International Core Journal of Engineering Volume 6 Issue 11, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0023

Research on Object Tracking and Target Recognition Based on Kalman Filter and YOLOV3

Xueming Zhai, Jilei Jia

School of computer, North China Electric Power University, Baoding, 071000, China.

Abstract

In order to solve the existing problems in moving target recognition and tracking, YOLOV3 is used to detect the target to be tracked in the current frame, and Kalman filter is used to predict the next position and the size of the bounding box according to the position of the current target. The improved Hungarian algorithm is used to correlate and match the data according to the intersection ratio and color histogram of the detected and predicted borders, and the target motion trajectory is obtained through continuous iteration of the system to complete the tracking. For the occluded target, a region based quality assessment network is introduced, which combines multiple high-quality detection images to recover the occluded part and improve the tracking accuracy.

Keywords

YOLOV3, Kalman filter, Hungarian algorithm.

1. Introduction

Image and video processing is an important category of computer application, which has become a research hotspot in military, public security, medical, detection and other fields, and shows a broader application prospect. Target tracking is generally divided into single target tracking and multi-target tracking. For single target tracking, there is a priori assumption. Therefore, even if the range is only framed in the initial position, a good tracking result can be obtained. However, the multi-target tracking, which is usually applied to pedestrian monitoring, is often a multivariable estimation problem. Therefore, in addition to object deformation and background interference, multi-target tracking also needs to solve the following problems: (1) automatic initialization and automatic termination of targets; (2) motion detection and similarity discrimination; (3) interaction and occlusion between targets; (4) recognition of lost targets.

In view of one or more of the above problems, researchers propose corresponding solutions, which can be divided into two categories: detection based data association algorithm and network minimum cost flow algorithm. In the former, multi-target tracking is regarded as a data association problem, and the track and detection between two consecutive frames are connected to form a longer trajectory. The multi-layer tracking framework proposed by Huang et al. [1] is a typical method of this kind of algorithm. It first forms a short track according to the detection between two adjacent frames, then carries out global correlation, and finally makes fine adjustment to the generated trajectory. The algorithm based on network minimum cost flow proposed by Milan et al. [2] transforms it into an energy minimization problem. Each detection is regarded as a node, and each node has corresponding energy. The purpose of the algorithm is to find the optimal solution of the energy function and form a tracking trajectory. With the rapid development of deep learning, many tracking algorithms are based on the above two categories of methods and add deep learning algorithm. The use of depth network can extract more robust image features, which makes the follow-up tracking method more accurate, and further improve the tracking accuracy. Wojke et al. [3] used the depth network to extract the detection and boundary box, and predicted the trajectory through the motion matching degree and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0023

appearance matching degree, and finally introduced the cascade matching method to track the long track; Chen et al. [4] used Kalman filter and improved target detection algorithm to separate the detection and classification, and then combined them to deal with the occlusion problem. Since most of the current research is based on detection tracking, the quality of the target detection algorithm can have an important impact on the tracking results, and how to coordinate the processing of data association and object occlusion also needs further planning.

2. Target Recognition Model

As a target detection network, yolov3 uses its multi-scale prediction mechanism to detect pedestrian targets with different sizes. Then, Kalman filter is used to predict the next position of the target according to the current tracking results, and calculate the intersection and merging ratio and color histogram of detection range and prediction range. The best matching is obtained by Hungarian algorithm according to the score, and the tracking track is obtained through continuous iteration.

2.1 Target Recognition Network Model

Yolov3 is improved on the basis of previous network, and target features are extracted by constructing deep residual network; Then, the anchor mechanism in the regional recommendation network is adopted, and the relative coordinate prediction is added to solve the problem of unstable model training and speed up the detection; at the same time, the feature pyramid structure (FPN) is introduced to enable the network to carry out multi-scale prediction and avoid missing detection of small objects. The structure diagram is shown in Figure1.

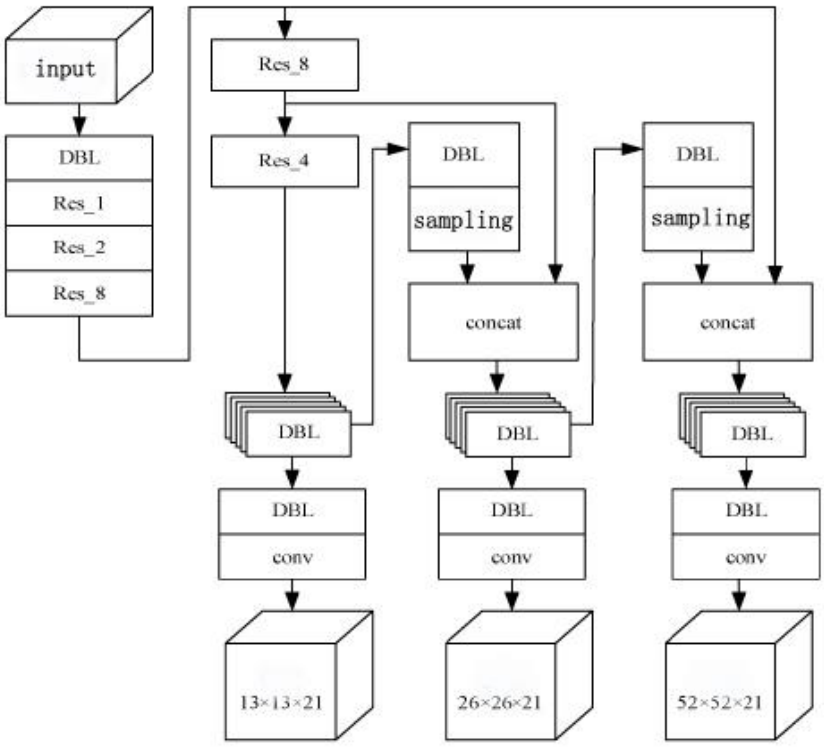

Figure 1. Overall network structure of yolov3

Firstly, the input image size is adjusted by DBL structure, and then the image features are extracted by multiple residual groups to obtain the feature map with multiple resolutions. Then, the feature map is unsampled and spliced with the original feature map. Then, the prediction results of 13 × 13, 26 × 26 and 52 × 52 scales are obtained by using the feature pyramid structure. In the image, the network can detect not only normal size targets, but also small objects, which is of great significance for multi-target tracking in dense scenes.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0023

2.2 Anchor Mechanism Model

The anchor mechanism was first proposed by Faster RCNN to select the bounding box of an object. Fast RCNN takes a combination of 3 scales and 3 aspect ratios in each sliding position, and generates 9 anchor points to select the bounding box. In view of the error caused by manual acquisition of anchor size, yolov3 adopts the improved k-means clustering algorithm and uses IOU score as the evaluation standard to select the most appropriate prior box. The details are as follows:

$$d_{centroid}^{box}=1\text{-}IOU_{centroid}^{box} \tag{1}$$

$d_{centroid}^{box}$ represents the distance from the prediction box to the cluster center, $IOU_{centroid}^{box}$ indicates the predicted box and the actual IOU score.

As shown in Figure2, CX and CY denote the number of grids from the first grid in the upper left corner where the central coordinate of the prediction box is located; bx, by, bw and bh represent the absolute position of the prediction box; tx, ty, tw and th represent the relative position of the prediction box; PW and pH represent the width and height of the prior box; σ (·) represents the Sig-moid function. The calculation method is as follows:

$$\begin{cases} b_x=\sigma(t_x)+c_x & b_y=\sigma(t_y)+c_y \\ b_x=p_w e_x^t & b_h=p_h e_y^t \end{cases} \tag{2}$$

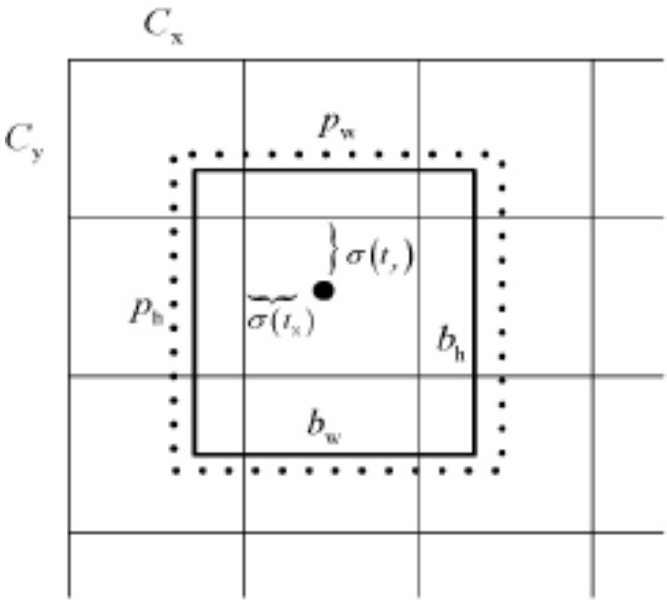

Figure 2. Relative coordinate prediction

3. Target Tracking Algorithm

After obtaining multiple detection targets in each frame by using yolov3, it is necessary to track the same target in consecutive frames and generate trajectories in turn. In this paper, the Kalman filter model is used to predict the position of the target in the next frame, and then the Hungarian algorithm [18] is used for data association. However, considering that if only the intersection ratio of the area is used as the matching basis when the positions between the targets are close, the original Hungarian algorithm is improved in this paper, and the color histogram is introduced for different similar targets Distinguish to make the result more accurate.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0023

3.1 Kalman Filter Algorithm

Kalman filtering algorithm (KF) is a process of obtaining the optimal estimation or optimal solution of the state of the system by using linear system equations by inputting and outputting some observation data and considering the influence of noise on the system.

This method can be regarded as a search algorithm prediction model, which can effectively predict the position of linear moving targets. When the algorithm is used to predict the next time, the state information of the target at the previous time should be considered. After the prediction, the model should be modified according to the actual output detected. As shown in Figure3.

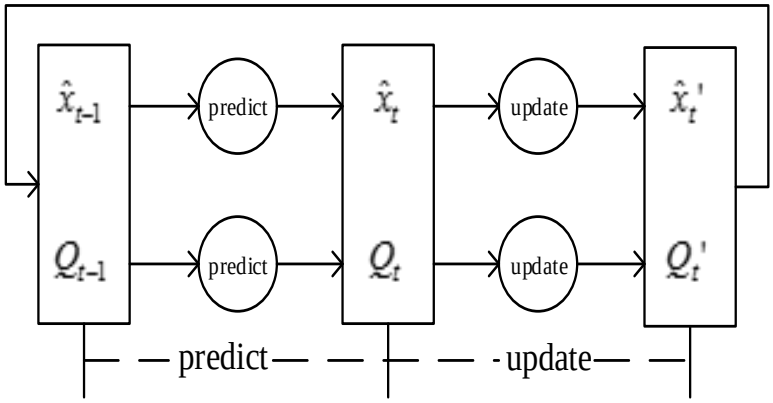

Figure 3. Schematic Diagram of Kalman Filter

In order to make the Kalman filter model work continuously, some parameters in the model need to be updated to ensure the real-time and accuracy of tracking. In the prediction stage, according to the state of the previous time, the target transfer matrix F and control matrix B are used to estimate the state of the next moment. The prediction formula of the state is shown in equation (3), where u represents the control quantity.

$$\hat{x}_t = F_t \hat{x}_{t-1} + B_t u_t \tag{3}$$

Suppose that Q is the covariance matrix of noise, then the matrix $\sum$ represents the transfer relationship between the uncertainties at each time. The specific calculation process is shown in formula (4).

$$\sum_t = F \sum_{t-1} F^T + Q \tag{4}$$

In the update stage of the target, the Kalman coefficient K is calculated by the state transition matrix $\sum$ and the observation matrix H, and the target state at the next moment is modified by the Kalman coefficient and the detected target state. The calculation of Kalman coefficient K is shown in formula (5), and the correction of target state is shown in formula (6).

$$K_t = \sum_t H^T (H \sum_t H^T + R)^{-1} \tag{5}$$

$$\hat{x}_t' = \hat{x}_t + K_t (y_t - H\hat{x}_t^-) \tag{6}$$

The above is the whole process of using Kalman filter algorithm to estimate the position of moving target. This method can reduce the search range in the target search strategy, so as to improve the matching efficiency.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0023

3.2 Improved Hungarian Algorithm

For data association in multi-target tracking, yolov3 first detects multiple targets m in the current frame, their coordinates and boundary box range, and Kalman filter estimates the target position in the current frame according to the tracking results of the previous frame, and obtains n prediction results, or N tracks. After all the detection and prediction results are obtained, the cross union ratio of the two areas is calculated first, and then the color histogram of the detected image is obtained. Then, the correlation matrix is generated by weighting the cross union ratio and histogram features, as shown in the following formula:

$$C= \alpha C(i,j)+ \beta(i,j) \quad (7)$$

Where: C is the intersection ratio of the two, which can be obtained by the pasteurization distance, and α and β are the weight coefficients, and the sum is 1. Finally, Hungarian algorithm is used to match the detection results with the prediction results to complete the data association and form the tracking track in multi frame images.

4. Experiment and Analysis

4.1 Experimental data

The data set is divided into training set and test set, including multiple pedestrian targets, and there are interaction and occlusion between targets. There are 11 videos in training set and test set respectively. Experimental environment: memory is 4GB, software is Python 3.6, GPU is RTX2080ti.

The evaluation criteria used in the experiment are tracking accuracy (TA), tracking accuracy (TP), the proportion of hit orbit hypothesis to the actual total orbit (MT), the proportion of missing target orbit to the actual total orbit (ML), the total number of tag switching (IDS) and the total number of false positives (FP).

4.2 Experimental Results and Analysis

In order to better analyze the performance of the algorithm, the two groups of experimental results are compared and analyzed. By testing the algorithm in different video sequences to analyze the experimental results of this algorithm in different application scenarios; at the same time, this algorithm is compared with other algorithms, so as to make further analysis and Research on the advantages and disadvantages.

The comparison of the algorithm in different sequences. In this paper, experiments are carried out on all sequences of the test set, and the data obtained are shown in Table 1.

Table1. Test results on different sequences

Sequence	TA/%	TP/%	MT/%	ML/%	IDS	FP
Venice-1	22.8	72.4	2.3	40.8	5	1364
KITTI-19	27.1	66.7	6.7	28.7	63	2358
KITTI-16	37.4	72.7	1.8	17.8	19	641
ADL-Rundle-3	39.5	72.6	11.6	33.9	28	1251
ADL-Rundle-1	26.9	71.8	28.5	28.4	30	3104
AVG-TownCente	20.3	69.2	4.9	44.5	136	1912
ETH-Crossing	27.8	74.9	4.2	65.1	6	93
ETH-Linthescher	15.4	72.2	4.8	78.3	11	1315
ETH-Jelmoli	42.7	73.1	24.7	31.2	25	1214
PETS09-S2L2	35.3	69.5	7.3	19.3	277	652
TUD-Crossing	73.3	73	69.5	15.2	173	42

According to the analysis of Table 1, from the average performance, the algorithm in this paper performs best on the sequence of TUD crossing, and the worst on Venice-1. Part of the reason is that

ISSN: 2414-1895 DOI: 10.6919/ICJE.202011_6(11).0023

the color contrast between pedestrian and background in Venice-1 is not obvious. The results show that the image quality has a great influence on the tracking results. For other video sequences, such as the tracking results will drift and the tracking accuracy will be reduced; or the image blur, such as kitt-16, will also affect the detection and tracking. In pets09-s2l2, the total number of label switching is the largest, because the number of targets is large and the color of pedestrian clothing is similar, so there is error in distinguishing different targets. In general, in addition to improving the tracking performance of the algorithm itself, we should also use fixed camera to shoot, or make motion compensation for mobile camera lens, and need to preprocess the acquisition sequence.

References

[1] Kaiqi Huang,Xiaotang Chen,Yunfeng Kang:Overview of intelligent video surveillance technology (Journal of Computer Science Press, China 2015), p.1093-1118. (In Chinese)

[2] Huang Kai-Qi,Ren Wei-Qi,Tian Tie-Niu.A review on image object classification and detection.Chinese Journal of Computers, 2014,37(6):1225-1240.

[3] Girshick R,Donahue J,Darrell T,et al.Rich feature hierarchies for accurate obje- ct detection and semantic segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA,2014:58 0-587.

[4] Wang L,Xu L,Min Y K, et al.Online multiple object tracking via flow and convolutional features[C]//IEEE International Conference on Image Processing. 2018.

[5] Shen J, Liang Z, Liu J, et al. Multiobject Tracking by Submodular Optimization[J]. IEEE Transactions on Cybernetics, 2018, PP(99):1-12.

[6] Nejhum S,Ho J,and Yang M H.Online visual tracking with histograms and articulating blocks[J]. Computer Vision and Image Understanding, 2010, 114(8): 901-914.

[7] Kwon J,Lee K M.Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive Basin Hopping Monte Carlo sampling[C]// Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009:1208-1215.