Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 3, No. 3, 2022

Aerial Traffic Statistics Based on YOLOv5+DeepSORT

Wei Liu[1], Lin Zhang[1, 2, *]

[1]College of Civil and Architectural Engineering, North China University of Science and Technology, Tangshan Hebei 063210, China
[2]Tangshan City Air and Ground Smart Transportation Key Laboratory, Tangshan Hebei 063210, China

Abstract: Traffic flow statistics, as an important part of intelligent transportation system, usually requires manual statistics, which is time-consuming and labor-intensive. In order to save manual labor and improve the statistical efficiency, this paper is based on the strategy of YOLOv5+DeepSORT to count the aerial traffic flow by UAV, and the results show that the statistical accuracy of this method is close to that of manual statistics, which has high practical value.

Keywords: Traffic Flow Statistics, Intelligent Traffic System, YOLOv5, DeepSORT, Aerial Drone Photography.

1. Introduction

As a crucial part of intelligent transportation system [1], the use of video for traffic statistics is a research hotspot. As a new data acquisition source, UAV has unique advantages in monitoring road traffic because of its characteristics of high mobility, large range, high resolution and strong adaptability, and plays a significant role in the development of transportation strategy. Therefore, traffic statistics based on UAV aerial video has become the development direction of intelligent transportation technology in the future, and has great development and application potential.

Traffic statistics based on drone aerial videos usually consist of two parts, namely vehicle detection and vehicle tracking. Traditional object detection algorithms such as Haar+Adaboost[2], Hog+SVM[3], etc. manually extract target features, and then use classification algorithms for classification and discrimination, these algorithms have slow detection speed, poor detection effect and high resource consumption, and manual extraction of features leads to frequent false detection and missed detection. With the continuous development of machine learning and computer computing power, traditional algorithms have been difficult to meet actual needs, so object detection has gradually shifted from traditional algorithms to methods based on deep learning.

There are two types of methods based on deep learning: One-Stage and Two-Stage. Two-Stage is characterized by obtaining candidate regions first, and then performing object detection, with high accuracy, the most representative of which is the R-CNN algorithm formed by combining Region Proposal and CNN in 2015 by GlRSHICK et al. [4]. Subsequent Fast-RCNN [5] and Faster-RCNN [6] were developed on the basis of the R-CNN algorithm. One-Stage is characterized by direct classification and regression of input targets, representing models such as SSD [7] and YOLO [8-10] series. At present, the YOLO model has been iterated to YOLOv5, released by Ultralytics, YOLOv5 contains a variety of models, such as YOLOv5s and YOLOv5m models, the main difference is that the depth and width of the network can be controlled to obtain models of different sizes. In order to balance accuracy with computational burden, this paper uses YOLOv5x as the benchmark model.

The target tracking algorithm can be divided into detection-based tracking and detection-free tracking according to the initialization method. Detection-based tracking requires a detector to detect the object in the image in advance, and then track the detected target, so the good or bad detection effect has a great impact on the tracking effect; Tracking without detection only requires manual annotation of tracked targets in the initial frame, and the algorithm is not flexible, and it cannot track different targets in subsequent frames. Object tracking can also be divided into online tracking and offline tracking according to how video frames are processed. Online tracking cannot process previous frame tracking results based on current frame information; Offline tracking can use the data before and after the current frame to obtain the global optimal solution, but it is not suitable for practical applications. In the vehicle tracking task of UAV aerial video, considering real-time and flexibility, detection-based online tracking is the closest method to practical application.

In this paper, the YOLOv5 object detection algorithm based on deep learning is used to detect vehicles, combined with the detection-based DeepSORT target tracking algorithm of online tracking, and the number of vehicles in UAV aerial video is completed by using virtual detection lines.

2. YOLOv5 Vehicle Detection Algorithm

YOLOv5 consists of four parts: input, backbone network, neck network and prediction. The network structure is shown in Figure 1.

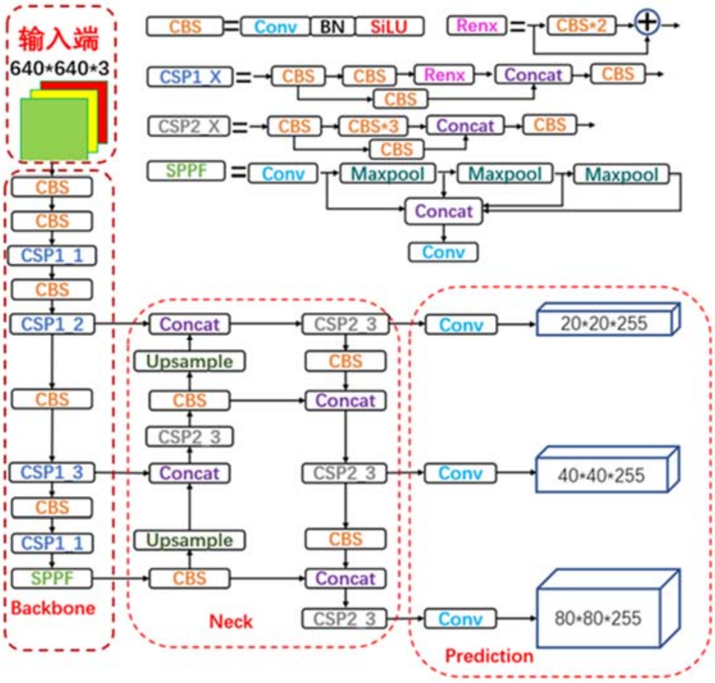

Figure 1. YOLOv5 network structure

2.1. Input terminal

The input end consists of three parts: mosaic data reinforcement, adaptive anchor frame calculation and adaptive image scaling. Mosaic data enhancement is to randomly zoom, randomly crop, randomly arrange and finally stitch and combine the 4 randomly selected pictures, which has the advantage of enriching the dataset and reducing the GPU. Adaptive anchor box calculation is mainly used to output prediction boxes on the basis of the initial anchor box when training for different datasets, and then compare with the real box to calculate the gap between the two. Adaptive image scaling adaptively zooms and fills many images with different aspect ratios according to the standard size to meet training requirements, reduce the amount of calculation, and improve the detection speed.

2.2. Backbone Network

The backbone network mainly contains CBS, CSP and SPPF three structures, CBS consists of convolution, batch normalization and SiLU activation function. The YOLOv5-6.0 version has two CSP structures, CSP1_X applied to the backbone part and CSP2_X to the Neck part, X represents the number of residual components (Resunit), and the network depth is controlled by the number of residual components. SPPF connects multiple MaxPool layers to achieve feature fusion at different scales.

2.3. Neck

The Neck network consists of a Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN). FPN upsamples high-level features from bottom to top, and enhances semantic information by fusing with low-level feature information. PAN adds top-to-bottom feature fusion on the basis of FPN, and uses the downsampling method to transmit low-level features to high-level information fusion, which strengthens the underlying information positioning ability. The combination of the two is conducive to the model to better learn features and enhance the sensitivity of the model to small targets.

2.4. Prediction

Considering the overlapping area, center point distance and aspect ratio between the prediction box and the target box, YOLOv5-6.0 Prediction uses CIOU-Loss as the target box loss function, which makes the prediction box regression faster and more accurate. At the same time, the prediction box is screened by non-maximum suppression (NMS) to remove the redundant prediction box.

3. DeepSORT Vehicle Tracking Algorithm

DeepSORT algorithm is improved on the basis of SORT, although the original SORT algorithm can effectively associate targets and can track in real time, but for multi-target tracking, the SORT algorithm can not take into account the content in the recognition box, can not guarantee the target identity switches (IDs). The DeepSort algorithm adds cascade matching and confirmation of new trajectories on the basis of Sort, reduces the number of ID switching, and processes the motion prediction and data correlation parts of the tracking problem through the Kalman filter and the Hungarian algorithm, respectively.

To track each vehicle detected by the YOLOv5 detector, define an 8-dimensional status vector X representing the vehicle state, see Equation (1):

$$X = (\mathrm{x}, y, g, h, x^{'}, y^{'}, g^{'}, h^{'}) \qquad (1)$$

where (x,y) represents the vehicle center coordinates; g and h indicate the vehicle bounding box aspect ratio and height, respectively; ($x^{'}, y^{'}, g^{'}, h^{'}$) represents the speed information corresponding to the vehicle (x, y, g, h) in the image coordinates.

The Mahalanobis distance between the predicted target frame obtained by the Kalman filter and the target frame of the current frame is calculated, and when its value is greater than the set threshold, it means that the predicted target frame is unreasonable, and it is reasonable and accurate to be less than the set value, so as to complete the tracking of the target. See Equation (2):

$$D^{(1)}(m,n) = (D_n - Y_{\mathrm{m}})^T S_m^{-1}(D_n - Y_{\mathrm{m}}) \tag{2}$$

where indicates the matching degree between the nth detected target and the m-track track; Ym represents the target prediction position of the mth trajectory; Dn represents the nth object detection bounding box; Sm represents the current frame covariance matrix predicted by the Kalman filter,

To improve IDs, the prediction accuracy is measured by using the minimum cosine between the predicted target and the feature vector of the predicted target contained in the trajectory as the degree of apparent match. See Equation (3) for the formula:

$$D^{(2)}(m,n) = \min\{1 - r_n^T r_k^{(m)} \mid r_k^{(m)} \in R_{\mathrm{m}}\} \tag{3}$$

In the Chinese formula, rn represents a feature vector corresponding to each detection block Dn, and rk represents the feature vector successfully associated with the last 100 frames. Weighted fusion of the two as the final measure, see Equation (4):

$$C(\mathrm{m,\ n}) = \lambda \mathrm{D}^{(1)}(m,n) + (1-\lambda)\mathrm{D}^{(2)}(m,n) \tag{4}$$

Finally, the Hungarian algorithm is used to detect whether a target of the current frame is a target of the previous frame. The cascade matching idea is adopted to avoid the loss of trajectory caused by the vehicle being blocked for a long time, and the Kalman filter prediction will lead to the problem of probability diffusion. The deepsort algorithm uses the above idea, and after accurately detecting the target, it accurately locates the target position in the next frame.

4. Experiments and Analysis

4.1. Experimental preparation

The Vehicle Detection and Vehicle Tracking dataset is a Visdrone2019 open-source aerial photography dataset. The dataset contains different traffic scenarios, including highways, intersections, T-junctions, etc., including different environmental backgrounds during the day and night. In this paper, a total of 6100 CAR, TRUCK and BUS sheets were extracted, including 4800 training sets and 1220 test sets. The experimental configuration environment is shown in Table 1, and the verification data is a ten-minute video of the intersection of Beixin Road and Xueyuan Road in Tangshan City taken by UAV, and the intersection of Beixin Road and Xueyuan Road is shown in Figure 2.

Table 1. Experimental configuration environment

Experimental environment	parameter
Deep learning framework	Pytorch
operating system	Windows 10
GPU model	NVIDIA GeForce GTX 1080
Programming tools	Pycharm
programming language	Python3.8
data set	VisDrone2019-DET

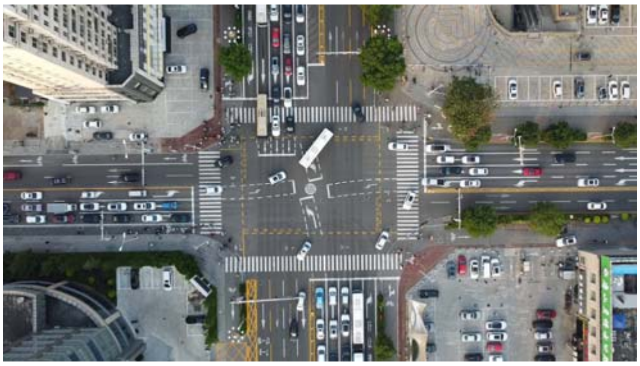

Figure 2. Intersection of North New Road and College Road

4.2. Analysis of experimental results

Intersection Name		Manual Statistics (units)	Algorithm Statistics (units)	Average Accuracy (%)
East import	Go straight	189	186	94.9
	Left	69	63	
	Turn right	75	71	
West import	Go straight	137	135	94.7
	Left	7	6	
	Turn right	36	36	
South import	Go straight	103	102	97.5
	Left	72	70	
	Turn right	54	52	
North import	Go straight	72	71	97.7
	Left	63	63	
	Turn right	18	17	

As can be seen from the above table, the average accuracy of the algorithm results of the east import algorithm at the intersection of Beixin Road and Xueyuan Road is 94.9% compared with the manual statistical results. Compared with the manual statistical results, the average accuracy rate of the western import algorithm reached 94.7%; Compared with the manual statistical results, the average accuracy rate of the South Import algorithm reached 97.5%; Compared with the manual statistical results, the average accuracy of the North Import algorithm reached 97.7%. On the whole, the statistical results of the algorithm in this paper are close to the manual statistical results, which has high practical value.

5. Conclusion

In this paper, the strategy of combining YOLOv5 object detection algorithm and DeepSORT multi-target tracking algorithm is used to realize the statistics of traffic flow at traffic intersections under UAV aerial video, and the experimental results show that the proposed algorithm shows high statistical accuracy and is practical. Although the model can run smoothly, it is still too burdensome for some devices with low computing power, and the model will be further compressed and lightweight and transplanted to embedded devices.

Acknowledgment

Fund: Natural Science Foundation of Jiangsu Province (20170157).

References

[1] Guan Jizhen. The development and evolution of intelligent transportation system and its intergenerational characteristics[J].Artificial Intelligence,2022(04):40-49.DOI:10.16453/j.cnki.ISSN2096-5036.2022.04.004.)

[2] VIOLA P，JONES M.Rapid object detection using aboosted cascade of simple features[C]//Proceedings of the2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[3] DALAL N，TRIGGS B.Histograms of oriented gradientsfor human detection[C]//2005 IEEE Computer SocietyConference on Computer Vision and Pattern Recognition (CVPR'05), 2005: 886-893.

[4] GlRSHICK R,DONAHUE J,DARRELL T,et al. Rich Feature Hierarchies for Accurate Object Detec-tion and Semantic Segmentation [CJ / / IEEE. Pro-ceedings of the lEEE Conference on Computer V1s1on and Pattern Recognition. New York: IEEE, 2014:580-587.]

[5] GIRSHICK R. Fast RCNN[J]In Proceedings of the IEEE International Conference on Computer Vision,2015,7: 1440-1448.

[6] REN S,HE K,GIR SHICK R, et al．Faster RCNN: towards real-time object detection with region proposal network [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39: 1137-1149.

[7] LIU W, ANGUELOV D, ERHAN D, et al.SSD: single shot multibox detector[C]//European Conference on Computer Vision, 2016: 21-37.

[8] REDMON J, FARHADI A.Yolov3: an incremental improvement[J].arXiv: 1804.02767, 2018.

[9] BOCHKOVSKIY A, WANG C Y, LIAO H Y M.Yolov4: optimal speed and accuracy of object detection[J].arXiv: 2004.10934, 2020.

[10] Ultralytics.YOLOv5[EB/OL].[2021].https：//github.com/ultralytics/yolov5.

[11] LIU Lei. Statistical research on intelligent traffic flow based on YOLO network[D].Xi'an University of Science and Technology,2019.)

[12] Zhao Kaidi. Design and implementation of vehicle and lane detection system based on UAV[D].Xidian University,2018.)

[13] HU Zhicheng,XIE Wei,WANG Jinsong,CUI Zhen. Research and application of video detection technology on vehicle flow[J].Digital World,2017(12):89-91.)