Multi-target Tracking Based on Deep Sort in Traffic Scene

Multi-target Tracking Based on Deep Sort in Traffic Scene

Chao Duan, Xingxing Li*

Department of electronics and information engineering, Guangzhou College of Technology and Business, FoShan, 528138, China.

Corresponding author e-mail: wslxx@jxstnu.edu.cn

Abstract. In this paper, we use the deep sort multi-target tracking algorithm to achieve multi-target tracking of pedestrians and vehicles in traffic scenes. In this paper, firstly, yolov4 is used to train the pedestrian and vehicle detection model in traffic scenes. Then, according to the detection frame predicted by yolov4, multi-target tracking is carried out for specific targets. The multi-target tracking algorithm uses deep Sort, which can be combined with yolov4, can achieve less ID switching in real-time reasoning and deal with the loss of occlusion, so as to achieve more stable tracking effect.

Keywords: deep Sort, multi-target tracking algorithm, traffic scenes, yolov4.

1. Introduction

Multi object tracking, namely multiple object Tracking (MOT), the main task is to give an image sequence, find the moving objects in the image sequence, and identify the moving objects in different frames, that is to say, given a certain accurate ID. of course, these objects can be arbitrary, such as pedestrians, vehicles, various kinds of animals, etc., and the most research is pedestrian tracking, because human is a non-rigid object, and real In international application, pedestrian detection and tracking has more commercial value. Multi target tracking is regarded as a data association problem. Cross detection results are correlated in video frame sequences. In order to solve the problem of data correlation, tracker uses a variety of methods to model the moving process and the appearance characteristics of moving objects. Sort algorithm is called simple online and real time tracking, For the current multi-target tracking [1], it depends more on its detection performance, that is to say, by changing the detector, it can improve 18.9%. Although the sort algorithm only uses the common algorithms such as Kalman filter and Hungarian algorithm, the detection performance can be improved by 18.9% Algorithm) can match the 2016 SOTA algorithm, and the speed can reach 260hz, 20 times faster than the former [2].

In target tracking, we don't use any appearance features of the tracked target, but only use the location and size of the detection frame for target motion estimation and data association, and there is no algorithm for re identification. Therefore, when the target is lost, it can’t be found, and only through detection to update the ID, which does not conform to the common sense of tracking algorithm, and needs to be improved Of course, this article is mainly about the pursuit of speed, rather than too much attention and error detection robustness. In the experiment, sort uses two detection models: fast RCNN of CNN based network and ACF of traditional pedestrian detection. In addition, in order to solve the problem of motion prediction and data association, two highly efficient algorithms, Kalman filter and Hungarian algorithm, are used.

Here we describe the object model, which represents and is used to propagate the target's identity to the next frame. Our approximate displacements between frames have a linear isokinetic model independent of the motion of other objects and cameras. The model of each target is as follows:

$$X = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T$$

Where u and v represent the horizontal and vertical coordinates of the target center respectively, s and r represent the size and proportion of the target detection frame. Note that the aspect ratio should be a constant. Therefore, the last three variables represent the next frame of prediction. When the detection is associated with the target, the detected boundary box is used to update the target state, and the velocity component is optimized by Kalman method. If no detection is associated with the target, only the linear velocity model is needed.

With the development of object detection in recent years, this tracking by detection algorithm has become more and more popular in mot. Previous algorithms, such as flow network formula and probability graph model, are global optimization problems in dealing with the whole process, but they are not suitable for online scenes, and the target identification must be available at each time step. More traditional are hypothesis tracking (MHT) and joint probabilistic data correlation filter (JPDAF). These methods perform frame by frame data association. Recently, these methods have been re recognized due to the improvement of detection technology.

In this paper, yolov4 target detection algorithm is used to train a pedestrian and vehicle detection model in traffic scene, and then the trained model is used to predict the detection box [3]. The coordinate position of each detection frame is input to deep sort, and the algorithm will output different ID of each target, so as to realize the end-to-end input and output results of multi-target tracking [4].

2. Principle

Sort algorithm uses simple Kalman filter to process the correlation of frame by frame data and Hungarian algorithm to measure the association. This simple algorithm achieves good performance at high frame rate. However, because sort ignores the surface features of the object to be detected, it can only be accurate if the uncertainty of the object state estimation is low. In deep sort, a more reliable measure is used to replace the association measure, and CNN is used to train in large-scale pedestrian data set to extract features, which has increased the robustness of the network to loss and obstacles.

In the state estimation of deep sort algorithm, an 8-dimensional space is used to describe the state (u, v, r, h, x^*, y^*, r^*, h^*), which respectively represents the position, aspect ratio, height of the bounding box center and the corresponding velocity information in the image coordinates. Then, a Kalman filter is used to predict the update trajectory. The Kalman filter adopts uniform model and linear observation model, and the observation variables are (u, v, r, h).

In the deep sort algorithm, each tracking target has its own moving track. For the track processing, it is mainly about when the track ends and when to generate a new track. Firstly, there is a threshold a for each track to record the time from the last successful match to the current time. When the value is greater than the preset threshold A_{max} thinks that the change track is terminated, intuitively speaking, the track that fails to match for a long time is considered to have ended. Then, in the process of matching, it is considered that new trajectories may be generated for the detection without successful matching. However, because these detection results may be some false warnings, the newly generated trajectory is labeled with the state "tenative", and then observe whether the continuous matching is successful in the following consecutive frames (three frames in the paper). If yes, it is considered that the new track is generated and labeled as "confirmed". Otherwise, it is considered as a false track, and the state is marked as "deleted".

2.1. Assignment Problem

In sort, we directly use Hungarian algorithm to solve the correlation between the predicted Kalman state and the new state. Now we need to combine the target motion and surface feature information, and fuse

these two similar measurement indicators. Deep sort uses Mahalanobis distance to evaluate the predicted Kalman state and the new state, as shown in the following formula:

$$d^{(1)}(i,j) = (d_i - y_i)^T S_i^{-1} (d_i - y_i)^T \tag{1}$$

Represents the motion matching degree between the j_{th} target and the i_{th} track, where S_i is the covariance matrix of the trajectory predicted by Kalman filter in the observation space at the current time, y_i is the predicted observation of the trajectory at the current time, d_j is the state of the j_{th} target (u, v, r, h).

Considering the continuity of motion, the target can be screened by the Mahalanobis distance. In this paper, 0.95 quantile of chi square distribution is used as the threshold $t^{(1)} = 0.4877$:

$$b_{i,j}^{(1)} = \sigma(d^{(1)}(i,j) \le t^{(1)}) \tag{2}$$

2.2. Appearance Metric

When the uncertainty of target motion is low, Mahalanobis distance is a good correlation measure, but in practice, for example, when the camera is moving, a large number of Mahalanobis distance can not be matched, which will make this metric invalid. Therefore, we integrate the second metric, for each BBox detection frame d_j. We calculate a surface feature descriptor r_j, $|r_j| = 1$, we will create a gallery to store the latest LK = the descriptors of 100 trajectories, i.e. $R_k = \{r_k^{(i)}\}_{k=1}^{L_k}$. Then we use the minimum cosine distance of the i_{th} and j_{th} trajectories as the second measure.

$$d^2(i,j) = \min(1 - r_j^T r_k^{(i)} \left| r_k^{(i)} \in R_i \right.) \tag{3}$$

Of course, we can also use a threshold function to express it:

$$b_{i,j}^{(2)} = \sigma(d^{(2)}(i,j) \le t^{(2)}) \tag{4}$$

We can combine these two scales as follows:

$$\begin{gathered} c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j) \\ b_{i,j} = \prod_{m=1}^{2} b_{i,j}^{(m)} \end{gathered} \tag{5}$$

Distance measure is good for short-term prediction and matching, while apparent information is more effective for long-term lost trajectory. The selection of super parameters depends on the specific data set. For example, for the data set with large camera motion amplitude, the motion matching degree is not considered directly.

Deep Sort also proposes a cascade matching strategy to improve the matching accuracy. The main reason is that when a target is occluded for a long time, the uncertainty of Kalman filter will be greatly increased, and the probability dispersion of continuous prediction will be caused. Assuming that the covariance matrix is a normal distribution, the variance of the normal distribution will become larger and larger if the continuous prediction is not updated The points far away from the mean Euclidean distance may obtain the same Mahalanobis distance as the points closer to the mean Euclidean distance. In the final stage, deep sort uses the IOU Association in the previous sort algorithm to match the unconfirmed and unmatched tracks with n = 1. This can alleviate large changes caused by apparent mutations or partial occlusion. Of course, there are advantages and disadvantages, which may lead to some new tracks being connected to some old tracks.

2.3. Deep Appearance Descriptor

In this paper, a deep convolution neural network is established to extract the feature information of the target, and the feature is projected to a unified hypersphere using L_2 normalization.

Table 1. Structure diagram of deep neural network.

Name	Patch Size/Stride	Output Size
Conv1	$3 \times 3/1$	$32 \times 128 \times 64$
Conv2	$3 \times 3/1$	$32 \times 128 \times 64$
Max Pool 3	$3 \times 3/2$	$32 \times 64 \times 32$
Residual 4	$3 \times 3/1$	$32 \times 64 \times 32$
Residual 5	$3 \times 3/1$	$32 \times 64 \times 32$
Residual 6	$3 \times 3/2$	$64 \times 32 \times 16$
Residual 7	$3 \times 3/1$	$64 \times 32 \times 16$
Residual 8	$3 \times 3/2$	$128 \times 16 \times 8$
Residual 9	$3 \times 3/1$	$128 \times 16 \times 8$
Dense 10		128
Batch and l_2 normalization		128

Deep The input of sort multi-target tracking algorithm is multiple target frames. Firstly, the multiple target frames are encoded with DNN to produce feature descriptors. The target of the next frame is matched with the feature descriptor to output the result of the tracking box. The results include the coordinate information of the tracking box, but the original information of the target frame is removed, including the confidence and target ID information When the proportion of IOU is greater than a certain threshold, the confidence level and target ID information of the original target payment are retained. If multiple IOU is greater than the threshold value, the information of the target box with the largest proportion of IOU is retained.

3. Experiments

In this paper, multi-target tracking is mainly composed of two parts. Firstly, yolov4 target detection is used to determine the location of the target, and then the deep sort multi-target tracking algorithm is used to lock the tracking target on the basis of the detection frame, and each target is given a different ID. finally, the final result is obtained through the fusion of the target frame and the tracking box IOU. First, the effect of yolov4 target detection is shown in Figure 1.

Figure 1. Results of Target detection.

In Figure 1, three types of targets are detected in traffic scenes using yolov4. One is pedestrian, one is vehicle, and the other is cyclist. Most of the targets are detected and the detection frame of target is determined. Then use deep sort to input the target frame, and fuse IOU and detection frame to get the final tracking box, as shown in Figure 2.

Figure 2 Multi target tracking results of deep sort.

Figure 2 shows the results of deep sort multi-target tracking. Each detected object is given a different ID. moreover, most vehicles and pedestrians rarely switch their tracking ID during the moving process. The tracking ID is always moving with fixed targets, and the algorithm can achieve a real-time speed of 35 frames per second when the number of targets is less than 30.

4. Conclusion

In this paper, the deep sort multi-target tracking algorithm is applied to the traffic scene, and yolov4 is used to detect three kinds of targets: travelers, vehicles and cyclists. Then, deep sort is used to input the target frame, and the tracking box containing confidence and target ID information is obtained by fusion of IOU and detection frame. Each detected object is given a different ID by using deep sort, and most vehicles and pedestrians have little switching of tracking ID in the moving process, and the algorithm can achieve real-time processing speed.

Acknowledgments

This work was financially supported by fund project, that is, Young Talents in Higher Education of Guangdong, China, (No. 2019KQNCX232 and No. 2019KQNCX231).

References

[1] Maher A , Taha H , Zhang B . Realtime multi-aircraft tracking in aerial scene with deep orientation network[J]. Journal of Real-Time Image Processing, 2018, 15(3):495-507.
[2] B G C A , Francisco Luque Sánchez b, B S T , et al. Deep learning in video multi-object tracking: A survey - ScienceDirect[J]. Neurocomputing, 2020, 381:61-88.
[3] Bochkovskiy A , Wang C Y , Liao H Y M . YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.
[4] Wojke N , Bewley A , Paulus D . Simple Online and Realtime Tracking with a Deep Association Metric[J]. 2017:3645-3649.