Summary of Multi-object Tracking Methods Combined with Deep Learning

Di Su *

School of Computer Science and Technology, Donghua University, Shanghai, China

* Corresponding Author Email: 231380117@mail.dhu.edu.cn

Abstract. The essence of the multi-target tracking (MOT) task is to identify and position multiple targets in the video sequence and ensure the continuity of the target identity in the time dimension, so as to produce the comprehensive path of each target in the whole video. Due to advancements in deep learning, multi-target tracking technology has undergone substantial enhancement, and is applicable to a variety of realistic application scenarios, such as autonomous driving, video surveillance, and action analysis, to meet diverse needs. On the basis of extensive literature research, according to the multi-target tracking algorithm framework, the current mainstream multi-target tracking algorithm is divided into multi-target tracking, joint detection, multi-target tracking and end-to-end multi-target tracking based on attention mechanism and introduce all kinds of representative algorithms in the framework of feature embedding. This paper examines the technical characteristics, advantages and disadvantages, and applicable environment of each type of algorithm in detail and shows the experimental results of each algorithm on the standard dataset MOT 17, so as to aid researchers in selecting suitable models for research.

Keywords: MOT; Computer vision; Deep learning.

1. Introduction

Multi-target tracking (MOT), one of the fundamental technologies in contemporary computer vision, finds extensive application in intelligent security, autonomous driving, video surveillance, and other contexts. MOT combines the domains of deep learning, pattern recognition, and conventional machine learning. The essential problem is to complete the precise positioning of the target by learning to determine the reliable characteristics of the target of interest and maintain its correct identity in the subsequent video sequence and produce the comprehensive path of the target on the whole time series [1].

In the early stage, the multi-objective tracking task was mostly conducted through traditional image processing, and complex correlation optimization methods such as particle filtering, multiple hypothesis tracking, Markov decision and joint probability data correlation were widely used in traditional MOT algorithms. However, these methods have large errors in predicting target locations and are less robust in environments with occlusion and similar targets [2]. The advancement of deep learning in recent years has led to significant advancements in multi-goal tracking technologies. To provide a straightforward and effective target tracking effect, the SORT technique first predicts the target motion state using the Kalman filter and then correlates the data using the Hungarian algorithm. DeepSORT The method is improved on the basis of SORT method, which is more accurate in tracking the movement state of the target, and the tracking results are more coherent, but the problem of pedestrian identity switching still occurs frequently. JDE combines detection and tracking tasks, which greatly improves the inference speed of the algorithm. In recent years, with the rise of the attention mechanism in the field of computer vision, researchers began to apply it to the target tracking algorithm to obtain global feature information. The addition of attention mechanism significantly improves the accuracy and robustness of multi-target tracking algorithm in complex environment [3].

However, in different situations, there are different requirements for algorithms. For example, for the single surveillance camera scenario, where the crowd is not dense and the target is less blocked, the simple application process, high calculation efficiency and high real-time performance; for the

shopping mall, subway station, airport and other places, the method for more efficient and stable scenes, such as traffic flow monitoring, sports events and other scenes.

Based on the multi-target tracking algorithm framework, this paper breaks down the aforementioned issues into three distinct paradigms: detection-based tracking (Track-by-Detection), joint detection and feature embedding (JDE) tracking, and Transformer-based end-to-end multi-target tracking. It also introduces and analyzes the features of the popular algorithm. Lastly, a comparison and analysis of different algorithms' performance on the MOT 17 dataset was conducted.

2. Algorithm introduction

2.1. Multi-target tracking based on the detection (TBD)

TBD paradigm is one of the most widely used algorithmic frameworks in the field of multi-target tracking. SORT, DeepSORT, and ByteTrack are the mainstream algorithms under the TBD paradigm. Supplementary Note, SORT algorithm itself does not use deep learning, but its dependent external detector usually relies on deep learning, and SORT algorithm is the basic algorithm of TBD, so SORT algorithm is also included in the scope of discussion.

2.1.1. SORT algorithm

In 2016, SORT was proposed by Alex Bewley et al, with the goal of achieving a balanced real-time and precision [4]. SORT algorithm mainly uses the motion information of the target for target association, does not rely on the feature information of the target, so the computational amount is low, can better adapt to the real-time requirements, and SORT structure is simple, clear algorithm, easy to understand and implement. The SORT algorithm consists of four parts: target detection, Kalman filter, Hungarian algorithm and tracking management. All targets in the image are first detected using an external target detector (Faster-RCNN). Next, the Kalman filter is used to predict the trajectory of the detected target in the following frame. The intersection between the detected target position and the predicted target trajectory (Intersection over Union, or IoU) is calculated, and the best match is found using the Hungarian algorithm. Finally, the Kalman filter's status is updated based on the matching results, and the undetected target and the newly detected target are managed. If a target is not detected, the counter of its tracker increases, deleted when the counter exceeds a certain threshold (i. e., frames are not detected continuously); if a new target is detected or the existing tracker no longer matches any detection result, a new tracker is generated. Repeat the above steps until the video ends to achieve multi-goal tracking. Although SORT algorithm has good tracking performance, there are also some limitations, such as relying only on movement information and ignoring the appearance information, SORT is prone to ID switching problems in scenes with dense target or occlusion, which will lose tracking; and insufficient robustness, in the case of serious occlusion or complex target movement mode, the effect of SORT may not be ideal.

To compensate for the limitations of SORT, some improved algorithms have been proposed, such as the DeepSORT algorithm.

2.1.2. DeepSORT algorithm

In 2017, the DeepSORT method was proposed by Wojke et al. In order to address the issues of target identification mismatch and loss tracking of the SORT method in complicated circumstances, DeepSORT is an enhancement of the classic SORT algorithm that incorporates deep learning feature extraction to create target association [5]. The DeepSORT algorithm successfully raises the accuracy and resilience of multi-target tracking by fusing motion data with appearance features.

The composition of DeepSORT is roughly the same as that of SORT, but DeepSORT introduces deep feature extraction on the basis of SORT using Kalman filter to obtain motion information for target correlation, which adds appearance features. It uses a pre-trained deep convolutional neural network (usually ResNet or Inception) to extract features from the detection box, extracting a fixed-length feature vector from each target as its unique appearance representation. In this way, even if the

appearance difference between the targets is small, the algorithm can more effectively distinguish between them, significantly reducing the ID Switch (identity mismatch) problem. In addition, DeepSORT combines the Hungarian algorithm with Mahalanobis distance and cosine distance to achieve the target association. The Hungarian algorithm compares the detection results with the prediction box, the cosine distance is used to quantify the target's appearance feature similarity, and the Mahalanobis distance is used to estimate the spatial similarity of the detection box and the prediction box. Compared to IoU alone, this method of integrating motor and appearance information is more reliable in complex or dense settings.

Under these optimizations, the tracking accuracy of DeepSORT is significantly improved compared with SORT, but due to the introduction of deep feature extraction operation, the running speed of the algorithm is greatly reduced, and the computational overhead is increased.

2.1.3. ByteTrack algorithm

In 2021, the ByteTrack algorithm was proposed. Based on the SORT structure, it adopts a new strategy to distinguish between high confidence and low confidence detection boxes to optimize the for the association method of the target. The power of the YOLOX detection algorithm enables ByteTrack to effectively use low-confidence detection boxes [6]. ByteTrack using the position measure matching strategy proposed by the SORT algorithm, a hierarchical matching was performed using the Hungarian algorithm according to the confidence of the YOLOX detection results. First, the high confidence detection box is matched with the existing tracking track to ensure the high confidence matching of the clearly visible target, and then the remaining low confidence detection box and the unmatched tracking track are twice matched to make up for the missed or occlusion association and increase the tracking continuity and recall rate of the target. This improvement ensures high precision while greatly improving tracking recall, allowing ByteTrack to perform well with high computational efficiency as well as in complex environments. Therefore, ByteTrack is an excellent choice in scenarios where high real-time needs and depth features cannot be used.

2.2. Joint detection and feature embedding for multiple-object tracking

JDE is a multi-target tracking method that integrates target detection and feature embedding in a single network. Its representative algorithms include JDE algorithm, FairMOT algorithm, Centertrack algorithm and CStrack algorithm.

2.2.1. JDE algorithm

The JDE algorithm is an efficient multi-objective tracking (MOT) method. The JDE algorithm first jointly learns the detector and embedding model in a single network, realizing the close fusion of detection and feature extraction. This means that the detector will also extract the feature vector of the target for the cross-frame correlation, which helps to achieve the identity consistency of the target [7]. The JDE algorithm will first input the video frame to the same neural network for simultaneous object detection and feature extraction. The network recognizes all the target objects in each frame and generates a detection box for the target. While generating the detection box, the network extracts a fixed length feature vector (i. e. embedding vector) representing the appearance characteristics of the target for subsequent target associations. Then the JDE algorithm will use the detected target location information and feature embedding to associate across frames. When the target positions are close, but the appearance features are different, the algorithm will maintain the target identity consistency by comparing the feature embedding vectors. After the association, JDE will update and manage the target track information, and finally output the track result.

In contrast to the conventional method of distinct processing, detection, and feature extraction, JDE eliminates the need for an extra feature extraction network and enhances tracking performance in real time. The JDE also uses a multi-task loss function to train the target detection and the ReID task together to better perform the data association in a complex environment. JDE uses a relatively light network structure and reduces the computational amount by sharing feature extraction modules.

Therefore, JDE has high computing efficiency in practical application, which is suitable for scenarios requiring real-time and limited resources.

2.2.2 FairMOT algorithm

FairMOT is an algorithm for multi-objective tracking (MOT), which was proposed in 2020 [8]. The traditional MOT algorithm usually uses two different networks for target detection and feature extraction, respectively, which easily leads to the detection and feature extraction on different feature layers, and then affects the accuracy of the association. However, FairMOT performs target detection and feature extraction in the same network and uses the same feature layer, so as to achieve fair performance between the detection and feature extraction tasks. On the other hand, FairMOT is based on CenterNet architecture and adopts the center point detection method, that is, the location and size of the target by detecting the central point and key features. ReID (Re-Identification) features are extracted at the center point for cross-frame correlation. Feature extraction is achieved through lightweight network modules, so it can better distinguish between different goals and maintain real-time performance.

FairMOT has something in common with JDE, both perform simultaneous detection and ReID feature extraction in the same network. However, FairMOT has made improvements based on JDE, especially in terms of fairness on the shared feature layer, making the detection and extraction more refined, which is slightly better than JDE in dense occlusion scenes.

2.2.3. Centertrack algorithm

CenterTrack is a multi-object tracking (MOT) algorithm based on center point tracking [9]. This algorithm combines object detection and tracking tasks in a single end-to-end neural network to predict the position and trajectory of the current frame target by using the center point position of the previous frame target. CenterTrack can not only perform efficient target detection, but also realize cross-frame correlation through the displacement information of the central point. The main innovation of CenterTrack is to simultaneously predict the displacement of the target position and trajectory, so as to achieve the unified processing of detection and tracking. This method does not require an additional association algorithm, and directly uses the position change of the center point to complete the association, and thus has a small computational amount and high accuracy. In addition, complex scenes, especially with dense targets or occlusion, can provide more stable tracking effect.

In contrast to the dependence of JDE and FairMOT on appearance feature extraction, CenterTrack does not require ReID features, but directly associate across frames based on the offset of the target position. As a result, CenterTrack is simpler, but may be less accurate than JDE) in scenes with similar target looks (such as JDE and FairMOT. Therefore, CenterTrack is more suitable for environments with high real-time requirements, unobvious target appearance features or limited resources.

2.2.4. CStrack algorithm

CSTrack is a kind of multi-object tracking (MOT) algorithm, which introduces cross-similarity computerized in object detection and identity association [10]. The main goal of CSTrack is to improve the accuracy and stability of multi-target tracking through a more robust feature matching strategy.

The key innovation of CSTrack is to use cross-frame similarity to improve tracking reliability. In traditional MOT methods, data association usually depends on in-frame similarity and position overlap, but in complex scenes, for example with dense targets or severe occlusion, single-frame similarity calculation is difficult to guarantee identity retention. CSTrack Cross-frame similarity is introduced. By calculating the similarity of targets between different frames, the identity can be maintained correctly when partial occlusion or posture changes. Therefore, CSTrack is very suitable for applications in complex targets, frequent occlusion, and highly dynamic scenarios, such as traffic monitoring, shopping mall security, crowd analysis, etc. It can maintain high tracking accuracy and identity consistency in dense scenes, occlusions and more interactions.

2.3. End-to-end multi-target tracking based on the attention mechanism

End-to-end multi-target tracking algorithm based on attention mechanism (usually based on Transformer architecture) is a multi-target tracking method that combines target detection and tracking tasks in an overall framework. Such algorithms use the self-attention mechanism and sequence modeling ability in Transformer to globally model the complex relationships between targets in video sequences, so as to complete the tracking and association while detecting targets. Representative algorithms are TrackFormer algorithm, Transtrack algorithm and TransMOT algorithm.

2.3.1. TrackFormer algorithm

TrackFormer is a Transformer-based multiple objective tracking (MOT algorithm) [11]. TrackFormer The self-attention mechanism and sequence modeling ability of Transformer architecture are used to directly generate target tracks from video frames, abandoning the need for detection and tracking separation processing in the traditional MOT algorithm, thus realizing an end-to-end multi-target tracking framework. The core of TrackFormer is the encoder-decoder architecture based on Transformer. The encoder beds the extracted image features into the encoder to generate a series of image feature representations. The self-attention mechanism of the encoder can capture the global context information between different targets in the image, thus enhancing the feature expression. The decoder uses two types of queries to handle the detection and tracking tasks of the input frame. One is the detection query (Detection Queries), which is used to detect new targets in the video frame. The decoder generates a detection box for the new target according to the output of the encoder. The second is the track query (Track Queries), which is used to track the track of a tracked target in the previous frame. In each frame, the decoder generates the target position of the current frame based on the trajectory query, thus maintaining the identity consistency of the target. In each new frame, TrackFormer's Transformer decoder automatically updates the trajectory of each target based on the trajectory query. TrackFormer Instead of requiring IoU, matching or ReID features in traditional MOT methods, but directly establishing associations between them across frames through the self-attention mechanism. The trajectory query updates based on the target state of the previous frame, thus effectively maintaining target identity in occluded and complex scenes.

TrackFormer With the help of the powerful features of Transformer, researchers unify the detection and tracking in a single model, achieving continuous target tracking and identity retention through the self-attention mechanism. This makes it very suitable for scenarios with complex target interactions, such as public safety monitoring, traffic flow monitoring, and domains where multiple moving targets need to be tracked in real time. It can maintain high tracking accuracy in scenes such as occlusion and target density.

2.3.2. Transtrack algorithm

TransTrack is a multi-objective tracking (MOT) algorithm based on Transformer [12]. Like TrackFormer, it adopts the Transformer encoder-decoder structure to model the relationship between targets through the self-attention mechanism, so as to realize the joint modeling of detection and tracking. However, TransTrack introduced two types of query, "tracking query" and "new target query" in the decoder, which were used for the detection of tracked targets and new targets respectively, reducing the confusion between the two and further distinguishing the management of old and new targets. The separation of tracking queries and new target queries makes TransTrack more identity-preserving, especially reducing ID Switch in occluded or dense target scenes. In addition, TransTrack also combines the DETR detection framework (DEtection TRansformer) to take advantage of its anchorless frame (anchor-free) detection to improve the detection accuracy. Therefore, in scenarios with dense occlusion and frequent target, the dual query mechanism of TransTrack is more stable and better identity retention.

2.3.3. TransMOT algorithm

TransMOT The structure is similar to TrackFormer algorithm, but TransMOT uses convolution feature extraction combined with Transformer endeode structure, which improves the feature extraction ability [13]. And also introduced more detailed trajectory management and recovery mechanisms. The trajectory management in TrackFormer is more basic, mainly relying on track query to update the target position in the current frame. But in the case of transient loss of targets, TrackFormer has no dedicated mechanism to retain these trajectories, and the model tends to treat targets as new targets when they are reappeared. This leads to an increase in ID Switch, especially more pronounced in scenes with dense targets or severe occlusions. The TransMOT provides a dedicated trajectory management mechanism to keep the occlusion, entry, and out of view. TransMOT Continue to track those lost targets over time and try to relate when they reappear, which significantly reduces the risk of target loss. Therefore, TransMOT has higher robustness and automatic association ability.

3. Experimental result

The performance of the various methods on the MOT 17 test set is shown in Table 1.

Table 1. Performance of multiple algorithms on the MOT 17 test set.

Methods	MOTA	IDF1↑	MOTP↑	IDP↑	IDR↑	MT↑	ML↓
SORT	43.1	39.8	69.5	90.7	49.0	12.5	42.3
DeepSORT	75.4	62.6	63.5	-	-	-	-
ByteTrack	80.3	77.3	69.0	95.0	85.2	53.2	14.5
JDE	64.6	55.8	60.3	-	-	-	-
FairMOT	73.7	72.3	60.4	92.8	74.5	36.8	24.8
CenterTrack	67.8	64.7	49.5	95.6	71.6	34.6	24,6
CSTrack	74.9	72.6	-	95.0	79.7	41.5	17.5
TrackFormer	65.0	63.9	-	92.9	80.7	47.3	10.4
TransTrack	74.5	63.9	-	90.5	84.7	55.3	10.2
TransMOT	76.7	75.1	-	-	-	51.0	16.4

MOT17 contains a range of different scenarios, including city streets, shopping centers, sidewalks, and more. Each scene contains a different density of pedestrians, with a large number of occlusion and pedestrian overlap. It can be seen from the table that the ByteTrack overall performance is optimal, especially in MOTA (multi-target tracking accuracy), IDF 1 (which measures the consistency of target identity) and identity matching, with high MT and low ML value, proving that it can accurately and continuously track targets in complex scenarios. TransTrack and TrackFormer also performed better in most of the indicators, especially in reducing target loss and maintaining identity consistency. In general, the end-to-end multi-target tracking algorithm based on the attention mechanism performs better than the other two paradigms for this difficult and diverse tracking task.

4. Conclusion

This review summarizes multi-object tracking methods combined with deep learning, dividing the mainstream multi-target tracking algorithm into multi-object tracking, joint detection and feature embedding, and end-to-end multi-target tracking based on attention mechanisms. For different algorithm frameworks, this paper examines its characteristics, advantages and disadvantages, and compares the experimental performance of various methods on MOT 17 standard dataset. Taken together, deep learning-based multi-object tracking techniques have made significant progress in target recognition and trajectory retention, especially when handling complex scenes (e. g., occlusion and target density). However, existing algorithms still face some challenges, such as improving real-time performance in a resource-constrained environment and reducing target identity switching in

complex occlusion situations. Future work could consider the introduction of an adaptive learning mechanism to adjust parameters to scene changes in real time and improve tracking stability. By further optimizing the deep learning model and the data association strategy, the multi-target tracking technology will be able to be applied to more practical scenarios.

References

[1] H. Liu, Multi-object Tracking Algorithm Based on Deep Learning, Jiangnan University, DOI:10.27169/d.cnki.gwqgu, 002649 (2023).

[2] X. Bai, Research on multi-object tracking method, Taiyuan University of Science and Technology, DOI:10.27721/d.cnki.gyzjc, 000549 (2023).

[3] F. Du, Research on the visual target tracking method based on the attention mechanism, And Harbin Institute of Technology, DOI:10.27061/d.cnki.ghgdu, 000239 (2021) .

[4] A. Bewley, Z. Ge, L. Ott, et al, Simple online and realtime tracking//2016 IEEE international conference on image processing (ICIP), IEEE, 3464-3468 (2016).

[5] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, 2017 IEEE international conference on image processing (ICIP), 3645-3649 (2017).

[6] Y. Zhang, P. Sun, Y. Jiang, et al, Bytetrack: Multi-object tracking by associating every detection box//European conference on computer vision, Cham: Springer Nature Switzerland, 1-21 (2022).

[7] Z. Wang, L. Zheng, Y. Liu, et al, towards real-time multi-object tracking//European conference on computer vision, Cham: Springer International Publishing, 107-122 (2020).

[8] Y. Zhang, C. Wang, X. Wang, et al, Fairmot: On the fairness of detection and re-identification in multiple object tracking. International journal of computer vision. 129, 3069-3087(2021).

[9] X. Zhou, V. Koltun, P. Krähenbühl, Tracking objects as points//European conference on computer vision, Cham: Springer International Publishing, 474-490 (2020).

[10] C. Liang, Z. Zhang, X. Zhou, et al, Rethinking the competition between detection and reid in multiobject tracking. IEEE Transactions on Image Processing. 31, 3182-3196(2022).

[11] T. Meinhardt, A. Kirillov, L. Leal-Taixe, et al, Trackformer: Multi-object tracking with transformers//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 8844-8854 (2022).

[12] P. Sun, J. Cao, Y. Jiang, et al, Transtrack: Multiple object tracking with transformer, arXiv preprint arXiv: 2012, 15460 (2020).

[13] P. Chu, J. Wang, Q. You, et al, Transmot: Spatial-temporal graph transformer for multiple object tracking//Proceedings of the IEEE/CVF Winter Conference on applications of computer vision, 4870-4880 (2023).