Studies on Vehicle Object Detection and Tracking in UAV Aerial Data

Ting Cao[1], Xinrong Zhang[1], Penghui Wang[2(✉)], and Chenle Wang[3]

[1] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China

[2] Key Laboratory of Road Construction Technology and Equipment, Ministry of Education, Chang'an University, Xi'an 710064, China
wangpenghui@xaut.edu.cn

[3] School of International Engineering, Xi'an University of Technology, Xi'an 710048, China

Abstract. Studies on multi-object detection and tracking for vehicles are important research topics and have been widely used in various fields such as autonomous driving, anomaly detection, traffic monitoring, and intelligent transportation system. Because of the interference factors such as long-time occlusion and heavy traffic flow, there are two major limitations to accurate vehicle tracking: missed detections and false detections. In this paper, for the demand of accuracy and real-time vehicle object detection and tracking, the YOLO v5 and Deepsort are combined to realize object detection and tracking in unmanned aerial vehicle (UAV) data. The object detector is trained and validated on its own dataset of traffic collected from highways; the open-source vehicle deep feature training dataset is used to train for tracker weights. Finally, the experiments verify the vehicle multi-object detection and tracking function in the case of heavy traffic flow with different vehicle types. The proposed method in this paper has certain theoretical significance and practical application value in the field related to multi-object detection and tracking of moving vehicles.

Keywords: UAV · Vehicle object · Detection and tracking · YOLO v5 · Deepsort

1 Introduction

Vehicle detection and multi-object tracking, as an important research topic in the field of computer vision, is a key technology in many intelligent application areas such as automated vehicle driving and traffic flow detection [1]. Due to the increasing number of automobiles, congestion, and car accidents are inevitable. Therefore, studying an intelligent transportation system with high practicality is necessary. Compared with ordinary ground inspection equipment, UAV aerial photography has the advantages of wide coverage, low cost, and flexibility. In recent years, researchers have made a lot of efforts in vehicle detection and multi-object tracking.

For object detection: early object detection models were built as an ensemble of hand-crafted feature extractors such as Viola-Jones detector [2], Histogram of Oriented

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 574, pp. 431–437, 2024.
https://doi.org/10.1007/978-3-031-65123-6_32

Gradients (HOG) [3], etc. However, traditional methods suffer from a lack of objecting in the selection of sliding window regions, and the inability of manually designed features to adapt to diversity goals and changes. Deep learning is characterized by learning the data representations [4]. Convolutional neural network (CNN) represents one such deep architecture that is most popular for learning with images and video. In 2012, Geoffrey Hinton's proposed AlexNet [5] model achieved remarkable results in the ImageNet image classification competition, which opened the curtain of convolutional neural networks applied to computer vision. In 2014, Girshick et al. proposed the R-CNN [6] model, which is the pioneering work of object detection using deep learning, and then continuously improved the algorithm. However, the disadvantage is the poor real-time performance. R. Joseph et al. proposed the YOLO (You Only Look Once) [7] model in 2015, which views the object detection task as a regression problem and can quickly recognize objects. Subsequently, researchers have proposed models such as YOLO v2, YOLO v3, and YOLO v4 [8]. In 2020, YOLO v5 was introduced which uses Cross Stage Partial Network (CSP) to reduce computational cost and outperforms YOLO v4 in terms of accuracy and speed. Real-time is important for vehicles that are in motion at all times. Therefore, in this thesis, the detection algorithm of YOLO v5 is chosen as the basis for the next step of vehicle tracking.

For object tracking: since object tracking relies on object detection, the above novel methods developed by large tech companies Facebook [9], Google, etc., are also a part of object tracking. In 2016, Alex Bewley [10] et al. proposed a new convolutional neural network-based object tracking algorithm, which is known as the SORT algorithm. After that, the authors proposed Deepsort [11] algorithm based on SORT, which can more accurately match objects between different frames, which greatly reduces the number of object ID switches in SORT. Zhou et al. [12] have developed a simultaneous MOT process, with the disadvantage of not considering that the object will reappear. Because the Deepsort tracking algorithm can meet the real-time performance of vehicle tracking for immediate motion in traffic scenarios and avoid ID switching memory overhead, which increases the accuracy and robustness of the algorithm, this thesis chooses Deepsort based tracking algorithm to accomplish the vehicle tracking task.

In summary, in this paper, based on UAV aerial images, the YOLO v5 detector and the Deepsort tracker are combined to improve the tracking accuracy and robustness in complex scenes. And the effectiveness of the proposed method is verified by comparative analysis of the homemade dataset.

2 Methodology

In this section, the design ideas of YOLO v5 detector and Deepsort tracker are first given, after that vehicle object tracking experiments are conducted based on UAV images and the results are presented. Figure 1 shows the methodology framework in this paper. It has two key parts: (1) YOLO v5 is used to realize vehicle detection; (2) Deepsort is used to realize vehicle tracking.

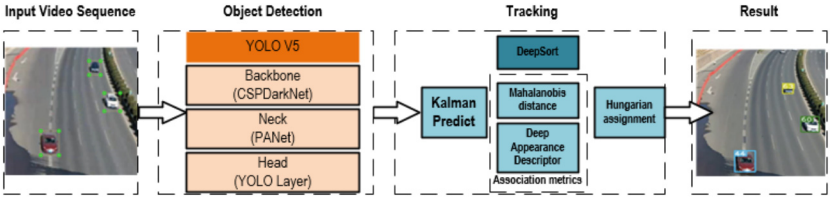


Fig. 1. Methodology Framework.

2.1 Object Detector Based on YOLO V5

The structure of a deep learning(DL) based object detection model generally consists of five parts: input, trunk, neck, head, and output. Among them, the backbone network is responsible for extracting features, the neck network is responsible for extracting more complex features, and the head network calculates the output results. First, the input part: the image is processed to fulfill the format requirements of the network structure for the input. This part mainly includes data reading, data preprocessing, and data enhancement. Second, the backbone part: in YOLO v5, the backbone network adopts CSPDarknet53 as its backbone network, which is mainly responsible for extracting relevant feature information from the input image. In the backbone network, the model gradually reduces the feature map size for subsequent processing. Third, neck part: in the YOLO v5 model, the neck network adopts the SPP structure and introduces the PANet model, which enhances the multi-scale nature of the feature extraction, thus improving the accuracy of object detection. Its four, head part: in the YOLO v5 model, the head network is mainly responsible for realizing the tasks of object classification and position regression. Fifth, the output part: it mainly consists of two parts: the screening mechanism and the bounding box reduction. In the screening mechanism, low-accuracy prediction results are proposed according to the threshold set in advance, and the detection results are optimized and screened based on the NMS algorithm.

2.2 Object Tracker Based on Deepsort

The processing of Deepsort object tracker system is mainly carried out in the order of video frames, which includes the following steps: firstly, read the position of each object ID detected frame in the current video frame and the depth features of the image object in each detected frame (which need to be extracted by themselves in practical use). Then, the object frames are filtered by confidence, i.e., the features of the object frames with lower confidence are removed. Next, non-maximal suppression of the recognized frames is performed to eliminate duplicate localization of multiple object detection frames for a single object. The prediction step is then performed and the Kalman filter is used for the prediction of the object location information. Once the prediction is complete, the accurate effectiveness of the tracking must be ensured, which requires the design of an implementation update function for real-time adjustment of the Kalman tracker parameters and feature set. In addition, it is necessary to calculate the mutual matching between the disappeared old object and the new tracked object, which first needs to match the detected results with the results obtained from the tracking prediction in order

to recognize the confirmed state tracker and the unrecognized state tracker. Finally, the design of the constructor for the Deepsort class itself is completed, with model paths, maximum Mahalanobis distance, minimum confidence, maximum t Intersection Over Union (IOU) distance, and whether or not to use cuda acceleration as members, so as to realize the call to this function after the object detection is realized.

2.3 Dataset

The dataset consists of two parts, the first part is the traffic situation dataset collected by ourselves, which is used for the training and testing of YOLO v5 model. The second part is from the public dataset, which is used to study the deep appearance features of Deepsort.

The vehicles are categorized into five classes, car, taxi, minivan, truck, and motorcycle, which are used to train the object detection model. Regarding the second part of the dataset, it is used to train the deep appearance model of the vehicles to facilitate the Deepsort algorithm to accomplish the vehicle tracking task better. This open-source dataset captures images of 769 different vehicles through surveillance videos. These images were fixed to 128x256 width and height and named according to a certain format, the file name consists of the following parts: the first 4 bits have the same name as the directory where the image is located, the middle 2 bits represent the camera ID, the next 5 bits represent the tracking ID, and the last 4 bits represent the image serial number.

3 Experiments and Analysis

After completing the data preprocessing, the model needs to be configured. According to the performance comparison of each YOLO v5 pre-training model, the yolov5m.pt model has good accuracy and speed and is well suited to the lightweight deployment requirements of this project.

In this paper, the migration learning method is used while modifying the nc parameters to adapt to the new dataset. Some compression processing methods are used, such as setting the input image size to --img 640, --batch 4, etc. After 50 training sessions, the model parameters that will be generated after each epoch training are saved. During the training process, the model's loss value is constantly optimized and reduced. According to the indicators constantly adjusted and optimized, and finally get the optimal solution parameters of network model training.

Detector effect analysis: this paper analyzes the three evaluation indexes from Precision, Recall, and mAP. Tested using the trained model, Fig. 2 demonstrates the model detection results, and the Precision-Recall graph and mAP values obtained are shown in Fig. 3.

In order to quantitatively evaluate the model constructed in this paper, the test results of the model are tested. It can be seen that mAP @ 0.5 and mAP @ 0.5: 0.95 are 0.5567 and 0.4919, respectively. Compared with the results of the official YOLO v5 model, it can be seen that the value of mAP @ 0.5 is relatively high, and the training effect is considerable. Considering the simplicity of the data set and the experimental

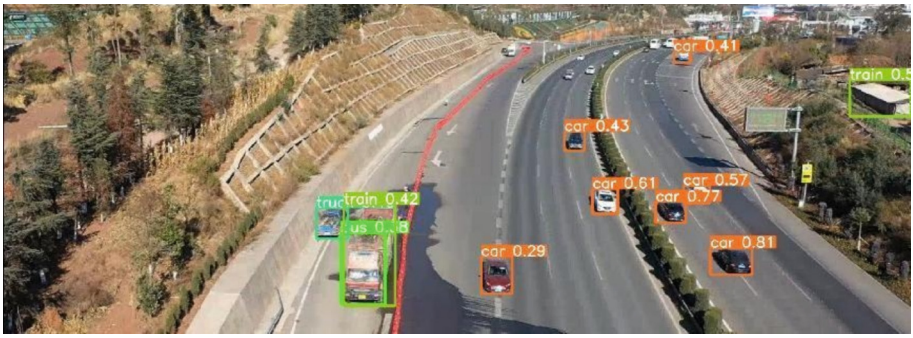
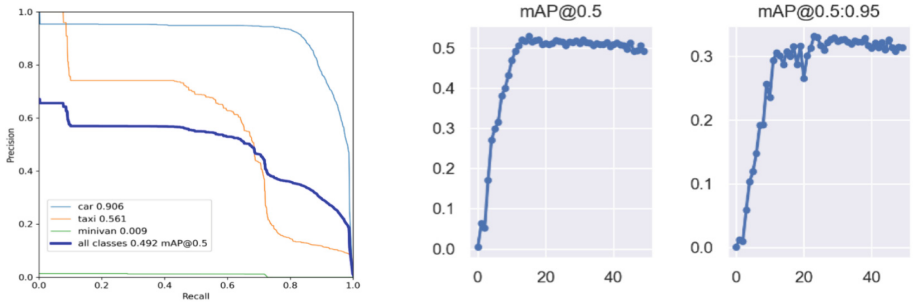


Fig. 2. Diagram of detection results



(a) **(b)**

Fig. 3. (a) Precision-Recall plot (b) Training result mAP plot

development environment, 50 pieces of training can achieve this effect. Therefore, the weight file can be used as the weight file of the object detection function in this paper and meet the requirements of traffic video. Then, the video composed of the test set is tested. Numerically, the test results of the model are more accurate.

Tracker effect analysis: first of all for the extracted vehicle depth feature effect analysis, the depth feature in the completion of the training of the total_loss function value remains stable, stable at about 4, indicating that the training effect is good. As shown in Fig. 4 below.

Finally, the overall system of vehicle detection and tracking in this paper is tested using the previously trained YOLO v5 model vehicle detection weights file and Deepsort vehicle tracking weights file. The system accomplished vehicle detection and tracking in the test set of videos as shown in Fig. 5.

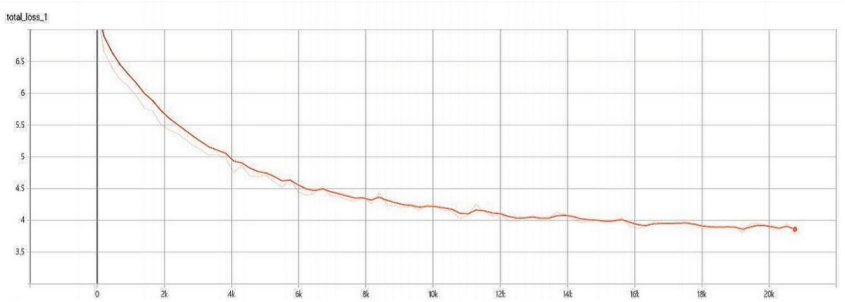


Fig. 4. Variation of total_loss function for tracker deep feature training

(a) **(b)**

Fig. 5. Vehicle tracking detection effect of (a) the current frame and (b) the next frame.

4 Conclusion

This paper mainly focuses on the characteristics of vehicle movement in traffic scenes from UAV data, a detection and tracking network for traffic safety is introduced based on YOLOv5 object detection to improve the accuracy of Deepsort based vehicle tracking algorithm. With the intelligent transportation system, YOLO v5 object detector and Deepsort tracker are combined to establish the vehicle detection and tracking system in UAV aerial images. Finally, after training with the urban traffic dataset and vehicle appearance dataset, the proposed system is implemented respectively, and their effects are also analyzed and verified. The results show that this detection and tracking system could effectively meet the needs of traffic tracking.

Acknowledgment. This paper was supported in part by the Open Project of Key Laboratory of Road Construction Technology and Equipment Ministry of Education(Chang'an University) under Grant 300102252510, in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01–03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, 2022JQ-264 and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474.

References

1. Zaidi, S.S.A., Ansari, M.S., Aslam, A., et al.: A survey of modern deep learning based object detection models. Digital Signal Process. **126**, 103514 (2022)
2. Viola, P., Jones, M.: Robust real-time object detection. Int. J. Comput. Vision **4**(34–47), 4 (2001)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 20–25 June 2005 (2005)
4. Pal, S.K., Bhoumik, D., Chakraborty, D.B.: Granulated deep learning and Z-numbers in motion detection and object recognition. Neural Comput. Appl. **32**(21), 16533–16548 (2020)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
6. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
7. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
8. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:200410934 (2020)
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
10. Bewley, A., Ge, Z., Ott, L., et al.: Simple online and realtime tracking. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). IEEE (2016)
11. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP). IEEE (2017)
12. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV, pp. 474–490. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_28