusion detector is added after the base network and is trained with softmax loss. To improve the generalization ability of the occlusion detector, the data augmentation metric is adopted during network training. We flap and crop the object, change the brightness, contrast, sharpness and saturation of the images with a certain probability. Finally two components are integrated together to get the final model.

2.3 Association Procedure

The association between tracklets and object detections could be formulated as an assignment problem, We adopt a modified version of confidence-based association metric [1] to solve this problem.

Affinity Computation. The representation of tracklet T_i^t and detection D_j^t at frame t is defined as below:

$$T_i^t =\{P_i^{t-d:t}(x,y,w,h), A_i^q(FB,UB,DB,LB,RB), conf_i, K_i(m,p)\} \tag{2}$$

$$D_j^t =\{x,y,w,h,F_j(FB,UB,DB,LB,RB), Olabel\} \tag{3}$$

where $P_i^{t-d:t}(x,y,w,h)$ is the positions and shapes of the objects from frame $t-d$ to frame t. $K(m,p)$ is a kalman motion model and m, p denote the mean and covariance matrix respectively. At frame $t+1$, $K_i(m,p)$ predicts the object's position $P_i^{t+1}(x,y,w,h)$ and calculates the motion and shape affinity as Eqs. 4 and 5, where D_j^{t+1} is the j-th object in frame $t+1$. Once the tracklet is associated with new detections, the detected bounding box is employed to update $K(m,p)$. Besides, $K(m,p)$ is also adopted to estimate positions for missed objects.

$$sim_{mot}\left(T_i^{t+1}, D_j^{t+1}\right) = e^{-w_1((\frac{P_i^{t+1}(x)-D_j^{t+1}(x)}{D_j^{t+1}(w)})^2+(\frac{P_i^{t+1}(y)-D_j^{t+1}(y)}{D_j^{t+1}(h)})^2)} \tag{4}$$

$$sim_{shp}\left(T_i^{t+1}, D_j^{t+1}\right) = e^{-w_2(\frac{\left|P_i^{t+1}(h)-D_j^{t+1}(h)\right|}{P_i^{t+1}(h)+D_j^{t+1}(h)}+\frac{\left|P_i^{t+1}(w)-D_j^{t+1}(w)\right|}{P_i^{t+1}(w)+D_j^{t+1}(w)})} \tag{5}$$

$A_i^q(FB,UB,DB,LB,RB)$ is a queue which stores part-based deep appearance feature vectors in q frames. $F_j(FB,UB,DB,LB,RB)$ is the appearance feature vectors of detection D_j, $Olabel$ is the occlusion label. The largest cosine distance between corresponding feature vectors in F_j and A_i^q queue is regarded as appearance similarity. When D_j is non-occluded, FB feature vector is employed for similarity computation and A_i^q would be updated by five types of feature vectors. When D_j is part-occluded, the maximum similarity of four divided parts would be employed. The corresponding feature vector which is employed for similarity computation would be adopted to update A_i^q, and when D_j is severe-occluded, the appearance similarity would not be adopted and A_i^q would not be updated. During experiment, parameter q and d are set to 6 as most occlusions in MOT dataset last for less than 6 frames. Two linear SVMs are trained to fuse two or three types of affinities in severe-occlusion and other occasions, and yield the final affinity in range of [0,1].

Association Procedure. A simple Hungarian algorithm is employed to obtain the global optimum based on affinity matrix. An affinity threshold τ_1 is set to filter unreliable associations whose affinity score is lower. During association, the tracklets with long length and high association affinities in previous frames should be more reliable and associated first. So each tracklet is modeled with a confidence score $conf_i$ which is calculated as Eq. 6, where sim_k is the association score in previous steps. A confidence threshold τ_2 is set to divide the tracklets into high confidence tracklets and low confidence tracklets. The association procedure is performed on them hierarchically and is summarized in Algorithm 1.

$$conf_i = \frac{\sum_{k=2}^{length(T_i)} sim_k}{length(T_i) - 1}(1 - e^{-w_3 * length(T_i)}) \tag{6}$$

Algorithm 1. The Association Procedure

```
Input:
    The set of object detections in the current frame $D = \{1, ..., N\}$; The set of trajectories associated up to current frame $T = \{1, ..., M\}$;
1: Divide the tracklets into high confidence tracklets $T^h$ and low confidence tracklets $T^l$ according to the confidence threshold $\tau_2$;
2: Calculate the affinity matrix between $D$ and $T^h$, associate them with hungarian algorithm, remove unreliable association whose score is below $\tau_1$
3: Use the same procedure as step 2 to associate $T^l$ and unassociated detections.
4: Update the tracklet confidence using Equ.6, update the kalman filter and the appearance queue, remove the tracklets which have been unassociated for more than $T_{max}$ frames;
5: Calculate the IOU affinity matrix between unassociated detections in consecutive frames and generate new tracklets if three detections in successive frames are associated;
```

3 Experiment

3.1 Evaluation Metrics

A good tracker should find correct numbers of objects and associate them with correct tracklets when a new frame arrives. At the same time, a good tracker should also track each object consistently and overcome the mismatch phenomenon. Based on the above criteria, most trackers adopt MOTA as main metric to evaluate their trackers' performance, which is calculated as below:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \tag{7}$$

In above formula, FN indicates the number of missed objects, FP indicates the number of false positives, IDSW indicates the number of mismatches. However, in most cases, the number of FN is one order higher than FP and two order higher than IDSW. This means the reduction of IDSW is of little significance for the improvement of MOTA. In addition, a mismatch should not be treated equal with a FP. With recent development of the precision of detectors, the number of FP and FN has dropped a lot, so we appeal for more attention to the consistency of trajectories. The score of MOTA is a good indicator of the tracking accuracy, but not capable of evaluating the consistency, so we adopt ID switch rate, ID precision, ID recall and IDF_1 introduced in [19] to evaluate the consistency of the trajectories. IDF_1 is calculated by matching trajectories to the ground-truth so as to minimize the sum of discrepancies between corresponding pairs. Unlike MOTA, it penalizes ID switches over the whole trajectory fragments with wrong ID, and can evaluate how well computed identities conform to true identities [19].

Besides above evaluation metrics, following common metrics are also adopted to evaluate our tracker comprehensively:

MT: Mostly tracked targets [2]. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.

ML: Mostly lost targets [2]. The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.

MOTP: Multiple Object Tracking Precision [2]. The misalignment between the annotated and the predicted bounding boxes.

3.2 Thresholds Selection

To obtain robust affinity threshold τ_1 and confidence threshold τ_2, we test our tracker with grid search method on MOT16 train dataset. The relationship between MOTA and two thresholds is shown in Fig. 3. We set τ_1 to 0.4 and set τ_2 to 0.3 for the rest experiments. The Fig. 3 also demonstrates that adopting confidence-based association metric could improve tracking accuracy.

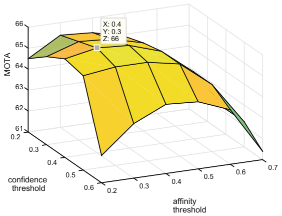


Fig. 3. Thresholds selection on MOT16 train dataset

3.3 Runtime

To investigate the feature extraction speed of our part-based deep appearance network comprehensively, we test our network and other trackers which adopt deep appearance model and take image patches as inputs on the same platform. The feature extraction speed is tested on a Quadro M4000 GPU and Intel E5V3 CPU and shown in Table 1. Dan and Pdan denote our full-part and part-based deep appearance network respectively. Compared with other trackers, our deep model gets faster speed with smaller batch size, and there is just a minor speed loss for the added part model. The speed for confidence-based association is not very fast and is about 5.16 fps, which is mostly owning to the large number of objects, but our part-based deep appearance network could be transplanted to other association metric conveniently.

Table 1. The speed and consumption for feature extraction

App model	Lmp [23]	AMIR [20]	DeepSort [24]	Dan	Pdan
Batch size	16 patches	16 patches	16 patches	1 frame	1 frame
Speed(fps)	2.47	2.42	16.19	20.75	19.10

3.4 Experiment Result

Table 2 shows the tracking results on MOT16 test dataset, Hist means histogram appearance model, and Dan-OD denotes full-part deep network without the guidance of occlusion detector for appearance model update. Trackers marked with * adopt same detections supplied in [26]. The results show that adopting part-based deep appearance network and occlusion detector could improve tracking accuracy and consistency obviously. Compared with histogram appearance model, the ID switches reduce from 1014 to 762, both ID precision and ID recall have a certain improvement. The reduction of mismatches also increases the rate of MT, this means our tracker is more capable of getting consistent and long trajectories.

Table 2. Tracking results on MOT16 test Dataset with private detector

Trackers		MOTA↑	IDSW↓	IDF$_1$ ↑	IDP↑	IDR↑	MOTP↑	MT↑	ML↓
KDNT* [26]	Batch	**68.2**	**933**	**60.0**	**66.9**	**54.4**	**79.4**	**41.0%**	**19.0%**
MCMOT-HDM [15]	Batch	62.4	1394	51.6	60.7	44.9	78.3	31.5%	24.2%
IOU [4]	Batch	57.1	2167	46.9	59.8	38.6	77.1	23.6%	32.9%
DeepSort* [24]	Online	61.4	781	62.2	72.1	54.7	79.1	32.8%	**18.2%**
Sort* [3]	Online	59.8	1423	53.8	65.2	45.7	**79.6**	25.4%	22.7%
EAMTT-16 [21]	Online	52.5	910	53.3	**72.7**	42.1	78.8	19.0%	34.9%
COMOT+Hist*	Online	58.7	1014	59.9	62.7	**57.3**	77.8	30.2%	18.3%
COMOT+Dan-OD*	Online	60.3	957	61.0	66.5	56.3	78.0	33.1%	18.4%
COMOT+Dan+OD*	Online	61.1	873	61.4	68.4	56.0	78.3	32.9%	18.7%
COMOT+Pdan*	Online	**62.8**	**762**	**62.6**	71.5	55.7	78.3	**34.9%**	18.3%

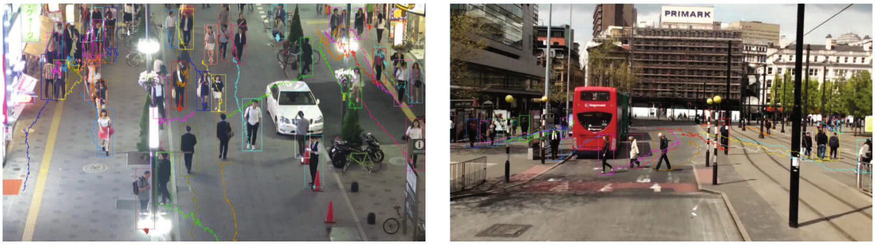
Table 3. Overall performance on MOT17 test dataset with public detections

Tracker		MOTA↑	IDSW↓	IDF$_1$↑	IDP↑	IDR↑	MT↑	ML↓
FWT-17 [11]	Batch	**51.3**	2648	**47.6**	63.2	**38.1**	**21.4%**	**35.2%**
MHT-DAM [14]	Batch	50.7	**2314**	47.2	**63.4**	37.6	20.8%	36.9%
IOU17 [4]	Batch	45.5	5988	39.4	56.4	30.3	15.7%	40.5%
EAMTT-17 [21]	Online	42.6	4488	41.8	59.3	32.2	12.7%	42.7%
GM-PHD [8]	Online	36.4	4607	33.9	54.2	24.7	4.1%	57.3%
COMOT(ours)	Online	**46.8**	**2, 121**	**49.2**	**68.7**	**38.3**	**15.3%**	**39.1%**

Table 4. Tracking results on MOT17 test dataset based on different public detections

Trackers		DPM [9]		FRCNN [18]		SDP [25]	
		MOTA↑	IDSW↓	MOTA↑	IDSW↓	MOTA↑	IDSW↓
FWT-17 [11]	Batch	**46.4**	833	**48.2**	780	59.4	1035
MHT-DAM [14]	Batch	44.6	**593**	46.9	**742**	**60.6**	**979**
IOU17 [4]	Batch	35.2	1272	44.9	1509	56.3	3207
EAMTT-17 [21]	Online	32.0	1244	42.3	1569	53.6	1675
GM-PHD [8]	Online	24.5	2155	39.3	920	45.2	1532
COMOT(ours)	Online	**36.0**	**756**	**45.3**	**618**	**59.1**	**747**

Tables 3 and 4 demonstrate the overall performance and the separated results based on different detectors on MOT17 benchmark respectively. The MOT17 benchmark provides three detection results: the DPM [9], FasterRCNN [18] and SDP detector [25]. As most trakers in MOT ranking list are anonymous submissions, we select trackers with explicit source for comparison. As demonstrated in Table 3, our tracker achieves competitive performance compared with other online trackers, both the consistency and accuracy gain a significant improvement. Compared with the FWT-17 [11] tracker, our tracker yields higher IDF$_1$ score and lower ID switch rate, this demonstrates our trajectories are more consistent. The overall accuracy of our tracker is lower than FWT-17, this is mostly due to our poor performance on DPM weak detections, and it is the inherent inferiority between online association and batch association. The batch methods take into consideration the frames in future time steps. Some sampled trajectories are shown in Fig. 4, and the numbers following '#' denote the frame numbers.



(a) Pedestrian street #240

(b) Moving bus #650

Fig. 4. Sampled trajectories in MOT17 benchmark.

4 Conclusion

In this paper, we propose a part-based deep network which employs ROI pooling method to extract part-based appearance feature to overcome the part-occlusion problem. An occlusion detector is proposed to predict the occlusion degree and guide the procedure of similarity fusion and appearance update. Extensive experiments show our tracker is more capable of getting consistent and long trajectories. Both the consistency and accuracy are competitive on MOT benchmark.

References

1. Bae, S.H., Yoon, K.J.: Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1 (2017)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. Eurasip J. Image Video Process. **2008**(1), 246309 (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: IEEE International Conference on Image Processing, pp. 3464–3468 (2016)
4. Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (2017)
5. Chari, V., Lacoste-Julien, S., Laptev, I., Sivic, J.: On pairwise cost for multi-object network flow tracking. CoRR, abs/1408.3304 (2014)
6. Chen, X., Qin, Z., An, L., Bhanu, B.: Multiperson tracking by online learned grouping model with nonlinear motion context. IEEE Trans. Circuits Syst. Video Technol. **26**(12), 2226–2239 (2016)

7. Chopra, S., Hadsell, R., Lecun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 539–546 (2005)
8. Eiselein, V., Arp, D., Ptzold, M., Sikora, T.: Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors. In: IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 325–330 (2012)
9. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
10. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
11. Henschel, R., Lealtaix, L., Cremers, D., Rosenhahn, B.: A novel multi-detector fusion framework for multi-object tracking. Eprint arXiv:1705.08314 (2017)
12. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_58
13. Kieritz, H., Becker, S., Hubner, W., Arens, M.: Online multi-person tracking using integral channel features. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 122–130 (2016)
14. Kim, C., Li, F., Ciptadi, A., Rehg, J.M.: Multiple hypothesis tracking revisited. In: IEEE International Conference on Computer Vision, pp. 4696–4704 (2015)
15. Lee, B., Erdenee, E., Jin, S., Nam, M.Y., Jung, Y.G., Rhee, P.K.: Multi-class multi-object tracking using changing point detection. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 68–83. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_6
16. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: MOT16: a benchmark for multi-object tracking. Eprint arXiv:1603.00831 (2016)
17. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE Trans. Pattern Anal. Mach. Intell. **36**(1), 58–72 (2013)
18. Ren, S., Girshick, R., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137 (2017)
19. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
20. Sadeghian, A., Alahi, A., Savarese, S.: Tracking the untrackable: learning to track multiple cues with long-term dependencies. Eprint arXiv:1701.01909 (2017)
21. Sanchez-Matilla, R., Poiesi, F., Cavallaro, A.: Online multi-target tracking with strong and weak detections. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 84–99. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_7
22. Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1442–68 (2014)
23. Tang, S., Andriluka, M., Andres, B., Schiele, B.: Multiple people tracking by lifted multicut and person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3539–3548 (2017)

24. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: IEEE International Conference on Image Processing, pp. 3645–3649 (2017)
25. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: Computer Vision and Pattern Recognition, pp. 2129–2137 (2016)
26. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: POI: multiple object tracking with high performance detection and appearance feature. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 36–42. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_3