Every detection box for multiple object tracking based on GAM

Yang Li*, Dianwen Wang, Zhejun Li, Xiao Xiao, Zhe Fu, Junlin Zhang
Traffic Control Technology Co., Ltd., Beijing 100070, China

ABSTRACT

Multiple object tracking (MOT) has become a very important task in the field of rail transit. Recently, the most effective method in MOT tasks is tracking by detection. Because there are many pedestrians under a single camera in the railway station, the occluded pedestrians will generate detection boxes with low-score. Once these low-score bounding boxes are mistakenly classified as background by the detector, the traditional tracking model may discard these boxes, resulting in poor tracking effect. In order to solve this problem effectively, this paper proposes a novel model with Generic Association Mechanism (GAM), which can pay attention to every detection box instead of only focusing on the high-score ones. This method utilizes the similarity between these low-score detection boxes and trajectories to restore true objects and filter out the background detection boxes. More attention can be focused on these low-score detection boxes through this method, which contribute the most to the tracking effect. This method can efficiently avoid the loss of true objects and fragmented trajectories of occluded objects. Finally, we test this proposed model with other models on several public datasets, and the results show that the proposed method achieves more significant performance improvement.

Keywords: Multiple object tracking, rail transit, generic association mechanism

1. INTRODUCTION

Multiple object tracking (MOT) is a challenging task in the field of rail transit. Since the passenger flow in the rail transit station is dense and the direction of passengers is changeable, the objective of this task is to accurately present the trajectory of passenger flow in the rail transit field. This task has many applications in practical scenarios, such as passenger flow analysis, intelligent driving, passenger flow prediction, passenger search and so on.

Recently, tracking-by-detection (TBD) paradigm has been widely used in the field of multiple-object tracking. Most of the existing methods[1-3] attempt to integrate detection and ID embedding extraction in a unified network to address the problem: Each frame of image in the video is used as the current input of the model. The detection part of the model first detects the target position to get its bounding box, and then the model conducts correlation analysis between the detected target features and the existing trajectory, and finally chooses to associate it to the existing trajectory or create a new trajectory. However, in the rail transit scene, the occluded pedestrians are prone to have low-score detection boxes, which will lead to these boxes with low scores than the threshold to be mistaken for background boxes and discarded. This situation is common in the field of rail transit, which will lead to poor tracking effect.

To solve this problem, it is a new perspective to focus on every detection box. Therefore, we propose a novel model with Generic Association Mechanism (GAM). The model framework is shown in Figure 1. This GAM model can process occluded pedestrians scenes. This work is based on FairMOT[1] model, which is the most widely used one-shot model. After each frame is processed in the detection step, these targets in a single video frame are located. Then the GAM model will pay attention to every detection box in the frame in order to recover the misclassified bounding boxes, such as fake background. The GAM model has two unique processing capabilities: When the score of the detection box is relatively high, the GAM model will match them according to their similarity with the trajectory as usual; When the low-score detection box comes, it will eventually use these boxes to repair these broken tracklets as many as possible. Thus, the GAM can perform more attention on low-score detection boxes than traditional method. In addition, although the GAM model only uses all the detection boxes with different scores when matching trajectories, it can also achieve great results when dealing with the discontinuity of tracking paths caused by the reappearance after occlusion. This result fully demonstrates motion cues can help in the process of scene association involving occlusion. In conclusion, the innovation of this paper is mainly in three parts:

* yanggglee@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125060V
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661751

The encoder-decoder model reads each frame of picture from the video and inputs it into the network in turn. Then the model generates detection boxes for this image. After that, the GAM improves temporal consistency by associating every detection box.

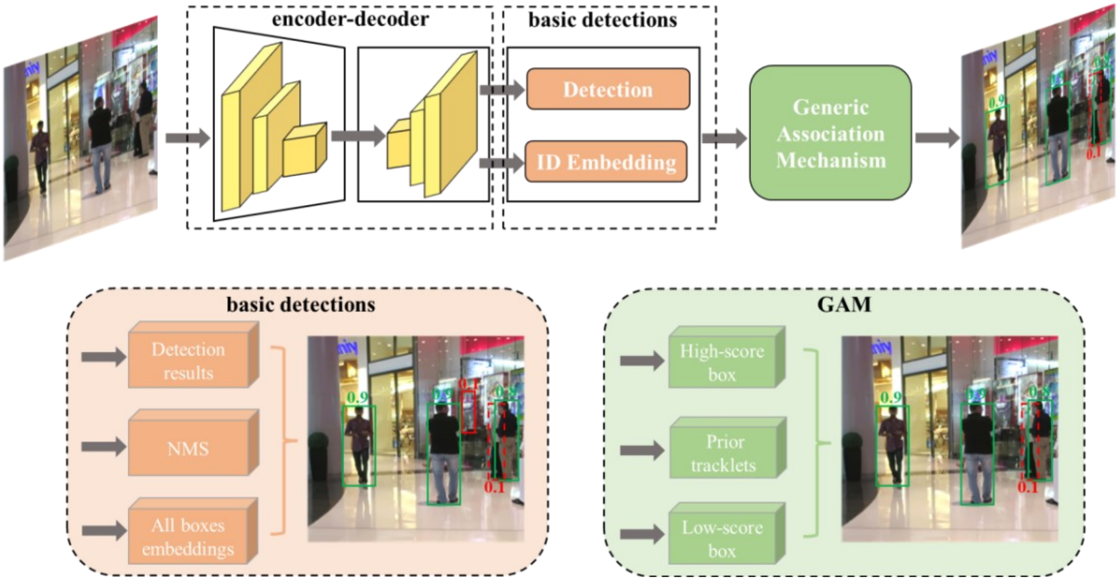

Figure 1. The model framework with the Generic Association Mechanism.

- We present a model with Generic Association Mechanism (GAM) for multiple-object tracking, which can ensure restore the misclassified bounding boxes.
- We only make full use of the detection boxes in the matching process, without adding attention mechanism or other modules, thus maintaining the running speed of the model.
- We verify this GAM model on two public datasets: MOT17 and MOT20, the results show that our method can achieve favorable gains.

2. RELATED WORKS

In recent years, the research in the field of multiple object tracking has developed rapidly. Scholars have put forward different innovation points and derived different models. These models are mainly divided into two types of methods, namely two-step structure and one-shot structure.

The two-step structure approach follows the idea of tracking after detection, where the detector first predicts the bounding boxes of objects and then links to tracklets by an association network. These methods, such as SORT[4] and DeepSORT[5], mainly focus on improving the accuracy of correlations. However, the ID of each candidate box is extracted and embedded through an isolated ReID network, which brings huge computation cost to the tracking task. To deal with this issue, scholars put forward one-shot structure.

The one-shot structure approach follows the idea of integrating detection and ID embedding, which has drawn great attention. They can run at a quasi real-time speed by sharing features. Tong et al.[6] first proposed a model by adding fully connected layers to a two-stage detector. This model can generate detection boxes and the corresponding ID embedding at the same time, so as to jointly handle the detection and Reid tasks. Due to the continuous improvement of the detection performance of models in the field of object detection, more and more tracking models use more advanced object detection modules in order to obtain higher tracking performance, such as RetinaNet and YOLO series detectors. However, when the tracking path is discontinuous due to occlusion in continuous image frames, the detector may discard some false low-score detection boxes, resulting in poor tracking effect. In order to maintain the persistence of trajectories, this paper presents the model with Generic Association Mechanism.

3. METHOD

This section is divided into two parts. First, we outline the general framework structure of multiple object tracking and briefly comb the tracking process. Finally, we introduce the proposed model with Generic Association Mechanism.

3.1 Framework overview

Our proposed model is conducted on FairMOT tracker which is a variant of the widely used one-shot structure. When the video frame $\boldsymbol{v}$ comes, the encoder-decoder extractor ξ will first process it. And then the model generates the feature X_t as follows: $X_t = \xi(v)$. Then X_t is picked up by the basic detection module to simultaneously predict detection results and ID embeddings as follows:$\left[B_t^{det}, X_t^{id}\right] = \varphi(X_t)$, where $B_t^{det}$are the detection results (including one map $P_t^{det}$for foreground probabilities and the others R_t^{det} for raw boxes). X_t^{id} denotes ID embedding. The detection results B_t^{det} are processed by NMS to generate the basic detection D. Each box in D corresponds to an embedding in X_t^{id}. We denote A_t^{id} as a set that contains embedding of all boxes in D. Then, the boxes D and the ID embedding A_t^{id} are utilized to associate with prior tracklets by GAM. Finally, the model will output the bounding boxes and IDs of the tracks T of the current image as the final result.

3.2 Model with GAM

The input of GAM is divided into three parts: detection boxes D, scores S and the Kalman Filter KF. The model also needs to be given three thresholds S_{high}, S_{low} and S_{track} (S_{track}>S_{high}>S_{low}). S_{high} and S_{low} are the threshold to determine whether the score of the detection box meets the requirements of the GAM model and S_{track} is the threshold to determine whether the tracking score meets the requirements of the GAM model. After the GAM model processes the current frame $\boldsymbol{v}$, it obtains the target trajectory T. At the same time, the GAM model will give the bounding box of the target and its identity information. The structure of the proposed GAM model is shown in detail in Figure 2.

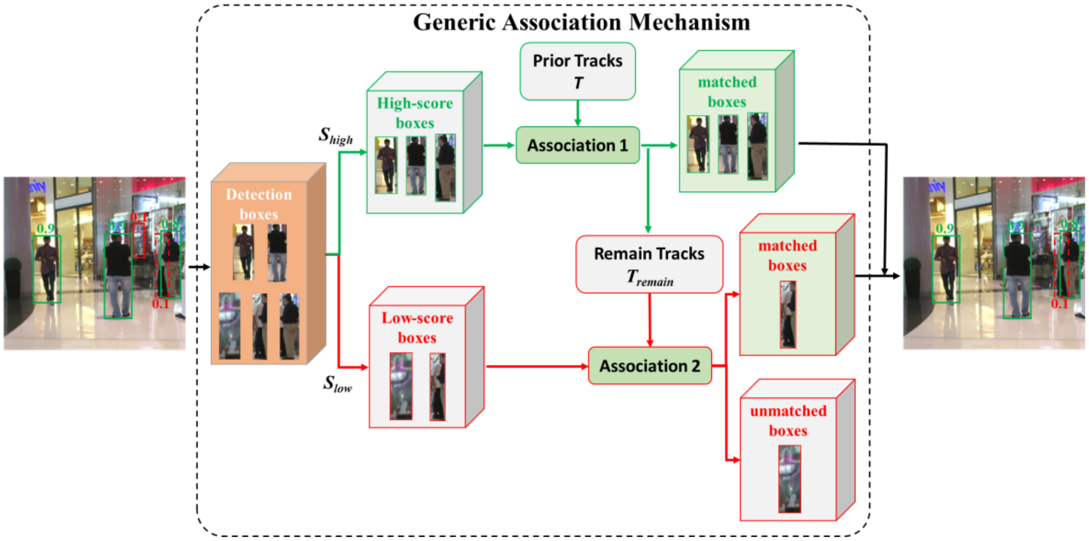

Figure 2. Method framework diagram.

Detection boxes and prior tracks are fed into the GAM. The GAM model first matches the high-score boxes according to the relationship between the previous trajectory. After that, the GAM model matches the low-score boxes with the remain trajectories according to a specific relationship as the second association. Finally, all the matched boxes generate the final detection result.

When the input comes, the GAM model will select three parts D_{high} , D_{low} and D_{remain} from all input boxes to be detected according to thresholds S_{high} and S_{low}. The specific operations are as follows: High score detection boxes D_{high} store boxes with a score higher than threshold S_{high}, because these boxes can easily identify their target identity and trajectory. The low score detection box D_{low} stores the boxes whose score is between threshold S_{low} and S_{high}, because these boxes can also match the appropriate trajectory according to the specific rules when performing secondary matching. For boxes with a score below the threshold S_{low}, further processing will be carried out in the subsequent session. For these separated parts, the GAM will use KF to predict the next position of each box in its matched trajectory T.

In the first stage of GAM module processing, GAM matches the boxes in the high score detection box D_{high} with all trajectories T. When matching, GAM first calculates the similarity between the detection box D_{high} and the prediction box of trajectory T according to the IoU. And then the GAM uses the Hungarian algorithm to complete the matching according to the similarity. In addition, the GAM retains the unmatched detection in D_{remain} and the unmatched tracks in T_{remain}. The D_{remain} will be used after these two associations and the T_{remain} will be used in the second association.

In the second stage of GAM module processing, GAM matches the boxes in the low score detection box D_{low} with remaining tracks T_{remain}. The difference between this association and the previous appearance feature methods is that it still uses IOU as the similarity. However, the box with low score detection is generated because the object display is incomplete due to serious occlusion, which makes the credibility of object features very low. After this matching, low score detection boxes incorrectly classified as the background will be matched to the remaining track T_{remain} at the first stage. The remaining low score detection boxes that are not matched are basically the real background. The GAM will delete these boxes and keep the unmatched tracks in $T_{re\text{-}remain}$. Then, the GAM will put the unmatched tracks $T_{re\text{-}remain}$ into T_{lost}. Like other trace processing methods, the GAM model deletes tracks in T_{lost} that have not been updated for a long period of time. This longer time period is generally used as an adjustable parameter. Otherwise, the lost tracks T_{lost} in T will be retained.

For the detection boxes D_{remain} generated after the first stage of processing, the GAM model will initialize it into new tracks. Specifically, an adjustable parameter K is defined to represent the number of frames. If the detection box in D_{remain} has a higher score than S_{track} and the number of consecutive frames is greater than K, the GAM model will initialize a new track for it.

At this point, the processing of the GAM ends.

4. EXPERIMENTS

We compare the proposed GAM model with other methods on two public available datasets and show the experimental results. Finally, we analyze the experimental results qualitatively and quantitatively.

4.1 Datasets and Metrics

We use MOT17[7] and MOT20[8], which are two commonly used public datasets in the field of multiple object tracking. MOT20 has a denser target for data images. Because of the high ridership in the rail transit sector, we chose MOT20 to test more crowded scenarios.

We employ the CLEAR metric, including MOTA, MT, ML, IDs and IDF1 to evaluate different aspects of the tracking performance. As the main indicator of MOT, MOTA (Multiple Object Tracking Accuracy) pays more attention to the detection performance. IDF1 evaluates the identity preservation ability and focuses more on the association performance. It means the ratio of correctly identified detections over the average number of ground-truth and computed detections. The Most Tracked ratio (MT) for the ratio of most tracked (> 80% time) objects and the Most Lost ratio (ML) for most lost (< 20% time) objects are used to show the consistency of tracklets.

4.2 Implementation details

The basic detector of our model is YOLOX and its pre-trained model in COCO is used as the initialization weights. During the model training, we use the SGD. And we set its weight decay to 5e-4 and momentum to 0.9. We initially set the learning rate and the batch size of the model to 5e-4 and 8. The adjustable parameter K is defined to 30.

4.3 Experimental results

We test our proposed model with a Generic Association Mechanism (hereinafter referred to as GAM) on two datasets MOT17 and MOT20. In addition, we choose several advanced models (hereinafter referred to as TransTrack[9], CSTrack[2], FairMOT[1], SOTMOT[10], CorrTracker[11]) in the field of multiple object tracking as comparison, and the experimental results of these models are the results written in their respective papers. The final comparison results are shown in Table 1, where the best score for each indicator is highlighted in bold.

Table 1. Different Results of different methods on MOT17 and MOT20 datasets.

Dataset	Method	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓
MOT17	TransTrack	75.2	63.5	**55.3%**	**10.2%**	3603
	CSTrack	70.6	71.6	37.5%	18.7%	3465
	FairMOT	73.7	72.3	43.2%	17.3%	3303
	SOTMOT	71.0	71.9	42.7%	15.3%	5184
	CorrTracker	76.5	73.6	47.6%	12.7%	3369
	GAM (ours)	**79.8**	**76.6**	52.4%	13.3%	**2785**
MOT20	TransTrack	65.0	59.4	50.1%	13.4%	3608
	CSTrack	66.6	68.6	50.4%	15.5%	3196
	FairMOT	61.8	67.3	68.8%	**7.6%**	5243
	SOTMOT	68.6	71.4	64.9%	9.7%	4209
	CorrTracker	65.2	69.1	66.4%	8.9%	5183
	GAM (ours)	**74.5**	**72.8**	**69.5%**	9.8%	**2105**

4.4 Results analysis

As can be seen from Table 1, our model GAM has significantly improved in MOTA and IDF1 metrics compared with other models. The improvement of these two metrics is mainly attributed to two association adopted by the GAM model to improve the utilization rate of the low score detection boxes. It also shows that our model significantly improves the final tracking effect. Other methods will simply discard these boxes information resulting in poor tracking effect. After all, these low-score detection boxes may contain some misclassified targets, not just the background. However, our GAM model has no improvement in the metric of MT and ML compared with other models. The reason is that although we take misclassified boxes out of these low-score detection boxes, it may not match previous tracks, resulting in the generation of a new track. We are also innovating in this direction to expect better results. Figure 3 shows the tracking results of our GAM model on these datasets.

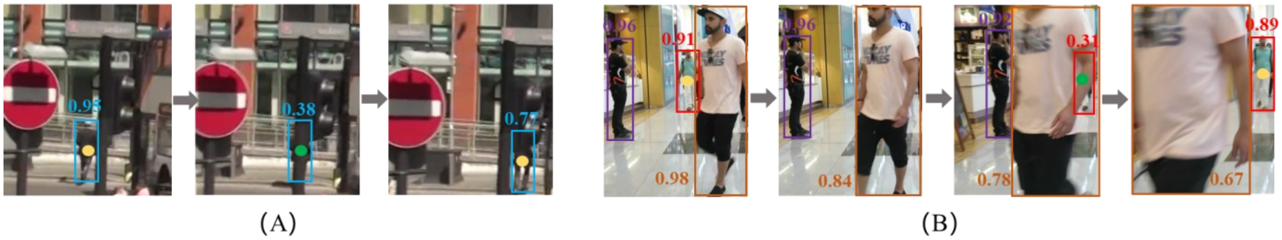

Figure 3. Qualitative results of GAM.

We select several sequences from these datasets and show the ability of GAM to handle occlusion cases. Figure 3a is the situation where people are blocked by objects, and Figure 3b is the situation where crowds block each other. Detection boxes with the same color represent the same target. The detection boxes with yellow circles and green circles represent high score boxes and low score boxes, respectively.

5. CONCLUSIONS

In this paper, a multiple object tracking model based on General Association Mechanism is proposed by studying the method of combining each detection box. This method can ensure that every detection box can fully play its role, and has been tested on two public datasets with several existing advanced methods, achieving better tracking effect. However, our model also needs to be improved: the model will slightly lose some traces in the long-distance frame, which is also a common problem in the current multiple object tracking field. We will continue to make further innovations in this field. In a word, GAM can provide a new way to improve the tracking effect of multiple-object data.

ACKNOWLEDGMENTS

This work is partially supported by National Key R&D Program of China (ID: 2020YFB1600705).

REFERENCES

[1] Zhang, Y., Wang, C., Wang, X., et al., "FairMOT: On the fairness of detection and re-identification in multiple object tracking," arXiv preprint arXiv:2004.01888, (2020).
[2] Liang, C., Zhang, Z., Lu, Y., et al., "Rethinking the competition between detection and REID in multi-object tracking," arXiv preprint arXiv:2010.12138, (2020).
[3] Wang, Z., Zheng, L., Liu, Y., et al., "Towards real-time multi-object tracking," ECCV 2020, 16, 107-22(2020).
[4] Bewley, A., Ge, Z., Ott L., et al., "Simple online and realtime tracking," IEEE in ICIP, 3464-68(2016).
[5] Wojke, N., Bewley, A. and Paulus, D., "Simple online and realtime tracking with a deep association metric," IEEE in ICIP, 3645-49(2017).
[6] Tong, X., Shuang, L., et al., "Joint detection and identification feature learning for person search," IEEE Conference on CVPR, 3415-24(2017).
[7] Milan, A., Leal-Taixe, L., Reid, I., et al., "MOT16: A benchmark for multi-object tracking," arXiv preprint arXiv:1603.00831, (2016).
[8] Dendorfer, P., Rezatofighi, H., Milan, A., et al., "MOT20: A benchmark for multi object tracking in crowded scenes," arXiv preprint arXiv:2003.09003, (2020).
[9] Sun, P., Jiang, Y., Zhang, R., et al., "Transtrack: Multiple-object tracking with transformer," arXiv preprint arXiv:2012.15460, (2020).
[10] Zheng, L., Tang, M., Chen, Y., et al., "Improving multiple object tracking with single object tracking," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2453-62(2021).
[11] Wang, Q., Zheng, Y., Pan, P., et al., "Multiple object tracking with correlation learning," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 3876-86(2021).