Multiple Object Tracking Challenge Technical Report for Team MT_IoT

Feng Yan, Zhiheng Li, Weixin Luo, Zequn jie, Fan Liang, Xiaolin Wei, Lin Ma
Meituan Inc.
{yanfeng05, lizhiheng03, luoweixin, jiezequn, liangfan02, weixiaolin02, malin11}@meituan.com

Abstract

This is a brief technical report of our proposed method for Multiple-Object Tracking (MOT) Challenge in Complex Environments. In this paper, we treat the MOT task as a two-stage task including human detection and trajectory matching. Specifically, we designed an improved human detector and associated most of detection to guarantee the integrity of the motion trajectory. We also propose a location-wise matching matrix to obtain more accurate trace matching. Without any model merging, our method achieves 66.67^2 *HOTA and* 93.97^1 *MOTA on the DanceTrack challenge dataset.*

1. Method

1.1. Overview of the proposed method

The schematic diagram of out method is illustrated in Fig. 1. We presented a two-stage tracking-by-detection structure including human detector and trace matching module. For the detection phrase, we adopted the YOLOX [6] for the object-intensive detection task. Then the boxes with low scores are fed into the Trajectory-based Patching module to determine which detection can be retained. All trustworthy detection will be matched with existing traces subsequently.

1.2. Human Detector

We adopted the YOLOX [6] as the human detector. For the tracking task, most false detection are caused by occlusion between targets in multi-object scenarios. To solve this problem, we introduce two public datasets Crowdhuman [8] and MOT20 [5] to enhance the robustness of the detector for multi-object scenes. After pretrained on both datasets, the human detector greatly improved the detection of obscured targets, further enhancing the continuity of the motion trajectory.

1.3. Trace Matching Module

Trajectory-based Patching Method Inspired by Byte Track [9] and OC-Sort [4], we believe detection boxes with low scores also contribute to securing the continuity of trajectory. We proposed a trajectory-based patching method to capture useful detection boxes from ones with low confidence scores. The method first utilize boxes with high scores to match with existing trajectories. Then for the unmatched traces, we try to correlate the last detection of each trajectory with the remaining boxes with low scores. Specifically, the method calculates IoU(Intersection over Union) scores between the last detection of each trace and each box with low score, then the box with highest IoU score is considered to be the best match. If there still exist unmatched traces, the prediction from the Kalman filtering will be retained as a short-term observation. Furthermore, we also impose the limitation to existing traces. When the confidence score is lower than the threshold, the trajectory will be considered to be untrustworthy and won't be correlated with any box in current observation. In most scenarios, occlusion is often accompanied by a slow decrease in detection score. The proposed patching method helps to maintain the trace coherence in challenging situations such as target occlusion.

Location-wise Matching Matrix In the dancing scene, the height of dancers change noticeably with the distance away from the camera. The height change in successive frames can provide clues to the dancer's location, and helps to remove unreasonable matches caused by overlapping targets. For the Hungarian matching algorithm, each element of matching matrix represents the distance calculated based on the IoU score between each trajectory and detection box. In the dance tracking task, we utilize the height of detection box to calculate the cost distance. Since the height of the human varies slightly between adjacent observations, the location-wise matching matrix drives the trajectory to find the best match among boxes of similar height. In addition, the horizontal movement of the dancer is usually violent than the vertical movement. Calculating distances based on height can smooth out the noise caused by rapid movement.

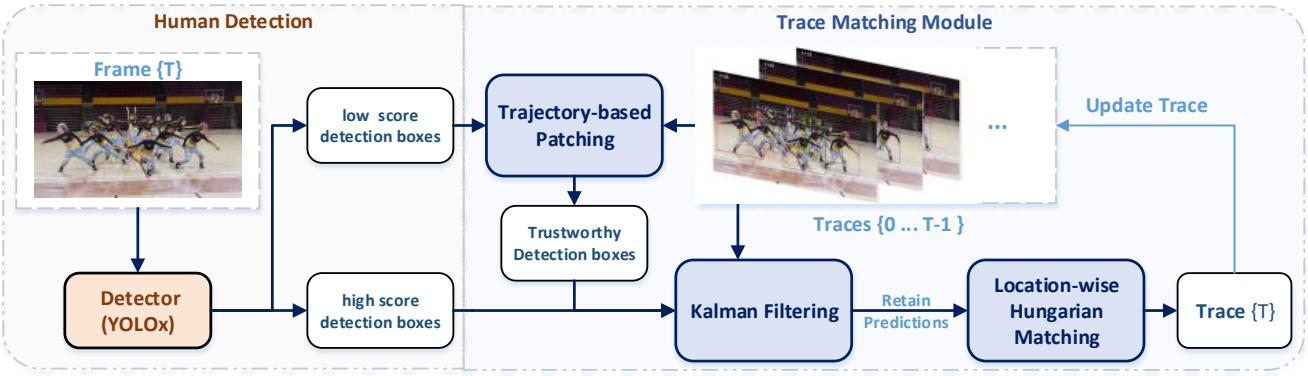

Figure 1. Overall architecture of our proposed method.

Table 1. Ablation study for pretraining on different datasets, where Crowd and MOT presents Crowdhuman and MOT20 dataset respectively.

Datasets	HOTA	DetA	AssA	MOTA	IDF1
-	55.41	80.99	38.04	91.80	55.47
Crowd	57.76	83.27	40.19	93.36	57.15
Crowd + MOT	58.82	83.33	41.63	93.44	58.93

Table 2. Ablation study for the proposed Trajectory-based Patching(TP) method with different IoU algorithms.

Datasets	HOTA	DetA	AssA	MOTA	IDF1
w/o TP	58.82	83.33	41.63	93.44	58.93
w/ TP(DIoU)	60.13	-	-	-	-
w/ TP(GIoU)	60.12	-	-	-	-
w/ TP(CIoU)	60.36	82.96	44.00	93.95	64.00

Table 3. Ablation study for different matching matrix.

Cost Matrix	HOTA	DetA	AssA	MOTA	IDF1
Area-based IoU	60.36	82.96	44.00	93.95	64.00
Height-based IoU	66.66	84.14	52.95	93.97	70.60

2. Experiment Results

In this section, we present the results of ablation experiments for the pretrained human detector, trajectory-based patching method and location-wise matching matrix. We also display two unproductive cases with corresponding explanations, which may contribute to tracking tasks in other scenarios.

2.1. Human Detector Pre-Training

We introduced two public datasets Crowdhuman [8] and MOT20 [5] to boost the performance of the human detector in occlusion scenes. After pre-training, the baseline is trained and tested on DanceTrack dataset as shown in TABLE 1. It can be seen from the table that pre-training provides a significant improvement in the tracking performance of the model.

2.2. Trajectory-based Patching Method

In this section, we present the experiment results of the proposed Trajectory-based Patching method. We also conducted the ablation study for different IoU algorithms including D-IoU [10], G-IoU [7] and C-IoU [10] as shown in TABLE 2.

2.3. Location-wise Matching Matrix

We compare the effects of different matching matrices on tracking performance as shown in TABLE 3. We conducted the ablation experiments with area-based and height-based IoU scores, respectively. As shown in the table, the proposed location-wise matching method provides a significant improvement in both tracking and detection performance on the DanceTrack dataset.

2.4. Unproductive cases

In this section, we present two attempts to boost the tracking performance including BoT-Sort [1] and model ensemble.

BoT-Sort BoT-Sort proposes the camera-motion compensation(CMC) to alleviate the effects of tracker movement. The CMC obtains the camera transformation matrix by aligning adjacent frames to achieve more accurate fore-

Table 4. Ablation study for camera-motion compensation(CMC).

Models	HOTA	DetA	AssA	MOTA	IDF1
w/o CMC	58.82	83.33	41.63	93.44	58.93
w/ CMC	56.91	80.92	40.13	91.65	57.72

Table 5. Ablation study for model ensemble.

Models	HOTA	DetA	AssA	MOTA	IDF1
Model A	55.41	80.99	38.04	91.80	55.47
Model B	58.82	83.33	41.63	93.44	58.93
Merge AB	49.55	77.18	31.96	87.20	47.03

ground and background matching. However, we found that considering camera motion slightly degrade the performance in dance tracking task. For the DanceTrack dataset, most videos were captured with no obvious changes in perspective. After considering camera motion compensation, small detection offsets of overlapping targets may result in matching to the wrong trajectory, leading to a slight degradation of tracking performance.

Model Ensemble We explored the effect of model ensemble on tracking performance. For each ensemble, We mix the tracking results of different models and filter out the repeated predictions by the Non-Maximum Suppression (NMS) algorithm. As shown in TABLE 5, tracking performance is not improved after model ensemble. For the dance tracking task, mix results from different models tends to add more negative samples which can't be filtered out by NMS, leading to a degradation of tracking performance.

3. Conclusion

In this paper, we presented a two-stage tracking-by-detection architecture with improved human detector for dance tracking task. We also proposed the Trajectory-based Patching method and Location-wise Matching Matrix to achieve better performance. If you want to run the detection trace model on ultra-low hardware, you can also refer to the other two articles [3] [2], which are currently used in real projects.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 2

[2] Qiong Bai, Jinming Xin, Ming Yan, Yu Wang, Erpeng Li, and Sanjun Zhao. An optimized mask-guided mobile pedestrian detection network with millisecond scale. In *2020 Chinese Automation Congress (CAC)*, pages 4975–4980. IEEE, 2020. 3

[3] Qiong Bai, Jingmin Xin, Hu Ye, Qinjie Wang, Peiwen Shi, and Sijie Liu. An efficient pedestrian detection network on mobile gpu with millisecond scale. In *2019 Chinese Automation Congress (CAC)*, pages 3195–3199. IEEE, 2019. 3

[4] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022. 1

[5] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 2

[6] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1

[7] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 2

[8] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1, 2

[9] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 1

[10] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12993–13000, 2020. 2