A Technical Report for SportsMOT Track on Multi-actor Tracking

Jiahao Wang*, Chang Meng, Donghao Li, Hao Wang, Yuting Yang, Licheng Jiao, Fang Liu

School of Artificial Intelligence, Xidian University

*Corresponding to: jh_wang1024@163.com

Abstract

This technical report introduces our solution to SportsMOT Track on Multi-actor Tracking, DeeperAction Challenge at ECCV 2022. Our solution is based on Detection-Based Tracking, which detection and tracking are implemented separately. We use DeepSORT and OC-SORT as trackers to track players based on the results of Cascade-RCNN-S101 and ATSS-SwinL. After linearly interpolating the result, We finally achieve 74.899 HOTA on the test set of SportsMOT.

Keywords *Mutli-object tracking, Tracking-by-detection, Re-identification, DeepSORT.*

1. Introduction

Multi-object tracking (MOT)[1, 2] is a fundamental task in computer vision, aiming to estimate objects ($e.g.$, pedestrians and vehicles) bounding boxes and identities in video sequences. As a mid-level task in computer vision, multiple object tracking grounds high-level tasks such as pose estimation, action recognition, and behavior analysis. It has numerous practical applications, such as visual surveillance, human computer interaction and virtual reality. These practical requirements have sparked enormous interest in this topic.Most existing MOT works can be grouped into two sets, depending on how objects are initialized: Detection-Based Tracking (DBT) and Detection-Free Tracking (DFT). DBT is more popular because new objects are discovered and disappearing objects are terminated automatically. DFT cannot deal with the case that objects appear. However, it is free of pre-trained object detectors.

Prevailing human-tracking MOT datasets mainly focus on pedestrians in crowded street scenes ($e.g.$, MOT17/20) or dancers in static scenes (DanceTrack). In spite of the increasing demands for sports analysis, there is a lack of multi-object tracking datasets for a variety of sports scenes where the background is complicated, players possess rapid motion and the camera lens moves fast. SportsMOT provides a large-scale multi-object tracking dataset, consisting of 240 video clips from 3 categories ($i.e.$, basketball, football and volleyball). It contains rich and complicated sports scenes.The objective is to only track players on the playground (i.e., except for a number of spectators, referees and coaches) in various sports scenes. It is much more challenging in: 1) Irregular motion trajectories (different from MOT17 or 20 datasets); 2) The appearance characteristics (same uniforms) of players on the same team during the competition are extremely similar. Therefore, tracking players in sports scenarios requires powerful detection models and targeted tracking strategies.

We adopt Cascade-RCNN-ResNest101[3, 4] and ATSS-Swin-L[5, 6] as our detection baseline. Firstly, we use the bbox information of the dataset to train the detectors and generate detection boxes for players. Then After processing these results with DeepSORT[7] and OC-SORT[8], we get the tracking IDs of players. Finally, we linearly interpolate the results to obtain the final tracked trajectory.

The main contributions of this paper are summarized as

• Use different detectors and trackers for different types of sports scenes.

• Implement two interpolation methods: simple interpolation and fragment interpolation.

We empirically validate the effectiveness and generality of the proposed method on multiple challenging benchmarks and achieve a good performance.

2. Related Works

DeepSORT[7] is a computer vision tracking algorithm for tracking objects while assigning an ID to each object. DeepSORT is an extension of the SORT (Simple Online Realtime Tracking) algorithm. DeepSORT introduces deep learning into the SORT algorithm by adding an appearance descriptor to reduce identity switches, Hence making tracking more efficient. To understand DeepSORT, First Let's see how the SORT algorithm works.

DeepSORT uses a better association metric that combines both motion and appearance descriptors. DeepSORT can be defined as the tracking algorithm which tracks objects not only based on the velocity and motion of the object

but also the appearance of the object.

For the above purposes, a well-discriminating feature embedding is trained offline just before implementing tracking. The network is trained on a large-scale person re-identification dataset making it suitable for tracking context. To train the deep association metric model in the DeepSORT cosine metric learning approach is used. According to DeepSORT's paper, "The cosine distance considers appearance information that is particularly useful to recover identities after long-term occlusions when motion is less discriminative." That means cosine distance is a metric that helps the model recover identities in case of long-term occlusion and motion estimation also fails. Using these simple things can make the tracker even more powerful and accurate.

Observation-Centric SORT (OC-SORT)[8] points out its limitations from the use of Kalman filter. These limitations play even bigger roles when the tracker fails to gain observations for supervision - likely caused by unreliable detection, occlusion, or fast and non-linear target object motion. Current motion models in MOT typically assume that the object motion is linear in a small time window and needs continuous observations, so these methods are sensitive to occlusions and non-linear motion and require high framerate videos. OC-SORT is robust to occlusion and non-linear object motion while still being simple, online, and realtime. This method is motivated by both analytical and empirical findings and focuses on leveraging observations more confidently in the interaction with Kalman filter. In many experiments on multiple popular tracking datasets, OCSort significantly outperforms the state of the art. It's especially significant for multi-object tracking under severe occlusion and on objects with dramatic non-linear motion.

3. Experiments

In this section, we detail the details of the methodology used in the competition, including algorithms, training strategies, data and tricks/post-processing.

3.1. Algorithms

3.1.1 Detector

Cascade-RCNN-ResNest101 and ATSS-Swin-L are chosen as detectors, and both detectors use pretrained models from the coco dataset. To facilitate the training and the reproduction of the results, we train the above two models on the MMDetection framework.

3.1.2 ReID

resnet50_b32x8[9] was chosen as ReID[10], and imagenet2012 was used as pretrained weights. To fully excavation the appearance features of the data, we set the feature dimension to 256. The relevant ablation experiments are shown in Table 1.

Table 1. A comparative study of ReID ablation experiments.

Method	Detector	ReID(out channel)	HOTA
DeepSort	Cascade-RCNN-S101	128	71.639
DeepSort	Cascade-RCNN-S101	256	72.51

3.1.3 Tracker

We use DeepSORT and OC-SORT as trackers for the inference test set. The inference parts are all done using the MMTracking framework.

3.2. Training Strategies

3.2.1 Detector

For Cascade-RCNN-ResNest101, SGD[11] was used as the optimizer, The initial learning rate was set to 0.02 and the epoch was set to 24. For ATSS-SwinL, AdamW[12] was used as the optimize. The initial learning rate was set to 0.00005 and the epoch was set to 36. In the training phase, we all use a multi-scale training strategy, and the image scales are set to (1280, 720) and (1280, 640). In the inference phase, we set the image size to (1280, 720), and use soft_nms with setting iou_thr to 0.7.

3.2.2 Tracker

For DeepSORT, match_iou_thr is set to 0.1, num_tentatives is set to 5, and num_frames_retain is set to 100. The ReID parameter match_score_thr is set to 5.

For OC-SORT, init_track_thr is set to 0.7, match_iou_thr is set to 0.1, num_tentatives is set to 3, vel_consist_weight is set to 0.3, vel_delta_t is set to 3 and num_frames_retain is set to 100.

3.3. Data

In the training phase, the training set and the validation set are fused to participate in the training. In the inference phase, the test set is divided into three categories: basketball, football, and volleyball. Among them, DeepSORT is used to track basketball and volleyball, and OC-SORT is used to track football. Combine the two results as the preliminary result.

3.4. Tricks/Post-Processing

Since only detection boxes with high confidence are tracked, there are many false negative detection boxes, resulting in the low performance of HOTA. To reduce the effect of confidence thresholds, we tried two simple interpolation methods. 1) Linearly interpolate trajectories whose total missing frames do not exceed 40; 2) Interpolate each

Table 2. Overall Optimization Program Results. Where AS represents ATSS-SwinL, and CRS represents Cascade-RCNN-Resnest101.

Method	Detector	ReID(out channel)	Dataset	Interpolation	HOTA
DeepSORT	Cascade-RCNN-S101	128	TrainSet+ValSet	No	71.639
DeepSORT	Cascade-RCNN-S101	256	TrainSet+ValSet	No	72.51
OC-SORT	Cascade-RCNN-S101	-	TrainSet+ValSet	No	70.431
DeepSORT+OC-SORT(football)	Cascade-RCNN-S101	-	TrainSet+ValSet	No	73.707
DeepSORT+OC-SORT(football)	Cascade-RCNN-S101	-	TrainSet+ValSet	Yes	74.301
DeepSORT	ATSS-SwinL	256	TrainSet+ValSet	No	73.507
DeepSORT+OC-SORT(football)	ATSS-SwinL	-	TrainSet+ValSet	No	74.173
DeepSORT-AS+OC-SORT-CRS(football)	-	-	TrainSet+ValSet	No	74.399
DeepSORT-AS+OC-SORT-CRS(football)	-	-	TrainSet+ValSet	Yes	74.899

trajectory only between frames with 10 or fewer missing frames. Although the interpolation method increases the number of false positive samples, it dramatically reduces the number of false positive False negative samples, allowing us to achieve an improvement of around 0.5 on the test set.

Table 2 is our overall optimization table for the competition scheme. By optimizing the detector and feature extractor, data association method, post-processing and other methods, we won the second place in DeeperAction Challenge at ECCV 2022(track3).

4. Conclusions

We analyze the dataset under different motion scenarios and then use different trackers on three types of sports scenes. Especially OC-SORT works better on football data. It's effective for multi-object tracking under severe occlusion and on objects with dramatic non-linear motion in sports scenes. We verify the effectiveness of our solution through extensive experiments and achieve good performance.

5. Acknowledgement

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would first like to thank my laboratory(IPIU) and tutors, for their valuable guidance throughout my studies. You provided me with the tools that I needed to choose the right direction and successfully complete my competition. I would particularly like to acknowledge my teammate, for their wonderful collaboration and patient support. Finally, I would not have been able to get in touch with this competition without the support of the organizer, MCG Group, who provided a good competition environment and reasonable competition opinions.

Thanks to the support of the National Natural Science Foundation of China (No.62076192), Key Research and Development Program in Shaanxi Province of China (No.2019ZDLGY03-06), the State Key Program of National Natural Science of China (No. 61836009), the Program for Cheung Kong Scholars and Innovative Research Team in University (No. IRT_15R53), The Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No. B07048), the Key Scientific Technological Innovation Research Project by Ministry of Education, the National Key Research and Development Program of China, and the CAAI Huawei MindSpore Open Fund.

References

[1] Yingkun Xu, Xiaolong Zhou, Shengyong Chen, and Fenfen Li. Deep learning for multiple object tracking: a survey. *IET Computer Vision*, 13(4):355–368, 2019. 1

[2] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, 2021. 1

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 1

[4] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 1

[5] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020. 1

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[7] 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. 1

[8] 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, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2

[10] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2

[11] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 2

[12] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 2