Min-Cost Network Flow and Trajectory Fix for Multiple Objects Tracking

Jiasheng Tang[1]*, Xiong Xiong[1]*, Chenwei Xie[1], Yanhao Zhang[1], Pichao Wang[1],
Fan Wang[1], Fei Du[1], Liang Han[2], Yun Zheng[1], Pan Pan[1], Hao Li[1]
[1]Alibaba Group, [2]Stony Brook University
{jiasheng.tjs,moxiong.xx,eniac.xcw,yanhao.zyh,pichao.wang,
fan.w,dufei.df,zhengyun.zy,panpan.pp,lihao.lh}@alibaba-inc.com liang.han@stonybrook.edu

Abstract

*We proposed our approach MCF-PA, which is based on Min-Cost network Flow (MCF)[14] optimization in an offline Multiple Objects Tracking (MOT) manner, together with a trajectory-level post association. We mainly follow the tracking-by-detection paradigm, by using a two-step association strategy on the provided detection and segmentation results. First, a classical network flow formulation of MOT is used with our designed cost functions. To deal with the detection failures due to occlusions and clutter, we train a regressor based on provided detection confidence scores to get a better estimation of the target existence. Then a pairwise classifier based on GBDT is trained to obtain the transition costs for MCF association, where multiple factors are taken into account, including time gap, appearance feature, bounding box IOU, box size and position. Noted that the appearance feature is fine-tuned on the Kitti[4] and MOTS[12] training data. In the second step, we propose to reason over the entire set of trajectories globally and generate final tracks by a post association step. All proposed tracklets from MCF step are clustered hierarchically based on appearance features, to be further connected to form longer trajectories. A SOT tracker with discriminative score is employed here to tackle challenges like occlusions, out-of-view, etc. We show a significant improvement over the benchmark in both sMOTSA and IDF1 on Kitti and MOTS benchmarks, and finally rank **2nd** on Track 3 competition.*

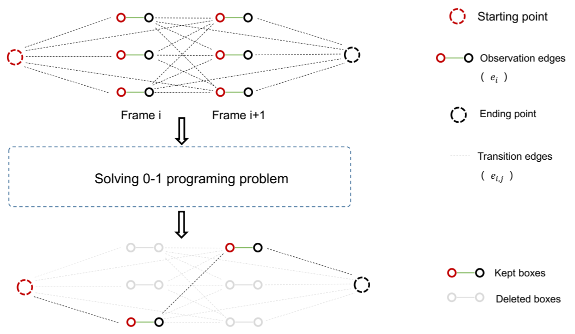

Figure 1. Pipeline of graph construction

1. Introduction

With the tremendous progress in object detection community, nowadays multiple-objects-tracking (MOT) techniques are mainly focused on the tracking-by-detection strategy, i.e. the major task is to associate or connect detected bounding boxes from different frames to form trajectories. Based on the setup, there are two different paradigms[3]: 1) online tracking where the current frame detections are immediately associated to historic trajectories or detections without looking at the future frames. 2) offline tracking where the frames are processed in a batch or even with the whole video sequence, and the decision of association is delayed until more frames are available. Intuitively, offline tracking methods would achieve better performance as it utilizes information both from the past and in the future, and we will focus on offline tracking in this paper.

In offline tracking, with all detection results from multiple frames, it is natural to build a graph structure on top of them, with each node representing a detection box and the edges representing their potential connection/association. Figure 1 demonstrates a typical graph structure for modeling data association step with a batch of detection results. As a well-formulated optimization approach, Min-Cost network Flow[14] has been widely used in offline tracking problem to solve the best matching between detections based on such a graph structure.

However, we find that MCF still suffers from a few draw-

*The authors contribute equally to this work.

backs while we analyze the produced trajectories. We then introduce a Post-Association module, inspired by SQE[6], by using a self-evaluation view for trajectory analysis. We estimate the distribution of appearance features of each trajectory, to get a better idea about its consistency, and further identify the ones that could be improved. Finally the trajectories are fixed by adopting a Single-Object-Tracking model as its response score.

2. The MCF-PA pipeline

We first briefly revisit MCF algorithm used in MOT, and discuss several drawbacks that would limit its performance. Then we come up with a simple method called **trajectory fix** for post association (**PA**).

2.1. Min-Cost network Flow

We adopt a tracking-by-detection framework where detections are given and the data association step is performed by MCF algorithm in an offline manner. Trajectory management is inherent in the optimization phase: both initialization and termination are represented as **Enter/Exit** edges and counts cost for the loss function, and linking boxes is determined by the **observation** edges and **transition** edges where our regressor and classifier show off their power.

We follow the traditional formulation of multi-object tracking as the min-cost flow optimization problem. Given a video with multiple targets, our goal is to track K moving targets in a “detect and track” manner. The inputs contain a set of candidate object detections provided by detectors. A measure of correspondence between detections are required. The tracking problem is thus formulated as a joint optimization problem of simultaneously selecting detections and connections between targets, which can be modeled by a **M**aximum **a P**osteriori estimation (**MAP**) objective. The MAP optimization can be cast with specific constraints encoding the network structure of the tracks.

$$\arg\min_{\mathbf{x}} \sum_i C_i x_i + \sum_{ij \in E} C_{ij} x_{ij} \tag{1}$$

with the constraints $\sum_j x_{ji} = e_i = \sum_j x_{ij}$ and $\sum_i x_{it} = K = \sum_i x_{si}$. The formulation encodes the joint selection of K tracks using the binary indicator $x_i \in \{0, 1\}$, which is 1 when the detection i is selected in some track. $x_{ij} \in \{0, 1\}$ is a binary variable, which is 1 when detection i and j are connected as the same track. The index i ranges over all possible detections. c_i denotes the observation cost of selecting detection i, which represents the negative detection confidence. c_{ij} represents the negative of the correspondence strength between detections i and j. The set of possible connections between detections is represented by E and could be a subset of all pairs of detections.

Structure of a flow conservation is encoded by the constraints that x_{ij} can take the value 1 only if both x_i and x_j is 1 with each detection belongs to at most one tracks and exactly K tracks are selected. This optimization problem with relaxed integer constraint can be solved efficiently using existing min-cost network flow algorithm.

2.2. Trajectory fix as post association

Directly optimizing MCF is challenging, thus some trade-offs have to be made. For example, it is usually time-consuming to optimize for a whole video with a large amount of objects, so some work like muSSP[13] tried to alleviate this problem by re-formulating the optimization problem. But in practice, most still optimize over a window-size (e.g.20) batch of frames instead of the whole sequence, leading to a sub-optimal solution. Additionally, as we showed before, the affinity score is calculated with box-level similarity which can easily make trajectory broken during occlusion, and LMP[11] adds lifted multi-cut edges on graph to capture long-range appearance changes but also greatly increases computation time.

Instead, we use a simple but effective approach for compensation of MCF. After tracklets are generated, we analyze trajectory-level information such as appearance feature to re-link short trajectories and also split mis-connected trajectory. It should be noticed that, if we use this module specifically after MCF, we can tune it by using a more aggressive threshold for linking boxes inside graph to produce more fragmented trajectories. Thus we only need the re-linking function of our PA module. This would reduce false positives.

The re-linking function is implemented in a simple way. We use a SOTA single object tracking (**SOT**) method called **SPM** tracker for dealing with discrimination. This tracker gives us a better response score when occlusions happen compared with the ReID model and any other SOT approaches. Concretely, a reference trajectory (**Ref Traj**), generated by loose ReID similarity with no time overlaps, is for querying from a candidate trajectories set. Nearly 5 boxes would be randomly selected from a **Ref Traj** and sent to **SPM** as templates. The tracker then generates 5 response scores for every box from candidate. We simply average among those response scores from 5 templates on a whole trajectory as the final affinity score. Threshold larger than 0.7 can be linked. After that, we get a final set for linking those reference and candidate trajectories. Since there could be one trajectory simultaneously selected by two **Ref Traj**. An optimal matching problem like Hungarian algorithm can be also applied to link final results.

3. Experiments

Our main focus is the association step, which is the mainly developed issue in the MOT community. As for detections, we simply tested the given detections and masks are sufficient enough for usage. Thus for all three tracks,

Method	sMOTSA	IDF1	MT	TP	FP	FN	Recall	Precision	IDS	Frag
ReMOTS	**69.9**	75.0	**248**	**28270**	819	**3999**	**87.6**	97.2	388	621
PTPM	68.8	68.5	244	28108	1084	4161	87.1	96.3	368	560
PT	66.8	67.3	234	27215	1059	5054	84.3	96.3	370	629
MCFPA	66.2	**76.4**	235	26516	849	5753	82.2	96.9	216	**449**
Lif_TS	65.3	75.2	216	26143	879	6126	81.0	96.7	**149**	457
IA-MOT	64.1	65.7	218	27069	1003	5200	83.9	96.4	1054	1341
USN	63.7	62.8	226	26430	1038	5839	81.9	96.2	764	1015
GMPHD_SAF	61.8	64.3	214	**477**	819	3999	76.5	**98.1**	524	770
TrackR-CNN	40.6	42.4	127	19628	1261	12641	60.8	94.0	567	868

Table 1. Track 1: MOTS challenge.

we use the pre-computed detections and masks as our input of the association module.

During the association step, we train three models in MCF: 1) A logistic regression model is for judging the existence of a box with classification score of detector. The given score of a box is just the classification score of which class it belongs to, which is not enough for judging existence or confidence of a box. It would be better if an IoU score is provided. 2) A **GBDT** model outputs the affinity score for two boxes. As we stated, features used for GBDT are appearance, box Iou, mask IoU, box size, time difference. 3) Here, a ReID model for modeling the appearance serving as one feature for GBDT. With 4-fold cross validation training the first two models on 4 MOTS sequences, we can reach **66.1%** sMOTSA on MOTS training sequences with public detections given by Track 3. The ReID model is built with **ResNet-50**[5] as backbone by using strong optimizing tools [2, 9, 7, 8, 1]. For the pedestrian ReID model, The model is first trained on public available ReID datasets and then fine-tuned on MOT17[10] and Kitti. For MOT, we half split the 7 training set for training and validation to avoid over-fitting. As for Kitti, the training and validation set are already provided. As for the car ReID model, we didn't fine-tune on Kitti, which made the result a little worse on Kitti car leaderboard.

Several simple but effective experiments show that our MCF could already reach top-level tracking results. And with PA module, about 40% to 60% improvement on ID switch could be got on MOTS. We also demonstrate our PA module on MOT17 dataset. For easily explain our results of PA module, the experiment is down on MOT17 training set. Noted that there's no training on this dataset for PA module, thus the results are compelling. We use MCF as the first step, and got 71.2% IDF1. Then we apply PA for trajectory fix, about 11.3% absolute improvement is got.

Some results also show our approach is the top-level one especially for ID-level preservation: we ranked the best IDF1 score and Frag rate, 2nd best for ID switches on the MOTS challenge, as shown in Table 1. For Kitti, see 2 and 3, with no other input messages such as LiDAR and GPS, we also get a good result. Table 4 shows the result of tracking-only challenge, which proves that with good detections, MCF-PA can archive a good performance. Due to the limited space, we just summarize parts of trackers. For whole leaderboard of this competition, please ref to the official MOTS website.

4. Conclusion

We demonstrate that MCF is still the SOTA tracking method with careful training. It should be noticed that the PA module is not only suitable for MCF but for all tracking methods. With this module, we can set MCF with a more aggressive parameter to start a new trajectory, instead of linking boxes thus can make less mistakes on false positives. PA module can relink those broken trajectories with trajectory-level analysis which compensate with the box-level matching in MCF and further directly improves the performance. It is a good plugin for MOT.

References

[1] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. 2017.

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. 2017.

[3] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] 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.

[6] Yanru Huang, Feiyu Zhu, Zheni Zeng, Xi Qiu, Yuan Shen, and Jianan Wu. Sqe: a self quality evaluation metric for

Method	**sMOTSA**	**MOTSA**	**MOTSP**	**MOTSAL**	**MODSA**	**MODSP**	**MT**	**ML**	**IDS**	**Frag**
PointTrack++	82.80	92.60	89.70	93.30	93.30	92.10	89.50	1.20	270	584
LIFTS	79.60	89.60	89.00	89.90	89.90	91.40	79.10	2.90	114	532
PointTrack	78.50	90.90	87.10	91.80	91.80	89.70	90.80	0.60	346	645
Lif_TS	77.50	88.10	88.30	88.60	88.60	90.90	79.60	2.70	183	569
MCFPA	77.00	87.70	88.30	89.10	89.10	90.80	82.90	0.60	503	724
HRNt	75.80	86.50	88.30	88.10	88.10	90.70	78.10	2.90	599	703
GMPHD_SAF	75.40	86.70	87.50	88.20	88.20	90.10	82.00	0.60	549	874
MOTSFusion	75.00	84.10	89.30	84.70	84.70	91.70	66.10	6.20	201	572
TrackR-CNN	67.00	79.60	85.10	81.50	81.50	88.30	74.90	2.30	692	1058

Table 2. The detailed results for car in KITTI MOTS leaderboard.

Method	**sMOTSA**	**MOTSA**	**MOTSP**	**MOTSAL**	**MODSA**	**MODSP**	**MT**	**ML**	**IDS**	**Frag**
PointTrack++	68.10	83.60	82.20	84.80	84.80	94.00	66.70	4.80	250	521
MCFPA	67.20	83.00	81.90	84.30	84.30	93.80	67.00	3.00	265	484
LIFTS	64.90	80.90	81.00	81.90	81.90	93.60	61.50	8.90	206	577
F	64.60	80.00	82.00	83.30	83.30	93.90	62.60	5.20	681	731
GMPHD_SAF	62.80	78.20	81.60	80.40	80.50	93.70	59.30	4.80	474	696
USN	62.80	77.00	82.40	79.60	79.60	94.20	52.60	7.40	547	734
TES	62.20	76.60	82.40	80.10	80.10	94.10	53.70	5.90	741	974
PointTrack	61.50	76.50	81.00	77.40	77.40	93.80	48.90	9.30	176	632
PointTrack(MF)	59.40	73.50	81.50	74.20	74.20	94.10	47.40	15.60	150	481
MOTSFusion	58.70	72.90	81.50	74.20	74.20	94.10	47.40	15.60	279	534
TrackR-CNN	47.30	66.10	74.60	68.40	68.40	91.80	45.60	13.30	481	861

Table 3. The detailed results for Pedestrian on KITTI MOTS leaderboard.

name	**car_k**	**ped_k**	**ped_m**	**score_t**
IA-MOT	76.4	64	69.4	69.8
MCFPA	77	67.2	66.1	69.1
TPM-MOTS	75.8	67.3	66.6	69.075
ReMOTS	72.6	64.6	67.9	68.25
GMPHD_SAF	76.2	64.3	64.3	67.275
Lift_TS	77.5	55.8	65.3	65.975
SRF	71.4	60.9	60	63.075
KQD	74.4	61.8	57.3	62.7
USN	72.1	59.3	59.5	62.6
YLC	62.3	57.2	59.1	59.425
SI	68.5	55.5	56.2	59.1
FK	64.1	54.5	54.3	56.8
PredTrack	-32.6	-49.3	50.8	4.925

Table 4. The detailed results for Track 3. car_k denotes KITTI MOTS car; ped_k denotes KITTI MOTS ped; ped_m denotes MOTS20 ped; score_t denotes total score.

parameters optimization in multi-object tracking. *arXiv preprint arXiv:2004.07472*, 2020.

[7] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[8] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 2019.

[9] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61, 2019.

[10] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[11] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3548, 2017.

[12] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019.

[13] Congchao Wang, Yizhi Wang, Yinxue Wang, Chiung-Ting Wu, and Guoqiang Yu. mussp: Efficient min-cost flow algorithm for multi-object tracking. In *Advances in Neural Information Processing Systems*, pages 423–432, 2019.

[14] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.