A Real-time Multi-target tracking method based on Deep Learning

Sitong Sun1*, Yu Wang1 and Yan Piao1
1School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
*Corresponding author’s e-mail: 2019100551@mails.cust.edu.cn

Abstract. In view of complex model and poor real-time performance of current multi-target tracking algorithms, a real-time online multi-target tracking method based on deep learning is proposed. Firstly, the detector is used to detect the target in video image space and obtain its detection frame. After that, the position, coordinates and motion of the next frame of the target are predicted by the Kalman filter. Then Complete Intersection over Union (CIoU) is used as the distance measure to calculate the overlap between the detection box and the prediction box, and further generate the similarity matrix. Finally, the Hungarian algorithm finds optimal matching in the bipartite graph formed by the two boxes, so as to realize the data association between multiple targets. Experimental evaluation shows that this method finishes real-time and online multi-target tracking well.

1. Introduction
Multi-target tracking is a significant and basic problem in the field of computer vision. It is applied to digital management, national defense and security, video analysis and autopilot. Multi-target tracking is to discover interesting objects and get their trajectories in an image space, then keep monitoring them all the time. Commonly used multi-target tracking methods can be split into offline mode and online mode. Batch tracking is offline tracking, it can also make use of the information from subsequent frames. So it can get several frames at a time and then obtain the global optimal solution in these frames. But it is not suitable for real-time computer vision tasks. At the same time, with the enhancement of deep learning network performance, target detection methods have also made great progress. Therefore, most of the current research is based on the detection of the tracking framework.

Detection-based tracking framework commonly used in multi-target tracking can achieve good results in most application scenarios. There are some global optimization batch algorithms for video such as multi-hypothesis tracking (MHT) and joint probabilistic data association (JPDA) [1], which perform data association based on frame. However, when there is high uncertainty of object allocation, the quantity of computation of these methods will increase exponentially with the number of tracking objects, and encounter difficulties in the processability and realizability of the calculation. Later Kim et al. and Rezatofighi et al. have improved MHT and JPDA methods respectively. Although the tracking effect has improved, there is still decision delay and it is not suitable for online tracking. Therefore, this paper uses a tracking-by-detection method, that first gets the detection frame through the detector in the video space and Kalman filter to predict and Hungarian method to correlate the data frame by frame. Affected by the application of YOLOv4 detection algorithm using CIoU [2] as the loss function, this paper chooses CIoU with higher robustness as the correlation metric and evaluates it on the MOT benchmark test.
2. Relevant work

The multi-target tracking method under the detection-based tracking framework divides it into two main tasks that do not disturb each other: target detection and data association. Firstly, the target detector trained offline is used to detect each frame in the video. Then the problem of multi-target tracking is regarded as a data association problem based on detection results, and the detection results from different video frames belonging to the same target are correlated to form the target trajectory. For the data association algorithm, it can be divided into batch frame association and frame-by-frame association. The target tracking algorithm of batch frame correlation is represented by network flow algorithm and graph partitioning algorithm. The multi-target tracking algorithms for frame by frame correlation include MHT, JPDA and Nearest Neighbor Data Association (NNDA).

Network flow algorithm makes the data association problem as a minimization problem by using network representation and easy to calculate objective functions. Zamir et al. proposed a method named Generalized Minimum Clique Graphs (GMCP) [3]. Firstly dividing the input video into several fragments and using the global method of GMCP to generate tracks to form the track segment, and then the tracks in all the clips are combined to get the complete track of the target. The NNDA takes the correlation door as the subspace of the image search space, and selects the target closest to the center of the correlation door as the match in the correlation door, and all other targets are not considered, and they are regarded as the match of other tracking targets. The Joint Probability Data Association algorithm (JPDA) does not need any prior information about the tracking target. It makes permutation and combination assumptions of all detection results within the correlation gate, and then calculates their associated probabilities with each target to construct the joint probability value.

In the case of frame by frame input, Hungarian matching, greedy matching and recursive neural networks are commonly used models for sequence prediction, and the correlation measurement function used includes space-time correlation and other methods. Visualization methods for spatiotemporal correlations have been widely studied. The basic methods proposed include detecting interdependencies using the Intersection over Union (IoU) or adding velocity models by using the Kalman filter. Lots of online tracking methods build a model of the target's appearance, movement, etc, or to build a global model through online learning to help detect the correlation between the track and the trace. When only one-to-one modeling is considered for matching, the global optimal solution can be used, such as the Hungarian algorithm adopted by Geiger et al. The algorithm first forms the trajectory by associating the detection on the adjacent frames, in which the correlation matrix is formed by the combination of geometric and appearance cues, and then uses geometric and appearance cues to associate the tracks again. Inspired by this two-step association approach, the association model is simplified here into a single phase.

3. Target tracking method

A single hypothesis tracking framework composed of Kalman filter and Hungarian algorithm is used for target trajectory prediction and preliminary matching in our paper. The measurement scale of complete intersection over union and merge ratio is used to match the unconfirmed target trajectory which has been initially screened, filter out the detection box that does not meet the conditions, and then make the global optimal selection.

3.1. Motion estimation

Kalman filter as a typical linear system equation of state, is an algorithm that can estimate the system state optimally by means of observation results. In fact, estimating the target position by Kalman filter through two processes: prediction and correction. In the prediction stage, Kalman filter predicts the position of targets in current frame according to the position of targets in previous frame. In the update stage, Kalman filter uses the detection result information of the current frame to correct the prediction stage to get predicted position, and then obtains a new estimate that is closer to the real position. Both of these processes are expressed by corresponding mathematical formulas.
(1) Predicting part
State prediction variance:

\[ X_k^- = \phi_{k-1,k} X_{k-1}^+ + w_{k-1} \]  

(1)

Error covariance prediction equation:

\[ P_k^- = \phi_{k-1,k} P_{k-1} \phi_{k-1,k}^T + Q_k \]  

(2)

(2) Update the part
Gain equation:

\[ K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \]  

(3)

State modification equation:

\[ X_k^+ = \phi_{k,k-1} X_k^- + K_k (Z_k - H_k \phi_{k,k-1} X_k^-) \]  

(4)

Error covariance modification equation:

\[ P_k^+ = (1 - K_k H_k) P_k^- \]  

(5)

The prediction part is responsible for using the last state \( X_{k-1} \) and error covariance \( P_{k-1} \) to predict the current time state \( X_k^- \) and error covariance \( P_k^- \), then get a priori estimation. The correction part is responsible for correcting the priori estimation \( K_k \).

We use a linear constant velocity model to transfer the identity information of targets to the motion and representation of the next frame. Let the state of each target model be as follows:

\[ \hat{X} = (u, v, h, \hat{u}, \hat{v}, \hat{h})^T \]  

(6)

Where \( u, v, R \) and \( H \) represent the central coordinates, aspect ratio and height of the outer rectangle of the target respectively. \( Z \) is the observation state of the next frame predicted by Kalman filter.

3.2. Data association
The similarity matrix is constructed by using ClIoU as a metric to express the size of overlap between the prediction box and the detection box. In SORT algorithm, IoU is used as a measure of the distance between the two boxes. However, IoU is only a simple comparison of the detection box area and the predicted box area. This method can not well express the degree of adjacency and disinter section between the two boxes, as shown in figure 1 below. The detection box is represented by a blue box, the prediction box is represented by a green box, and the red box is the smallest circumscribed rectangle of the two boxes.

![Figure 1. Schematic diagram of overlap between detection box and prediction box](image-url)

When the detection box and the prediction box do not intersect, that is IoU=0, which can not reflect the distance between the two boxes. The loss function is not derivable and can not optimize the situation.
where two boxes do not intersect. When the two bounding boxes have both same size and IoU value, then IoU can not distinguish the difference between the two intersecting situations. As shown in Figures (a), (b) and (c) respectively.

In order to improve the problem of non-coincidence of IoU boundary boxes, Rezatofighi et al. increased the measure of intersection scale and proposed Generalized Intersection over Union (GIoU) [4], as shown in figure (d). However, when the size of the detection box is same and the tracking box is located inside it, no matter where it is, which is always equal to the difference set of the detection box. Therefore, the GIoU values of the three states in Figure 1 are equal, and when GIoU is reduced to IoU, it is impossible to distinguish the relative position relationship.

However, when the prediction box is inside the detection box and the size of the detection box is same, the difference between the two boxes is same. So the GIoU values of the three states are also same, and the GIoU is reduced to IoU, which can not distinguish the relative position relationship, as shown in figure (e), figure (f) and figure (g).

To try to relieve and solve to a certain extent problem of minimizing the normalized distance between the detection box and the prediction box and the more accurate regression when the detection box and the prediction box overlap, this paper uses CIoU as the distance measure between the two frames. In previous studies, CIoU was used as a regression loss function in advanced target detection algorithms and improved the performance of the detector to some extent. As a distance measure, CIoU considers the distance information of center point and scale information of the aspect ratio of the bounding box on the basis of IoU and GIoU. CIoU is designed as a distance metric as follows:

\[
    \text{CIoU} = \frac{A \cap B}{A \cup B} - \frac{D_{\text{d2}}} {D_{\text{d1}}} - \frac{D_{\text{d2}} - D_{\text{d1}}}{D_{\text{d1}}} - \frac{v}{1 - \text{IoU}} + \frac{v}{v + 1}
\]

\[
    v = \frac{\frac{4}{\pi^2} (\arctan \frac{w_{\text{dec}}}{h_{\text{dec}}} - \arctan \frac{w_{\text{tra}}}{h_{\text{tra}}})^2}{D_{\text{d2}} - D_{\text{d1}}}
\]

Among them, A and B are respectively expressed as the distance between the detection frame and the tracking box, \(D_{\text{d1}}\) is the distance between the two central points of two boxes, \(D_{\text{d2}}\) is the diagonal distance of minimum outer rectangle of the detection frame and the tracking box, and \(v\) is a parameter to measure the consistency of aspect ratio.

The Hungarian algorithm deals with the optimal allocation problem by solving the similarity matrix. In addition, when the target overlap is less than the threshold \(\delta_{\text{thr}}\), the allocation is rejected directly.

4. Experimental results and analysis

To try to fully demonstrate the practicability of our method, experiments are carried out on 2DMOT2015, MOT16 and MOT17 data sets. Several challenging pedestrian tracking sequences are selected such as frequent occlusion, crowded scenes and sequences located at different viewing angles. In order to quantitatively evaluate multi-target tracking algorithms, several metrics are usually used to reflect the performance of different aspects of multi-target tracking algorithms. Some classical evaluation indicators are number of false detections (FP), number of missed detections (FN), number of times an ID switches to a different previously tracked object (IDs) [5], multi-object tracking accuracy (MOTA) and multi-object tracking precision (MOTP) [6].

In 2DMOT2015, the TUD-Stadtmitte dataset with low viewing angle and serious occlusion, the TUD-Campus dataset looking sideways at pedestrians and the PETS09-S2L1 dataset with frequent target occlusion in high-speed nonlinear mode are selected. The experimental results show that the effect is the best when the value of \(\delta_{\text{thr}}\) and IoU are both setting to 0.3. The experimental data conclusions are shown in Table 1.
Table 1. Tracking results based on MOT15 dataset sequence.

| Tracker | Sequence     | MOTA↑ | MOTP↑ | FP↓ | FN↓ | IDs↓ |
|---------|--------------|-------|-------|-----|-----|-----|
| Sort    | TUD-Stadtmitte | 55.2  | 65.6  | 157 | 332 | 27  |
|         | TUD-Campus    | 45.4  | 70.6  | 32  | 134 | 10  |
|         | PETS09-S2L1   | 60.3  | 71.4  | 960 | 677 | 139 |
|         | TUD-Stadtmitte | 60.1  | 65.7  | 96  | 336 | 29  |
| Ours    | TUD-Campus    | 49    | 72.4  | 20  | 153 | 14  |
|         | PETS09-S2L1   | 67.2  | 71.7  | 570 | 740 | 156 |

MOT16 is extended on the basis of 2DMOT2015, deleting some relatively simple video sequences and adding some new video sequences. The video environment is more complex, and the number of pedestrians, occlusion and camera movement are more comprehensive, in which the detector used by MOT16 is DPM. The evaluation measures used in this article are the average of the sequence measures contained in MOT16. When the value of δ is 0.3 and the value of IoU is 0.2, the tracking effect of both is the most obvious. The experimental data conclusions are shown in Table 2.

Table 2. Tracking results based on MOT16 dataset sequence.

| Tracker | MOTA↑ | MOTP↑ | FP↓ | FN↓ | IDs↓ |
|---------|-------|-------|-----|-----|-----|
| Sort    | 25.3  | 77.8  | 14198 | 67446 | 1091 |
| Ours    | 26.3  | 77.9  | 12643 | 67882 | 1065 |

MOT17 is extended on the basis of MOT16, in order to explore the performance of multi-target tracking algorithm in the case of inputting different quality detection results, the organizer extends the single detection result provided by MOT16 benchmark to three different groups of detection results, using DPM, Faster-RCNN and SDP target detectors respectively, which is more accurate than the detection level in MOT16. For the SDP detector, the tracking effect of the two detectors is the most obvious when the value of δ is 0.1 and the value of IoU is 0.3. For Faster-RCNN detector, the effect is best when δ and IoU are both 0.2. The experimental data conclusions are shown in Table 3.

Table 3. Tracking results based on MOT17 dataset sequence.

| Tracker | Detection | MOTA↑ | MOTP↑ | FP↓ | FN↓ | Recall↑ |
|---------|-----------|-------|-------|-----|-----|---------|
| Sort    | SDP       | 85.1  | 86.8  | 1041 | 9177 | 86.2    |
|         | FRCNN     | 82    | 88.3  | 1437 | 10382 | 84.1    |
| Ours    | SDP       | 85.2  | 86.8  | 1211 | 9018 | 86.3    |
|         | FRCNN     | 82.1  | 88.3  | 1725 | 10162 | 84.3    |

From the above experimental results, we can see that whether the environment is too complex or not, the effect of the detector is generally good, and the tracking effect has been improved. Among them, when using the detector with high precision, the detected MOTA value is more than 80%, which has a good tracking result. As shown in figure 2.

Figure 2. Part of the results of our algorithm are shown on MOT16 train dataset

5. Conclusion

Multi-target tracking from the perspective of real-time performance of various indicators, this paper uses complete intersection and merging as a distance measure to improve the accuracy of trajectory matching, and uses the improvement of the quality of the nearest detection algorithm to predict the results of detected targets by Kalman filter. Then the similarity matrix of the detection frame and the tracking box is obtained. Finally, the matching between the detection box and the tracking box is realized by the Hungarian algorithm, which does not meet the direct refusal allocation of the threshold...
requirement. Combined with the experimental results, we can see that CIoU, as a distance measure, has a good effect on the tracking problem.

References
[1] Bewley A, GE Zongyuan, Ott L, et al. Simple online and realtime tracking [C]// 2016 IEEE International Conference on Image Processing (ICIP), September 25-28, 2016, Phoenix, AZ, USA. IEEE, 2016:3464-3468.
[2] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU Loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020. 3, 4.
[3] Zamir A R, Dehghan A, Shah M. GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs[M]. Springer-Verlag, 2012: 343-356.
[4] Rezatofighi H, Tsoi N, Gwak JY, et al. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.
[5] Yuan L, Chang H, Nevatia R. Learning to associate: HybridBoosted multi-target tracker for crowded scene[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2009.
[6] Keni B, Rainer S. Evaluating multiple object tracking performance: The CLEAR MOT metrics[J]. EURASIP Journal on Image and Video Processing, 2008, 2008(1).