# A Vehicle Trajectory Extraction Method for Traffic Simulating Modeling

Chen Chen[1,a], Nver Ren[2,b], Maoying Li[3,c*], Xu Cheng[4,d]

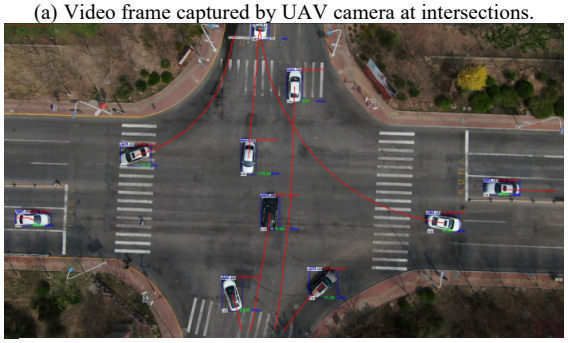
[1]China Automotive Technology and Research Center Co., Ltd. Tianjin, China
[2]China Automotive Technology and Research Center Co., Ltd. Tianjin, China
[3]China Automotive Technology and Research Center Co., Ltd. Tianjin, China
[4]China Automotive Technology and Research Center Co., Ltd. Tianjin, China

**Abstract**—Vehicle driving behavior in urban traffic environment is interactive and complex. In order to simulate the driving behavior of real vehicles, it needs to rely on a large number of real vehicle trajectory data. However, making the simulated model more consistent with the real situation, a large vehicle trajectory dataset is needed. Therefore, this paper proposes a method to automatically extract accurate vehicle trajectory from video captured by UAV camera. The proposed computer vision method obtains the trajectory by detecting the vehicle and then tracking the trajectory, and calculates the relevant parameters of the vehicle, such as position, speed and direction. Therefore, the extracted data can be used as real driving behavior data for simulation modeling. The experimental results show that the proposed method is effective and reliable.

## 1 INTRODUCTION

Due to the interaction between traffic participants, vehicle driving behavior in urban traffic environment is very complex. At present, most of the research on traffic vehicle simulation focuses on the motion model and lays particular stress on the relevant theoretical research. However, the interaction between vehicles and other traffic participants in complex situations is not fully considered. In order to reproduce or simulate traffic scene, traffic vehicle simulation must rely on detailed data of vehicle driving behavior. In other words, the analysis of a large number of vehicle trajectory data can be used to model the driving behavior of vehicles in complex environments. Such vehicle trajectory data are usually derived from specific infrastructure and equipment, at very high cost and in very small quantities. For example, the NGSIM dataset [1] contains about 9000 tracks for highway driving scenarios, but the highway is a very limited environment with very little interaction between vehicles. Therefore, in order to obtain the vehicle position, speed, trajectory, direction and other data required for vehicle simulation modeling in complex traffic situations, this paper proposes a computer vision method to automatically extract the required vehicle driving behavior data from traffic video. By this way, data can be obtained in a way that is low cost and easy to process on a large scale.

(a) Video frame captured by UAV camera at intersections.



(b) Renderings of trajectory extraction at intersection (the blue box is the vehicle object detection result, and the red line is the trajectory).

Figure 1. Renderings of the method for automatic extraction of accurate vehicle trajectory by UAV video.

Using computer vision and deep learning, this paper presents a end-to-end method to automatically extract accurate vehicle trajectory from video captured by unmanned aerial vehicle(UAV) camera. Figure 1 shows the screenshot of the traffic vehicle video and the renderings of the corresponding vehicle trajectory extraction method. The advantages of traffic video with UAV aerial view are as follows:

[a]chenchen@catarc.ac.cn, [b]rennver@catarc.ac.cn, [d]chengxu@catarc.ac.cn
[*] Corresponding author: [c*] limaoying@catarc.ac.cn

- The vehicles photographed from the UAV aerial view are discrete and there is no occlusion, which is conducive to the vehicle tracking.
- The center of the bounding box obtained by object detection can be considered as the center of gravity of the vehicle, and there is no need to re-estimate the position of the center of gravity of the vehicle.

UAV aerial view can successfully collect the video data of traffic vehicle driving behavior. The proposed method has the advantages of mass processing data, stability and reliability. And this method has positive and important significance for promoting the development of traffic simulation.

## 2 RELATED WORK

### 2.1 Object Detection

Since 2012, when Krizhevsky et al. won the title of ILSVRC for the AlexNet model designed by Convolutional Neural Networks(CNN), the model based on deep CNN has become one of the first choice in the field of object detection. And R-CNN[2] proposed by Ross Girshick et al. in 2014 completely combines selective search, CNN feature extraction and SVM object classification, which lays a foundation for the extensive application of deep learning in object detection. In 2015, Ross Girshick et al. improved on R-CNN and proposed Fast R-CNN[3]. Fast R-CNN has a great improvement in speed and accuracy compared with R-CNN. The ROI pooling structure effectively solves the problem that images of different sizes need to be enlarged to the same size. Meanwhile, classification loss and border regression loss were combined with unified training. Later, in order to solve the time bottleneck of candidate box extraction, Faster R-CNN[4] proposed RPN to extract the candidate box, and the detection speed is greatly improved.

Above methods are divided into two stages: region extraction and target detection. Although the detection speed has been greatly improved after optimization, the speed is not very ideal. YOLO[14], SSD[11] and other algorithms integrate the two stages of region extraction and target detection. They transform the candidate box position problem into the regression problem of coordinate offset, and add the coordinate regression part into the classification network for calculation. The introduction and implementation of this idea, on the premise of no great loss of detection accuracy, can improve the detection speed of images to more than 40FPS. The object detection method used in this paper is currently relatively popular YOLOv4[5].

### 2.2 Multiple Object Tracking

Tracing has traditionally been a matter of tracking points of interest through space and time. This changed with the rise of powerful deep networks. Early trackers were simple, fast, and reasonably robust. However, they were liable to fail in the absence of strong low-level cues.

Nowadays, tracking is dominated by pipelines that perform object detection followed by temporal association, also known as tracking-by-detection[8-10]. These models rely on a given accurate recognition to identify objects and then link them up through time in a separate stage. Tracking-by-detection leverages the power of deep-learning-based object detectors and is currently the dominant tracking paradigm.

An off-the-shelf object detector [4,5] first finds all objects in each individual frame. Tracking is then a problem of bounding box association. SORT[8] tracks bounding boxes using a Kalman filter and associates each bounding box with its highest overlapping detection in the current frame using bipartite matching. DeepSORT [6] augments the overlap-based association cost in SORT with appearance features from a deep network. More recent approaches focus on increasing the robustness of object association. Xu et al. [10] take advantage of the spatial locations over time. BeyondPixel [12] uses additional 3D shape information to track vehicles.

## 3 VEHICLE TRAJECTORY EXTRACTION METHOD

We present an end-to-end traffic vehicle trajectory extraction framework based on UAV perspective. In order to extract parameters such as vehicle trajectory, direction and speed, the algorithm framework proposed in this paper is shown in Figure 2.

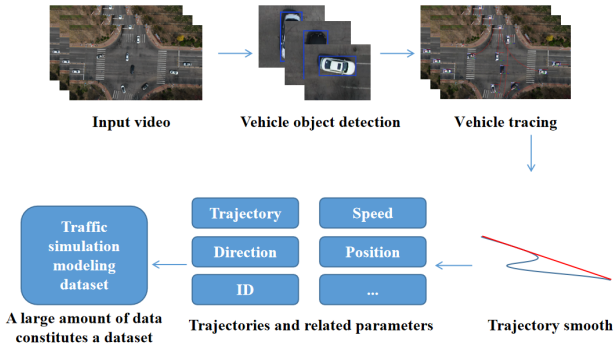



Figure 2. Traffic vehicle trajectory extraction framework.

According to the traffic vehicle trajectory extraction framework in Figure 2, it is necessary to detect vehicles through object detection of the collected traffic video, and then track the detected vehicles to obtain vehicle trajectories and relevant parameters after smoothing. It can be seen that in the vehicle trajectory extraction method, vehicle object detection is the most basic, and the result of object detection directly affects the accuracy of vehicle tracking.

### 3.1 Camera Calibration

In order to obtain the driving behavior information of the vehicle in the real world, it is necessary to calibrate the UAV camera and convert the pixel coordinates into the world coordinates.

The relationship between the three-dimensional point $M=[X,Y,Z]^T$ in the world coordinate system and the two-dimensional point $m=[u,v]^T$ in the pixel coordinate system is as follows:

$$s\tilde{m} = A[R \quad t]\tilde{M} \tag{1}$$

where, s is the scaling factor, A is the internal parameter matrix of the camera, [R t] is the external parameter matrix of the camera, $\tilde{m}$ and $\tilde{M}$ are the homogeneous coordinates corresponding to m and M respectively.

### 3.2 Vehicle Object Detection

The dynamic changes of vehicles and the aerial bird's eye view bring challenges to object detection. For this reason, we use YOLOv4[5], a popular algorithm in the field of object detection, to train the VisDrone dataset.

YOLO treats the object detection problem as a regression problem. The image is transmitted quickly through the network directly. YOLOv2[15] introduces batch regularization and anchors, and YOLOv3[16] introduces multi-scale prediction. YOLOv4 adopts the best optimization strategy in the field of CNN, and has different degrees of optimization from data processing, backbone network, network training, activation function, loss function and other aspects. YOLOv4 modified the model to make training using a single GPU more effective and adaptive, including CBN, PAN, SAM, etc. Comparison of the YOLOv4 and other state-of-the-art object detectors, YOLOv4 runs twice faster than EfficientDet with comparable performance and Improves YOLOv3's AP and FPS by 10% and 12%, respectively.

### 3.3 Vehicle Tracking

In order to link the object detection results of all frames in the video, multiple object tracking of vehicles is also required. We need to track the vehicle and give each vehicle a unique ID. However, the object is easily affected by truncation, occlusion and motion, which leads to the instability of trajectory tracking results. The aerial bird's eye view by UAV can better avoid the occurrence of occlusion. In this paper, we use DeepSORT[6], which is commonly used in industry, to track vehicles.

DeepSORT is an improved algorithm based on the SORT algorithm. SORT uses the Hungarian matching algorithm to perform recursive Kalman filter and frame-by-frame data association in the image space, and correlation metrics to calculate bounding box overlap rates. DeepSORT introduces the cosine distance between the performance features of the prediction box and the detection box as a measurement method. Combined with the Markov distance, the final incidence matrix can better represent the distance between the predicted location and the detection location, thus making the tracking result more stable.

Therefore, we use the center of the bounding box as the center of gravity of the vehicle. The vehicle trajectories can be obtained by connecting center of gravity of the vehicle with the same ID of multiple frames together.

### 3.4 Trajectory Smooth

In order to obtain a smooth vehicle trajectory, all observations obtained over a long period of time can be used to estimate the state of the system at each moment during the period. In this paper, Rauch-Tung-Striebel(RTS) smoothing [7] was used to conduct post-processing on the trajectory. It includes two steps of forward recursion and backward recursion, in which the process of forward recursion is consistent with Kalman filter algorithm, and the latter can further reduce the fluctuation of estimated results.

### 3.5 Parameter calculation

The driving behavior parameters of the vehicle are calculated for simulation modeling. For example, the vehicle direction is the tangent direction of the vehicle trajectory, and the vehicle speed is the trajectory length divided by time. And vehicle driving data is the basic data in traffic simulation modeling.

## 4 EXPERIMENTS

The vehicle trajectory extraction method proposed in this paper is applied to the video of different complex traffic scenes captured by UAV to explore its usability and accuracy.

### 4.1 Datasets and Evaluation Metrics

*1)Datasets.* The VisDrone dataset covers traffic data from the UAV perspective of 14 different cities in China from north to south, enabling extensive evaluation and research of visual analysis algorithms on the UAV platform. In this paper, YOLOv4 was used to train the VisDrone dataset to adapt to vehicle object detection under the UAV.

*2)Evaluation metrics.* We use the official evaluation metrics to Evaluation tracking accuracy. The common metric is multiple object tracking accuracy(MOTA)[13]:

$$MOTA = 1 - \frac{\sum_t (FP_t + FN_t + IDSW_t)}{\sum_t GT_t} \tag{2}$$

where $GT_t$, $FP_t$, $FN_t$, and $IDSW_t$ are the number of ground-truth bounding boxes, false positives, false negatives, and identity switches in frame t, respectively. When objects are successfully detected, but not tracked, they are identified as an identity switch (IDSW). In our studies, we report false positve rate (FP), false negative rate (FN), and identity switches (IDSW) separately. The formula is as follows:

$$FP = \frac{\sum_t \mathrm{FP_t}}{\sum_t \mathrm{GT_t}} \tag{3}$$

$$FN = \frac{\sum_t \mathrm{FN_t}}{\sum_t \mathrm{GT_t}} \tag{4}$$

$$IDSW = \frac{\sum_t \mathrm{IDSW_t}}{\sum_t \mathrm{GT_t}} \tag{5}$$

We also report the Most Tracked ratio (MT) for the ratio of most tracked (> 80% time) objects and Most Lost ratio (ML) for most lost (< 20% time) objects.

### 4.2 Implementation Details

In YOLOv4, the default hyper-parameters are as follows: the training steps is 500,500; the step decay learning rate scheduling strategy is adopted with initial learning rate 0.01 and multiply with a factor 0.1 at the 400,000 steps and the 450,000 steps, respectively; The momentum and weight decay are respectively set as 0.9 and 0.0005.

The default input resolution for vehicle tracking images is 3840 × 2160. We resize and pad the images to 416 × 416. We only output tracklets that have a confidence threshold of 0.5 or higher. And non-maximum suppression threshold is 0.4.

### 4.3 Framework Applicability

Trajectories can be extracted from these different videos using the method presented in this paper. Figure 3 shows traffic vehicle trajectories in different scenes. After testing and analyzing 120 traffic scenarios, the proposed vehicle trajectory extraction framework is stable and reliable, and has strong applicability.

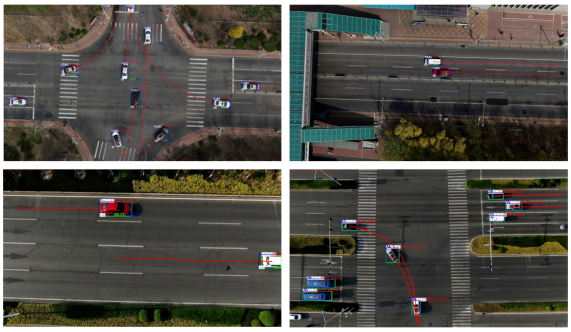


Figure 3. Traffic vehicle trajectories in different scenes.

### 4.4 Tracking Accuracy

The tracking accuracy corresponds to the tracking algorithm's ability to track the correct vehicle. In order to verify the advantages of our framework, we compare two other methods:

*1)Faster R-CNN + DeepSORT:* The one-stage object detection algorithm YOLOv4 is replaced by the two-stage object detection algorithm Faster RCNN.

*2)YOLOv4 + Kalman filter:* The Kalman filter predicts each object's future state through an explicit motion model estimated from its history. It is the most widely used motion model in traditional real-time trackers. We use the popular public implementation from SORT.

Therefore, in order to assess the tracking accuracy, this paper manually marks the track from the upper-left video shown in Figure 3 and calculates the relevant indicators, as shown in Table 1. The most important indicator is MOTA. ↑ indicates that higher is better, ↓ indicates that lower is better.

TABLE I. TRACKING INDEX STATISTICS OF VEHICLE TRAJECTORY EXTRACTION METHOD

| Method | Evaluation Metrics | | | | | |
|---|---|---|---|---|---|---|
| | ***MOTA*↑** | ***MT*↑** | ***ML*↓** | ***FP*↓** | ***FN*↓** | ***IDSW*↓** |
| Faster R-CNN + DeepSORT(1) | 58.6 | 26.8 | 24. | 998 | 269 | 123 |
| YOLOv4 + Kalman filter(2) | 55.7 | 25.3 | 24.9 | **8878** | 27540 | 1673 |
| Ours | **60.8** | **30.1** | **23.9** | 10899 | **20785** | **507** |

Table 1 lists the results on the UAV video. Our methods improves MOTA by 2.2 and 5.1 point (3.6% and 8.4% relative improvement) over Comparison method 1 and 2, respectively. And our method successfully reduce the number of identity switches, reducing from 1232(method 1) to 507. It can be seen that the MOTA index is higher, the number of track interrupts or identity switches is less, and the tracking stability is stronger.

## 5 Conclusion

In this paper, we proposed an end-to-end traffic vehicle trajectory extraction framework from UAV video. Traffic video by UAV camera as input, it uses popular YOLOv4 object detection algorithm to detect vehicles. And it uses multiple object tracking algorithm DeepSORT to track vehicle trajectories. Finally, it will smooth trajectory and calculate parameters, such as vehicle location, direction, speed.

Experiments show that the framework is stable and reliable, and can handle mass traffic videos. To some extent, this framework can be used to obtain the vehicle driving behavior data required for traffic simulation modeling.

## References

We would like to thank the company leaders and colleagues for their support to this paper.

## References

1. U. D. of Transportation, "NGSIM Next Generation Simulation," 2007.
2. Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 580–587, 2014. 2, 4.
3. Girshick R. Fast R-CNN[J]. Computer ence, IEEE International Conference on Computer Vision (ICCV), pages 1440–1448, 2015. 2.
4. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
5. Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.

6. Wojke N, Bewley A, Paulus D. Simple Online and Realtime Tracking with a Deep Association Metric[J]. 2017:3645-3649.
7. Simon D. Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches. New York, NY, USA: Wiley-Interscience, 2006.
8. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple Online and Realtime Tracking[J]. arXiv, 2016.
9. Tang S, Andriluka M, Andres B, Schiele B. Multiple People Tracking by Lifted Multicut and Person Re-identification[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
10. Xu J, Cao Y, Zhang Z, Hu H. Spatial-Temporal Relation Networks for Multi-Object Tracking[J]. 2019.
11. Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. European Conference on Computer Vision (ECCV), pages 21–37, 2016. 2.
12. Sharma S, Ansari J A, Murthy J K, Krishna K. Beyond Pixels: Leveraging Geometry and Shape Cues for Online Multi-Object Tracking[C]// IEEE International Conference on Robotics and Automation. IEEE, 2018.
13. Leal-Taixé, Laura, Milan A, Schindler K, et al. Tracking the Trackers: An Analysis of the State of the Art in Multiple Object Tracking[J]. 2017.
14. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2016. 2.
15. Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition(CVPR). IEEE, 2017:6517-6525.
16. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.