2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP) | 978-1-6654-7857-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSP54964.2022.9778590

Pedestrian Angle Recognition Based on JDE Multi-object Tracking Algorithm

Chunde Yang[1], Tongyao Xu[2], Minghong Lv[3], Yi Yu[4], Xiaoming Jiang[*]
[1,2]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[3,4,*]School of Bioinformatics, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[*]tongyao.xu@outlook.com

Abstract. **We often execute multi-object tracking algorithms and pedestrian angle recognition algorithms independently when we solve complex problems that need to track multiple pedestrians and identify their angles simultaneously. Firstly, the multi-target tracking algorithm is used to locate the target and determine the target's identity. Then the pedestrian angle recognition model is used to recognize the pedestrian angle. This strategy is complicated and inefficient since the two models repeatedly extract target features. This paper realizes the combination of multi-object tracking algorithm and pedestrian angle recognition algorithm in response to this problem. Specifically, we added a new classification branch based on the JDE multi-object tracking algorithm. It enables the model to perform pedestrian angle recognition while detecting and embedding feature extraction. Finally, experiments show that the improved algorithm can perform pedestrian tracking and pedestrian angle recognition simultaneously. The algorithm's ability to track large and medium targets is the same as the JDE multi-object tracking algorithm. Compared with the Hydraplus model, the improved algorithm has a higher recognition accuracy. Besides, the algorithm can achieve real-time performance.**

Keywords: surveillance scene; JDE multi-object tracking algorithm; pedestrian angle recognition; real-time

I. Introduction

Multi-object tracking tasks and pedestrian angle recognition tasks are common tasks in computer vision. The purpose of the multi-object tracking task is to output the trajectory of all targets in the video. The purpose of the multi-label classification task is to output multiple attributes of the target in the image. Multi-object tracking tasks and pedestrian angle recognition tasks are closely related, and they often appear together. In the complex surveillance video field, we sometimes need to track multiple pedestrians and classify each pedestrian's attributes. However, these two different computer vision tasks are often studied independently, and few people combine them. Therefore, how to combine the multi-target tracking task with the pedestrian angle recognition task has become an urgent problem to be solved.

At present, the most straightforward method is to use a tracking model to detect the location of all targets and the ID of each target appearing in the video, and use another model to classify the attributes of each pedestrian appearing in the video. Although this method can solve this problem, this method has some disadvantages. Since the tracking model and the classification model do not share the feature extraction part, the issue of poor real-time performance often occurs. To solve this problem, we propose a model to simultaneously complete the detection of all targets in the video, the extraction of apparent features, and the recognition of pedestrians. Different from the method of using the tracking model and the recognition model separately, our method achieves the purpose of improving the inference speed of the network model by sharing the feature extraction part of the tracking model and the angle recognition model. Compared with tracking methods, our method can recognize pedestrian angles. Compared with pedestrian angle recognition methods, our method is capable of multi-object tracking. In addition, our method almost meets the real-time requirements.

The overview of our work is as follows:

(1) We introduce an algorithm that combines multi-object tracking and pedestrian angle recognition. The algorithm uses a model to simultaneously extract the location of all targets in the video, the appearance features of each target, and the angle recognition information.

(2) We conducted analysis on the algorithm from overall architecture, detection and embedding feature extraction, and pedestrian angle recognition.

(3) Compared with the JDE multi-object tracking algorithm, our algorithm has the same ability to track large and medium targets. Compared with the Hydraplus pedestrian attribute recognition algorithm, our algorithm has a higher angle recognition ability. Besides, compared with the two-stage method, our method is real-time.

II. Related work

At present, the mainstream method in the field of multi-object tracking is detection-based tracking, which has four stages, including object detection, appearance feature and motion information extraction, affinity calculation, and data association.Among the detection-based multi-object tracking algorithms, the most representative one is the DeepSort[1] multi-object tracking algorithm, which is proposed by Wojke et al.

In recent years, there have been some methods to integrate the detection network with the ReID feature extraction network, such as the JDE multi-object tracking algorithm, which is proposed by Wang et al[2]. Based on the YOLO[23] model, this algorithm adds a branch to extract embedded features. They combined the object detection step and the embedded feature extraction step to achieve the trade-off between model speed and accuracy. Subsequently, Zhan et al proposed the FairMot[3] multi-object tracking algorithm and pointed out that

the anchor-based detection framework has the problem of misalignment of anchor and object. P. Bergmann et al.[4] proposed Tracktor++ multi-object tracking algorithm, which only uses detectors to perform tracking tasks. They modified the structure of RFCN[5], and used the target area of the previous frame as a candidate area generated by the RPN network to regress to obtain the position of the target in the current frame.Zhang et al.[6] proposed an FFT multi-object tracking algorithm, which adds an optical flow branch to Tracktor++ to improve tracking performance.

Huang et al.[7] proposed the MIFT multi-object tracking algorithm, which improved the motion model and data association part of Tracktor++. For the motion model, MIFT integrates the Kalman filter algorithm with the ECC model. Experiments prove that the fused Kalman filter and the ECC model can improve 1.4 MOTA compared to the separate Kalman filter and the ECC model. Zhou et al.[8] proposed the CenterTrack multi-object tracking algorithm, which implements a tracking mechanism similar to Tracktor++ based on the CenterNet[9] detector.

In addition, there are some multi-object tracking algorithms based on single-target trackers. Chu et al.[10] proposed the FAMNet multi-object tracking algorithm, which assigns a single target tracker to each object for tracking. The algorithm consumes a lot of computing resources and has poor real-time performance.

III. Improved JDE tracking algorithm

A. Architecture overview

This study modified the Feature Pyramid Network Structure (FPN)[11]. Specifically, we subtracted the branch for predicting small targets from the FPN, and added two branches for predicting the classification information of large targets and medium targets from the FPN. Figure 1 shows the improved network structure. In the model inference stage, each frame of picture is up-sampled by 1/8, 1/16, and 1/32 to obtain three feature maps of different sizes. Input the feature map of the smallest size into the classification branch and the detection/embedding feature extraction branch, respectively, to obtain the prediction information of the large target. The feature map of the smallest size is fused with the feature map of the medium size, and then the fused feature map is input into the classification branch and the detection/embedding feature extraction branch, respectively, to obtain the prediction information of the small target.

Finally, the prediction result obtained by each branch includes four parts, as shown in Figure 2:

(1) Classification information (target, non-target): 2A*W*H;

(2) Regression information (bounding box coordinates): 4A*W*H;

(3) Feature embedded information (pedestrian re-identification): 512*W*H;

(4) Classification information (pedestrian angle recognition): 6*W*H;

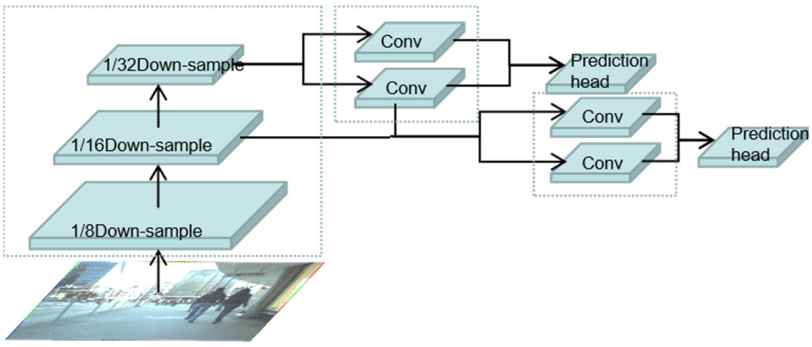

Figure 1. Network architecture. The network has two branches for predicting large targets and medium targets, and each branch includes two sub-branches for predicting classification information and detection/embedded features.

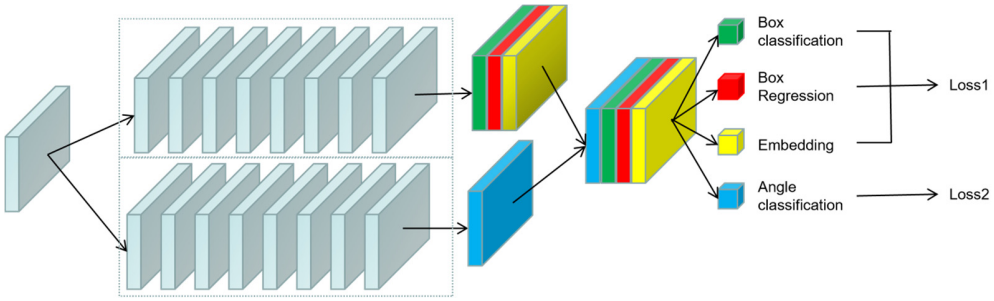

Figure 2. Prediction head. The prediction head shows the final output information of the model, including classification information, detection information, embedded features, and angle recognition information.

Since the pedestrian attribute recognition dataset has no location information, we use the location information predicted by the detection branch to guide the training of the classification branch. Firstly, this paper select the position with the greatest confidence from the prediction results generated by the detection branch. Then, this paper find the corresponding classification result in the classification branch through the position with the greatest confidence. Finally, this paper use BCEWithLogits to calculate the loss between the found prediction result and the real label. In the training stage, this paper use the tracking dataset to train the detection and embedding feature extraction branch. Then, this paper fix the parameters of the backbone network and the detection branch. At the end, this paper use the prediction results produced by the detection branches to guide the training of classification branches.

B. Detection and embedding feature extraction

Like the JDE multi-object tracking algorithm, the detection part uses the cross-entropy loss function to classify the target and the non-target and uses the smooth-L1 loss to calculate the regression loss of the bounding box. Besides, this paper use triplet loss to calculate embedded feature extraction loss.

C. Multi-label classification

This paper uses a separate tracking dataset and a classification dataset instead a composite dataset, which contains both classification information and tracking information. The two reasons are as follows. First, because the classification information is coupled with the tracking information in the composite dataset, it is difficult for us to extend our model to other classification tasks. Secondly, due to the classification dataset is far more than the tracking dataset in our reality world, we can make full use of the existing classification dataset. Since the classification dataset has no position information, we use the position predicted by the tracking branch as the classification target's position when the classified image is input. This paper calculates classification loss by the real label and the selected classification information predicted by the position. This paper classifies human attributes into six categories, and uses BCEWithLogits to calculate the classification loss.

IV. Experiment and result analysis

A. Datasets and evaluation criteria

In this paper, we created a multi-label multi-category dataset based on the existing pedestrian re-identification dataset and pedestrian attribute recognition dataset, which contains 23537 pictures. In this datasets, each picture contains two labels (face angle, body angle), and each label has three Categories (FrontFace, SideFace, BackFace; FrontBody, SideBody, BackBody). Finally, this paper uses the new pedestrian attribute recognition dataset to training model.

B. Experimental results

This paper's experimental part is divided into three parts: comparison with JDE multi-object tracking algorithm, comparison with HydraplusNet pedestrian attribute recognition algorithm, and visual analysis.

Because the improved algorithm removes the branch of tracking small targets and the parameters of tracking large and medium targets branch are exactly the same as the JDE tracking algorithm, our tracking algorithm has the same ability to track large and medium targets compared with the JDE tracking algorithm. Since the model adds two classification branches, the improved algorithm is slower than the JDE tracking algorithm. Despite this, our algorithm still achieves real-time performance. Using the video provided by the MOT16 dataset to test and select the best result, the model inference speed reached 23.73FPS. The hardware equipment used is NVIDIA RTX2080Ti.

Deng et al used average accuracy(mA) to evaluate multi-attribute classification algorithms. Firstly, they calculate the accuracy of the positive sample and the accuracy of the negative sample for each attribute. Then, they calculate the average accuracy of the positive and negative samples. We use the above evaluation criteria to test our model, and the results are shown in Table 1:

Table 1. Comparison with HydraplusNet

	Attribute1	Attribute2	Attribute 3	Attribute 4	Attribute 5	Attribute 6	mA
HydreaplusNet	0.95	0.91	0.91	0.93	0.85	0.91	0.91
ours	0.96	0.90	0.92	0.95	0.95	0.95	0.94

The above evaluation criteria treat each attribute independently, ignoring the relationship between the attributes in the multi-attribute recognition problem. Li et al referred to the above evaluation standards as label-based evaluation standards and used sample-based evaluation standards to solve this problem. Example-based evaluation standards include four indicators: Accuracy, Precision, Recall, and F1. We use the above evaluation criteria to test our model, and the results are shown in Table 2:

Table 2. Comparison with HydraplusNet

	Accuracy	Precision	Recall	F1
HydraplusNet	0.86	0.90	0.89	0.89
ours	0.90	0.92	0.92	0.92

Using 1088*608 size video for testing, it is found that the model maintains the tracking performance of JDE algorithm for medium and large targets, the recognition effect is better, and the frame rate reaches 23.73FPS.

Figure 3. Experimental effect. The ID, face angle, and body angle of each target appearing in the video are given.

We use GradCam to visualize the areas of interest of the model on the front face, side face, back face, front face, side face, and back face respectively. The result is shown in Figure 4.

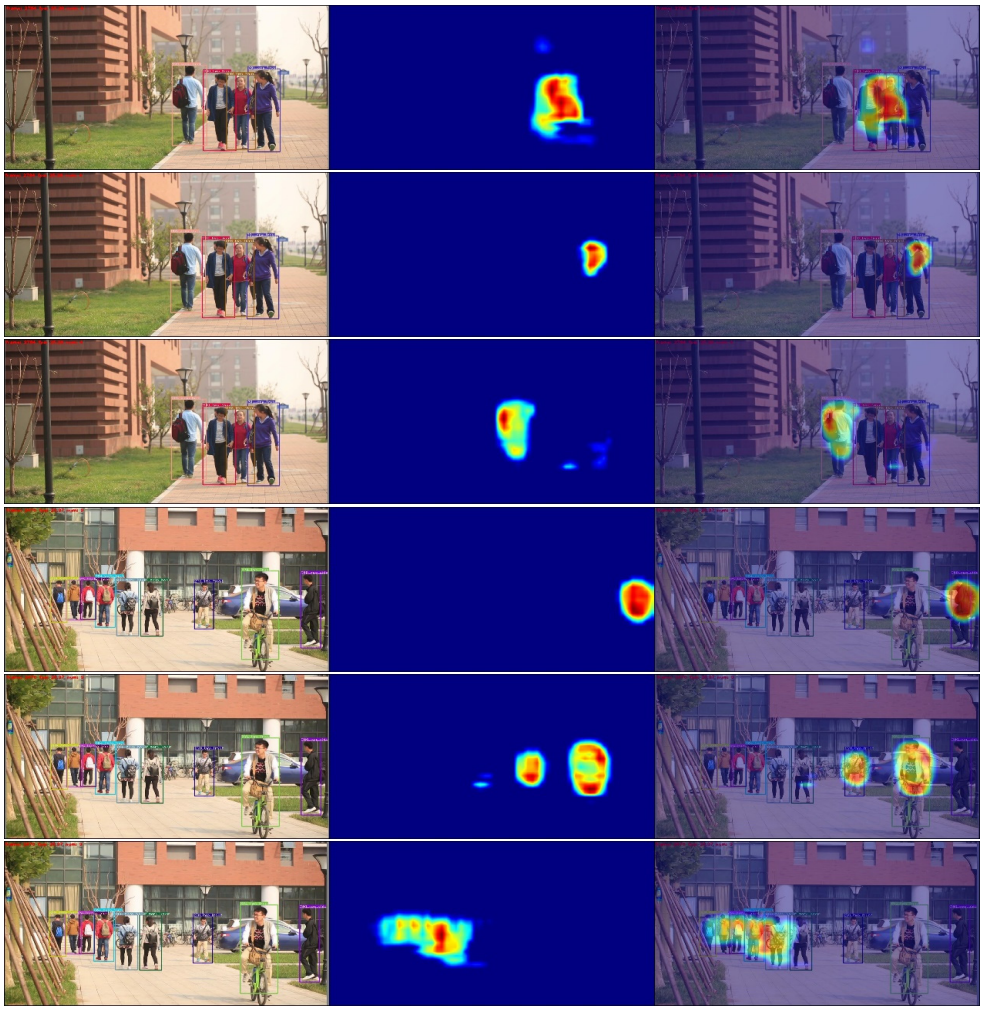

Figure 4. Illustration of visual analysis. 1-3 show the model's response to the front face, side face, and back face, respectively, and the 4-6 pictures show the model's response to the front body, side body, and back body.

V. Conclusion

Aiming at the large storage space of surveillance video and the difficulty of browsing and retrieval, this paper combines the multi-object tracking task with the classification task and improves the JDE multi-object tracking algorithm. it can recognize the pedestrian angle while completing the tracking task. After analysis and experiments, the following conclusions are drawn:

(1) The improved algorithm has the same ability to track large and medium targets as the JDE multi-object tracking algorithm.

(2) The improved algorithm has higher recognition accuracy.

(3) The improved algorithm can basically meet the real-time performance and can achieve the purpose of practical application.

References

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

[2] Wang Z , Zheng L , Liu Y , et al. Towards Real-Time Multi-Object Tracking[M]. Springer, Cham, 2020.

[3] Wang Z , Zheng L , Liu Y , et al. Towards Real-Time Multi-Object Tracking[M]. Springer, Cham, 2020.

[4] Zhan Y, Wang C, Wang X, et al. A Simple Baseline for Multi-Object Tracking[J]. arXiv preprint arXiv:2004.01888, 2020.

[5] Bergmann P, Meinhardt T, Leal-Taixe L. Tracking without bells and whistles[C]//Proceedings of the IEEE international conference on computer vision. 2019: 941-951.

[6] Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016: 379-387.

[7] Zhang J, Zhou S, Chang X, et al. Multiple Object Tracking by Flowing and Fusing[J]. arXiv preprint arXiv:2001.11180, 2020.

[8] Huang P, Han S, Zhao J, et al. Refinements in Motion and Appearance for Online Multi-Object Tracking[J]. arXiv preprint arXiv:2003.07177, 2020.

[9] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//European Conference on Computer Vision. Springer, Cham, 2020: 474-490.

[10] Chu P, Ling H. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 6172-6181.

[11] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.