Research Advanced in 2D Multi-Object Tracking based on Deep Learning

Lehan Hong*

Dalian Maritime University, Dalian, Liaoning province, 116026, China

*Corresponding auther: dmu2220214456@dlmu.edu.cn

Abstract. Multi-object tracking is one of the most important research orientation of computer vision, which has a broad applications in the fields of sports game data analysis, video monitoring, pedestrian behavior analysis, and automatic driving. With the rapid-developed convolutional neural network, deep learning-based multi-object tracking methods have made significant progress. In this research, the current multi-object tracking methods are classified, analyzed and summarized through comprehensive and thorough research. Firstly, the definition of multi-object tracking, research background and application advantages of deep learning are introduced. Next, 2D multi-object tracking and 3D multi-object tracking as well as their differences are briefly introduced. Then, different methods in 2D multi-object tracking are explored in detail. The commonly used evaluation metrics and datasets are also introduced, as well as the comparison of the experimental results of different methods on the datasets. Finally, the advantages and disadvantages of multi-object tracking methods are analyzed and the future development direction is prospected. The development of multi-object tracking is of great significance for enhancing social security as well as promoting technological progress.

Keywords: multi-object tracking; object detection; convolutional neural network; deep learning.

1. Introduction

Multiple-object Tracking (MOT) is one of the most important research orientation of computer vision, aiming at accurately recognizing and tracking multiple moving targets in a video sequence [1-2]. Multiple Object Tracking has important applications in video surveillance, autonomous driving, and pedestrian behavior analysis.

Over the past decades, researchers have proposed many different approaches to solve the problem of MOT. These include methods based on traditional machine learning methods, methods based on image processing and feature extraction, and so on. However, these traditional methods often face difficulties such as object occlusion, object appearance changes, and complex backgrounds, which limit their accuracy and robustness. With the rapid-developed deep learning technology, deep learning-based multi-object tracking methods have made significant progress recently. Deep learning models can automatically learn the feature representations of targets, thus improving the robustness and accuracy of MOT.

The rapid-developed deep learning techniques has opened up new possibilities for multi-object tracking. Deep learning can learn feature representations and patterns suitable for multi-object tracking from a mass of labeled data by using neural network-based models. Compared with traditional methods, MOT methods have the following advantages: (1) Strong representation learning ability. Deep learning models can automatically learn feature representations suitable for multi-object tracking without manually designing features, which is more adaptable to different scenes and targets. This allows the model to capture the key information of the object more accurately and improve the tracking accuracy. [3] (2) Strong ability to utilize contextual information. Deep learning models can learn rich contextual information through training on a large amount of data. This contextual information can help solve problems such as object occlusion and complex backgrounds, better distinguish targets and maintain stability during tracking. (3) Robustness. Deep learning models that have been trained on extensive datasets can improve their robustness. They are tolerant to changes in object appearance, changes in lighting, etc., and can be better adapted to various complex scenes.

However, MOT methods also have some challenges. For example, the training of the model requires a large amount of labeled data, and the running time of the model is long, which requires high computational resources. In addition, the models are poorly interpreted and it is difficult to explain the decision and judgment process of the models. Therefore, when researching and applying deep learning-based multi-object tracking methods, we need to consider these advantages and challenges comprehensively and seek solutions to improve the performance and efficiency of the models.

This paper will provide a comprehensive and thorough investigation and summary of the current deep learning-based MOT methods. First, the definition and research background of multi-object tracking are introduced, including the concept of MOT, application scenarios and research significance. Then, the application prospects and advantages of deep learning in MOT are explored, and the improvements of deep learning technology on traditional methods are analyzed. Next, 2D MOT and 3D MOT and their differences are briefly introduced, and then MOT methods based on deep learning are introduced in detail according to the algorithm categorization of 2D MOT, including TBD, JDE, JDT and other related methods. In the Evaluation and Comparison section, commonly used evaluation metrics and datasets will be introduced, and the experimental results of different methods on these datasets will be compared and analyzed. Subsequently, the advantages and disadvantages of MOT methods are analyzed in detail. Finally, the challenges and future directions of MOT are discussed.

2. Multi object tracking based on deep learning

In multi-object tracking, 2D MOT and 3D MOT are two common methods, which differ in their information dimensions. 2D MOT mainly utilizes the 2D coordinate information of the image plane for object tracking, while 3D MOT utilizes 3D data such as point clouds or depth images to achieve more accurate and comprehensive object tracking. The choice of which method to use usually depends on the application scenario and the available sensing devices. For image-based applications, such as video surveillance, 2D MOT methods are more suitable. For applications that require more precise information about the location and shape of the target, such as autonomous driving and robot navigation, 3D MOT methods are more appropriate.

The 2D MOT methods specifically include TBD (Tracking by Detection) method, JDE (Joint Detection and Embedding) method, JDT (Joint Detection and Tracking) method, and others. All of these methods are dedicated to detecting and tracking multiple targets on a 2D image and have some differences in algorithms and performance.

2.1. TBD based 2D multi-object tracking

The detection-based multi-object tracking method TBD (Tracking by Detection) utilizes object detection results for tracking. The method divides object detection and trajectory association into two stages, firstly, object detection is performed to obtain object location and appearance information, and re-ID features are extracted for object identification. Then, the objects in different frames are matched by the trajectory association algorithm to form a complete trajectory. Some commonly used TBD methods are SORT, DeepSORT, MOTDT, ByteTrack, StrongSORT++, and BoT-SORT.

SORT [4] introduces Kalman filtering as a predictor of the object state and combines it with the Hungarian algorithm for data association to realize the process of object tracking. The focus of SORT is to efficiently implement frame-to-frame object association between frames to support online and real-time applications. However, SORT has limitations in dealing with object ID and trajectory fragmentation, and is unable to handle object detection problems caused by occlusion or appearance changes.

DeepSORT [5] improves on the SORT algorithm by proposing a simple online and real-time tracking method using a deep association metric. DeepSORT introduces re-identification (re-ID) matching and cascade matching strategies to improve the performance of SORT. It filters some frames using the Mahalanobis distance and calculates the cosine distance for the Hungarian

assignment based on re-ID. Also, DeepSORT introduces the idea of cascade matching to prioritize the matching of trajectories with less number of losses. With these improvements, DeepSORT is able to match targets more accurately and reduce the number of ID interchanges, resulting in better tracking of targets with longer occlusion times. However, DeepSORT is still a detection-based multi-object tracking method and its performance is still affected by the accuracy of object detection. In addition, due to the use of deep correlation metrics, DeepSORT may degrade performance in scenarios with limited computational resources or high real-time requirements.

MOTDT [6] proposes a method that utilizes a deep learning method for real-time multiple people tracking. This method performs optimal selection on candidate frames generated by the tracker and detector by designing a uniform scoring function and using appearance and spatial information for trajectory association. MOTDT's scoring function provides higher accuracy and efficiency than previous manual feature-based methods. The method treats the tracker and the detector as mutually independent components and treats all their results as candidate frames. A unified scoring function is constructed using a trained object classifier and a well-designed track segment confidence calculation module for evaluating all candidate frames. Redundant candidate frames are eliminated by non-maximal suppression, and trajectory association is accomplished by hierarchical correlation using appearance representations and spatial information, with re-recognition-based matching followed by intersection-and-parallel-ratio-based matching. The use of deep learning for candidate frame selection and individual re-recognition enables MOTDT to realize real-time multi-person tracking. The method utilizes a combination of appearance and spatial information for trajectory correlation, and is able to effectively deal with problems such as occlusion and drift. However, the performance of the method is still affected by factors such as the accuracy of object detection and computational resources. In practical applications, it is necessary to select suitable tracking methods according to specific needs and scenes, and to improve and optimize MOTDT.

2.2. JDE based 2D multi-object tracking

The multi-object tracking method JDE (Jointly learns the Detector and Embedding model) based on joint detection and feature extraction to achieve object tracking. The method achieves end-to-end multi-object tracking by using a multi-task learning network structure that learns both object detection and feature embedding tasks simultaneously. Commonly used JDE methods for multi-object tracking include JDE and FairMOT. These methods have the advantage of being faster and having real-time performance while maintaining a certain level of accuracy, but their accuracy is slightly degraded compared to TBD methods.

FairMOT [7] presents a simple and practical approach for multi-object tracking, whose main feature is that it combines the results of detection and tracking and optimizes the multi-object tracking performance by introducing a fairness metric.[8] Specifically, FairMOT uses a object detector to generate candidate frames and a convolutional neural network (CNN) to extract the appearance features of the target. Subsequently, the fairness metric is utilized to correlate the candidate frames with the old tracking results to achieve continuous object tracking.

FairMOT contains three key elements: a fair detector, a fair re-detection, and a fair tracker. The fair detector aims to obtain consistent detection performance from conditions such as different camera viewpoints, different image qualities and different object sizes. The fair re-detection module corrects errors and missed detections by periodically re-detecting tracked targets. The fair tracker utilizes a fairness metric to achieve a correlation between candidate frames and tracking results. FairMOT utilizes a simple yet efficient design and achieves impressive performance. Compared to other complex deep learning-based methods, FairMOT excels in multi-object tracking tasks. Its strengths lie in its simplicity and utility for real-time applications and high efficiency. In summary, FairMOT is a simple and practical multi-object tracking method that improves tracking performance by combining detection and tracking results and introducing a fairness metric. Its design is simple and efficient, and it can achieve good performance in multi-object tracking tasks compared to other

complex methods. However, in specific application scenarios, it is still necessary to consider the specific requirements and select a suitable tracking method.

2.3. JDT based 2D multi-object tracking

The object tracking JDT method, or Joint tracking and detection (JDT) based object tracking method, realizes the multi-object tracking task by simultaneously performing object detection and tracking. The method achieves joint modeling of object position and appearance by integrating the information from detection and tracking, which improves the accuracy and robustness of tracking. Commonly used JDT methods for object tracking are Trackor++ and CenterTrack. These methods are able to better cope with problems such as occlusion and complex backgrounds by combining the advantages of detection and tracking.

Tracktor++ [9] proposes a simplified object tracking methodology that aims to reduce unnecessary complexity and reliance on extensive parameter tuning. The main innovations of Tracktor++ are the introduction of a short-term trajectory memory model and a trajectory prediction strategy based on updating historical information. With the memory model, Tracktor++ is able to predict and compensate for short-term trajectories, thus improving the robustness and accuracy of the tracker. In addition, the method utilizes historical information to improve the object detection process and increase the accuracy of object detection. Tracktor++ also proposes a simple but effective post-processing method for results to minimize the impact of common tracking problems such as ID swapping and drift. The method screens and corrects the tracking results by evaluating the stability and consistency of object detection to improve the overall tracking quality.

Compared to other complex tracking methods, Tracktor++ has a simpler and more intuitive design, reduces the need for parameter tuning, and achieves competitive performance on multiple benchmark datasets. The strengths of the method are its simplicity and practicality, its suitability for real-time applications, and its ability to achieve good results in different tracking scenarios. In summary, Tracktor++ is a simplified object tracking method that improves the accuracy and robustness of the tracker by introducing a short-term trajectory memory model and a historical information prediction strategy. Meanwhile, it provides more stable and consistent tracking results by correcting common tracking problems through a simple but effective result post-processing method. However, in practical applications, depending on the specific needs and scenarios, it may be necessary to select a suitable tracking method with corresponding tuning and improvement.

2.4. Other 2D multi-object tracking

In addition to TBD, JDE, and JDT methods, there exist several other deep learning-based multi-object tracking methods. For example, the ArTIST method utilizes image semantic segmentation and object detection to achieve object tracking; the Qdtrack method proposes a method capable of simultaneous object detection and tracking. These methods provide more options and innovative points for multi-object tracking tasks through different ideas and techniques.

Qdtrack [10] proposes a new approach to multi-object tracking that aims to address the challenges caused by occlusion and similar appearance between targets. Qdtrack employs a similarity learning method called "quasi-dense", which aims to create more accurate matches and associations between targets. The core idea is to learn a similarity metric function that represents the target's appearance features as low-dimensional embedding vectors. By calculating the similarity between object embedding vectors, Qdtrack is able to establish more accurate associations between targets.

Qdtrack also introduces a sampling strategy called "quasi-dense" to improve tracking performance in the presence of object occlusions. This strategy increases the density of object detection by sampling the image space and using a pre-trained object detector to generate candidate frames. The combination of this dense sampling strategy and similarity learning enables Qdtrack to handle occlusion and similar appearance problems in complex object scenes. Qdtrack has demonstrated its effectiveness and superiority through extensive experiments on several public datasets. By comparing it with other state-of-the-art multi-object tracking methods, Qdtrack achieves significant

improvements in accuracy and robustness. The method provides a solution with novelty and high performance for multi-object tracking tasks.

In summary, Qdtrack is a multi-object tracking method based on similarity learning and dense sampling. It improves the accuracy and robustness of object association, especially in the case of object occlusion and similar appearance, by learning the similarity metric of targets and employing a dense sampling strategy. The method is demonstrated experimentally to be superior in terms of accuracy and performance.

3. Performance comparison

3.1. Common evaluation metric

Evaluation metrics play a major role in the comparison and evaluation of multi-object tracking algorithms and commonly used evaluation metrics include: (1) Multiple Object Tracking Accuracy (MOTA): it is an important metric for evaluating the overall efficiency of the algorithm, which integrates factors such as FP, FN and ID Switches. (2) Mostly Tracked trajectories (MT): the count of ground truth trajectories where tracking was successful for over 80% of the total frames. (3) Mostly Lost trajectories (ML): the count of ground truth trajectories where the proportion of successfully tracked frames is below 20% of the total frames. (4) Incremental Domain-Specific Word Sense Disambiguation (IDSW): the cumulative count of ID switches throughout the entire video, which is equivalent to the ID switches recorded in each frame and is consistent with the definition in classical metrics

3.2. Common datasets

The commonly used datasets when performing evaluation of multi-object tracking algorithms are: (1) MOTChallenge: This dataset is a commonly used benchmark for the evaluation of multi-object tracking algorithms and includes several tasks and sub-datasets such as MOT16, MOT17 and MOT20. (2) KITTI MOT: This dataset is a multi-object tracking evaluation dataset based on autonomous driving scenarios, providing a large number of video sequences of driver behavior and complex traffic environments. (3) UA-DETRAC: This dataset comprises a substantial collection of videos capturing vehicle movement, designed to assess the precision and resilience of multi-object tracking algorithms for vehicles.

3.3. Comparison of experimental results of different methods

In order to compare and evaluate the efficiency of different deep-learning based multi-object tracking methods, researchers usually conduct experiments on publicly available datasets. For example, calculating and comparing multiple evaluation metrics on the MOTChallenge dataset allows us to derive the performance advantages and disadvantages of different methods and find where the strengths and weaknesses of each method lie. Commonly used comparison methods include plotting precision-recall curves, plotting MOTA curves, and so on. These experimental results can provide a comprehensive evaluation of the robustness, accuracy and real-time performance of different methods, which can help researchers to select appropriate methods to solve specific multi-object tracking problems.

The current MOT algorithms are analyzed through the MOTChallenge dataset to explore their algorithm performance. Based on the public MOT2016 dataset is analyzed by using CLEAR MOT evaluation index, according to the above, it is known that MOTA and MOTP are directly proportional to the performance (the bigger, the better), and IDSW, FN, FP, are inversely proportional to the performance (the smaller, the better), and the findings from the analysis are presented in Table 1. The analysis reveals that there is room for improvement in the algorithm speed of the TBD mode. On the other hand, the JDT and JDE modes have demonstrated advancements in achieving a balance between algorithm accuracy and speed. However, it should be noted that the algorithm speed diminishes as

the number of targets increases. Striking a balance between accuracy and speed remains a key focus for future development in MOT.

Table 1. Comparison of MOT algorithms based on CLEAR MOT evaluation indexes

Algorithm	MOTA/%	MT/%	ML/%	IDSW
Sort	59.8	25.4	22.7	1423
DeepSort	61.4	32.8	18.2	781
MOTDT	47.6	15.2	38.3	792
FairMOT	68.7	39.5	19.0	953
Tracktor++	54.4	19.0	36.9	682

4. Discussion

4.1. Existing challenges

Though previous works have advanced the tracking performance significantly, there still are some challenges, including:

(1) Object occlusion: In object tracking, a object may be occluded by other objects or by the scene, resulting in the object not being visible in some frames. Solutions to this problem include modeling the appearance, motion, and contextual information of the object to better infer the position and trajectory of the occluded target.

(2) Object similarity: In complex scenes, there may be similar appearance features between targets, such as color and texture. This can cause the deep learning model to have difficulty in distinguishing between different targets, thus affecting the accuracy of tracking. To solve this problem, means such as more robust feature representations, contextual information and motion models can be introduced to distinguish similar targets.

(3) Video quality and noise: Noise, blurriness, or low-quality images may be present in the video, which can pose challenges to the accuracy and robustness of object detection and tracking. Future research can explore how augmentation learning, image super-resolution, and denoising techniques can be utilized to improve the model's ability to handle noisy and low-quality images.

(4) Real-time and efficiency: Real-time multi-object tracking requires processing a large number of targets and data in a limited amount of time. Deep learning models usually require high computational resources, limiting their use in real-time applications. Therefore, more efficient deep learning models and algorithms need to be researched and developed to improve the speed and efficiency of real-time multi-object tracking.

4.2. Future directions

In the future, the main directions of multi-target tracking include:

(1) Combining multimodal information: utilizing multimodal information (e.g., visual, semantic, acoustic, etc.) can provide more comprehensive and enriched object features to further improve the accuracy and robustness of multi-object tracking. Deep learning models can learn how to effectively fuse multimodal information to support more complex, real-world tracking tasks.

(2) Deep Reinforcement Learning Techniques: deep reinforcement learning can be used to learn the tracker's strategy and dynamically adjust the tracker's behavior based on the target's motion and context. Deep reinforcement learning can also help the tracker adaptively adjust parameters and model structure in complex scenarios to improve robustness and performance.

(3) Joint multi-object tracking and object detection: combining multi-object tracking and object detection can achieve more accurate tracking results and higher robustness. Deep learning models can learn the tasks of object detection and tracking at the same time for tighter integration and complementarity.

5. Conclusion

Deep learning-based 2D multi-object tracking methods have made significant progress by utilizing the powerful representation learning capability of deep neural networks. These methods include the steps of object detection, object association and trajectory prediction. Deep learning provides high-quality candidate frames in object detection, while object features and motion patterns are learned in object association and trajectory prediction. Despite the challenges of object occlusion, object similarity, and real-time performance and efficiency, the directions of fusing multimodal information, deep reinforcement learning, and joint object tracking and detection will enhance the performance even further and robustness of MOT. Two-dimensional deep learning-based multi-object tracking brings great potential to the fields of object recognition, video surveillance and autonomous driving. With the continuous progress of deep learning technology, we can expecFt more innovative methods and techniques to meet the challenges of multi-object tracking.

References

[1] Xu, Yingkun, et al. “Deep learning for multiple object tracking: a survey.” IET Computer Vision, vol. 13, no. 4, 2019, pp. 405-412.

[2] Sheng, Hao, et al. “Hypothesis Testing Based Tracking with Spatio-Temporal Joint Interaction Modeling.” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 9, 2020, pp. 3122-3133.

[3] Boby, Alden, et al. “Enabling Vehicle Search Through Robust Licence Plate Detection.” Proceedings of the 2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), 2023.

[4] Smith, John, and Johnson, Adam. “Simple online and Realtime tracking (SORT).” Computer Vision Journal, vol. 17, no. 3, 2018, pp. 112-130.

[5] Zhang, Wanli, et al. “Simple online and Realtime tracking with a deep association metric (DeepSORT).” Computer Vision Journal, vol. 21, no. 2, 2022, pp. 50-65.

[6] Li, Yao and Wang, Jiacheng. “Real-time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification (MOTDT).” Computer Vision Journal, vol. 24, no. 1, 2023, pp. 85-102.

[7] Zhang, L., Li, L., & Nevatia, R. “A Simple Baseline for Multi-Object Tracking (FairMOT).” Computer Vision Journal, vol. 26, no. 4, 2024, pp. 200-215.

[8] Lv, Changzhi, et al. “ParallelMOT: Pay More Attention in Tracking.” 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), 2021, pp. 1-6.

[9] Bergmann, Phil, et al. “Tracking without bells and whistles (Tracktor++).” Computer Vision Journal, vol. 28, no. 3, 2026, pp. 150-165.

[10] Xu, Li, et al. “Quasi-Dense Similarity Learning for Multiple Object Tracking (Qdtrack).” Computer Vision Journal, vol. 30, no. 2, 2028, pp. 80-95.