Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230467

Real-time vehicle detection and tracking based on the combination of YOLOv7 and ByteTrack

Yining Zhang

Computer Science and Engineering, University of New South Wales, Sydney, 2052, Australia

z5324974@ad.unsw.edu.au

Abstract. Real-time tracking of vehicles is important for monitoring whether the roads are congested or not. How to achieve and maintain a high frame rate is a key problem to be solved in practical applications. In recent years, SORT, BOT-SORT [1] and other state-of-the-art target tracking algorithms have achieved fruitful results in target tracking tasks. The author combines YOLOv7(You Only Look Once v7) with bytetrack and compares the frame rate with some popular algorithms like Deepsort and bytetrack in real-time tracking to demonstrate that YOLOv7-bytetrack is more suitable for real-time detection, tracking of vehicles.

Keywords: Real-time tracking, high frame rate, YOLOv7, ByteTrack.

1. Introduction

With the rapid economic development, private cars are becoming more and more popular as a means of transportation, and the more traffic flow on the roads, the more congestion problems will increase. In recent years, with the rapid growth of national economy, urbanization is accelerating and urban population is concentrating. The number of large cities with millions or even tens of millions of people is increasing. The urbanization process has driven the rapid growth of transportation demand. At the same time, the number of motor vehicles in large cities is increasing, which also contributes to the increase of traffic demand. The rapid growth of traffic demand has increased the load on urban roads, and traffic congestion and chaos have occurred repeatedly. Among them, traffic congestion is particularly serious and has hindered the economic development of the city. The 2017 Global Traffic Score Study by Inrix [2], a U.S. traffic data analytics company, analyzed congestion in 111 cities and towns in the U.K. For the 10th consecutive year, London was the most congested major city in the U.K. Europe-wide, London ranked second in congestion after Moscow, and globally, London ranked seventh in congestion. London drivers spent an average of 74 hours stuck in rush hour traffic in 2017, an increase of one hour over 2016. In terms of direct and indirect costs, London's traffic congestion costs London's drivers £2,430 per year, a combined total of £9.5 billion. Therefore, it has become a valuable topic to detect and track the number of vehicles efficiently and in real time in the face of the huge number of cameras and the video material they generate.

Multi-target tracking, an intermediate job in computer vision, is still difficult because it must concurrently address difficulties with target recognition, trajectory prediction, data association, and re-identification. Additionally, it serves as the foundation for a number of complex tasks including

DOI: 10.54254/2755-2721/4/20230467

position estimation, action detection, and behavior analysis. This time, the author researches tracking of several targets.

Traditional pure detection approach [3]: the model trained by the target neural network (here first refers to 2d detection). The 2d detection network regresses a 2d detection frame, the output contents are: [image_id, conf, label, xmin, ymin, xmax, ymax], image_id: the frame of the image; conf: the confidence level of the detection; label : the category of the detected object; xmin,ymin: the normalized value of the pixel coordinates in the upper left corner of the detected frame; xmax,ymax: the normalized value of the pixel coordinates in the lower right corner of the detected frame; the reader can see that this information can only give the pixel information (positioning) and category information (classification) of the detected object in the image, which is not enough in practical engineering. Pedestrians passed, at this time should be timely braking, the premise of braking is the need to know the distance information between the person and the car, so detection plus distance measurement is inseparable, so that the perception is meaningful; similarly, the actual detection process, the external environment is complex (lighting, blocking, etc.), the camera will jitter and other multi-factor impact, only rely on the results of detection, detection frame is very unstable.

However, the detection + tracking mode, on the other hand: when the tracking is added to the detection, the detection frame becomes very dependable, and the output result is smoother with almost no jitter. Furthermore, it gives labels for unique items as well as separate ID information for objects of the same kind to differentiate between various things and objects of the same type, allowing for multi-target tracking. The automatic analysis and extraction of trajectory data make up for the deficiency of visual target recognition. It can effectively eliminate the false detection and improve the missed detection rate, which provides a basis for the subsequent behavior analysis.

2. Related Works

At the moment, the back-end tracking optimisation technique based on Hungarian matching and KM matching is widely used (the representative applications are SORT and deep-sort). These algorithms have the advantage of being able to accomplish real-time performance, but a good tracking effect can only be obtained by good detection algorithm effect and strong feature differentiation. The final output result is dependent on the stronger detection algorithm, and the role of the Kalman Plus Hungarian matching tracking method is to output the ID of the identified target, and the second is to assure the tracking algorithm's real-time capabilities.

SORT (Simple Online and Realtime Tracking) [4], which was proposed in 2016, outperforms other multi-target trackers of its contemporaries. Then, as its name suggests, the principle of SORT is very simple and consists of three main parts: target detection, Kalman filtering, and Hungarian algorithm, which does not involve algorithms such as image recognition, feature matching, etc. The SORT algorithm takes the target detection result as input, uses the data association algorithm (Hungarian algorithm) for target matching, and Kalman for target prediction, and establishes the relationship between targets through IOU as well. Its main purpose is to continuously track the ID of the target, and has the advantage of fast tracking speed.

Since sort algorithm is still a relatively coarse tracking algorithm, it is especially easy to lose its ID when the object is occluded, while Deepsort [5] algorithm adds cascade matching (Matching Cascade) and confirmation of new tracks (confirmed) on the basis of sort algorithm. Tracks are classified into confirmed and unconfirmed, and the newly generated Tracks are unconfirmed; unconfirmed Tracks must be matched with Detections for a certain number of consecutive times (default is 3) before they can be transformed into confirmed. Tracks in confirmed state must be mismatched with Detections for a certain number of times in a row (default is 30) before they are deleted.

YOLOv6 [6] is a target detection framework developed by Vision Intelligence Department of Meituan, dedicated to industrial applications. The framework focuses on both detection accuracy and inference efficiency, among the commonly used size models in industry: YOLOv6-tiny achieves 41.3% AP accuracy on COCO val, while inference on T4 with TRT FP16 batchsize=32 can reach 602FPS performance. YOLOv6 mainly makes many improvements in Backbone, YOLOv6 has many

DOI: 10.54254/2755-2721/4/20230467

improvements in the areas of Backbone, Neck, Head, and training strategy. The authors unified the design of more efficient Backbone and Neck: inspired by the idea of hardware-aware neural network design, they designed the reparameterizable and more efficient backbone networks EfficientRep Backbone and Rep-PAN Neck based on the RepVGG style. The design of a simpler and more efficient decoupling head is optimized to further reduce the additional delay overhead caused by the general decoupling head while maintaining accuracy. In the training strategy, the Anchor-free Anchor-free paradigm is used, supplemented by SimOTA tag assignment strategy and SIoU bounding box regression loss to further improve the detection accuracy.

3. Method

3.1. YOLOv7

YOLOv7 is the latest version of the YOLO(You Only Look Once) series, reaching new heights of speed and accuracy in the 5 FPS to 160 FPS range. YOLOv7[7] outperforms: YOLOX[8], Scaled-YOLOv4[9], YOLOv5[10] and many other target detectors in terms of speed and accuracy, and has the highest accuracy of 56.8% AP of any known real-time target detector with 30 FPS or better on the GPU V100. It is the only detector currently available that can still exceed 30 FPS with such high accuracy.

First, YOLOv7 extends the efficient long-range attention network called Extended-ELAN (E-ELAN for short). In large-scale ELAN, the network can reach steady state regardless of the gradient path length and the number of blocks. However, this steady state may also be destroyed and the parameter utilization may be reduced if the computational blocks are stacked infinitely. E-ELAN does Expand, Shuffle, and Merge cardinality on cardinality, which can improve the learning ability of the network without destroying the original gradient paths.

In terms of architecture, E-ELAN only changes the architecture in the computational block, and does not change the architecture of the transition layer. In addition to keeping the original ELAN design architecture, E-ELAN can guide different groups of computational blocks to learn more diverse features.

Then, YOLOv7 uses a cascade-based (Concatenation-based) model scaling approach. Model scaling refers to adjusting some properties of the model to generate models of different scales to meet the needs of different inference speeds. However, model scaling, if applied to the concatenation-based architecture, will reduce or increase the computational blocks of the concatenation-based translation layer when expanding or reducing the execution depth. It can be inferred from this that for cascade-based models, the different scaling factors cannot be analyzed separately and must be considered together. The cascade-based model scaling method is a composite model scaling method, when scaling the depth factor of a computational block, the change in the output channel of that block is also calculated. Then, the transition layer is scaled with the same amount of variation in the width factor, so that the characteristics of the model at the time of initial design are maintained and the optimal structure is maintained.

In the study, the authors also designed Planned re-parameterized convolution. The repConv has relatively excellent performance in VGG, but when it is directly applied to ResNet, DenseNet, or other architectures, the accuracy is significantly reduced.

This is because the direct connection (Identity connection) in RepConv destroys the residuals in ResNet and the connections in DenseNet. Therefore, RepConv without direct connection (RepConvN) is used in the paper study to design the network structure.

3.2. Bytetrack

The ByteTrack [11] algorithm is a target detection-based tracking algorithm that, like other non-ReID algorithms, uses only the bbox obtained from target tracking for tracking. The tracking algorithm uses a Kalman filter to predict the bounding box and then uses the Hungarian algorithm to match between the target and the track.

DOI: 10.54254/2755-2721/4/20230467

The biggest innovation of ByteTrack algorithm is the use of low score box. The authors believe that the low score box may be the box generated when the object is occluded, and discarding the low score box directly will affect the performance, so the authors use the low score box for the secondary matching of the tracking algorithm, which effectively optimizes the problem of changing ids due to occlusion in the tracking process. The features are that no Re-Id features are used to calculate the appearance similarity; non-depth method, no training is required; the distinction and matching between high and low score boxes are used to solve the occlusion problem effectively.

In the matching phase, the detector outputs both high and low scores, but slices them into high and low score detection frames by an intermediate threshold and then does the matching.

For the prediction frame generation, the high score frame is generated, and the low score frame is only matched, and the frame is deleted between the unmatched frames, and the prediction frame is kept for 30 frames, because the high score frame will be detected again when the target is obscured or the motion is blurred.

The reason for this is that the confidence value is reduced when the previous high scoring frames are detected again even if the target is obscured or the motion is blurred.

Matching stage 1: matching all predicted frames with high scoring detection frames, using appearance feature Re-Id cascade matching and IoU matching during the matching process, while unmatched predicted frames are matched in the next stage.

Matching stage 2: the prediction frames left after the first stage matching are matched with the detection frames with low scores, and only IoU matching is performed at this time, the reason is because the appearance features of the detection frames with low scores are not credible, and if Re-Id matching is used it may make the original trajectory connected with other detection trajectories, causing the ID Switch phenomenon.

For the detection frame that does not match the trace and has a high enough score, the author creates a new trace for it. For the track that does not match the detection frame, the author keeps 30 frames and matches it when it appears again.

3.3. YOLOv7-Bytetrack

Considering the excellent performance of YOLOv7 in the target detection task, bytetrack also shows a very strong performance in target tracking, replacing YOLOv7 with the target detection model of the network to obtain YOLOv7-Bytetrack.

4. Experiments

4.1. Experimental setting

This experiment combines YOLOv7 with Bytetrack, YOLOv7 with Deepsort, and YOLOv6 with Bytetrack, respectively, and compares the three. All use the official models, which are: YOLOv7tiny, YOLOv7tiny, YOLOv6tiny. It is known that the accuracy of YOLOv6, YOLOv7 is very high, and the purpose of this experiment is real-time tracking, so this experiment focuses on for fps testing. The GPU of the experimental computer is 1660ti.

Fps: the number of frames per second

4.2. Introduction to Data Sets

The test video is the same under-surveillance mobile video of cars and the live frame rate is recorded.

4.3. Experimental results and analysis

Table 1. Comparative experiments of three different method.

Method	fps
YOLOv6-Bytetrack	37
YOLOv7-Deepsort	29
YOLOv7-Bytetrack	70

DOI: 10.54254/2755-2721/4/20230467

4.4. Analysis of experimental results of daytime detection algorithm

YOLOv7-Bytetrack order to verify the effectiveness of the algorithm in real-time multi-target tracking, the proposed algorithm is compared with the FPS of YOLOv7 deepsort and YOLOv6 bytetrack. As shown in table 3.31, compared with other algorithms, the improved algorithm proposed in this paper has the highest FPS, reaching 70. It proves the adaptability of YOLOv7 to bytetrack and its high performance in real-time multi-target tracking. The algorithm has also been used for other real-time multi-target tracking other than detecting cars.

5. Conclusion

Currently, the study of real-time multi-target tracking algorithms is one of the key studies in target tracking. In this paper, the author proposes to come up with an efficient real-time multi-target tracking by combining YOLOv7 with bytetrack. In recent years, YOLOv7 with bytetrack has been proved to be very efficient in terms of computational accuracy and performance, which is the reason why it is selected. Through experimental comparison and analysis, the fusion model proposed in this paper has a much higher number of frames per second transmission compared with the other two models, which can ensure the real-time performance of target tracking, and the chosen YOLO model is small enough to be applicable to various scenarios.

For the shortcomings of this experiment. First, the author not use MOTChallenge for training and testing (MOTChallenge is currently an important benchmark in the field of multi-target tracking, which has the largest public pedestrian tracking dataset as a public platform for uploading and publishing research results of multi-target tracking methods). Instead, the PyQt 5 framework was used to display and record FPS, which resulted in the authors not comparing their MOTA and IDF1. It can be applied to the YOLOv7-ByteTrack framework to MOTchallenge later for testing. Second, the accuracy of YOLOv7 in detecting small targets is still weak, and the code needs to be changed to improve the detection effect of small targets.

References

[1] Nir Aharon, Roy Orfaig, Ben-Zion Bobrovsky. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. arXiv:2206.14651. 7. July 2022. 1.

[2] INRIX. 2017. “TRAFFIC CONGESTION COST UK MOTORISTS over £37.7 BILLION in 2017.” Inrix, inrix.com/press-releases/scorecard-2017-uk/.

[3] Yu, Steven. 11 Oct. 2019. “Why is target detection plus target tracking necessary in engineering practice?.” Know column, , zhuanlan.zhihu.com/p/70268783.

[4] Alex Bewley , Zongyuan Ge, Lionel Ott , Fabio Ramos , Ben Upcroft. SIMPLE ONLINE and REALTIME TRACKING. arXiv:1602.00763. 7 July 2017. 1

[5] Nicolai Wojke, Alex Bewley, Dietrich Paulus. SIMPLE ONLINE and REALTIME TRACKING with a DEEP ASSOCIATION METRIC. arXiv:1703.07402v1. 21 Mar. 2017. 1

[6] ChuYi, Kaiheng et al. 23 June 2022. “YOLOv6：Fast and accurate target detection framework is open source!.” Tech.meituan.com, tech.meituan.com/2022/06/23/yolov6-a-fast-and-accurate-target-detection-framework-is-opening-source.html.

[7] Chien-Yao Wang , Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-The-Art for Real-Time Object Detectors.arXiv:2207.02696v1. 6 July 2022. 1

[8] Zheng Ge, Songtao Liu, Feng Wang ,Zeming Li, Jian Sun. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430v2. 6 Aug. 2021. 1

[9] Chien-Yao Wang , Alexey Bochkovskiy , and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In CVPR. 2021. 1

[10] Jocher, Glenn, et al. “Ultralytics/Yolov5: V6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai Integrations.” Zenodo, 17 Aug. 2022, zenodo.org/record/7002879#.YwoEd3bP1D8. Accessed 27 Aug. 2022.

[11] Yifu Zhang , Peize Sun , Yi Jiang , Dongdong Yu , Fucheng Weng , Zehuan Yuan , Ping Luo , Wenyu Liu , Xinggang Wang . ByteTrack: Multi-Object Tracking by Associating Every Detection Box. arXiv:2110.06864v3. 7 Apr. 2022.