Estimation of Vehicle Pose and Position with Monocular Camera at Urban Road Intersections

Citation: Yuan, Jin-Zhao, et al. "Estimation of Vehicle Pose and Position with Monocular Camera at Urban Road Intersections." Journal of Computer Science and Technology, vol. 32, no. 6, Nov. 2017, pp. 1150–61.

As Published: http://dx.doi.org/10.1007/s11390-017-1790-3

Publisher: Springer US

Persistent URL: http://hdl.handle.net/1721.1/113347

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Yuan JZ, Chen H, Zhao B *et al.* Estimation of vehicle pose and position with monocular camera at urban road intersections. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(6): 1150–1161 Nov. 2017. DOI 10.1007/s11390-017-1790-3

Estimation of Vehicle Pose and Position with Monocular Camera at Urban Road Intersections

Jin-Zhao Yuan[1], Hui Chen[1,*], Bin Zhao[2], and Yanyan Xu[3,*]

[1] *School of Information Science and Engineering, Shandong University, Jinan 250100, China*

[2] *Beijing Xuanlongding-xun Technology Co., Ltd., Beijing 100041, China*

[3] *Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge 02139, U.S.A.*

E-mail: yuanjinzhaosdu@163.com; huichen@sdu.edu.cn; zhaobin_cn@mail.sdu.edu.cn; yanyanxu@mit.edu

Received June 24, 2017; revised September 12, 2017.

Abstract With the rapid development of urban, the scale of the city is expanding day by day. The road environment is becoming more and more complicated. The vehicle ego-localization in complex road environment puts forward imperative requirements for intelligent driving technology. The reliable vehicle ego-localization, including the lane recognition and the vehicle position and attitude estimation, at the complex traffic intersection is significant for the intelligent driving of the vehicle. In this article, we focus on the complex road environment of the city, and propose a pose and position estimation method based on the road sign using only a monocular camera and a common GPS (global positioning system). Associated with the multi-sensor cascade system, this method can be a stable and reliable alternative when the precision of multi-sensor cascade system decreases. The experimental results show that, within 100 meters distance to the road signs, the pose error is less than 2 degrees, and the position error is less than one meter, which can reach the lane-level positioning accuracy. Through the comparison with the Beidou high-precision positioning system L202, our method is more accurate for detecting which lane the vehicle is driving on.

Keywords vehicle pose and position estimation, road sign detection, homograph matrix, road intersection, urban environment

1 Introduction

Precise estimation of vehicle ego-localization is currently one of the key subjects of smart vehicle research. With a better knowledge of location and pose, the central control unit of the smart vehicle could react accurately when facing emergencies. Normally, global positioning system (GPS) is a mature and widely used technique for localization in advanced driver assistance systems (ADAS)[1]. However, GPS can only manage an accuracy of 3∼5 meters and has a risk of signal loss which is serious in the urban environments. Meanwhile, the accuracy of 3∼5 meters can neither determine the lane nor distinguish the position of the front and rear cars. The fixed radio base station can locate the moving vehicles[2], but it highly depends on the infrastructure and is costly. With the development of the automatic driving technology, different kinds of autonomous driving cars with very expensive sensors have already driven on the road. They are sonar[3], laser scanners[4], inertial measurement units (IMU), cameras[5], position and orientation sensors[6] or various combinations of the above[7-8]. Although these sensors can reach a high accuracy, they still cannot perform well for complex traffic scenes at intersections due to congested traffic conditions with serious vehicle occlusions and indeterminate lanes. Meanwhile, various combinations of sensors, which are even more expensive than the vehicle itself, are not applicable as widely as GPS.

High-precision localization using cameras can re-

Regular Paper

Special Section of CAD/Graphics 2017

This work was supported by the Key Project of National Natural Science Foundation of China under Grant No. 61332015 and the Natural Science Foundation of Shandong Province of China under Grant Nos. ZR2013FM302 and ZR2017MF057.

*Corresponding Author

duce the system cost significantly, but this usually needs high-precision digital maps[9-11] or pre-constructed image databases for matching[12-13]. Uchiyama *et al.*[12] achieved the ego-localization using a matching database and a binocular camera, where the synchronization and the stereo matching of binocular vision are too difficult to guarantee the positioning accuracy. Wong *et al.*[13] used an on-board monocular camera and an image sequence database to determine the actual location of the vehicle on road. Their algorithm relies on a large and complex database, including the scenes along and around the road. Any change of the road scene will affect the accuracy of positioning.

Up to the present, vision-based lane detection is already used to localize the moving vehicle. Two classic methods are found in the literature: the feature-based techniques[14-15] and the model-based techniques[16-17]. Although these methods can detect lanes well in the case of slight occlusion, they are less feasible in the situation of traffic congestion, and they are not able to estimate the direction of the lanes either. Nedevschi *et al.*[18] proposed an ego-localization method using landmarks at the intersections, but it fails when the landmark is covered by the crowded vehicles.

There is an intensive literature on estimating rotation and position of a camera from the points of the space corresponding to its image points, often referred to as Perspective-n-Point (PnP) problem. However the methods presented in [19-20] require matching 2D-3D pairs and selecting non-coplanar points as the prerequisite, often involving considerable amount of manual interactions.

The well-known 2D feature points matching algorithm, SIFT (scale-invariant feature transform)[22], has a high matching accuracy, but the calculation is time-consuming, which limits its application in real-time vehicle pose estimation for planar road sign as in our case.

In this paper, we propose a vehicle pose and position estimation method in the complex traffic scene at intersections using an on-board monocular camera and a simple pre-constructed database. Moreover, the GPS in our experiments is the only one used to provide a rough position of the vehicle before calculating the vehicle's pose and position. The contributions of our paper mainly include: 1) a pose and position estimation method using monocular camera is proposed for driver-assistance and autonomous driving; 2) a rectangle object detection method is proposed to detect the road sign accurately in cases of weak light and partial occlusion; 3) an accurate road sign's vertex extraction algorithm based on Hough transform is proposed to calculate the exact plane homograph matrix, by which the vehicle's pose and position at the intersection can be obtained accurately.

This paper is organized as follows. In Section 2, we give a brief overview of the proposed method. The structure of the database is introduced in more detail in Section 3. The detailed process of road sign detection and the vertex extraction algorithm based on Hough transform are described in Section 4 and Section 5, respectively. The process of pose and position estimation is shown in Section 6. We discuss the experimental results in Section 7 before the conclusions in Section 8.

2 Overview of Proposed Method

This paper presents a method for vehicle pose and position estimation in complex traffic scenes at urban road intersections with the consideration of traffic jam and lane occlusion. The estimation of vehicle pose and position is realized by using the road sign ahead of the intersection, which has two main characteristics: 1) a blue color background; 2) a rectangular shape with a relatively fixed range of width-to-height ratio, as shown in Fig.1.

Fig.1. Road signs at different road intersections.

Usually, every intersection has one road sign ahead for a fixed direction in China. As current GPS facilities have an accuracy of less than five meters, we can determine which intersection the vehicle is traversing. The GPS information is associated with the information of the road sign stored in the pre-constructed database. In this way, the road signs can be targeted by the GPS information. After confirming the intersection, the rectangular road sign will be detected from the input images recorded by the moving on-board camera using three constraints. Then, the coordinates of four vertices can be obtained using proposed extraction algorithm based on Hough transform. We calculate a plane homograph matrix between the actual size of the corresponding road sign in the real world and the four vertices detected above. The on-board camera's pose

and position can be calculated with the plane homograph matrix in the road sign coordinate system. The flow chart of our method is shown in Fig.2.

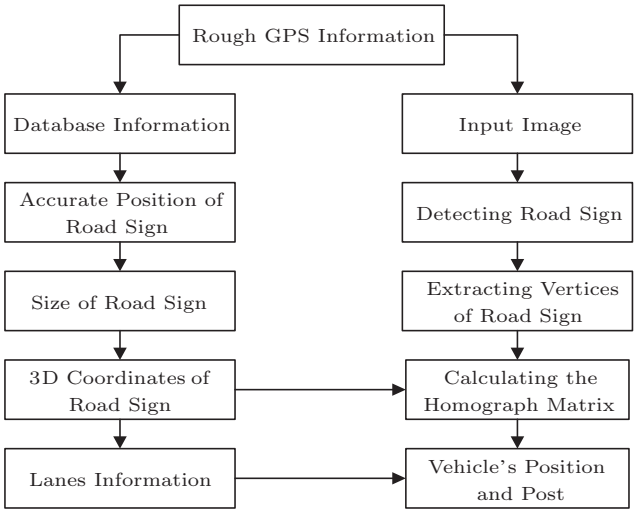


Fig.2. Flow chart of the proposed method.

3 Database Construction

In this paper, the database consists of three parts associated with road signs located at all intersections. The first part is the accurate positions of road signs measured by high-precision RTK-GPS Compstar CC20[23]. The second part is the sizes of the road signs. The third part is the lane information including the width and the index near the road sign. Roads at different intersections have different numbers and widths of lanes. For simplicity, we define the road sign coordinate system as the world coordinate system, and let the original point be fixed on the center of the road sign. Therefore, the size of the road sign can be converted to the 3-dimensional (3D) coordinates under the world coordinate system. The number and width of the lanes can be quantified as the coordinates under the road sign coordinate system, as shown in Fig.3.

The complete database is a series of numbers associated with different road signs including the signs' accurate position, size, amount and width of lanes. These data are indexed by the accurate position. Some samples of the database are presented in Table 1.

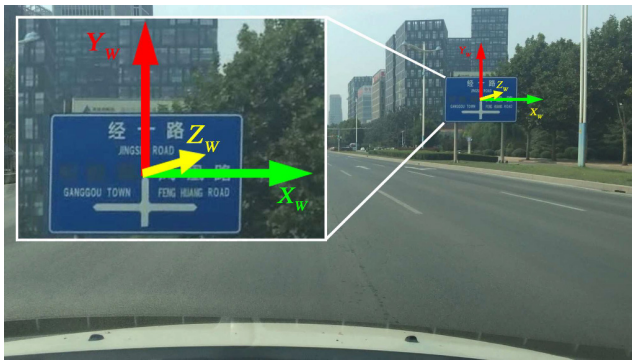


Fig.3. Road sign coordinate system (X_W, Y_W, Z_W).

4 Road Sign Detection

At present, sign detection methods mostly rely on the threshold segmentation based on different color spaces, or feature point extraction and matching algorithms like SIFT or SURF[24]. On this basis, the signs are classified and identified by machine learning methods like random forests[25] or SVM (support vector machine)[26]. The SIFT algorithm is time-consuming and requires a large storage of sign images at different distances in the database, increasing the overhead of the database greatly. The detection and classification methods based on SVM or other algorithms also need the corresponding color threshold segmentation preprocessing in the hue-saturation-value (HSV)[26], CIELUV (LUV)[27], or hue-saturation-intensity (HSI)[28] space. In this paper, to overcome the high computation cost, we do not classify or identify the target area. Instead, we introduce another two constraints to eliminate the interference area and obtain the target sign based on the threshold segmentation in the HSV space.

Table 1. Examples of Database from 3 Continuous Intersections on Jingshi Road

Position of Road Sign	Intersection (Direction)	Property of Lane-7	Position of Lane-7
E117.157776/N36.669938	Jingshi+Aotidong (West to East)	2, 3	(20, 24)
E117.133895/N36.665991	Jingshi+Shunhua (West to East)	1	(20, 24)
E117.147572/N36.666628	Jingshi+Fenghuang (West to East)	1	(20, 24)

Note: The road signs at the three intersections have the same size (width × height): 5 m × 3 m. For lane-1∼lane-6, the properties and positions of each level of the three intersections are the same: lane-1: property: 0, 4; position: (−11, −7.5); lane-2: property: 1, 2; position: (−7.5, −4); lane-3: property: 1; position: (4, 8); lane-4: property: 1; position: (8, 12); lane-5: property: 1; position: (12, 16); lane-6: property: 1; position: (16, 20). For properties, 0 means right turn, 1 means straight, 2 means left turn, 3 means u-turn, and 4 means bus lane.

In our method, when the GPS roughly informs the vehicle approaching a certain road sign in our database (e.g., x meters from the sign), the road sign detection module starts to detect this road sign from the current image recorded by the on-board camera. Meanwhile, the locating of target road signs is limited by three predefined constraints: 1) HSV threshold, 2) aspect ratio, and 3) area size.

First, the input image is converted from RGB to HSV color space for a better representation of real-light pixel color characteristics. As the color of the road signs in the urban environment of most countries is blue, we design the following constraints for the three channels of HSV:

1) a hue value $200 < H < 280$;

2) an intensity value $0.35 < V < 1$;

3) a saturation value $0.35 < S < 1$.

The input images are binarized according to the above thresholds, and then we perform the morphology processing to reduce the discontinuous regions. The peripheral contours of the candidate regions can be obtained from the binarized images. However, the above threshold range contains a lot of noise like the license plates, blue color billboards, vehicles, buildings, and the other various plane objects. Considering the above kinds of noise have different aspect ratios, and road signs always maintain a certain ratio, we set the second constraint with a ratio value $1 < r < 2$.

To improve the detection accuracy, the third constraint is introduced. Note that the area of the noise regions is much smaller than that of the target road sign, and the number of pixels in the region of the road sign is constrained to between 3 000 and 240 000 (the size of the input image is $1\,920 \times 1\,080$). Finally, the largest remaining region is selected as the target road sign.

The detection process and results in different scenarios are presented in Fig.4. The results imply that our detection method can be better adapted to the scenarios of noise, night, dawn, blocked by lamps, and blocked by vehicles. Under the night scenario, using the headlight can obtain a relatively good detection result within a limited distance. The road conditions only in mainland China are referred in this paper. With regard to the applicability in other countries and regions, it does only need to appropriately adjust the value of hue (H) in the HSV space according to the road conditions in these areas.

5 Accurate Extraction Algorithm of Road Sign's Vertices

In order to calculate the homograph matrix between the road sign plane in the real world and that in the

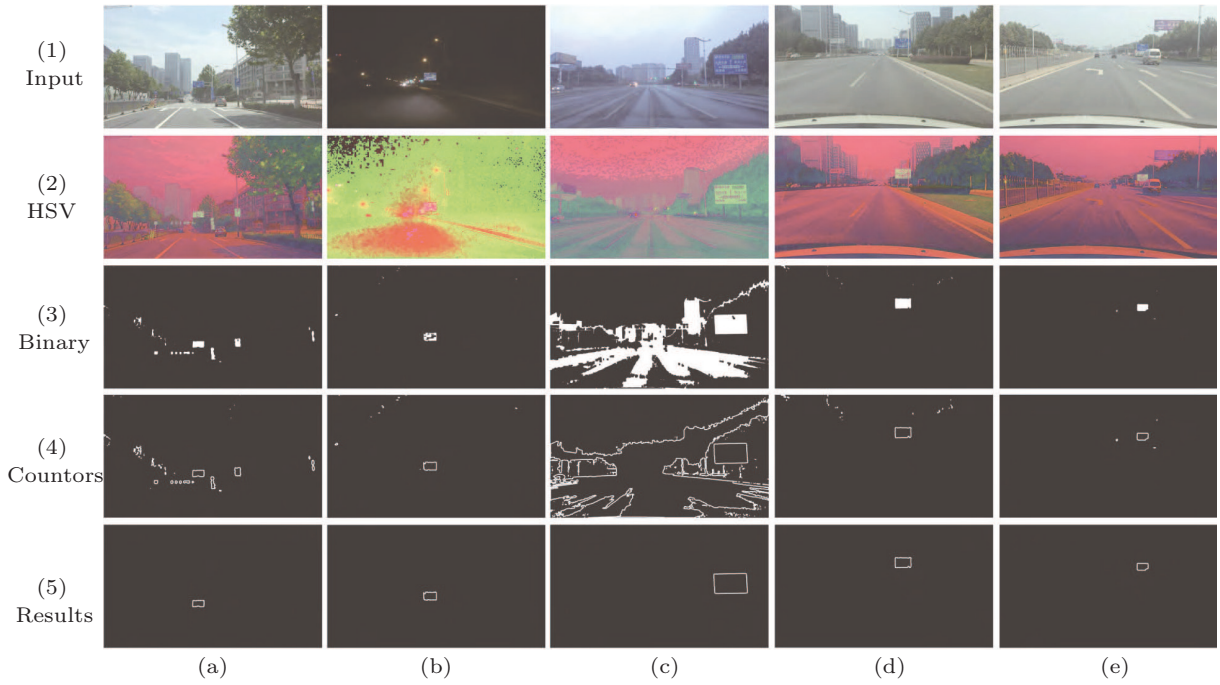


Fig.4. Detection process and results under different scenarios. (a) Noise. (b) Night. (c) Dawn. (d) Blocked by lamps. (e) Blocked by vehicles.

image, it is necessary to extract the four vertices' coordinates of the road sign in the image plane in the case that the vertices' 3D coordinates of the actual one have been known. The approximate contour of the road sign can be obtained by the detection step. However, since the contour is not a standard quadrilateral, the accurate vertex coordinates cannot be obtained directly. In order to obtain the accurate road sign's vertex coordinates of the input image, an accurate extraction algorithm of the road sign's vertices based on Hough transform algorithm is presented. The flow chart of the algorithm is shown in Fig.5.

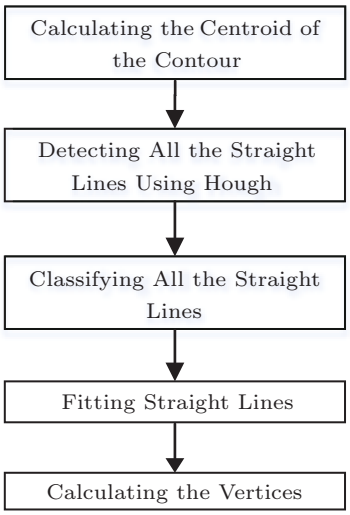


Fig.5. Flow chart of the accurate extraction algorithm of the road sign's vertices.

5.1 Straight Lines Detection

Firstly, we calculate the centroid coordinates (x_p, y_p) of the contour point set using (1).

$$x_p = \frac{\sum_N x_i}{N}, \quad y_p = \frac{\sum_N y_i}{N}, \tag{1}$$

where N is the amount of the pixels constituting the contour, (x_i, y_i) is the pixel's coordinate.

Secondly, all the straight lines are detected in the road sign peripheral contour using Hough transform with the polar resolution $\delta\rho = 0.5$. The detected lines can be classified to four categories including upper, lower, left and right lines sets centered on the centroid, as shown in Fig.6.

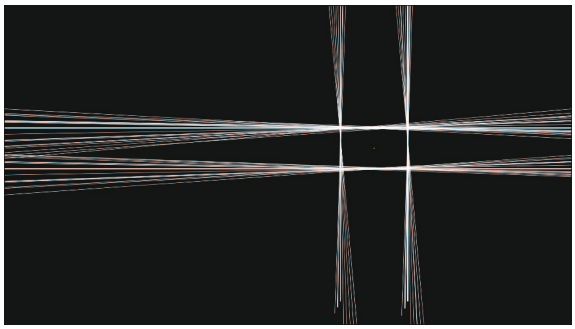


Fig.6. Straight lines detection of the contour.

5.2 Straight Lines Classification and Fitting

The lines detected by Hough transform can be represented using (2).

$$\rho = x \sin\theta + y \sin\theta, \tag{2}$$

where parameters ρ and θ denote the polar and the angle of the line in the Hough space, respectively. Then we substitute the abscissa x_p or ordinate y_p of the centroid point into each straight-line equation to calculate y_i' and x_i' using (3).

$$y_i' = \frac{\rho_i - x_p \cos\theta_i}{\sin\theta_i}, \quad x_i' = \frac{\rho_i - y_p \sin\theta_i}{\cos\theta_i}, \tag{3}$$

where parameters ρ_i and θ_i correspond to the polar and the angle of each line detected above, respectively. It is easy to determine the lines' classification by the slope k and the comparison of y_p and y_i' or x_p and x_i', where $k = -\cot(\theta)$. Afterwards, we traverse all the detected straight lines using the following conditions to determine the lines' classification.

• If $y_i' < y_p$ and $k < -1 \| k > 1$, the line belongs to the upper set.

• If $y_i' > y_p$ and $k < -1 \| k > 1$, the line belongs to the lower set.

• If $x_i' < x_p$ and $-1 < k < 1$, the line belongs to the left set.

• If $x_i' > x_p$ and $-1 < k < 1$, the line belongs to the right set.

With these conditions, the above method can quickly and accurately classify all detected lines to four groups (i.e., upper, lower, left, and right) and get a perfect result as good as the clustering algorithm which has a time complexity of $O(n^2)$. Then, the lines are associated with each group using parameters ρ and θ. Due to the invariant rotation, we directly average the values of ρ of the lines in each group. Moreover, we average the values of θ for the upper and lower groups. However, for the left and right groups, since the value of θ is

in two intervals with a large span, averaging the value of θ directly results in the deflection of line detection, as shown in Fig.7. In order to obtain accurate fitting value, θ of the left and the right group is rotated by $\pi/2$ counterclockwise firstly using (4) before being averaged, and rotated by $\pi/2$ clockwise after being averaged, as shown in (5).

$$\theta_{\text{left}} = \theta_{\text{left}} + \frac{\pi}{2}, \theta_{\text{right}} = \theta_{\text{right}} + \frac{\pi}{2}, \tag{4}$$

$$\theta_{\text{left}} = \frac{\sum_{N_{\text{left}}} \theta_i}{N_{\text{left}}} - \frac{\pi}{2}, \theta_{\text{right}} = \frac{\sum_{N_{\text{right}}} \theta_i}{N_{\text{right}}} - \frac{\pi}{2}, \tag{5}$$

where N_{left} and N_{right} correspond to the number of lines in the left and the right group respectively. ρ_{left} and θ_{left} are the fitting results of left lines, and ρ_{right} and θ_{right} are the fitting results of right lines.

The coordinates of the four vertices of the road sign can be easily obtained by the intersections of the four straight lines. The process and results of the proposed algorithm are shown in Fig.8. To verify the accuracy of the vertices coordinates calculated by the proposed algorithm, we visualize the error analysis by restoring the image with a planar perspective transformation matrix and performing a difference comparison. In doing so, an extended orthogonal image of the actual road sign is constructed, and the vertices' coordinates of the orthogonal road sign can be measured.

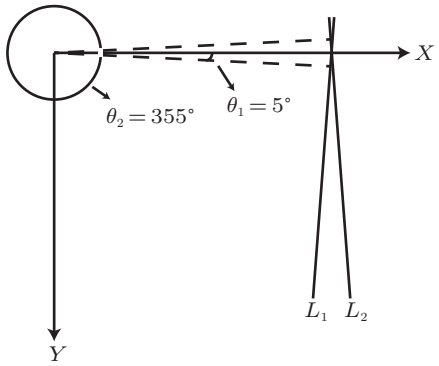


Fig.7. Example of line deflection for the left and right groups. θ_1 and θ_2 are the angles of lines L_1 and L_2 in one of the groups respectively.

With the corresponding vertices coordinates, the eight parameters in the perspective transformation matrix $\boldsymbol{M}$ can be calculated, which warps the orthogonal road sign into the one detected from the input image. The restored image can be obtained by warping the orthogonal one using the matrix $\boldsymbol{M}$. Fig.9(a) shows two groups of overlap comparison between the input image and the warped orthogonal image marked as a, b, c, and d. We can see that the road sign in the input image overlaps the orthogonal one well. Fig.9(b) gives the image difference between the input image and the orthogonal one. The result shows that $\boldsymbol{M}$ is accurate, and the vertices coordinates calculated by the proposed method have high precision.

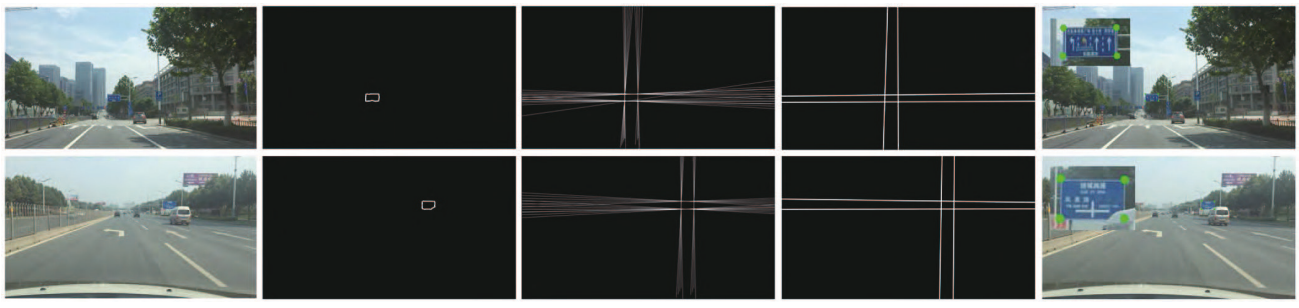


Fig.8. Pipeline and the results of traffic sign detection. (a) Input. (b) Contours. (c) Lines set. (d) Lines fitting. (e) Results.

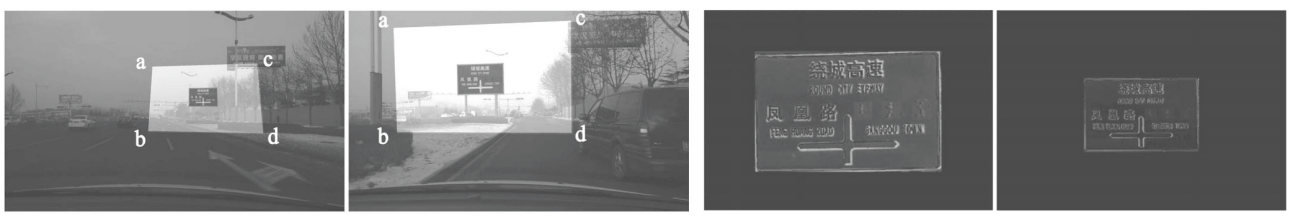


Fig.9. Overlap comparison. (a) Overlap between the input image and the warped orthogonal one. (b) Image difference (zoomed).

During the analysis of the continuous 100 images, the abscissa and the ordinate of the vertices calculated by the algorithm were compared with the real abscissa and ordinate respectively. As shown in Fig.10, the average error of the abscissa (blue points) and the ordinate (red points) is 2.592 and 2.734 pixels respectively.

Moreover, we compared the processing time per frame and false positive of the sign detection by SIFT, SVM, and the proposed method. 200 continuous frames were used for this comparison. The distance from these frames to the sign ranged from 50 to 100 meters and the image size was $1\,920 \times 1\,080$. As shown in Table 2, our method outperforms the methods using SIFT and SVM.

The proposed vertices extraction algorithm detects the four vertices by fitting all of the points in the contour, which ensures that the accuracy of the vertices' coordinates is high enough. Compared with SVM and other machine learning methods, the proposed detection method does not need much calculation and training, and it is simpler and more effective. Compared with the classic methods like SIFT and SVM, our method is much faster and more robust because of the priori-assumed matching of the four vertices.

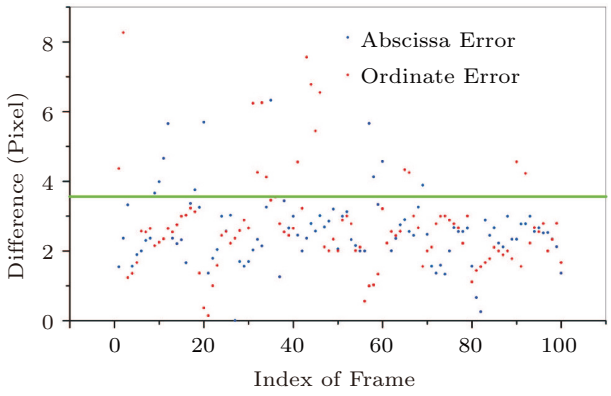


Fig.10. Error analysis of vertices' accuracy.

Table 2. Comparison of Road Sign Detection Results

Number of Images	Method	Average Processing Time (ms/frame)	Number of Correct Detection	Correct Rate (%)
200	SIFT	3 201.26	169	84.50
200	SVM	912.15	159	79.50
200	Our method	121.71	181	90.50

6 Estimation of Vehicle Pose and Position

As shown in Fig.3, the size of the road sign can be converted into 3D coordinates under the world coordinate system, where $Z_W = 0$. The four vertices of the road sign can be obtained by the method described in Section 4. With the corresponding four vertices, the plane homograph matrix between the actual road sign and the input image can be calculated by (6).

$$\boldsymbol{p} = \boldsymbol{H}\boldsymbol{P}_W, \tag{6}$$

where $\boldsymbol{H}$ is the homograph matrix, and $\boldsymbol{p}$ and $\boldsymbol{P}_W$ are vertices' homogeneous coordinates of the input image and the actual object plane, respectively.

6.1 Estimation of Vehicle Pose

$\boldsymbol{H}$ is a 3×3 matrix. Let

$$\begin{cases} \boldsymbol{H} = (\boldsymbol{h}_1\ \boldsymbol{h}_2\ \boldsymbol{h}_3) = s\boldsymbol{K}\,(\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{t})\,, \\ \boldsymbol{W} = (\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{t})\,, \end{cases} \tag{7}$$

where $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are rotation vectors, $\boldsymbol{t}$ is the translation vector, $\boldsymbol{K}$ is the camera's intrinsic parameters pre-calibrated using Camera Calibration Toolbox for Matlab, and s is a scale factor generated by the depth value under the camera coordinate system. We can get the rotation and translation matrix $\boldsymbol{W}$ using (6) and (7), as defined in (8).

$$\boldsymbol{r}_1 = \frac{\boldsymbol{K}^{-1}\boldsymbol{h}_1}{s}, \boldsymbol{r}_2 = \frac{\boldsymbol{K}^{-1}\boldsymbol{h}_2}{s}, \boldsymbol{t} = \frac{\boldsymbol{K}^{-1}\boldsymbol{h}_3}{s}. \tag{8}$$

$\boldsymbol{R}$ is structured as a 3×3 rotation matrix, i.e., $\boldsymbol{R} = (\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{r}_3)$. Since the rotation vectors are orthogonal to each other, $\boldsymbol{r}_3 = \boldsymbol{r}_1 \times \boldsymbol{r}_2$, where $\times$ represents the vector cross product. But putting the rotation vectors together simply cannot get the accurate rotation matrix, and the translation vector is also not accurate enough. In order to improve the accuracy, we need to perform singular value decomposition to $\boldsymbol{R}$ using $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$. Since $\boldsymbol{R}$ is an orthogonal matrix, $\boldsymbol{D} = \boldsymbol{I}$, where $\boldsymbol{D}$ means the sigular value decomposition matrix and $\boldsymbol{I}$ means the identical matrix, and then we have $\boldsymbol{R}' = \boldsymbol{U}\boldsymbol{I}\boldsymbol{V}^{\mathrm{T}}$, which is the camera rotation matrix we need. In order to describe the rotation matrix more intuitively, we convert $\boldsymbol{R}'$ to a 3×1 vector $\boldsymbol{r}$ using Rodrigues transform, and $\boldsymbol{r} = (\alpha\ \beta\ \phi)$. α represents the pitch angle of the vehicle, β represents the heading angle, and ϕ represents the roll angle. The three parameters can reflect the vehicle's pose veritably.

6.2 Estimation of Vehicle Position

The rotation and transformation matrix $\boldsymbol{W}$ is constituted by the rotation matrix $\boldsymbol{R}'$ and the translation matrix $\boldsymbol{T}$. $\boldsymbol{T}$ can be obtained using (9).

$$\boldsymbol{T} = (\boldsymbol{T}_1\ \boldsymbol{T}_2\ \boldsymbol{T}_3) = \boldsymbol{R}^{-1}\boldsymbol{W}, \tag{9}$$

where $\boldsymbol{T}$ is a 3×3 matrix, and $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, $\boldsymbol{T}_3$ are the column vectors of $\boldsymbol{T}$. The vehicle's translation vector $\boldsymbol{t}$ can be calculated by (10).

$$\boldsymbol{t} = (t_x\ t_y\ t_z)^{\mathrm{T}} = \frac{\boldsymbol{T}_3}{\|\boldsymbol{T}_1\|}, \tag{10}$$

where t_x, t_y, t_z represent the vehicle's position under the road sign coordinate system. By combining the lanes information pre-stored in the database, it is easy to obtain the vehicle's position on the road and which lane the vehicle is driving on.

Now, the six parameters α, β, ϕ, t_x, t_y, t_z have been obtained to determine the vehicle's pose and position. Compared with the classical PnP[20-21] algorithms which rely on non-coplanar control points and complex three-dimensional calculations, the four control points selected in our method are located on the same plane of the road sign, which simplifies the computation significantly by calculating the homograph matrix with the coplanar points instead of the non-coplanar points. Our pose and position estimation procedure is based on the theory of homograph matrix which is also used by Zhang's calibration method[29]. Our method extracts the plane target and control points automatically while Zhang's method requires manual interaction.

7 Experiments and Analysis

7.1 Experimental Environment

Experiments were conducted on the Jingshi Road which is an arterial road with bidirectional 14 lanes in Jinan, China. In order to verify the effectiveness of the proposed method, we chose four major intersections with a total of 24 lanes. The size of the road signs and the width of the lanes were measured to construct the database. The satellite image is shown in Fig.11, in which the green arrows represent the vehicle's driving path, and the blue rectangles represent the road signs with the size of 5 m × 3 m. The experiment camera Sony Exmor RS IMX145 was installed behind the windscreen of the moving vehicle, and the optical axis of the on-board camera was parallel to the vehicle's traveling direction. The camera was calibrated using Camera Calibration Toolbox for Matlab. The video was captured in RGB scale with a rate of 30 FPS and a size of 1 920 × 1 080 pixels. The vehicle with the camera and the virtual image of the intersection including the lanes' information are shown in the corners of Fig.11.

Fig.11. Satellite image of the experimental road.

The vehicle used in our experiments traveled along a fixed line in order to record the vehicle's trajectory. At the same time, we recorded the vehicle's real lateral position t_x on the road. The speed was controlled at 27 km/h, which is equal to 0.25 m/frame, and thus we can measure to obtain the real distance t_z to the road sign for every frame. The vehicle's actual heading angle β was recorded by the compass software. The parameters t_y, pitch angle α, and roll angle ϕ are ignored to analyze because they are negligible in the practical application.

The multi-sensor system used as reference is Beidou high-precision positioning system L202 including Beidou navigation, inertial measurement unit and difference modules[23]. L202 updates the positioning data every three seconds and provides lane-level positioning accuracy for vehicles. Fig.12 shows the equipment installed in the vehicle and the positioning data obtained by L202 for later comparison. Fig.12(c) is the zoomed positioning data point clouds of the white frame in Fig.12(b).

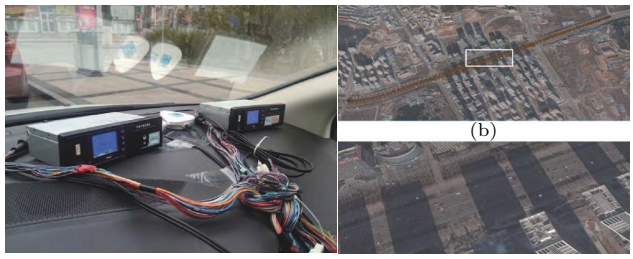


Fig.12. Beidou high-precision positioning system and output. (a) Equipment. (b) Bird's-eye view of positioning data. (c) Zoomed-in white frame in (b).

The results of the proposed method are shown in Fig.13. Some of the results calculated once per 10 frames were intercepted and shown in Fig.13(a), and the pose and position data are displayed on the top of the images. The results of the proposed method are plotted in the simulated road, as shown in Fig.13(b). It is obvious that the result matches the lanes well.

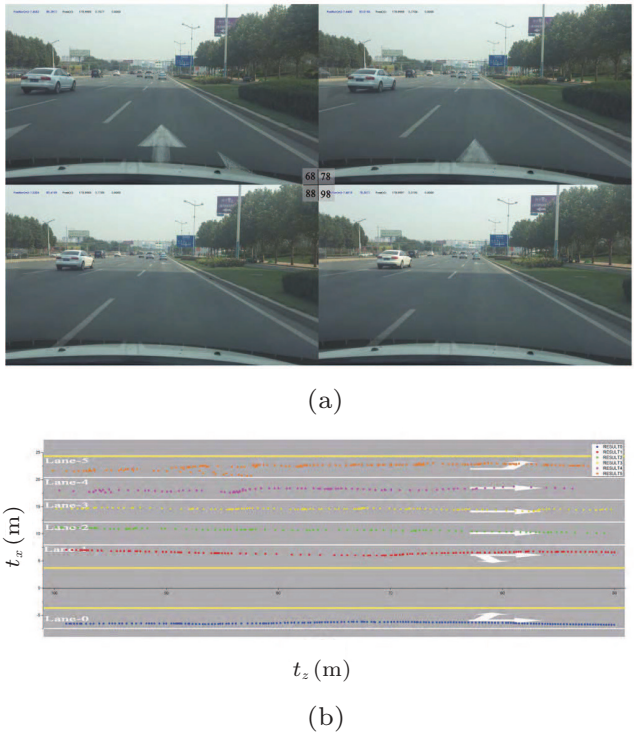


Fig.13. Results of the proposed method. (a) Road sign detection with vertices marked, pose and position data shown on the upper left corner. (b) Six groups of the results projected to the simulated road.

7.2 Experimental Data Analysis

To validate the results of the proposed method with ground truth, we selected 200 continuous frames ranged from 100 m to 50 m to analyze errors with the ground truth, as shown in Fig.14(a) for t_z. Fig.14(b) shows the error of parameter t_z for every frame against the ground truth. It indicates that t_z is more accurate when the distance to the road sign is closer. When the distance is larger than 150 m, the error of t_z may exceed 1 m. Since the road sign is too small to be detected, the estimation of pose and position is meaningless. Figs.14(c) and 14(d) show the estimated t_x and its error compared with the ground truth. The average error of t_x is less than 0.5 m. It demonstrates that the proposed method can achieve the lane-level positioning. Figs.14(e) and 14(f) show the analysis of the heading angle β. The smooth black curve in Fig.14(e) represents the ground truth. The changed curve direction indicates that the vehicle changes the lane. The purple curve presents the result of the proposed method, which shows that the pose estimation can reflect the actual driving attitude of the vehicle. Fig.14(f), the error analysis, shows that the angle error is between +2 degrees and −2 degrees.

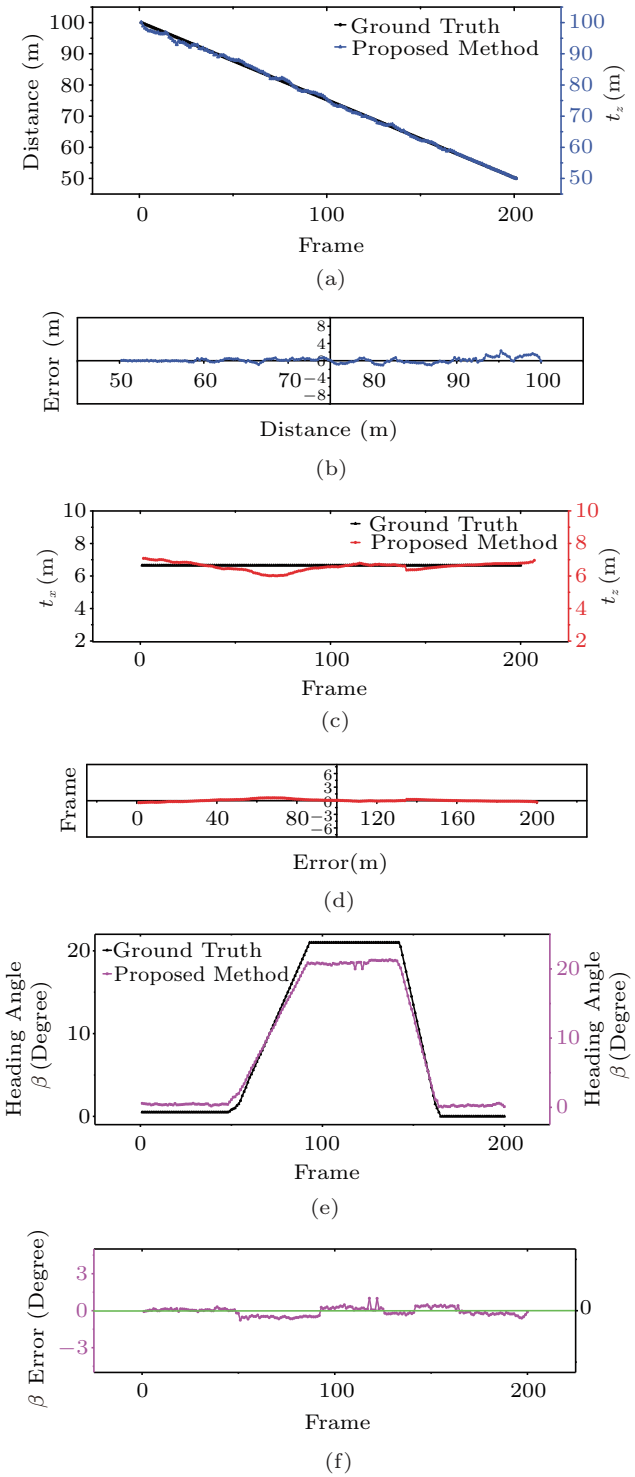
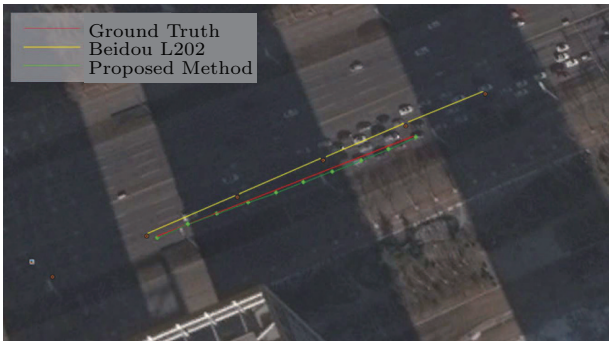


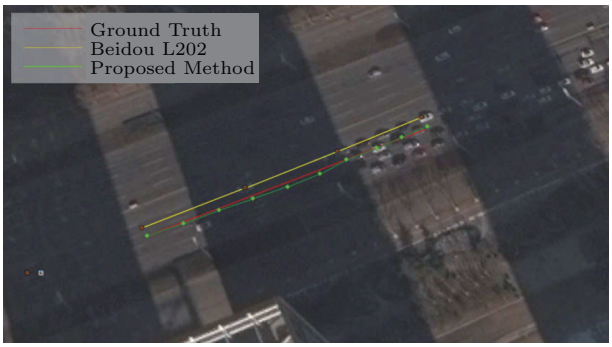
Fig.14. Results compared with ground truth and error analysis. (a) Result of t_z. (b) Error of t_z. (c) Result of t_x. (d) Error of t_x. (e) Result of t_y. (f) Error of t_y.

For a better comparison, we plot the results calculated by the proposed method and the results of Beidou L202 in the same satellite map, as shown in Fig.15. The red line is the ground truth, the yellow circular points are the results of Beidou L202, and

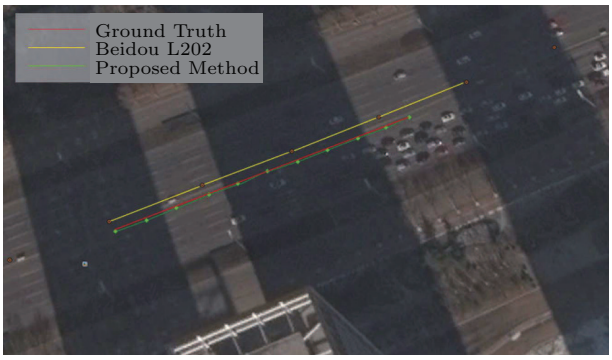
the green rhombus points are results of the proposed method, in which obtained data were calculated once per 10 frames. Figs.15(a)∼Fig.15(c) represent results in three different lanes respectively. It shows that Beidou L202 has relatively accurate positioning results when the satellites' signal is stable, but part of points still have an obvious deviation compared with the ground truth, and even some of the points are positioned on the opposite lane. Compared with the Beidou L202, the results of the proposed method are more consistent with the actual trajectory.



(a)



(b)



(c)

Fig.15. Comparison with Beidou L202 (distance of 100 to 20 meters to the road sign). (a) Results in lane-3. (b) Results in lane-5. (c) Results in lane-7.

7.3 Performance Analysis

Our algorithm is implemented on a PC with a 3.20 GHz Intel Core i5-3470 processor running Windows 10 operating system, using C/C++ and OpenCV library. The input image has a resolution of 1 920 × 1 080. The performance of the proposed method can be analyzed from two parts: 1) the road sign detection, and 2) the pose and position calculation. The average time of road sign detection and vertices extraction algorithm calculated from 300 images was about 122 ms, and the average time of the pose and position calculation was 23 ms. With the hardware acceleration or GPU parallel algorithm, the proposed method can effectively improve the efficiency and achieve real-time processing.

Most of the current navigation systems cannot inform which lane the vehicle is driving on. Associating the proposed system with the high-precision digital map, we can easily determine the vehicle's position in the intersection. The system can remind the driver in advance which lane the vehicle is driving on, especially at the busy intersections, important entrances and exits of the viaduct, where traffic jam and accidents are always caused by unwanted misjudgments. The accuracy of the navigation system can be further enhanced at the important intersections.

The current autonomous driving system cannot solve the problem of vehicles' ego-localization well in the occlusions common in intersections. The proposed method provides a new idea for ego-localization and can reduce the cost of the autonomous driving system effectively. Moreover, these expensive sensors cannot work very well facing the occlusions in the intersections. The proposed method will provide more stable and reliable assistance for the autonomous driving system.

Another potential application of the proposed method is virtual traffic stream scene simulation, which is the key for the cities' virtual reality (VR) and effective method to improve the occlusions in the intersections. The proposed method can obtain the vehicles' pose and position in the intersections. With the data entered into the VR system, we can get the virtual traffic flows in real time. This application could adjust the vehicles' amount in different lanes and the traffic light control with the data provided by the proposed method to improve the occlusions in the certain intersections.

8 Conclusions

An approach for vehicle pose and position estimation at city road intersections was proposed by using low-cost facilities: an on-board monocular camera and a common GPS with the presence of road sign ahead in front. The rough position of the vehicle provided by GPS is used for matching the road sign data in a pre-constructed database. The road sign is further detected with three constraints with the consideration of weak light and partial occlusion. We demonstrated that our method has a correctness of 90.50% or higher in sign detection within 150 meters, and is faster than the SIFT and SVM. It is noteworthy that our road sign detection method may not perform well enough in the late night, which would be a future research direction. The four vertices of the detected road sign with their corresponding ones in real world were used to calculate the planar homograph matrix which was resolved into three rotation and three translation vectors under the road sign coordinate system. These vectors were converted to the vehicle's pose and position on the road. The experimental results showed that, within 100 meters distance to the road signs, the pose error is less than 2 degrees, and the position error is less than one meter, which can reach the lane-level positioning accuracy. Experimental results also showed that our method is more accurate than the Beidou high-precision positioning system L202 at a distance of 100 to 20 meters to the road sign.

The prospect applications of our method include high-precision digital map navigation, autonomous driving system assistance, and virtual traffic stream scene simulation.

References

[1] Abbott H, Powell D. Land-vehicle navigation using GPS. *Proceedings of the IEEE*, 1999, 87(1): 145–162.

[2] Liu Y, Bai B. Research on GPRS vehicle location network service system. In *Proc. IEEE International Conference on Computer, Mechatronics, Control and Electronic Engineering*, Aug. 2010, pp.401-404.

[3] Choi B S, Lee J J. Mobile robot localization in indoor environment using RFID and sonar fusion system. In *Proc. IEEE International Conference on Intelligent Robots and Systems*, Oct. 2009, pp.2039-2044.

[4] Armesto L, Tornero J. Robust and efficient mobile robot self-localization using laser scanner and geometrical maps. In *Proc. IEEE International Conference on Intelligent Robots and Systems*, Oct. 2006, pp.3080-3085.

[5] Lategahn H, Schreiber M, Ziegler J *et al.* Urban localization with camera and inertial measurement unit. In *Proc. IEEE Intelligent Vehicles Symposium*, Jun. 2013, pp.719-724.

[6] Wahab A A, Khattab A, Fahmy Y A. Two-way TOA with limited dead reckoning for GPS-free vehicle localization using single RSU. In *Proc. the 13th IEEE International Conference on ITS Telecommunications*, Nov. 2013, pp.244-249.

[7] Wei L, Cappelle C, Ruichek Y. Camera/laser/GPS fusion method for vehicle positioning under extended NIS-based sensor validation. *IEEE Transactions on Instrumentation & Measurement*, 2013, 62(11): 3110-3122.

[8] Rezaei S, Sengupta R. Kalman filter-based Integration of DGPS and vehicle sensors for localization. *IEEE Transactions on Control Systems Technology*, 2007, 15(6): 1080-1088.

[9] Tuna G, Gulez K, Gungor V C *et al.* Evaluations of different simultaneous localization and mapping (SLAM) algorithms. In *Proc. the 38th IEEE Conference on Industrial Electronics Society*, Oct. 2012, pp.2693-2698.

[10] Chausse F, Laneurit J, Chapuis R. Vehicle localization on a digital map using particles filtering. In *Proc. IEEE Intelligent Vehicles Symposium*, Jun. 2005, pp.243-248.

[11] Peng J, EI Najjar E B, Pomorski M *et al.* Virtual 3D city model for intelligent vehicle geo-localization. In *Proc. IEEE International Conference on Intelligent Transport Systems Telecommunications*, Oct. 2009, pp.477-480.

[12] Uchiyama H, Deguchi D, Takahashi T *et al.* Ego-localization using streetscape image sequences from in-vehicle cameras. In *Proc. IEEE Intelligent Vehicles Symposium*, Jun. 2009, pp.185-190.

[13] Wong D, Deguchi D, Ide I *et al.* Single camera vehicle localization using SURF scale and dynamic time warping. In *Proc. IEEE Intelligent Vehicles Symposium*, Jun. 2014, pp.681-686.

[14] Song R, Chen H, Xiao Z, Xu Y, Klette R. Lane detection algorithm based on geometric moment sampling. *SCIENTIA SINICA Informationis*, 2017, 47(4): 455-467. (in Chinese)

[15] Lai A H S, Yung N H C. Lane detection by orientation and length discrimination. *IEEE Transactions on Systems Man & Cybernetics, Part B*, 2000, 30(4): 539-548.

[16] Lakshmanan S, Grimmer D. A deformable template approach to detecting straight edges in radar images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1996, 18(4): 438-443.

[17] Kaliyaperumal K, Lakshmanan S, Kluge K. An algorithm for detecting roads and obstacles in radar images. *IEEE Transactions on Vehicular Technology*, 2001, 50(1): 170-182.

[18] Nedevschi S, Popescu V, Danescu R *et al.* Accurate ego-vehicle global localization at intersections through alignment of visual data with digital map. *IEEE Transactions on Intelligent Transportation Systems*, 2013, 14(2): 673-687.

[19] Ansar A, Daniilidis K. Linear pose estimation from points or lines. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2003, 25(5): 578-589.

[20] Schweighofer G, Pinz A. Globally optimal $O(n)$ solution to the PnP problem for general camera models. In *Proc. British Machine Vision Conference*, Sept. 2008.

[21] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate $O(n)$ solution to the PnP problem. *International Journal of Computer Vision*, 2009, 81(2): 155-166.

[22] Lasota M, Skoczylas M. Recognition of multiple traffic signs using keypoints feature detectors. In *Proc. IEEE International Conference and Exposition on Electrical and Power Engineering*, Oct. 2016, pp.535-540.

[23] Teunissen P J, Odolinski R, Odijk D. Instantaneous BeiDou+GPS RTK positioning with high cut-off elevation angles. *Journal of Geodesy*, 2014, 88(4): 335-350.

[24] Ren F X, Huang J, Jiang R *et al.* General traffic sign recognition by feature matching. In *Proc. IEEE Image and Vision Computing New Zealand*, Nov. 2009, pp.409-414.

[25] Ellahyani A, Ansari M E, Jaafari I E. Traffic sign detection and recognition based on random forests. *Appl. Soft. Comput.*, 2016, 46: 805–815.

[26] Maldonado-Bascon S, Lafuente-Arroyo S, Gil-Jimenez P *et al.* Road-sign detection and recognition based on support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, 2007, 8(2): 264-278.

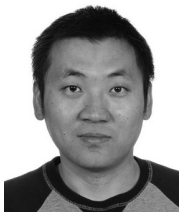
[27] Yuan Y, Xiong Z, Wang Q. An incremental framework for video-based traffic sign detection, tracking, and recognition. *IEEE Transactions on Intelligent Transportation Systems*, 2017, 18(7): 1918-1929.

[28] Fartaj M, Ghofrani S. Traffic road sign detection and classification. *Majlesi Journal of Electrical Engineering*, 2012, 6(4): 54-62.

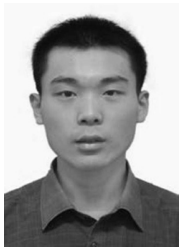
[29] Zhang Z. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2000, 22(11): 1330-1334.

Jin-Zhao Yuan is a postgraduate student at the School of Information Science and Engineering in Shandong University, Jinan. His research interests include autonomous driving, camera calibration, and 3D vision analysis.

Hui Chen is a professor at the School of Information Science and Engineering in Shandong University, Jinan. She received her Ph.D. degree in computer science from the University of Hong Kong, Hong Kong, in 2002, and her Bachelor's and Master's degrees in electronics engineering from Shandong University, Jinan, in 1984 and 1987, respectively. Her research interests include computer vision, 3D morphing and virtual reality.



Bin Zhao received his Master's degree in communication engineering from Shandong University, Jinan, in 2008. Currently, he is a CTO at Beijing Xuanlongding-xun Technology Co., Ltd. His research interests include virtual reality, computer vision, and data mining.



Yanyan Xu received his Ph.D. degree in pattern recognition and artificial intelligence from Shanghai Jiao Tong University, Shanghai, in 2015. Currently, he is a post-doctoral associate at Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, and a visiting scholar at Department of City and Regional Planning, UC Berkeley. His research interests include urban computing, data mining, and computer vision.