Multi-target tracking algorithm based on deep learning

Wenxiao Huo¹, Jiayu Ou², Tianping Li*

¹ School of Physics and Electronic Science, Shandong Normal University, Jinan, Shandong, China
² School of Physics and Electronic Science, Shandong Normal University, Jinan, Shandong, China
* School of Physics and Electronic Science, Shandong Normal University, Jinan, Shandong, China

Corresponding author’s e-mail: sdsdltp@sdnu.edu.cn

Abstract. With the continuous development of deep learning in multi-target tracking, the use of convolutional neural network for feature extraction has replaced the traditional feature extraction method, but the accuracy of target tracking needs to be improved. In order to further improve the accuracy of multi-target tracking, a new multi-target tracking algorithm based on RFB is proposed in this paper. The algorithm is mainly divided into three parts: multi-target detection, feature extraction and multi-target tracking. In the multi-target detection part, CenterNet was selected as the detection network to improve detection accuracy. In feature extraction part, RFBNET is combined with pedestrian re-recognition network to strengthen feature extraction capability. DeepSORT algorithm is used in multi-target tracking. Experimental results on MOT16 data set show that the proposed algorithm is more effective than other methods.

1. Introduction

Multi-target tracking is a hot research field in the world, and it has developed rapidly in recent years, and its application fields are more and more. In order to achieve better tracking effect, convolutional neural network is widely used in multi-target tracking. Voigtlaender[1] et al. proposed a multi-target tracking and segmentation method and used convolution for feature extraction. However, due to the inherent recognition domain, the convolution-based feature extraction mixes the foreground feature and the background feature. In this paper, a new multi-target tracking algorithm based on RFB is designed to improve the feature extraction ability of the network by using RFBNET, so that it can accurately extract the required feature information under the condition that the target is occluded, so as to prevent false detection of the model. Then, through data association, different targets can be better distinguished.

2. Multi-target tracking

In the aspect of multi-target tracking, although the DeepSORT [2] tracking algorithm has added appearance extraction features and reduced the number of ID switching on the basis of the SORT [3] tracking algorithm, the accuracy and ID switching performance still need to be improved, so we have made some improvements.
2.1. Target detection
In the process of multi-target tracking, detection and tracking are two complementary problems, and good detection can reduce tracking orbit drift. At present, the popular target detection algorithms based on convolutional neural network include Flyer R-CNN[4], SSD[5] and YOLOV3 [6]. However, while the convolutional neural network improves the detection accuracy, it also increases the depth and complexity of the network, which increases the computational burden. In order to improve the detection speed and maintain the detection accuracy, we choose CenterNet as the target detection network. CenterNet is an anchor-free target detection network. It has the advantages of small model, fast running speed and high precision.

2.2. A new pedestrian re-identification network using RFB
In the field of multi-target tracking, feature extraction is the premise of target matching, and it is also an essential part of tracking. Feature extraction network can extract the deep semantic features of the target from the image, and the extracted target features have very powerful expressive ability. Pedestrian reidentification network can alleviate the problem of ID switching in multi-target tracking, so we retain the feature extraction of Reid and add RFBNET on this basis to strengthen the learning of deep target features by CNN model.

2.2.1. R-Reid Appearance Feature Extraction Network
In this paper, the RFB module is combined in the Reid pedestrian re-identification network. By simulating the relationship between the size of RF and the eccentricity, the RFB enhances the discriminability of features and the robustness of the model, and carries out larger regional feature extraction while maintaining the same number of parameters. The feature extraction network in this paper is trained offline on Market-1501 dataset [7].

The network architecture of feature extraction is shown in Figure 1. The input image first goes through the RFB module to increase the receptive field of feature extraction, and then enters two 3×3 convolution and a pooling layer to reduce the dimension of vector and avoid overfitting. Finally, the extracted appearance features are output through six residual blocks.

2.2.2. RFB module
RFB is a multi-branching structure composed of convolutional layers of convolutional cores of different sizes. The network structure is shown in Figure 2. It adopts void convolution to expand the receptive field, which is helpful for detecting target information of different sizes. The module mainly includes the following parts: three 1x1 convolution, mainly used to reduce the amount of calculation and cross-channel information fusion; Three 3×3 void convolution layers expand the receptive field of the network; Then three features of different sizes are added to achieve the purpose of fusion of different features. Finally, the fused feature is sent into a 1×1 convolution, and its output is combined with the output concat of shortcut. Through activation function, the feature map is output.

Figure 1  R-Reid Appearance Feature Extraction Network
2.3. Date association

The DeepSORT tracking algorithm mainly uses motion information and appearance information to carry out data association. These two associations are described in detail below.

Motion information association: the Mahalanobis distance between the detection box and the tracking box is used to describe the degree of motion association.

\[
d^{(1)}(i, j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i)
\]  
(1)

Formula (1) represents the motion matching degree between the jth detection result and the ith track. Whereinto, \( S_i \) is the covariance matrix of the observation space at the current moment of the Kalman filter, \( y_i \) is the predicted observation quantity at the current moment, and \( d_j \) represents the observation quantity of the jth detection.

Appearance information association: Since camera motion invalidates the Mahalanobis distance measurement method, a second correlation method is introduced. R-Reid network is used to extract the unit norm feature vector \( r \), and then the minimum cosine distance between the ith tracker and the jth detection result feature vector of the current frame is taken as the apparent matching degree. The formula is as follows.

\[
d^{(2)}(i, j) = \min \left\{ 1 - r_j^T r_k^{(i)} \mid r_k^{(i)} \in R_i \right\}
\]  
(2)

A linear weighting of the above two measures was used as the final measure:

\[
c_{i, j} = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j)
\]  
(3)

3. Experiment

3.1. Date Set

The data set used in this paper is MOT16[8]. MOT16 data set is a data set proposed in 2016 that is specially used to measure the standard of multi-target tracking detection and tracking methods and is specially used for pedestrian tracking, including 7 training sets and 7 test sets. This data set exists in the form of video frames, from the first frame to the last frame of the video. There are multiple objects in constant motion and the scene is complex.

3.2. Comparative experiment of multi-target tracking algorithms

We evaluated the tracking performance on the sequence set provided by the MOT benchmark database, and evaluated the multi-target tracking algorithm against the MOT Challenge standard [9]. In this paper, classical SORT and DeepSORT methods were selected for comparison, and the results are shown in Table 1.
Table 1  Comparison of evaluation indexes of multi-target tracking algorithm

|       | MOTA↑ | MOTP↑ | MT↑  | ML↓ | ID↓ | FP↓ | FN↓ | Runtime↑ |
|-------|-------|-------|------|-----|-----|-----|-----|---------|
| SORT  | 59.8  | 79.6  | 25.4%| 22.7%| 1423 | 8698 | 63245 | 60Hz    |
| DeepSORT | 61.4  | 79.1  | 32.8%| 18.2%| 781  | 12852| 56668 | 20Hz    |
| Ours  | 63.2  | 79.4  | 30.2%| 19.5%| 675  | 9876 | 56503 | 30Hz    |

Indicators in Table 1 have the following meanings:
MOTA(↑): Multi-target tracking accuracy.
MOTP(↑): Multi-target tracking accuracy.
MT(↑): The number of tracking tracks that account for more than 80% of the real length of the target.
ML(↓): The number of tracks that the target tracking lost accounts for up to 20% of the real length.
ID(↓): The number of times the target ID has been switched.
FP(↓): the total number of false detection.
FN (↓): Total number of missed detection.
Runtime(↑): Frame frequency.

For the evaluation index with (↑), the higher the score, the better the effect; For the evaluation index with ↓, the lower the score, the better the effect.

As can be seen from Table 1, the running speed of the algorithm in this paper is 10Hz higher than that of DeepSORT. The tracking accuracy of multi-targets is 3.4% higher than that of SORT algorithm and 1.8% higher than that of DeepSORT algorithm. Compared with DeepSORT algorithm, the total number of false detection and the total number of missed detection are reduced by 2976 and 165. The number of ID switches decreased by 106 times.

3.3. Multi-target tracking algorithm rendering
In order to further demonstrate the effectiveness of the algorithm in this paper, part of the running effects are shown in Figure. 3, which are respectively frame 91, frame 123, and frame 188. From the figure, we can see that the target is continuously tracked and the tracking effect is good. The number on the target box in the picture represents the ID number assigned to each person. It can be seen from the figure that target No. 21 and target No. 32 can be correctly associated when they reappear after being fully blocked by target No. 4. In addition, target No. 40, which is relatively small in the distance from the camera, can also be correctly detected and tracked. Experimental results show that the proposed method can solve the target occlusion problem in tracking to a certain extent and reduce the ID switching frequency, which has certain practical value and research significance.

Figure 3  Tracking effect

4. Conclusion
In order to improve the accuracy and speed of multi-target tracking, a new multi-target tracking algorithm based on RFB is proposed in this paper. The algorithm takes CenterNet as the detection network, uses RFBNet in the ReID part to enhance the resolution and robustness of appearance feature extraction, and then realizes the tracking between targets through data association. Aiming at the problems in multi-target tracking, such as the change of pedestrian illumination, the proximity of background and target color, and the blocking of target, our method can track pedestrians more
effectively. After experimental evaluation, we have improved the tracking accuracy of multi-targets to a certain extent, and reduced ID switching.

Acknowledgments
The authors would like to thank the authors of Market-1501 dataset and MOT16 dataset for sharing their experimental data with us.

References
[1] Voigtlaender P, Krause M, Osep A, et al. MOTS: Multi-Object Tracking and Segmentation[C]2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.
[2] Wojke N, Bewley A, Paulus D. Simple Online and Realtime Tracking with a Deep Association Metric[J]. IEEE, 2017:3645-3649.
[3] Bewley A, Ge Z, Ott L, et al. Simple Online and Realtime Tracking[J]. 2016 IEEE International Conference on Image Processing (ICIP), 2016.
[4] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
[5] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. Springer, Cham, 2016.
[6] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
[7] Liang Z, Zhi B, Sun Y, et al. MARS: A Video Benchmark for Large-Scale Person Re-Identification[J]. Springer, Cham, 2016.
[8] Milan A, Leal-Taixé L, Reid I, et al. MOT16: A Benchmark for Multi-Object Tracking[J]. 2016.
[9] L Leal-Taixé, Milan A, Reid I, et al. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking[J]. 2015.