Re-Identification and Tracklet-Plane Matching for Multi-Object Tracking and Segmentation

Yuchen Yuan[*1], Xiangbo Su[1], Wei Zhang[1], Tao Wang[2], Wei Shi[3], Zhenbo Xu[4],
Mian Peng[1], Xiao Tan[1], Weiyao Lin[2], Hongwu Zhang[1], Shilei Wen[1], Errui Ding[1]
[1]Department of Computer Vision Technology (VIS), Baidu Inc., China
[2]Shanghai Jiao Tong University, China
[3]Beihang University, China
[4]University of Science and Technology of China, China
[*]Corresponding Author.

Abstract

As a significant extension to the widely applied multi-object tracking (MOT), multi-object tracking and segmentation (MOTS) introduces segmentation masks as additional information to the bounding box based tracking task, which brings more natural descriptions of the scene and offers a potential solution to the occlusions between objects. Nevertheless, the segmentation information raises new challenges to conventional tracking methods; effective utilization of pixel-wise masks thus becomes a vital topic in MOTS. In this paper, we focus on the tracking task itself with pre-defined detection and segmentation results in the scene, and propose a novel framework that integrates segmentation-based feature extraction, short tracklet construction, and tracklet-plane matching for long trajectories. To make our method robust for varied object scales, we also propose a depth-aware instance filtering strategy, which dynamically adapts the filtering threshold. During the 5th BMTT MOTChallenge Workshop of CVPR 2020, our method achieves the 2nd place on the MOTS20 leaderboard.

1. Introduction

As an important task of the computer vision community, multi-object tracking (MOT) aims to establish the temporal relationships of detected objects per frame and associate them into complete trajectories. In recent years, facilitated by the techniques in deep learning, various computer vision tasks have obtained major advances, *e.g*. classification [6], detection [10], and segmentation [3]. MOT, on the other hand, still remains challenging due to the severe occlusions and scale variations among target objects. As an alternative, multi-object tracking and segmentation (MOTS) has been proposed as a new concept [14], which introduces pixel-wise segmentation masks as additional information. In general, it is inevitable for bounding boxes to include background or parts of the overlapping objects (especially under scenes with heavy occlusions); such representation issue, however, is fairly solved with segmentation masks. Nevertheless, how to effectively utilize the segmentation mask as new feature sources is still a question that have not been fully addressed before.

In this paper, we propose a complete framework for MOTS that implements tracking with pre-defined bounding boxes and their corresponding segmentation masks per frame (denoted as ”candidates” in our following contents). Our framework consists of two major sections, namely tracklet construction (TC) and tracklet association (TA).

1.1. Tracklet Construction

With the confidence scores provided, we first establish a depth-aware instance filtering strategy to remove noisy candidates. After that, mean-value-padding is conducted of the regions that have not been covered by the segmentation mask within the bounding box, and a 1024-dimensioned appearance feature vector is extracted via a re-identification model. We then carry out Hungarian matching based on the weighted distance between the remaining candidates and existing tracklets. If an unmatched candidate have high enough confidence, it will initialize a new tracklet. And finally, we adopt the bottom y coordinate as the mask merge confidence, so that the affiliation of the overlapping region between any candidate pair can be determined.

1.2. Tracklet Association

A secondary filtering strategy is conducted to remove low-quality tracklets generated by the TC section. After that, the remaining tracklets are associated by the TPM algorithm [9], which includes tracklet-plane construction and in-plane tracklet matching. Long trajectories are then created by interlinking short tracklets. Optional post-processing (*e.g.* interpolation and smoothing) are conducted before outputting our final results.

2. Related Works

2.1. Person Re-Identification Methods

Person re-identification (ReID) is the task of associating images of the same person taken from different cameras or different occasions of the same camera. Recently, [11, 17] have made great progresses in ReID by designing more powerful networks for global feature extraction. The deep pyramid feature learning (DPFL) architecture [4] takes multi-scale representations into account, feature fusion optimized simultaneously by concurrent per-scale re-id losses and interactive cross-scale consensus regularisation in a closed-loop design. To make the extracted features more discriminative, the human body parts are exploited [18]. [13] introduce the binary segmentation mask to design a person region guided neural network, which highlights the human body and eliminates the background bias for ReID. And the PCB network [12] uniformly splits the image into several stripes and then learn discriminative feature within each part.

2.2. Multi-Object Tracking

As mentioned earlier, with the heavy occlusions and scale variations presented, MOT has become a challenging task in computer vision, among which object association is the key problem to be solved. Most existing methods tend to optimize a template-matching problem with diversified features, such as geometric and motion features [2], 2D convolutional neural network (CNN) features [16]. More advanced association methods have also been proposed, such as TNT [15] and TPM [9]. Recently, a new trend emerges that combines detection and tracking into an end-to-end pipeline, as such Tracktor++ [1]. In our framework, since the detection and segmentation results are pre-defined, the tracking process is still solved as a feature-extraction-and-object-association problem.

3. Methodology

The overall flow chart of our proposed framework is shown in Fig. 1. Each step is illustrated in detail as below.

3.1. Person ReID Appearance Feature Extraction

Before the training of our ReID model, we first conduct a depth-aware instance filtering to remove noisy candidates. Obviously, candidates far away from the viewpoint tend to be less reliable, and should have lower threshold in the confidence-based filtering. In practice, for a candidate with y as the bottom vertical coordinate, we set two endpoints y_l, y_h as well as two extreme thresholds th_l, th_h, so that the candidate has the filtering threshold th is set as:

$$th = \begin{cases} th_l, & if \;\; y \leq y_l, \\ th_l + (th_h - th_l)\dfrac{y - y_l}{y_h - y_l}, & if \;\; y_l < y < y_h, \\ th_h, & if \;\; y \geq y_h. \end{cases} \tag{1}$$

To enable appearance feature extraction of segmentation masks instead of bounding boxes, we conduct mean-value-padding to the regions not covered by the mask within its bounding box. Following existing state-of-the-art ReID works, we adopt ResNet50_IBN_a [8] pretrained on ImageNet [5] as our backbone model. Specifically, the last fully-connected (FC) layer is replaced by a new module as FC1-BN-FC2, where BN is a batch normalization layer. In testing, we extract the 1024-dimensioned feature before the FC2 layer as the appearance feature of the candidate.

3.2. Tracklet Construction

We use the tracklet-plane matching (TPM) [9] method to obtain trajectories of objects, which consist of two steps, *i.e.* tracklet construction and in-plane tracklet matching. In the tracklet construction step, for the candidates of adjacent frames, we acquire their appearance feature from the ReID model, as well as their motion information to establish their similarity matrix. The Hungarian algorithm[7] is then performed for candidate association. Specifically:

$$S_{to}(T, D) = A(T, D) + \lambda_s M(T, D), \tag{2}$$

where $A(T, D)$ and $M(T, D)$ represent the appearance similarity and motion similarity, respectively; λ_s is a balancing term. For an unmatched candidate, a new tracklet will be initialized if its confidence is higher than a pre-defined threshold th_{init}.

After that, the confidence of each tracklet is represented by averaging the confidence of all the candidates it associated. Another threshold $th_{tracklet}$ is applied to further filter out abnormal tracklets.

Note that in testing, the overlapping region between any candidate pairs must be assigned to one candidate. We perform such operation by evaluating the mask merge confidence of the candidates, which is empirically set as their bottom y coordinates. In other words, the larger bottom y a candidate has, the higher chance it is on the top.

3.3. Tracklet-Plane Construction

To eliminate the interference of noisy or missing candidates in the tracklets, a tracklet-plane matching method is developed then, which organizes related tracklets into planes and clarifies any association confusion caused by

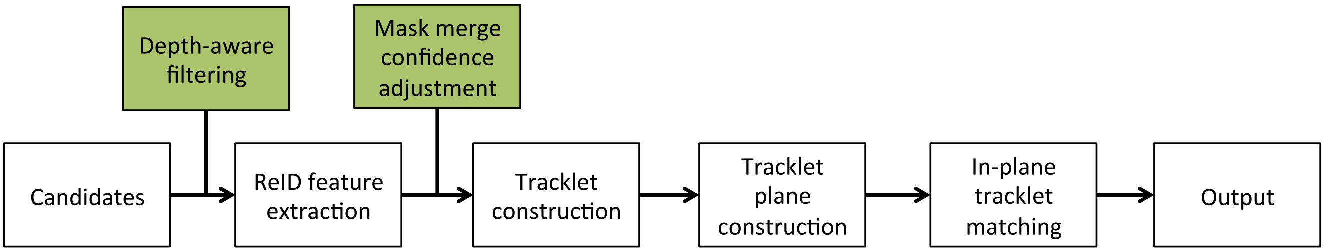
Depth-aware filtering
Mask merge confidence adjustment
Candidates
ReID feature extraction
Tracklet construction
Tracklet plane construction
In-plane tracklet matching
Output

Figure 1: Flow chart of our proposed framework, where white boxes stand for the main steps, and green boxes stand for the auxiliary steps.

noisy or missing candidates. The optimization function of tracklet-plane is:

$$(X^*, Y^*, n_p^*) = \underset{X,Y,n_p}{\arg\min}\, \Phi_1(X,Y,n_p) + \Phi_2(X,Y,n_p) + \lambda_p n_p,$$
$$s.t. \quad x_i^m, y_j^m \in \{0,1\}; \sum_{m=1}^{n_p} x_i^m \leqslant 1; \sum_{m=1}^{n_p} y_j^m \leqslant 1 \qquad (3)$$

where $x_i^m = 1$ or $y_j^m = 1$ indicates the end or the start of tracklet t_i is connected to the tracklet-plane P_m. n_t represents the number of tracklets and n_p represents the number of tracklet-planes. It is seen that the start and the end of each tracklet can only be connected to at most one tracklet-plane. $\Phi_1(X,Y,n_p)$ and $\Phi_2(X,Y,n_p)$ are the optimization terms for evaluating tracklet-plane qualities, which are defined as:

$$\Phi_1(X,Y,n_p) = -\sum_{m=1}^{n_p}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} 2x_i^m y_j^m W_i W_j S_{tt}(T_i,T_j), \qquad (4)$$

$$\Phi_2(X,Y,n_p) = \sum_{m=1}^{n_p}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} (x_i^m x_j^m + y_i^m y_j^m) W_i W_j S_{tt}(T_i,T_j), \qquad (5)$$

where W_i represents the importance of tracklet T_i, calculated as the average confidence score of the tracklet, and $S_{tt}(T_i,T_j)$ denotes the similarity between T_i and T_j, which is calculated by the similarity between the first or last three frames of the tracklets.

In this way, tracklets that are more likely to belong to the same trajectory will connect to different sides of the same tracklet-plane, while easily confused tracklets will not connect to the same side of a tracklet-plane, so that they will not interfere with each other in the following in-plane tracklet matching process.

3.4. In-Plane Tracklet Matching

We then perform tracklet-wise association on each tracklet-plane to obtain long trajectories. The in-plane matching can be modeled as:

$$Z^* = \underset{Z}{\arg\max} \sum_{m=1}^{n_p}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} x_i^m y_j^m z_{ij} W_i W_j S_{tt}(T_i,T_j),$$
$$s.t. \quad z_{ij} \in \{0,1\}; \sum_{j=1}^{n_t} z_{ij} \leqslant 1; \sum_{i=1}^{n_t} z_{ij} \leqslant 1 \qquad (6)$$

where $Z = \{z_{ij}\}, i = 1...n_t, j = 1...n_t$ represents the tracklet association status. $z_{ij} = 1$ means T_i and T_j are linked. Again, the Hungarian algorithm is applied to solve this equation. After that, interpolation, merging and deleting operations are further applied on the associated tracklets to obtain clean and coherent trajectories.

4. Experiments

4.1. Datasets

As required by the BMTT MOTChallenge, the MOTS20 and the KITTI-MOTS datasets [14] are adopted in our experiments.

4.2. Implementation Details

The parameters involved in our framework are set as below. Training set of ReID model: the training set of MOTS20 + KITTI-MOTS20; y_l: average bottom y of the entire training set; y_h: video-wise mid-vertical point; th_l: 0.3; th_h: 0.45; mean-value for padding: $[104, 117, 123]$ in BGR channel; λ_s: 0.5; th_{init}: 0.4; $th_{tracklet}$: 0.4.

4.3. Challenge 1: MOTSChallenge

The Challenge 1 of the BMTT MOTChallenge focuses on the MOTS20 dataset, and our complete framework introduced in section 3 is implemented for evaluation.

As shown in Table. 1, compared with other participating teams, our method achieves the 2nd place in the MOTS20 leaderboard with sMOTSA as 68.8.

4.4. Challenge 3: Tracking Only (MOTS+KITTI)

The same framework is applied on the KITTI-MOTS dataset as well, the results of which are then combined with

Table 1: MOTS20 leaderboard results.

Rank	**Team**	**sMOTSA**
1	ReMOTS	69.9
2	Ours	68.8
3	GMPHD_SAF	68.4
4	PT	66.8
	Baseline	40.6

Table 2: Tracking-only leaderboard results.

Rank	**Team**	**sMOTSA**
1	COSTA	69.9
2	MCFPA	69.1
3	Ours	69.075
4	ReMOTS	68.25
5	GMPHD_SAF	67.275

the results from section 4.3 as the final result for the tracking only task of Challenge 3. Note that we do not perform post-processing operations *e.g.* interpolation, merging or deleting due to the specific rule that no segmentation mask other than the pre-defined ones can be shown in the submission. As shown in Table. 2, our method achieves the 3rd place in the leaderboard of this challenge with sMOTSA as 69.075.

5. Conclusion

In this paper, we propose a complete framework for MOTS based on pre-defined bounding boxes and segmentation masks, which conducts straightforward feature extraction and effective tracklet-plane-matching-based object association. In the future, we will continue exploring more effective methods in MOTS to facilitate its applications in real-world scenarios.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 941–951, 2019. 2

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[4] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV*, pages 2590–2600, 2017. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[7] Harold W Kuhn. The hungarian method for the assignment problem. *NRL*, 1955. 2

[8] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2

[9] Jinlong Peng, Tao Wang, Weiyao Lin, Jian Wang, John See, Shilei Wen, and Erui Ding. Tpm: Multiple object tracking with tracklet-plane matching. *Pattern Recognition*, page 107480, 2020. 1, 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[11] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. pages 3820–3828, 2017. 2

[12] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 501–518, 2018. 2

[13] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[14] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 1, 3

[15] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 482–490, 2019. 2

[16] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[17] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385, July 2017. 2

[18] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2