The research of multi-target tracking based on improved YOLOv3

Jiawen Xu*, Xinbiao Lu and Chi Zhang
School of Energy and Electrical Engineering, Hohai University, Nanjing, China

*Corresponding author. Email: xjw0219@126.com

Abstract. Multi-target tracking based on detection is a hot topic in the field of computer vision. Aiming at solving the problem of inconvenient application of large network model, this paper proposes a multi-target tracking algorithm based on improved YOLOv3. We use MobileNetV3 to replace the deep network as the backbone of YOLOv3. Compared with Faster R-CNN and YOLOv3, the model of MobileNetV3-YOLOv3 is greatly reduced and the detection speed is improved. Compared with the MobileNetV1-YOLOv3, the accuracy is improved and FPS is higher. The improved YOLOv3 model is used in DeepSORT algorithm for multi-target tracking. The experimental results show that the algorithm used in this paper has the best detecting and tracking effect.

1. Introduction
Target tracking in surveillance security plays a very important role in the process of social development. It is the mainstream method to introduce deep learning into multi-target tracking since convolution neural network has powerful feature extraction to affect the effectiveness of the target tracking. Faster R-CNN[1] is the first algorithm to apply deep learning to the detection part of multi-target tracking, and then SSD[2] and YOLO[3] are tried in tracking, among which YOLOv3[4] algorithm is the best.

In order to get higher precision and accuracy, more complex network models have been proposed. But the deeper the network is, the more parameters and the calculation are. However, the computing and storage capacity of general embedded devices are limited, which requires a lightweight and flexible model. At present, miniaturized networks include SqueezeNet, ShuffleNet, MobileNet[5] and Huawei's GhostNet[6].

In the field of multi-target tracking, the traditional methods such as flow network generation[7] and probabilistic graphic models[8] are very popular, but they are not suitable for online scenarios. The MHT algorithm proposed by D.B.Reid[9] and JPDAF proposed by T.E.Formann[10] are also used in tracking multiple targets, but these methods are too computational. The famous algorithm is SORT[11], which has the advantages of high accuracy and real-time tracking, but it has shortcomings in tracking occlusion. On this basis, the DeepSORT[12] algorithm is proposed to extract the apparent features, which improves the tracking effect.

This paper proposes a multi-target tracking algorithm based on improved YOLOv3. We use the lightweight network as the backbone of the model, which ensures the tracking accuracy, further reduce the size of the model, improve the detection speed, and is more conducive to the deployment of mobile devices.
2. Multi-target tracking algorithm

2.1. Target detection algorithm YOLOv3

YOLOv3 is a continuous improvement on the basis of YOLOv1 and YOLOv2. The main network of YOLOv3 is Darknet-53, and the experimental effect on ImageNet is good. There are 252 layers in this model, so we consider using a lightweight network as the backbone of YOLOv3.

2.2. Improved YOLOv3

MobileNetV3[13] has fewer layers and is suitable for embedded devices, so we replace Darknet-53 with mobileNetV3 as the backbone network of YOLOv3. Taking MobileNetV3_small-YOLOv3 as an example, its model structure is shown in the Fig. 1:

Figure 1. Model structure of MobilenetV3_small-YOLOv3.

MobileNetV3 uses the depthwise separable convolutions of MobileNetV1 and the inverted residual with linear bottleneck of MobileNetV2, and introduces the lightweight attention mechanism based on the Sequeue and Exception structure[13]. In the Fig.1, MobileNetV3conv2d_BatchNormalization_NL is a convolution unit (NL is the activation function: H-swish or ReLU6 or LeakyReLU). Its main module, bottleneck (Bneck), uses MBH or MBR unit with $1 \times 1$ convolution to increase dimension, then uses $3 \times 3$ depthwise separable convolution, adds SE and introduces $\alpha$ to adjust weight to reduce channels.

After fusing the convolution kernel of the same scale, MobileNetV3-YOLOv3 outputs three kinds of feature maps. For this, nine sizes of anchor boxes in Table 1 are clustered:

| Feature map | 13*13  | 26*26  | 52*52  |
|-------------|--------|--------|--------|
| Anchor box  | (116x90) | (62x45) | (10x13) |
|             | (156x198) | (59x119) | (16x30) |
|             | (373x326) |          | (33x23) |

We predict the bounding box (bbox) by these prior bbox and the method of logistic regression.

2.3. Multi-target tracking algorithm DeepSORT

First, the object detection algorithm based on MobileNetV3_small-YOLOv3 is used as the detection of DeepSort. Then we filter bbox by confidence score and non_maximum suppression.

We use Kalman filter[14] to predict and update the target trajectory by formula(1):

$$ x' = Fx $$  \hspace{1cm} (1)  

$x = [c_x, c_y, r, h, \dot{c}_x, \dot{c}_y, \dot{r}, \dot{h}]$, $(c_x, c_y)$ represents center coordinates of the prior bbox on the feature map, $(r, h)$ respectively represents the length-width ratio and height, $(\dot{c}_x, \dot{c}_y, \dot{r}, \dot{h})$ represents the velocity information, and $F$ is the state transition matrix.

In the data association, the Mahalanobis distance[12] between the prediction bbox and the detection bbox is used to associate the motion information by formula(2):

$$ d^{ij}(i, j) = (d_j - y_j)^T S^{-1}_i (d_j - y_j) $$  \hspace{1cm} (2)
$d_j$ represents the position of the $j$-th detection bbox and $y_i$ represents the position of the $i$-th tracker, $S_i$ is the covariance matrix between them. When the Mahalanobis distance is less than the threshold ($b_{ij}^{(1)} = \|d^{(1)}(i,j) \leq r^{(1)}\}$, the matching is considered to be on.

When there is motion caused by other external factors, Mahalanobis distance will fail. So the cosine distance measure of apparent feature is introduced by formula (3):

$$d^{(2)}(i, j) = \min \{1 - r_j^T v_i | r_j^T \in R_i\}$$

$r_j$ is the feature vector of $d_j$, $R_i$ is the feature vector set of $y_i$. Similarly, We set the threshold ($b_{ij}^{(2)} = d^{(2)}(i, j) \leq t^{(2)}$). In this paper, I take it as 0.2. Multiple residual network modules are used to extract the feature vectors of bboxes.

DeepSORT makes weighted sum of Mahalanobis distance and cosine distance[12] in data association. I set $\lambda = 0$.

$$c_{ij} = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j)$$

IOU matching is performed for unconfirmed trackers, unmatched trackers and unmatched detection, and Hungarian algorithm[15] is used to assign them again. Cascade matching[12] is used to match the tracks with smaller disappearance time.

The whole flowsheet is shown in Fig.2.

Figure 2. DeepSORT based on MobileNetV3-YOLOv3.

3. Experimental results
The experimental framework of algorithm is TensorFlow, and the data set is VOC and the infrared dataset taken by me, with a total of 18952 images. The training and testing platform of the network is a workstation with NVIDIA RTX 2080ti 11GB GDDR6.
3.1. Comparative experiment of target detection algorithms
In this experiment, Faster R-CNN, YOLOv3, MobileNetV1-YOLOv3 and MobileNetV3-YOLOv3 are selected for comparison in Table 2.

Table 2. Comparison experiment of target detection algorithms.

| Algorithm            | mAP(%) | Model size(MB) | FPS(detection+tracking) |
|----------------------|--------|----------------|-------------------------|
| Faster R-CNN         | 89.5   | 158            | 10.7                    |
| YOLOv3               | 90     | 235            | 23                      |
| MobileNetV1-YOLOv3   | 86.48  | 92.5           | 53.5                    |
| MobileNetV3-YOLOv3   | 89.4   | 89.2           | 87                      |

It can be seen that under the similar accuracy, the model of our algorithm is greatly reduced, and the detection speed is greatly improved. Compared with MobileNetV1-YOLOv3, MobileNetV3-YOLOv3 has higher accuracy and FPS.

3.2. Experimental results of improved DeepSORT
The experiment is carried out with the infrared video taken by ourselves, and different frames are selected to check the effect of target detection and tracking.

![Figure 3. Experimental results of improved DeepSORT.](image)

Selecting frame 32, frame 140 and frame 172 respectively, it can be seen that the algorithm can effectively detect each target and track continuously after the target is occluded. Our algorithm can be effectively realize multi-target tracking.

4. Conclusion
In this paper, the lightweight network MobileNetV3 is used as the basic framework of YOLOv3. The model MobileNetV3-YOLOv3 as the detection part of DeepSORT to form the improved multi-target tracking algorithm, which is smaller and faster than other algorithms like Faster R-CNN and YOLOv3. Compared with MobileNetV1 as the backbone network, our algorithm has higher accuracy and faster detection speed. The experimental result shows that the algorithm can not only track multiple targets effectively, but also deal with the problem of ID switch after the target is occluded, which is more conducive to the implementation in the mobile devices.

Acknowledgement
This topic is supported by “the Fundamental Research Funds for the Central University”, project number is B200203208. I would like to express my sincere gratitude to my funder.

References
[1] S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks Proceedings of the Advances in Neural information Processing Systems (2015), pp. 91-99
[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed. C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector Proceedings of the European Conference on Computer Vision, Springer (2016), pp. 21-37
[3] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: unified, real-time object detection Proceedings of the IEEE conference on computer vision and pattern recognition (2016), pp. 779-788
[4] J. Redmon, A. Farhadi. YOLOv3: An Incremental Improvement[C]. Proceedings of Computer Vision and Pattern Recognition (CVPR). 2018:1-6

[5] Xue Juntao, Ma Ruohan, Hu chaofang. Deep learning algorithm for multi-target tracking based on mobilenet [J/OL]. Control and decision: 1-6 [2021-02-07] http://doi-org-s.vpn.hhu.edu.cn :8118/10.13195/j.kzyjc.2019.1424

[6] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu. “GhostNet: More Features From Cheap Operations,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 1577-1586, doi: 10.1109/CVPR42600.2020.00165

[7] J. Berclaz, F. Fleuret, E. Tu’retenk, and P. Fua. “Multi-ple object tracking using k-shortest paths optimization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 9, pp. 1806–1819, 2011

[8] B. Yang and R. Nevatia, “Multi-target tracking by on-line learning of non-linear motion patterns and robust appearance models,” in CVPR, 2012, pp.1918–1925

[9] D.B.Reid,“Analgorithmfortrackingmultiple targets,” IEEE Trans.Autom.Control,vol.24, no.6, pp.843–854,1979

[10] T.E.Fortmann,Y.Bar-Shalom,and M. Scheffe, “Sonar tracking of multiple targets using joint probabilistic data association,” IEEE J. Ocean. Eng.,vol.8,no.3, pp.173–184, 1983.

[11] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft. Simple online and realtime tracking. Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), IEEE (2016), pp. 3464-3468

[12] N. Wojke, A. Bewley, D. Paulus. Simple online and realtime tracking with a deep association metric.Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017), pp. 3645-3649

[13] A. Howard et al., "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140

[14] R.E. Kalma,A new approach to linear filtering and prediction problems.J. Basic Eng., 82 (1) (1960), pp. 35-45

[15] H.W. Kuhn. The hungarian method for the assignment problem.Naval Res. Logist. Q., 2 (1–2) (1955), pp. 83-97