VEHICLE TRACKING AND SPEED ESTIMATION FROM UNMANNED AERIAL VIDEOS

M. Shahbazi [2, *], S. Simeonova [1], D. Lichti [1], J. Wang [1]

[1] Dept. of Geomatics Engineering, University of Calgary, Calgary, T2N 1N4 Canada – (ssimeono, ddlichti, jwang4)@ucalgary.ca
[2] Centre de géomatique du Québec, Saguenay, G7H 1Z6 Canada – mshahbazi@cgq.qc.ca

Technical Commission II

KEY WORDS: Tracking-by-Detection, Deep learning, 3D Reconstruction, Ray Casting, Feature Tracking, Speed Estimation

ABSTRACT:

In this paper, a solution for vehicle speed estimation using unmanned aerial videos is described. First, convolutional neural networks and Kalman filtering using deep features are used for detecting and tracking vehicles. Then, a photogrammetric approach is developed for estimating the three-dimensional (3D) position of the tracked vehicles on the road, which allows determining their speed. No assumptions are made about either the 3D structure of the road (e.g., constraining it to be a planar surface) or the camera pose (e.g., restricting it to be stationary). Therefore, this solution applies to videos acquired by a moving unmanned aerial vehicle from complex road structures (e.g., multi-level highways). This solution is also robust to changes of viewpoint and scale, which makes it applicable to situations where cars undergo orientation and resolution changes as observed from the sky (e.g., in roundabouts). Experiments showed that a high detection accuracy could be achieved with an F1-score of 94.54%. Besides, the tracking technique performed well, with a multiple-object tracking accuracy of 89.8% at a speed of 11 frames per second on videos of 2720x1530 pixels. Vehicle positioning (and thus, speed estimation) could be performed with an average accuracy of 0.6 m.

1. INTRODUCTION

Automated traffic monitoring plays a significant role in Intelligent Transportation Systems (ITS) and the implementation of smart cities (Ismagilova et al., 2019; Zhu et al., 2019). Traditional methods of traffic monitoring are based on static sensing systems installed on stationary ground infrastructure, such as surveillance cameras, radar sensors, and loop detectors. With the advancements of computer vision and artificial intelligence techniques, surveillance cameras have gained unprecedented value for video-based traffic monitoring (Jain et al., 2019). Applications of video-based surveillance include vehicle re-identification (Lou et al., 2019), detecting traffic congestion (Ke et al., 2019), measuring the speed of vehicles (Hua et al., 2018), predicting traffic density (Chung and Sohn, 2017), detecting accidents (Thomas et al., 2018), and many more. While efficient, video surveillance systems are expensive to mount and maintain and have a limited field of view (Ke et al., 2017; Leitloff et al., 2014; Niu et al., 2018). Video-based traffic monitoring using Unmanned Aerial Vehicles (UAVs) is an alternate solution to stationary systems. UAV-mounted cameras can capture videos of the traffic-flow from various viewpoints to provide a thorough insight into road conditions (Biswas et al., 2019; Cao et al., 2011; Li et al., 2019).

A fundamental component of traffic monitoring is speed estimation (Buch et al., 2011; Karim and Dehghani, 2010; Moranduzzo and Melgani, 2014; Nemade, 2016). Speed estimation from stationary cameras is generally a less challenging task since the camera's exterior orientation parameters are fixed. However, in the case of UAV videos, the UAV moves continuously. Thus, one should detect not only the movements of the cars, but also the movements of the UAV. Besides, the movements detected in the image space must be transformed into displacements in the object space with a true scale so that the estimated speeds are accurate (Costa et al., 2020). A standard solution to this problem is to determine an approximate image-to-object scale factor (photo scale) using features of known length, e.g., road markings (Ke et al., 2017; Najiya and Archana, 2018) and the approximate length of the cars (Biswas et al., 2019; Li et al., 2019). Some studies assume that the scene can be approximated as a planar object, whose 3D model relative to the UAV is known (Feng et al., 2018). Although effective most of the time, this assumption fails at multi-level roads or the ones with considerable changes in slope.

In this paper, we present a solution for detecting, tracking, 3D positioning, and measuring the speed of vehicles from unmanned aerial videos. This approach is fast and can be applied at the video's frame rate (up to 11 frames per second on color videos of 2720x1530 pixels). The main differences of this approach with the existing literature are that 1) a deep-learning tracking-by-detection method is used to track the cars; 2) geometric calibration of the wide-angle camera of the UAV is performed to increase the accuracy of speed estimation; 3) no assumptions are made about the kinematics of the UAV; 4) no pre-assumption are made about the road structure, the road markings or the size of the vehicles.

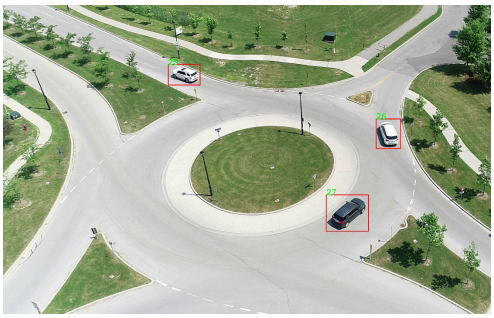
2. METHODOLOGY

2.1 Study area

The study area is a roundabout (Figure 1) located in Uxbrdige, ON, Canada. This location was selected since it was not a crowded area outside of rush hour. This way, the UAV could get closer to the road without disobeying the regulations of Transport Canada concerning the legal distance limit to pedestrians. In addition, our car (equipped with a real-time

* Corresponding author

kinematic (RTK) GNSS positioning system) could be driven in the test area to collect some ground-truth information.

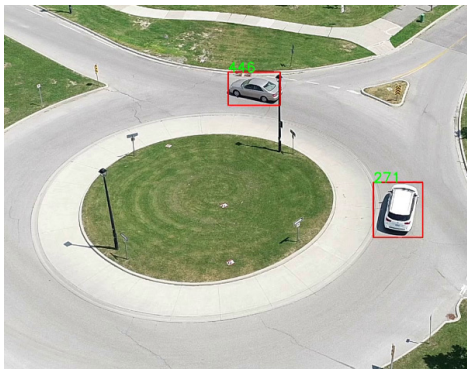


a

b

Figure 1. Sample frames of a video captured by the UAV as it first looked towards East (a) and then moved around and changed its viewpoint (b)

Roundabouts represent a challenging environment for UAV-based vehicle tracking due to the following reasons. First, when the traffic moves through a roundabout, the complete trajectory of the vehicle cannot be retrieved using traditional surveillance camera setups due to their limited field of view. Second, as a vehicle completes its maneuver through the roundabout, its appearance features change aggressively. An example of such changes in the viewpoint can be noticed for a vehicle with ID "271" in Figure 2. First, the car's back, then its side, then its front and again its side are observed in different frames of the video. If the UAV also moves and re-orients itself during data collection (as shown in Figure 1), then more severe appearance changes are expected. Therefore, it becomes exceptionally challenging to robustly track the vehicle and steadily re-identify it through its entire trajectory.



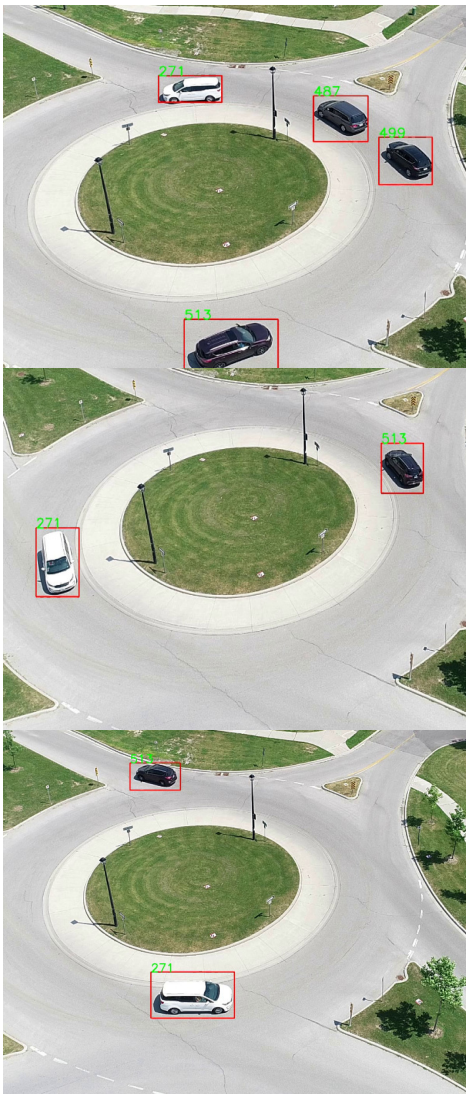


Figure 2. Trajectory of a vehicle around the roundabout causing severe changes in its appearance features

2.2 Camera calibration

The UAV used in our experiments was a DJI Phantom 4, equipped with a camera with a 4.5-mm lens capable of capturing 2K videos at 30 frames per second. In an offline procedure, using the test-field of Figure 3, the camera of the UAV was calibrated. A traditional pinhole perspective projection model augmented with three radial distortion coefficients, two tangential lens distortion coefficients, and two sensor affine distortion parameters was used in a free-network self-calibrating bundle adjustment (Shahbazi et al., 2017).

2.3 Vehicle detection and tracking in aerial videos

In this paper, our previous approach, based on deep-learning tracking-by-detection, is applied (Wang et al., 2019). This approach is intended to be invariant to substantial orientation and scale variations in the videos. The detection procedure is performed by a state-of-the-art object detector, You Only Look Once (YOLOv3). To track the vehicles, deep appearance features are used for vehicle re-identification, and Kalman filtering is used for motion estimation. The focus of this paper is

on speed estimation. A comprehensive discussion about the performance of the applied tracking-by-detection approach can be found in (Wang et al., 2019).

Figure 3. Camera calibration test-field

2.4 3D positioning and speed estimation

Assuming that a small number of well-distributed feature points on the road have known 3D coordinates and that the 3D mesh model of the road is available, a photogrammetric approach is suggested for determining the 3D position of each car in the scene. To achieve the two conditions mentioned above, first, before starting the tracking mission, the UAV flies a grid-like trajectory over the road. From the videos captured by the UAV during this flight, an adequate number of frames are extracted. Then, the frames are down-sampled to half their original sizes and are processed through a conventional structure-from-motion approach to generate a sparse point cloud of the road. Poisson reconstruction is applied next to create the 3D triangulated mesh model of the road. Amongst the sparse tie points reconstructed during this procedure, a small number of well-distributed and visually distinctive points are selected, and their 3D coordinates are stored. These tie points are referred to as salient road features.

In the first frame of the tracking video, the salient road features are manually detected. Since these features have known 3D coordinates, photogrammetric space resection can be applied to determine the exterior orientation parameters (EOPs) of the first frame. These features are then automatically tracked in the next frames of the video using a pyramidal implementation of Lucas Kanade feature tracking (Bouguet, 2000). Successfully tracked features are used for re-estimating and updating the EOPs of each frame.

Once the EOPs of a frame are known, the next step is to determine the 3D position of the vehicles detected in that frame. Each vehicle is detected with a rectangular bounding box. The center of the box is considered as the point whose 3D position relative to the road represents the location of the car. Since both the object point (vehicle) and the camera (UAV) are moving, the 3D position of the vehicle should be estimated only from a single frame, i.e., stereo or multiple views are unavailable for the conventional spatial intersection. To this end, a ray-tracing approach is applied for directly determining the 3D position of the car from its observation in one frame. The direction of the ray that originates from the perspective center of the camera and passes through the center of the car can be determined using Equation (1).

$$\mathrm{D}_{i,j} = \mathbf{R}_j \begin{bmatrix} x_{i,j} - c_x + \delta x_{i,j} \\ y_{i,j} - c_y + \delta y_{i,j} \\ -f \end{bmatrix} \quad (1)$$

In Equation (1), $\mathrm{D}_{i,j}$ is the direction of the ray originating from the perspective center of frame j and passing through vehicle i observed in frame j. $\mathbf{R}_j$ is the rotation matrix from the image space of image j to the object space. $(x_{i,j}, y_{i,j})$ are the detected/tracked location of vehicle i in frame j, expressed in the image principal coordinate system. The interior orientation parameters (IOPs) of the camera include (c_x, c_y, f), and (δ_x, δ_y) are the distortion corrections that are modeled as follows. All the distortion coefficients and the IOPs are estimated through the calibration procedure.

$$\begin{aligned} \delta_{x_{i,j}} &= (x_{i,j} - c_x)(k_1 r_{i,j}^2 + k_2 r_{i,j}^4 + k_3 r_{i,j}^6) \\ &\quad + p_1(r_{ij}^2 + 2(x_{ij} - c_x)^2) + 2p_2(x_{i,j} - c_x)(y_{i,j} - c_y) \\ &\quad + s_1(x_{ij} - c_x) + s_2(y_{ij} - c_y) \\ \delta_{y_{i,j}} &= (y_{i,j} - c_{y_m})(k_1 r_{i,j}^2 + k_2 r_{i,j}^4 + k_3 r_{i,j}^6) \\ &\quad + p_2(r_{i,j}^2 + 2(y_{i,j} - c_y)^2) + 2p_1(x_{i,j} - c_x)(y_{i,j} - c_y) \end{aligned} \quad (2)$$

where:

$r_{i,j} = \sqrt{(x_{i,j} - c_x)^2 + (y_{i,j} - c_y)^2}$

k_1, k_2, k_3 : radial lens distortion parameters

p_1, p_2 : decentering lens distortion parameters

s_1, s_2 : sensor affine distortion parameters

This ray hits one of the faces of the roads' triangular mesh, where the hit position defines the 3D location of the car, as shown in Figure 4. The Optimized Collision Detection (OPCODE) library is used to implement this ray-casting concept (Terdiman, 2003). Given the changes in the 3D position of a vehicle between two consecutive frames and the constant frame rate of the camera, the speed of the vehicle can be simply calculated.

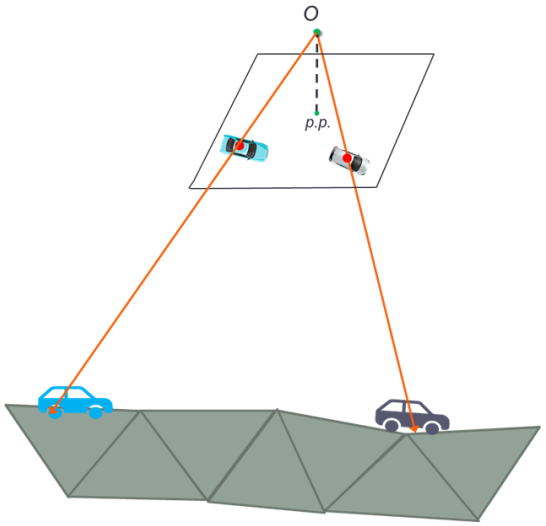
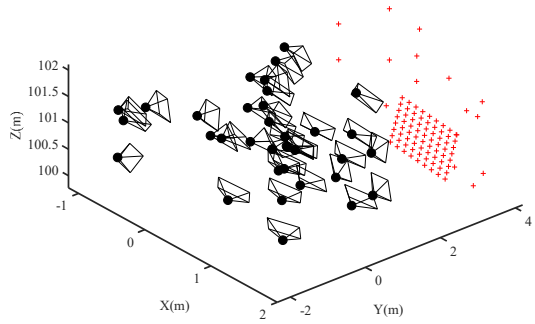


Figure 4. Concept of ray-casting to estimate the position of the vehicles observed in a frame with known EOPs

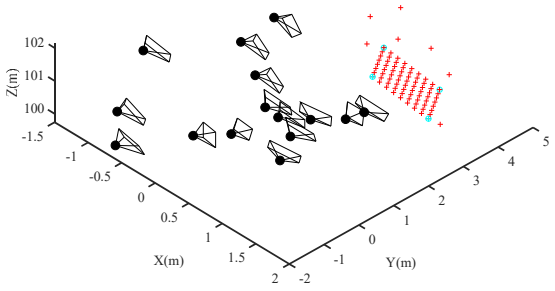
3. EXPERIMENTS AND RESULTS

To calibrate the camera, a total of 35 images (Figure 5.a) were captured. The test-field included 88 targets, and a total of 2660 observations were made in the calibration images. Also, 15

independent images were captured for check purposes (Figure 5.b), where the known corners of the checkerboard were used as checkpoints. The root-mean-square (RMS) and the standard deviation (StD) of the re-projection errors on the checkpoints were 0.16 and 0.08 pixels, respectively.



a



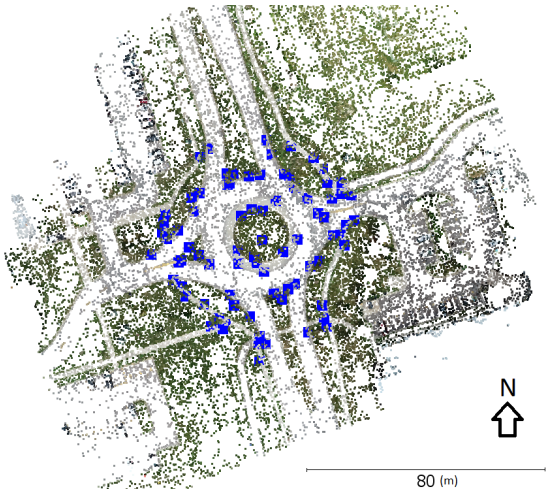
b

Figure 5. Network of (a) calibration images and (b) check images

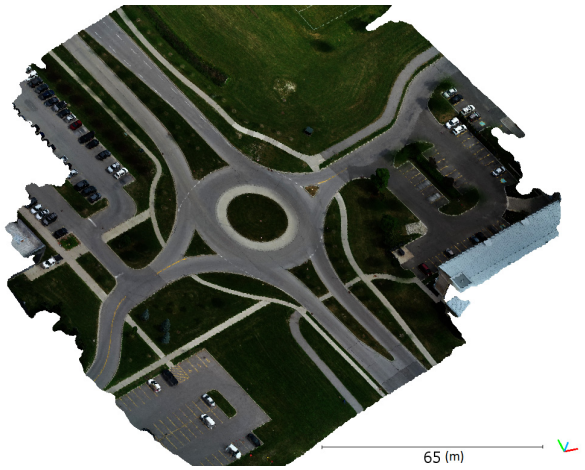
Before the UAV started its detection-and-tracking mission, it moved around the study area in a grid-pattern and captured videos. Using commercial software, Pix4D Mapper Pro, a conventional structure-from-motion technique was applied to these videos to reconstruct a low-density point cloud (Figure 7.a). From the point cloud, a triangular mesh model was generated using the open-source software, CloudCompare Stereo (Figure 7.b). In our experiment, to ensure that the 3D model of the road had a real scale, few Ground Control Points (GCPs) were established (as shown in Figure 6), whose coordinates were measured using RTK GNSS with an accuracy of 3-5 cm. In general, one can argue that external distance observations, e.g. the length of road markers, could also be used to set the scale of the 3D model. The quality of such methods for scale definition depends on the geometric configuration of the external observations as well as their accuracy. To detect and label the GCPs in images, the approach of Shahbazi et al. (2015) was applied.

Figure 6. Eleven GCPs established in the study area to ensure that the road's mesh model was generated with a correct scale

Certain features among the mapped 3D points were selected as the salient road features. These features are denoted in Figure 7.a using blue rectangles. When tracking these features in consecutive frames, the EOPs of the frame were updated only if 1) the features were moved more than two pixels in average; or 2) the total number of successfully tracked features fell below five features. Otherwise, the EOPs of the previous frame were assigned to the current frame. This condition allowed saving considerable computational time as space resection required an iterative, time-consuming adjustment.



a



b

Figure 7. The 3D model of the scene reconstructed before the tracking mission starts; a) 3D point cloud; b) 3D mesh

In total, two videos were captured for this study. Figure 8 and Figure 9 represent the position and orientation of the UAV in these test videos. One video was captured from approximately 40 meters above the road, and the other one was captured from roughly 80 meters above the road. This meant the approximate spatial resolution was 21 mm in the first video and 42 mm in the second video. The first video was 2 minutes long, and the second video was 14 minutes long. The videos were both captured with a frequency of 30 frames per second. Please note that, in Figure 8 and Figure 9, only every other 30 frames are displayed to avoid overloading the figures.

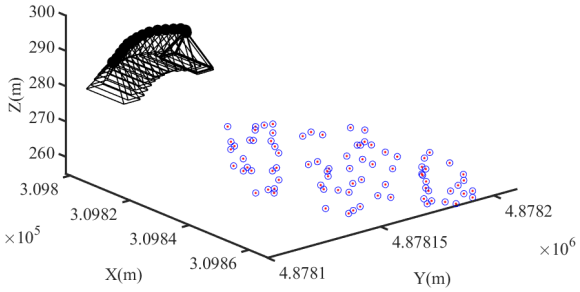


Figure 8. Orientation of the camera during capturing one of the tracking videos from ~40 m above the road. The points on the ground represent the salient road features

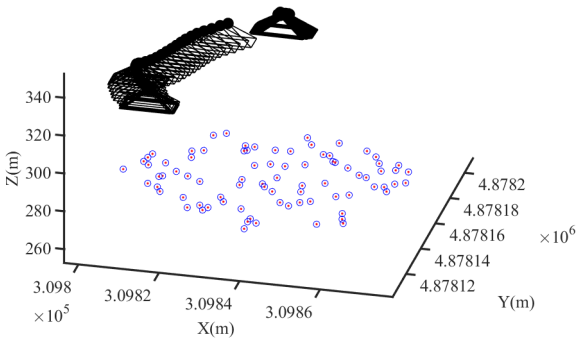
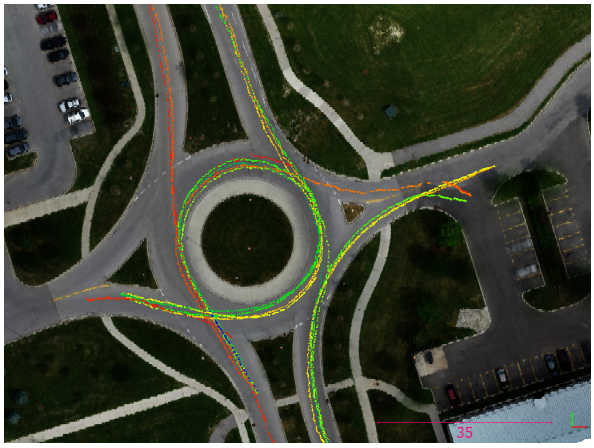


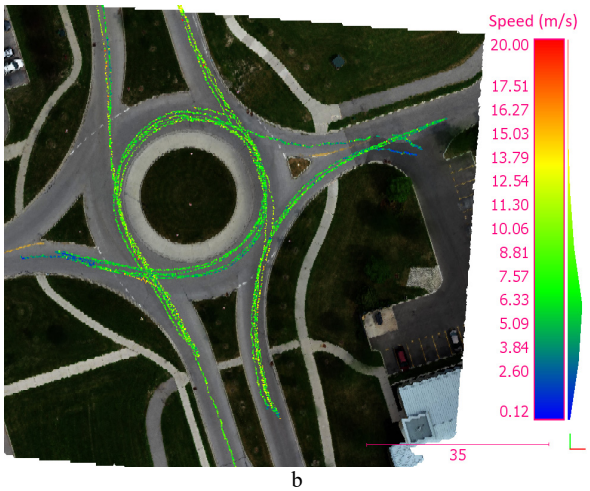
Figure 9. Orientation of the camera during capturing one of the tracking videos from ~80 m above the road

Vehicles were detected and tracked in these videos using the technique of Section 2.3. It should be mentioned that some frames from the low-altitude video were used in the training dataset of YOLOV3 network. As such, the performance of the tracking-by-detection algorithm was only evaluated on the high-altitude video. The precision, recall and F1-score of the detection algorithm in the high-altitude video were 100%, 89.64% and 94.54%, respectively. This meant that no other objects were wrongly detected as vehicles, but some vehicles were not detected at all (false negative rate of 10.34%). The tracking performance is presented by a measure called multiple-object tracking accuracy (MOTA). It is the most widely used metric that summarizes the tracking accuracy in terms of false negatives, false positives, and ID switches (i.e., when the same vehicle is wrongly identified as two different entities in consecutive frames). The MOTA in the high-altitude video was 89.8%.

Figure 10.a and Figure 11.a show the trajectories of all the cars that were visible and tracked in the test videos. From the 3D position of each vehicle in each frame, and with the knowledge of frames time interval (0.033 seconds), the ground speed of each vehicle could be determined (Figure 10.b and Figure 11.b).

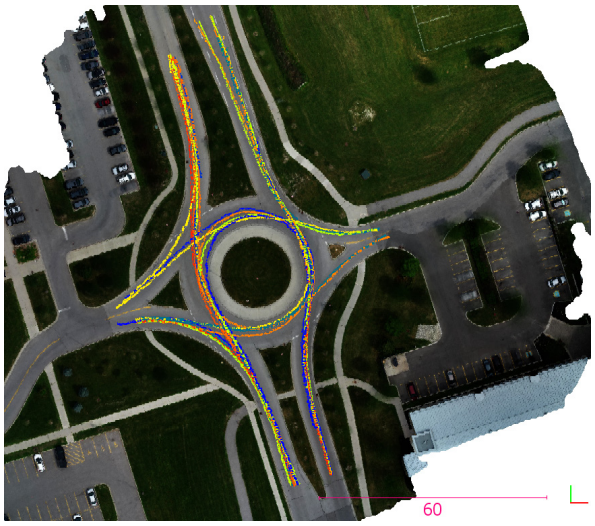


a

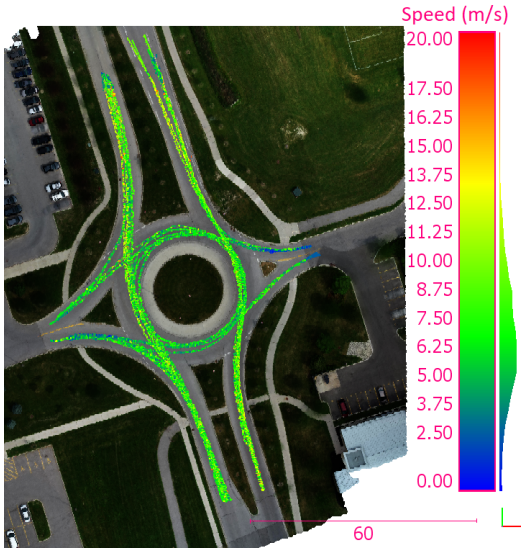


b

Figure 10. Position of the tracked cars observed in the low-altitude video; a) different colors correspond to different car IDs; b) located vehicles are colored by their speed.



a



b

Figure 11. Position of the tracked cars observed in the high-altitude video; a) different colors correspond to different car IDs; b) located vehicles are colored by their speed.

To estimate the accuracy of our 3D positioning method, we held a GNSS receiver outside the window of a car and drove around the roundabout while the video was captured. The position of the car was recorded every second (Figure 12). Once the video-based position of the car was computed, we determined the difference between the two positions (Figure 13 and Figure 14). Since the GNSS antenna was held approximately 0.5 m below the car roof and was also slanted, only the differences in the horizontal plane (2D positions) were considered for accuracy assessment. The distribution of the errors in the estimated position of the test vehicle compared to its measured position is presented in the histograms of Figure 15. In most cases, the positioning error was below 1.5 m. The mean and StD of the positioning were 0.619 m and 0.340 in the low altitude video, and they were 0.599 m and 0.378m in the high-altitude video. However, it should be noted that the GNSS antenna was located beside the window while the tracking positions correspond to the center of the car. Thus, a considerable portion of the positioning error could be naturally related to the difference in the location of the GNSS antenna and the car's center.

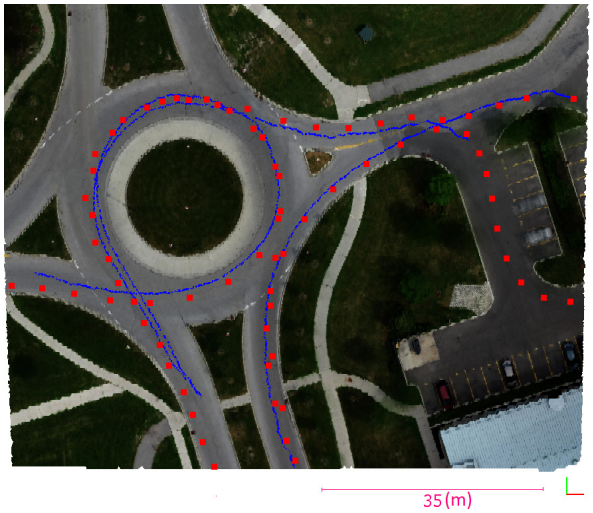


Figure 12. Estimated position of the test vehicle using the proposed vision-based approach in the low-altitude video (blue) and its measured position using real-time kinematic GNSS (red)

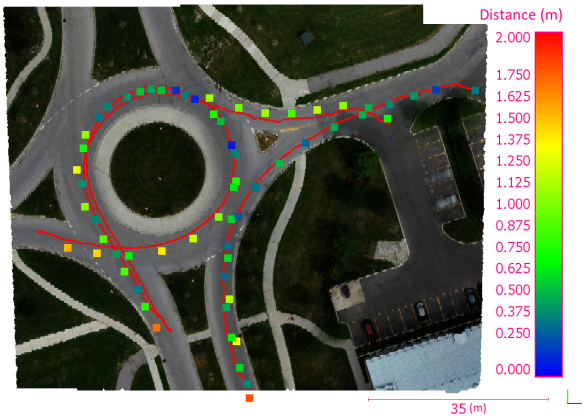


Figure 13. Distance of the estimated position of the test vehicle in the low-altitude video from its measured position using RTK GNSS; the trajectory includes two tours around the roundabout.

4. CONCLUSION

In this paper, the results of our approach for vehicle speed estimation from unmanned aerial videos were reported. The proposed method used the concepts of optical flow detection, space resection, and ray casting for tracking vehicles in 3D relative to the road structure. From the positioning results, the vehicles' speed could be estimated. Tracked positions in our experiment were compared to the ground-truth data measured via RTK GNSS positioning. The deviations of the estimated positions from the measured ones were mostly under 1.5 meters. While our study shows that UAVs are effective platforms for traffic monitoring, their commercial application in this context sounds futuristic. UAV-based traffic monitoring still faces fundamental challenges such as regulatory restrictions, lack of autonomy, low power endurance, and high sensitivity to environmental conditions like strong winds.

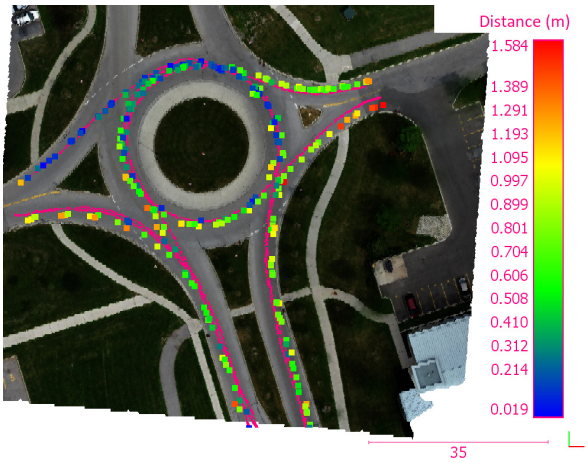


Figure 14. Distance between the estimated position of the test vehicle in the high-altitude video and its measured position using RTK GNSS; the trajectory includes seven tours around the roundabout.

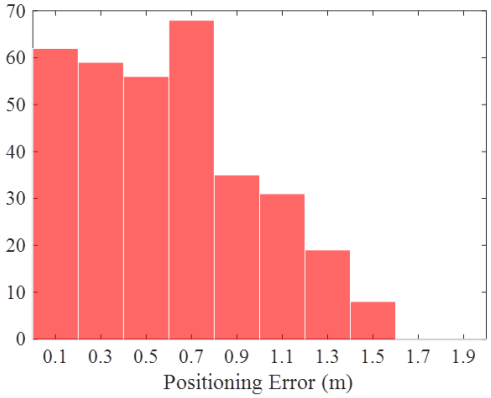


Figure 15. Histograms of vision-based positioning errors in both low-altitude and high-altitude videos

ACKNOWLEDGEMENTS

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as Canada's Tri-Council Agencies New Frontiers in Research Fund.

REFERENCES

Biswas, D., Su, H., Wang, C., Stevanovic, A., 2019. Speed estimation of multiple moving objects from a moving UAV platform. *ISPRS International Journal of Geo-Information*, 8(6), 259.

Bouguet, J.-Y., 2000. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm, Report, Microprocessor Research Labs, Intel Corporation.

Buch, N., Velastin, S.A., Orwell, J., 2011. A review of computer vision techniques for the analysis of urban traffic. *IEEE Transactions on Intelligent Transportation Systems*, 12(3), 920 – 939.

Cao, X., Wu, C., Lan, J., Yan, P., Li, X., 2011. Vehicle detection and motion analysis in low-altitude airborne video under urban environment. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(10), 1522 – 1533.

Chung, J., Sohn, K., 2017. Image-based learning to measure traffic density using a deep convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 19(5), 1670-1675.

Costa, L.R., Rauen, M.S., Fronza, A.B., 2020. Car speed estimation based on image scale factor. *Forensic Science International*, 310(May), 110229.

Feng, W., Ji, D., Wang, Y., Chang, S., Ren, H., Gan, W., 2018. Challenges on large scale surveillance video analysis, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 69–76.

Hua, S., Kapoor, M., Anastasiu, D.C., 2018. Vehicle tracking and speed estimation from traffic videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 153-160.

Ismagilova, E., Hughes, L., Dwivedi, Y.K., Raman, K.R., 2019. Smart cities: Advances in research- An information systems perspective. *International Journal of Information Management*, 47(August), 88-100.

Jain, N.K., Saini, R.K., Mittal, P., 2019. A review on traffic monitoring system techniques, *Soft Computing: Theories and Applications*, 569–577.

Karim, M.R., Dehghani, A., 2010. Vehicle speed detection in video image sequences using CVS method. *International journal of the Physical Sciences*, 5(17), 2555-2563.

Ke, R., Li, Z., Kim, S., Ash, J., Cui, Z., Wang, Y., 2017. Real-Time Bidirectional Traffic Flow Parameter Estimation from Aerial Videos. *IEEE Transactions on Intelligent Transportation Systems*, 18(4), 890-901.

Ke, X., Shi, L., Guo, W., Chen, D., 2019. Multi-Dimensional Traffic Congestion Detection Based on Fusion of Visual Features and Convolutional Neural Network. *IEEE Transactions on Intelligent Transportation Systems*, 20(6), 2157-2170.

Leitloff, J., Rosenbaum, D., Kurz, F., Meynberg, O., Reinartz, P., 2014. An operational system for estimating road traffic information from aerial images. *Remote Sensing*, 6(11), 11315-11341.

Li, J., Chen, S., Zhang, F., Li, E., Yang, T., Lu, Z., Li, J., Chen, S., Zhang, F., Li, E., Yang, T., Lu, Z., 2019. An Adaptive framework for multi-vehicle ground speed estimation in airborne videos. *Remote Sensing*, 11(10), 1241.

Lou, Y., Bai, Y., Liu, J., Wang, S., Duan, L.Y., 2019. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(8), 3794-3807.

Moranduzzo, T., Melgani, F., 2014. Car speed estimation method for UAV images. In *Proceedings of IEEE Geoscience and Remote Sensing Symposium*, 4942-4945.

Najiya, K. V., Archana, M., 2018. UAV Video Processing for Traffic Surveillence with Enhanced Vehicle Detection. In *Proceedings of IEEE International Conference on Inventive Communication and Computational Technologies*, 662-668.

Nemade, B., 2016. Automatic Traffic Surveillance Using Video Tracking. *Procedia Computer Science*, 79, 402-409

Niu, H., Gonzalez-Prelcic, N., Heath, R.W., 2018. A UAV-Based Traffic Monitoring System-Invited Paper, In *Proceedings of IEEE Vehicular Technology Conference*, 1-5.

Shahbazi, M., Sohn, G., Théau, J., Menard, P., 2015. Development and evaluation of a UAV-photogrammetry system for precise 3D environmental modeling. Sensors, 15(11), 27493-27524.

Shahbazi, M., Sohn, G., Théau, J., Ménard, P., 2017. Robust structure-from-motion computation: application to open-pit mine surveying from unmanned aerial images. *Journal of Unmanned Vehicle Systems*, 5(4), 126-145.

Terdiman, P., 2003. Optimized Collision Detection. http://www.codercorner.com/Opcode.htm (1 June 2018).

Thomas, S.S., Gupta, S., Subramanian, V.K., 2018. Event detection on roads using perceptual video summarization. *IEEE Transactions on Intelligent Transportation Systems*, 19(9), 2944-2954.

Wang, J., Simeonova, S., Shahbazi, M., 2019. Orientation- and scale-invariant multi-vehicle detection and tracking from unmanned aerial videos. *Remote Sensing*, 11(18), 2155.

Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2019. Big Data Analytics in Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 20(1), 383-398.