MONOCULAR VEHICLE 3D BOUNDING BOX ESTIMATION USING HOMOGRAHY AND GEOMETRY IN TRAFFIC SCENE

Yiqiang Chen[⋆] *Feng Liu*[†] *Ke Pei*[†]

[⋆] RAMS Reliability Technology Lab, Huawei Technology Co., Ltd.
[†] TTE-DE RAMS Lab, Huawei Technology Co., Ltd.

ABSTRACT

Video surveillance applications such as vehicle speed measurement and traffic condition monitoring are prevailing nowadays. Monocular 3D object detection by traffic surveillance cameras is one of the key means to achieve these functions. The methods in literature depend on camera calibration and require 3D annotation to learn a model. In this paper, we propose a novel vehicle 3D bounding box estimation method making use of the 3D-2D geometry consistency and homography transformation. Our method conducts vehicle 3D bounding box estimation using uncalibrated traffic cameras without requiring any 3D annotation and large computational resources. With quantitative and qualitative experiments, we show a comparable result to the state-of-that art methods and a faster execution time.

***Index Terms*—** Monocular 3D bounding box estimation, Homography, Traffic surveillance

1. INTRODUCTION

With the development of intelligent transportation systems (ITS), it is important to automatically obtain information such as positions, speeds, or classes of vehicles. These data can be further used for traffic signals control [1], traffic behavior analysis [2] and traffic state estimation [3]. In this context, camera vision algorithms such as vehicle and pedestrian detection, tracking, and re-identification, attracted considerable attention from both academia and industry.

However, in 2D detection and tracking applications, only two-dimensional positions and trajectories in the image space are extracted. Compared with 2D information on images, 3D spatial positions in the world coordinate system have obvious advantages in practical applications.

With the recent success of deep learning for computer vision applications, some deep Convolutional Neural Network (CNN) approaches have been proposed for monocular 3D object detection. For example, Mousavian *et al*. [4] used a deep CNN to estimate the 3D object orientation and 3D object dimensions, then produce 3D bounding boxes by combining with geometric constraints. Sochor *et al*. [5] estimates the vehicle contour and the directions to the vanishing points to reconstruct the 3D bounding box. Zhu *et al*. [6] performed the 3D vehicle detection by estimating the rotated bounding boxes in the bird's eye view (BEV) images generated from inverse perspective mapping. The drawbacks of these deep learning methods are the high computational cost and the dependence on 3D object annotations which are difficult to produce.

Unfortunately, most of the monocular 3D detection approaches cannot be directly applied to traffic surveillance cameras. The extracted 3D bounding boxes stand on image space, which requires the transformation from image space to the real world coordinate. Camera intrinsic and extrinsic calibrations address the purpose of physical measurements in the scene. However, due to the initial purpose of the video surveillance system, many already installed cameras and the archived video data are not calibrated.

The common solution consists of online calibrating surveillance cameras by extracting vanishing points on images [7]. The position and rotation of the camera can be further solved by the vanishing points [8, 9]. For intrinsic calibration, the focal length is calculated based on an intuitive assumption that vehicles have very similar width and height. The main limitation of this approach is another assumption that the majority of vehicles move in approximately straight, mutually parallel trajectories towards the vanishing direction. This is not always the case, e.g., at intersections.

Unlike existing methods, in this paper, we propose a novel vehicle 3D bounding box estimation approach without requiring any 3D annotation data for model training and camera calibration. In order to achieve 3D detection and the distance measurement in real-world coordinates with uncalibrated cameras, we propose to use homography mapping between the BEV plane and image plane to replace the calibration matrix for 2D image-real world connection. It is convenient to use the BEV image since it is parallel and linear to the world coordinate system. The homography and the scale of the transformed BEV images can be obtained conveniently using satellite images from public map services. Then our method leverages the prior of the vehicle sizes and the consistency between 3D and 2D detections established through homography transformation. The vehicle position and orientation in the BEV plane are solved by optimization

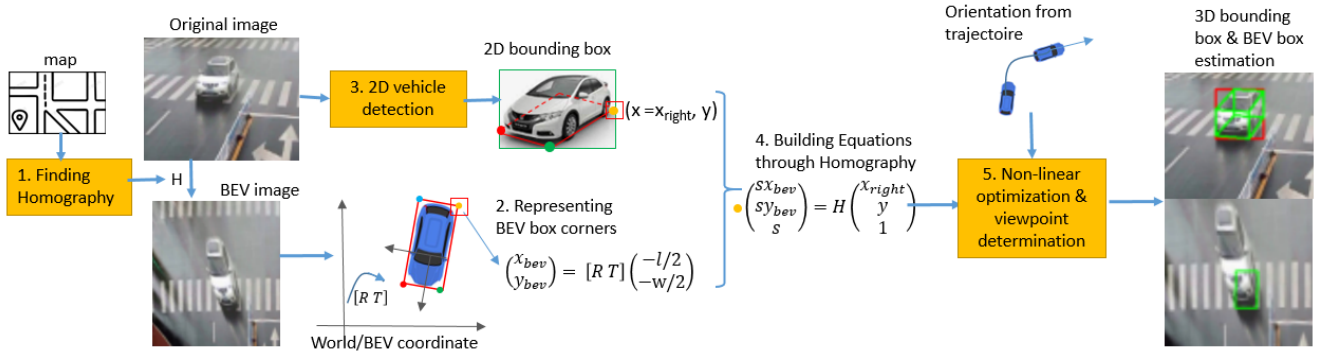

Fig. 1. Overview of our monocular 3D bounding box estimation pipeline. First, we find the homography matrix from the image plane to the BEV plane. Then we perform 2D vehicle detection on the image and build a BEV box representation. The equations based on BEV-2D detection box constraint and homography are established and the BEV box is further obtained by using non-linear optimization to solve the equations.

established on the consistency constraint. We experimentally show a promising 3D bounding box estimation result of our method and a faster execution time.

2. METHOD

An overview of the proposed method is shown in Fig 1. In this section, we first summarize the methods to build homography mapping to BEV plane and vehicle detection methods on traffic scenes as preliminaries. Then we present the proposed BEV-2D box geometry consistency-based 3D bounding box estimation method.

2.1. Mapping to BEV plane through homography

The planar homography relates to the transformation between two planes. Our goal is to find the homography transformation from the image plane to the BEV plane. This homography matrix is more accessible to obtain than camera calibrations and can be derived in several ways. In the case where camera calibration information is available, we can derive the homography matrix by projective mapping:

$$H_{bev}^{img} = MK \begin{pmatrix} R & T \end{pmatrix} Q_{4\times 3} \tag{1}$$

Where (x_{bev}, y_{bev}) denotes pixel on the BEV plane. (u, v) denotes pixel on the image plane. M is a matrix to define the scale, transformation and rotation between the BEV plane and the real-world coordinates. This can be defined freely by users. K is the intrinsic and R, T are extrinsic. Q is 4×3 matrix to reduce the height dimension in perspective transformation.

Unfortunately, in most circumstances, calibration is not available in traffic surveillance scenes. Some homography estimation methods are proposed in the state-of-the-art, for example, [10] extracts vertical vanishing point and ground plane vanishing line (horizon) in the image, then to further estimate homography to the BEV plane.

Otherwise, homography can be solved by using corresponding points of two planes by Direct Linear transformation (DLT). In this case, some human annotation work is required. This annotation can be correspondences of the image pixels and landmarks on maps. In the worst cases where even a map is also not available, we can make use of some road prior to define homography, for example, annotating a quadrilateral defined by lane endpoints that correspond to a rectangle on the BEV plane. The scale can be determined following road construction norms, such as the pre-defined road width, lane spacing distance.

2.2. 2D vehicle detection

Our approach necessitates combining with a 2D vehicle detector. Vision-based vehicle object detection is divided into traditional machine vision methods and deep learning methods. Traditional machine vision methods use vehicle motion [11] or background modeling [12] to separate it from a fixed background image. The use of deep CNN has achieved success in the field of vehicle object detection. CNNs have a strong ability to learn image features and can perform classification and bounding box regression tasks, e.g., Faster RCNN [13] and YOLO [14].

2.3. 3D bounding box estimation

First, we build a BEV bounding box representation. The four BEV corners can be expressed respectively as $(\pm\frac{l}{2}, \pm\frac{w}{2})$ in the vehicle-centered coordinate. Then the corners transformed in the BEV plane can be formulated as:

$$\begin{pmatrix} x_{bev} \\ y_{bev} \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \end{pmatrix} \begin{pmatrix} \pm\frac{l}{2} \\ \pm\frac{w}{2} \\ 1 \end{pmatrix} \tag{2}$$

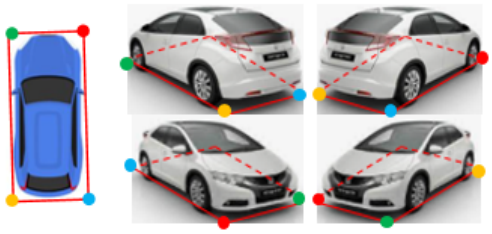

Fig. 2. The four possible BEV corner and 2D box side correspondence cases.

where θ is the rotation and t_x, t_y is the translation from vehicle center to the BEV plane origin. In our approach, we consider cars, vans, trucks, and buses as different categories. We assume the low variance of the vehicle dimensions of the same category. Similar to [8], we take the statistical average size of each category for car length l and width w.

Then we establish image plane 2D bounding box and BEV bounding box relation. We use the fact that the perspective projection of the 3 out of 4 corners of the BEV box touches respectively the left, right and bottom side of the 2D detection bounding box to build geometry consistency relation (see Fig 2). We make three equations following this constraint. For example, the projection of the right front corner of the BEV box touches the left side of the 2D bounding box. This derives the following equation:

$$X_{left} = \left[H_{bev}^{img} \begin{pmatrix} x_{bev} \\ y_{bev} \\ 1 \end{pmatrix} \right]_x \tag{3}$$

Where $[\cdot]x$ refers to the x coordinate of the homography projected point. X_{left} is the x coordinate of the left side of the detection bounding box. Similar equations can be derived for the other sides of the 2D bounding box with X_{right} and Y_{down}.

Then we can solve the t_x, t_y, θ with the three equations. Since the equations contain trigonometric functions, nonlinear optimization methods, e.g., the Levenberg-Marquardt algorithm could be applied. To ensure a favorable solution, we use the projection of the 2D box center position on the BEV image and the 2D center trajectory direction as optimization initial start point. The trajectory can be easily derived by detection association across frames.

There are four possible cases for the correspondence between the four BEV corners and the three 2D box sides (see Fig 2). The existing methods [4, 5, 15] usually predict the viewpoints from a learned model. In our approach, we propose to solve the equations for all the 4 possible cases. The wrong corner and box side correspondence relation leads to false orientation estimation. So among the four solutions, we choose the one whose orientation is the closest to the trajectory orientation as the final solution.

However, the solved orientation easily jitters due to unstable bounding box size. The trajectory is formed by the projected 2D centers. The orientation tangent to the trajectory suffers from the variation from the 2D box distortion under the BEV plane. Since the noise sources of these two cases are different, we found that it's effective to average the orientation tangent to the trajectory and the solved orientation as the final orientation to stabilize the result. The final BEV box corners can be further calculated using solved t_x, t_y and the averaged orientation. Finally, the projected 3D box is completed by approximating the projected box height by the vertical distance between the upper side of the 2D box and the last projected BEV corner.

2.4. Implementation details

In this work, we use the YOLOv3 [16] network for 2D vehicle detection. The detection model is trained on MS COCO dataset [17]. We use SORT [18] method for detection association across frames. In this experiment, for the very first frame that a vehicle appears in the image, we extract the trajectory direction from a pre-computed statistical direction field on the appearance position of the scene.

3. EXPERIMENTS

In this section, we provide an analysis of the proposed approach with empirical experiments. We conclude with qualitative as well as quantitative results on two different experiment settings on the Ko-PER dataset.

3.1. Dataset

We evaluate our approach on Ko-PER dataset [19]. Laserscanner and video camera data are gathered at a busy urban intersection. The 3D bounding boxes are annotated on 4833 frames of two cameras of different viewpoints of this intersection, which is suitable for evaluating our method. The camera calibrations and the map of the intersection are also provided in the dataset. We obtain the homography by annotating the correspondences between the map and the image pixels.

3.2. Quantitative results based on 2D detector

Following [6], we use the Average Precision (AP) as the evaluation metric. Two different criteria are applied for AP calculation. The first uses IoU with ground truth BEV boxes and the second uses the offset of the center predicted BEV box to the ground truth boxes. To notice that we exclude some regions on the BEV image where ground truth boxes are missing due to the limited laserscanner field of view. In the experiments, we ignore the detections of parked vehicles that are not annotated in the dataset.

We compare our approach to the previous methods in the literature on the Ko-PER benchmark. The evaluation result

Methods	Average precision(%) $IoU \geq 0.5$	$d \leq 0.5l$	fps
IPM detection [15]	65.67	71.96	-
dual tailed r-box [6]	**82.44**	91.20	29.6
ours	70.53	**92.16**	**36.4**

Table 1. 3D bounding box estimation result on Ko-PER compared to the literature. d is the distance between prediction and ground truth centers. For ours, the YOLOv3 is executed on GPU and the 3D estimation method is performed on CPU.

Methods	mean IoU	mean d (m)	mean d/l ratio
basline	0.260	2.029	0.428
ours	**0.706**	**0.466**	**0.095**

Table 2. 3D bounding box estimation result on Ko-PER compared to the baseline.

is shown in the Tab 1. Our approach outperforms [15] and gets a comparable result to [6]. The lower AP score for IoU criteria can be explained by the case with a large difference between the real vehicle dimensions and the used pre-defined size. In [6], a more complex dual-branched CNN is applied to extract features on both original and BEV images and further performs detection on combined features. Compared to these deep learning-based methods, the advantage of our approach is low computation resource demand. In terms of frame rate, our method executes 22% faster than [6]. The computation time can be even further reduced by replacing YOLO with a lighter method like background subtraction, making it easier to implement the 3D estimation application on the camera side. Furthermore, our approach does not require any training procedure and 3D annotations.

3.3. Quantitative results based on 2D ground truth

The previous section evaluated the combination of our approach and the YOLOv3. We isolate our approach for evaluation by using the 2D ground truth at the place of the 2D detection. Therefore, for each 2D ground truth, we output a BEV box. The AP is no longer suitable as an evaluation measure. So we adopt the mean IoU, the mean offset distance, as well as the mean offset and ground truth car length ratio as metrics. We set up a baseline method for comparison. The baseline took the projection of the 2D center as the BEV box center and used the trajectory tangent direction to determine the corners of the BEV box. The result is shown in Tab 2. Our approach achieves 43% and 1.6m improvements on mean IoU and mean offset distance compared to the baseline, which demonstrates the effectiveness of our method.

3.4. Qualitative results

In addition to Ko-PER, we perform tests on videos on the CIDAS dataset [20]. Fig.3 shows the qualitative estimation

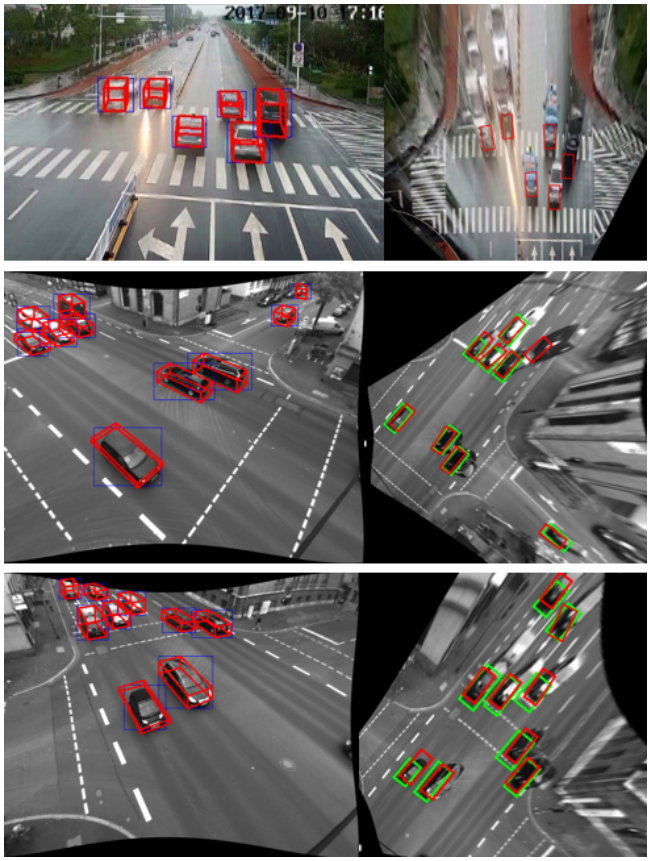

Fig. 3. Examples of 3D bounding box estimation. The first row is from the CIDAS dataset. The second and third rows are two different camera views in the Ko-PER dataset. YOLOv3 detection results are shown in blue. Our 3D bounding box and BEV box predictions are in red. The ground truth is in green.

results of bounding boxes on the BEV and 3D bounding box on the original image based on the YOLO detector. The parameter settings are the same for all experiments. However, different ways are applied to obtain homography matrix. For the CIDAS dataset, we use the lane width prior to annotate point correspondences. This shows that our methods can be applied without camera calibrations. In addition, our approach shows good generalization ability thanks to the use of a geometry constraint instead of a model learned from data.

4. CONCLUSION

In this paper, we introduced a 3D bounding box estimation method based on 2D object detection and homography without using 3D annotations and camera calibration. We leveraged the projected BEV box corners and the 2D detection geometry relations to resolve the vehicle 3D information. With quantitative and qualitative experiments, our approach shows a promising result and a faster execution time for the monocular 3D bounding box estimation task.

5. REFERENCES

[1] Stefan Lämmer and Dirk Helbing, "Self-control of traffic lights and vehicle flows in urban road networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 04, pp. P04019, 2008.

[2] Brendan Tran Morris and Mohan Trivedi, "Understanding vehicular traffic behavior from video: a survey of unsupervised approaches," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 041113, 2013.

[3] Mohammad Shokrolah Shirazi and Brendan Tran Morris, "Vision-based turning movement monitoring: count, speed & waiting time estimation," *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 1, pp. 23–34, 2016.

[4] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7074–7082.

[5] Jakub Sochor, Jakub Špaňhel, and Adam Herout, "Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance," *IEEE transactions on intelligent transportation systems*, vol. 20, no. 1, pp. 97–108, 2018.

[6] Minghan Zhu, Songan Zhang, Yuanxin Zhong, Pingping Lu, Huei Peng, and John Lenneman, "Monocular 3d vehicle detection using uncalibrated traffic cameras through homography," *arXiv preprint arXiv:2103.15293*, 2021.

[7] Markéta Dubská, Adam Herout, Roman Juránek, and Jakub Sochor, "Fully automatic roadside camera calibration for traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1162–1171, 2014.

[8] Markéta Dubská, Adam Herout, and Jakub Sochor, "Automatic camera calibration for traffic understanding.," in *BMVC*, 2014, vol. 4, p. 8.

[9] Sung Chun Lee and Ram Nevatia, "Robust camera calibration tool for video surveillance camera in urban environment," in *Proceedings of the IEEE CVPR Workshops*. IEEE, 2011, pp. 62–67.

[10] Syed Ammar Abbas and Andrew Zisserman, "A geometric approach to obtain a bird's eye view from an image," in *Proceedings of the IEEE/CVF ICCV Workshops*, 2019, pp. 0–0.

[11] Ya Liu, Yao Lu, Qingxuan Shi, and Jianhua Ding, "Optical flow based urban road vehicle tracking," in *2013 ninth international conference on computational intelligence and security*. IEEE, 2013, pp. 391–395.

[12] R Manikandan and R Ramakrishnan, "Video object extraction by using background subtraction techniques for sports applications," *Digital Image Processing*, vol. 5, no. 9, pp. 435–440, 2013.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[15] Youngseok Kim and Dongsuk Kum, "Deep learning based vehicle position and orientation estimation via inverse perspective mapping image," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 317–323.

[16] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[18] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[19] Elias Strigel, Daniel Meissner, Florian Seeliger, Benjamin Wilking, and Klaus Dietmayer, "The ko-per intersection laserscanner and video dataset," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014, pp. 1900–1901.

[20] Qiang Chen, Yong Chen, Ola Bostrom, Yehong Ma, and Eryong Liu, "A comparison study of car-to-pedestrian and car-to-e-bike accidents: data source: the china in-depth accident study (cidas)," Tech. Rep., SAE Technical Paper, 2014.