SCIENCE CHINA
Information Sciences

• MOOP •

February 2019, Vol. 62 024101:1–024101:3
https://doi.org/10.1007/s11432-018-9590-5

Vehicle tracking by detection in UAV aerial video

Shaohua LIU[1,2], Suqin WANG[3], Wenhao SHI[3], Haibo LIU[1],
Zhaoxin LI[4] & Tianlu MAO[4]*

[1]*School of Electronic Engineering, University of Beijing Posts and Telecommunications, Beijing* 100876*, China;*
[2]*Institute of Electronic and Information Engineering in Guangdong, University of Electronic Science and Technology of China, Dongguan* 523808*, China;*
[3]*School of Control and Computer Engineering, North China Electric Power University, Beijing* 102206*, China;*
[4]*Beijing Key Lab of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China*

Received 21 March 2018/Revised 26 July 2018/Accepted 5 September 2018/Published online 2 January 2019

Citation Liu S H, Wang S Q, Shi W H, et al. Vehicle tracking by detection in UAV aerial video. Sci China Inf Sci, 2019, 62(2): 024101, https://doi.org/10.1007/s11432-018-9590-5

With the maturity and popularity of UAV (unmanned aerial vehicle) technology [1–3], UAV video is becoming an effective supplement to the fixed monitoring video [4, 5]. In the aspect of traffic information acquisition, the advantages of UAV are obvious. UAV can fly not only between the buildings, but also on the freeway, and even into the tunnel, showing the unique flexibility and maneuverability. UAV can control its hovering position artificially and has a high angle shot to get more comprehensive and clearer video data. In some emergent situation, such as evacuation caused by typhoons and earthquakes, there are lots of countryside areas without fixed road monitoring cameras. UAV is undoubtedly the best choice to fast deploy and get traffic information at the front.

The shooting angle of UAV is from top to bottom, with the characteristics of large scene, different background, containing dozens of vehicles with different size and wide-range of speed. Although there are plenty of existing vehicle tracking methods, for UAV video, getting high tracking accuracy to support further traffic information analysis is still a big challenge.

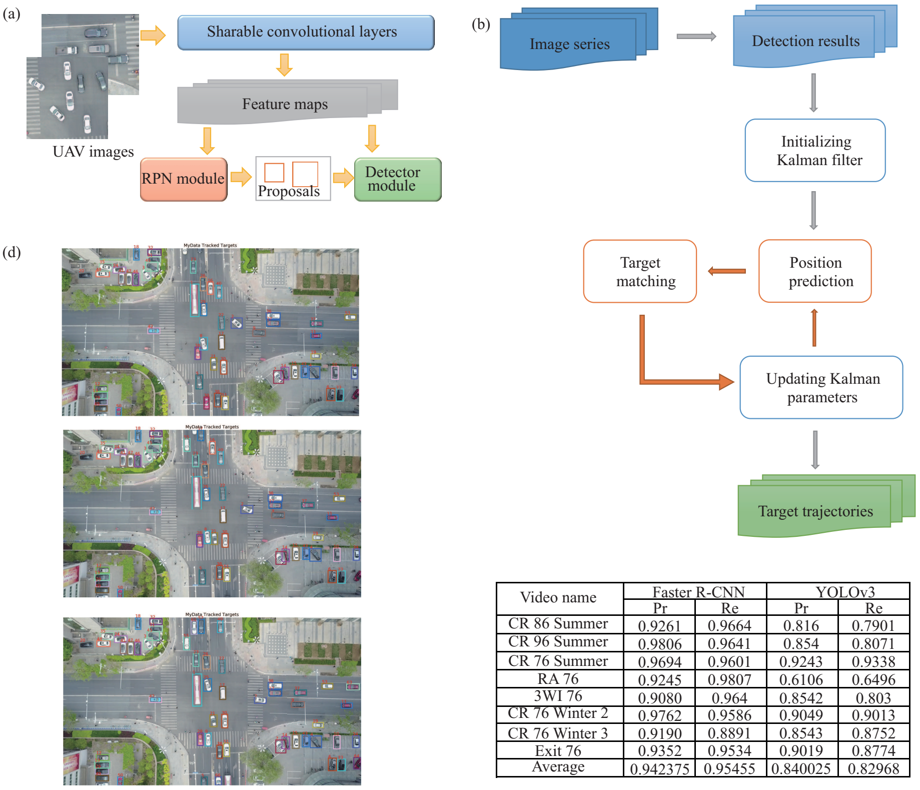
We introduce a tracking by detection method for vehicles in UAV video with high accuracy and robustness. We also made a dataset with 9 videos which were shot from different fly altitudes and captured varied traffic scenarios. Comparing with YOLOv3 method [6], our method gets better precision and recall enough to support further traffic information analysis.

Our tracking method has two stages. In the detection stage, we select faster R-CNN [7] for the reason of it does not depend on shifting scene and is very suitable for traffic scene with dozens of objects, and more importantly, it has a high accuracy. In the tracking stage, there are several object tracking algorithms here and abroad: vehicle tracking method based on Kalman filtering [8], vehicle tracking method based on mean shift, vehicle tracking method based on feature, on-line boosting tracking method. Our tracking method is developed partly on Kalman filtering method because of its relatively small storage and computational cost.

Detection algorithm. As a tracking by detection method, we firstly trained a faster R-CNN (region-based convolutional neural network) [7] to detect vehicles in an UAV Video. As shown in Figure 1(a), the whole training process could be divided into three steps:

(1) Extract feature maps from each UAV aerial image using sharable convolutional layers.

* Corresponding author (email: ltm@ict.ac.cn)



Video name	Faster R-CNN		YOLOv3	
	Pr	Re	Pr	Re
CR 86 Summer	0.9261	0.9664	0.816	0.7901
CR 96 Summer	0.9806	0.9641	0.854	0.8071
CR 76 Summer	0.9694	0.9601	0.9243	0.9338
RA 76	0.9245	0.9807	0.6106	0.6496
3WI 76	0.9080	0.964	0.8542	0.803
CR 76 Winter 2	0.9762	0.9586	0.9049	0.9013
CR 76 Winter 3	0.9190	0.8891	0.8543	0.8752
Exit 76	0.9352	0.9534	0.9019	0.8774
Average	0.942375	0.95455	0.840025	0.82968

(c)

Video name		CR 86 summer	CR 96 summer	CR 76 summer	RA 76	3WI 76	CR 76 winter 2	CR 76 winter 3	Exit 76
GT		89	110	183	43	96	134	159	119
Tracking result based on faster R-CNN	MT	87	107	179	42	93	131	151	115
	PT	2	1	4	1	3	2	5	3
	ML	0	2	0	0	0	1	3	1
	IDs	58	12	138	5	37	13	96	103
	MOTA (%)	90.47	95.98	92.66	95.73	91.11	94.92	81.96	94.41
	MOTP	0.011	0.011	0.009	0.01	0.01	0.01	0.014	0.011
Tracking result based on YOLOv3	MT	77	91	174	33	79	124	150	92
	PT	6	12	7	6	13	5	6	24
	ML	6	7	2	4	4	5	3	3
	IDs	63	33	25	41	12	72	70	91
	MOTA (%)	61.31	67.08	86.39	24.68	69.35	80.14	72.70	76.22
	MOTP	0.172	0.167	0.163	0.145	0.189	0.152	0.163	0.177

(e)

Figure 1 (a) Working flow of the faster R-CNN; (b) working flow of tracking by detection; (c) detection results; (d) visualized tracking results of a test video; (e) tracking results.

(2) Generate thousands region proposals on each feature map using the RPN (region proposal network) module in the faster R-CNN by a sliding window mechanism.

(3) Train the RPN module and the detector module in fast R-CNN with shared features.

In the training process, region proposals are generated by a sliding window mechanism, thus each of them could be regarded as a candidate to be considered. They could be marked for the purpose of training via comparing with labeled samples.

After training, the faster R-CNN works in such a way: the RPN module tells where to pay attention and the detector answers what's that.

Tracking algorithm. To utilize Kalman filtering as a tracking method, we construct a Kalman filter motion model including equations of state and observation.

During the tracking process, the time between the adjacent images is very short. So we could assume that the motion of vehicle in a unit time is a uniform motion. Then, the system state and the observed value are linear. The equation of state is

$$S(t) = A(\Delta t)S(t-\Delta t) + \omega(t-\Delta t), \qquad (1)$$

where $S(t)$ represents the state of the system at time t, $A(\Delta t)$ expresses the state transform matrix within Δt, $\omega(t)$ indicates the estimation error. We use four dimensional vectors, containing the vehicle's position and velocity, to represent the system state and the estimation error.

In the UAV images, only the position can be observed for the state vector. So the equation of observation is

$$O(t) = H(t)S(t) + e(t), \tag{2}$$

where $H(t)$ is the observation matrix, $O(t)$ is the observation vector, $e(t)$ is the observation error.

After constructing the motion model with equations of state and observation, we could track vehicles by the Kalman filtering algorithm based on detection results as shown in Figure 1(b).

Experiments. We used a Dajiang "Mavic" UAV to get videos for our experiments. It shot from different fly altitudes, 76, 86, and 96 m, and got videos of varied traffic scenarios including several crossroads, a three-way intersection, a roundabout and a high way exit. We used one of the videos, which was shot from 76 m height on a crossroad with totally 1314 UAV images, as training examples. The rest of 8 UAV videos, totally over 11 min, are also labeled as benchmarks data set.

In Figure 1(c) we compared the detection results of our method with YOLOv3 using the recall (Re) and the precison (Pr) as the evaluation criterions with IoU (intersection over union) threshold being 0.7. Results show that although we use one video for training, the recall rate and the detection precision of faster R-CNN are not affected by altitude and scene changes, which proves the stability and robustness of it. And whether recall or precision, faster R-CNN get more than 10% higher than YOLOv3.

Figure 1(d) gives the visualized tracking results of a video. And in Figure 1(e), we compared the tracking results of our method with YOLOv3. The GT means the number of groundtruth trajectories. The MT means percentage of GT trajectories which are covered by tracker output for more than 80% in length. The ML means percentage of GT trajectories which are covered by tracker output for less than 20% in length. The PT means GT-MT-ML. IDS means the total of number of times that a groundtruth trajectory is interrupted in tracking result. Multi-object tracking accuracy (MOTA): summary of overall tracking accuracy in terms of false positive, false negatives and identity switches [9]. Multi-object tracking precision (MOTP): summary of overall tracking precision in terms of bounding box overlap between ground-truth and reported location [9]. We could see that, based on the accurate detection results, our method gets better results than tracking based on YOLOv3. More and dynamic results could be found in our attached video.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2016YFC0802500), National Natural Science Foundation of China (Grant No. 61532002), the 13th Five-Year Common Technology pre Research Program (Grant No. 41402050301-170441402065), and Science and Technology Mobilization Program of Dongguan (Grant No. KZ2017-06).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wang K L, Ke Y J, Chen B M. Autonomous reconfigurable hybrid tail-sitter UAV U-Lion. Sci China Inf Sci, 2017, 60: 033201

2 Li P, Yu X, Peng X Y, et al. Fault-tolerant cooperative control for multiple UAVs based on sliding mode techniques. Sci China Inf Sci, 2017, 60: 070204

3 He W, Huang H F, Chen Y N, et al. Development of an autonomous flapping-wing aerial vehicle. Sci China Inf Sci, 2017, 60: 063201

4 Kanistras K, Martins G, Rutherford M J, et al. A survey of unmanned aerial vehicles (UAVs) for traffic monitoring. In: Proceedings of International Conference on Unmanned Aircraft Systems, 2013. 221–234

5 Xiao J J, Cheng H, Sawhney H, et al. Vehicle detection and tracking in wide field-of-view aerial video. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010. 679–684

6 Redmon J, Farhadi A. Yolov3: an incremental improvement. 2018. ArXiv:1804.02767

7 Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015. 91–99

8 Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking. In: Proceedings of IEEE International Conference on Image Processing (ICIP), 2016. 3464–3468

9 Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J Image Video Process, 2008, 2008: 246309