IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 12, 2019, accepted December 18, 2019, date of publication December 23, 2019,
date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961762

Visual Navigation Features Selection Algorithm Based on Instance Segmentation in Dynamic Environment

XIAOKAI MU[1], BO HE[1], (Member, IEEE), XIN ZHANG[1], TIANHONG YAN[2], XU CHEN[1], AND RUI DONG[1]

[1]College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China
[2]School of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China

Corresponding author: Bo He (bhe@ouc.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0301400, and in part by the National Natural Science Foundation of China under Grant 51379198.

ABSTRACT Ego-motion estimation, as one of the core technologies of unmanned systems, is widely used in autonomous robot navigation, unmanned driving, augmented reality and other fields. With the development of computer vision, there has been considerable interest in ego-motion estimation with visual navigation. One of the core technologies in Visual navigation is using the matching feature points between consecutive image frames to estimate pose. Since the feature-based method performed under the assumption of a static environment, it susceptive to the dynamic targets. Visual navigation in the dynamic environment has become an important research issue. This paper proposed a practical and robust features selection algorithm of visual navigation which avoids using the feature points on dynamic objects. Firstly, according to the instance segmentation of deep neural network, the objects are classified into potential dynamic and static categories. Subsequently, the matching features on the potential moving objects are used to update vehicle state respectively, meanwhile, the relevant reprojection error of other feature points on the background could be calculated. Eventually, the result of whether the target is moving or not will be judged by the reprojection error, and the features on dynamic targets are removed. To illustrate the effectiveness of the features selection method in the dynamic environment, the proposed algorithm is merged into an MSCKF based on trifocal tensor geometry, and it has been evaluated in a public dataset. Experimental results demonstrated the effectiveness of the proposed method.

INDEX TERMS Ego-motion estimation, visual navigation, features selection, instance segmentation, reprojection error.

I. INTRODUCTION

Accurate ego-motion estimation plays a vital role in autonomous robot navigation. In the last few years, there has been considerable interest in ego-motion estimation based on visual navigation [1], [2]. As the sensor of visual navigation, camera could obtain sufficient information about the surrounding environment. However, the ego-motion estimation based on monocular camera has the scale uncertainty question [3], [4]. The dual-camera system, which calculates the parallax of two images to measure the distance of object, can solve this problem [5], [6], whereas the dual-camera system has high computational complexity, and a relatively long baseline is required for medium and remote measurements. The inertial measurement unit (IMU) is another solution to solve the scale problem. Since the IMU is able to get real scale information that the single camera could not be, the visual-inertial odometry (VIO) can improve both the reliability and precision of navigation [7]–[11]. Because the feature information will be added into state vector in traditional extended Kalman filter framework of VIO [12], the computational complexity will increase over time, and it is not suitable for large scale environments. One approach to overcome this question is using multi-state constraint Kalman

The associate editor coordinating the review of this manuscript and approving it for publication was Seung-Hyun Kong.

filter (MSCKF) [13] to remove the feature position from state vector, and this method uses sliding windows to establish the constraints of features among multiple poses, which is more suited to ego-motion estimation of vehicles.

One of the core technologies in visual navigation is matching the feature points between consecutive image frames. The mismatches are inevitable during the feature matching. Random sample consensus (RANSAC) [14], [15] is usually used to address this issue, but it may also fail in the complex dynamic environment, especially when there are a large number of moving objects in the picture. In the sequence image, the optical flow is generated if movements exist in the image. A proper way is using the optical flow to distinguish moving targets [16]. However, the accuracy of this method is relatively low. Another method is using depth information to estimate inappropriate feature points [17]. Meanwhile, it increases the computation complexity. As a choice, employing the image classification algorithm to obtain the moving objects could help to eliminate the outliers [18]. Whereas, the traditional image classification method has relatively low efficiency and is challenging to meet the requirements of a high dynamic environment. A practical and robust features selection algorithm is necessary.

Motivated by the outstanding performance of deep-learning [19], [20], several relevant visual navigation technologies in dynamic environment have been proposed. A single shot multibox detector is constructed to detect dynamic objects with prior knowledge, and selection tracking algorithm is proposed to eliminate the interference from dynamic objects [21]. However, the detect results expressed by the least surrounding boxes can not obtain the full contour information. The efficiency of this method needs to be further improved. Another dynamic SLAM method combines semantic segmentation network with moving consistency check method by optical flow to reduce the impact of dynamic objects [22]. It is a pity that the performance in outdoor scenes of this work was not shown. Additionally, a visual SLAM system which adds the capabilities of dynamic object detection and background inpainting has been proposed to achieve precise navigation [23]. Also, the method could be used in dynamic scenarios with monocular, stereo and RGB-D configurations. Nevertheless, compared with the other two, the performance of this algorithm in the monocular system decreases obviously.

From the reviews above, the existing technologies can not realize the precise ego-estimation using the monocular camera in the dynamic outdoor environment. This paper proposed a features selection algorithm based on deep-learning for VIO, which could be performed in dynamic outdoor environment. On the basis structure of visual inertial navigation algorithm, deep learning is introduced to segment potential dynamic targets. A features excluding strategy is used to eliminate the features on moving objects. By removing these feature points, the proposed method could realize the utilized maximized features and achieve precision visual navigation in dynamic environments. The proposed features selection algorithm will be merged into an MSCKF based on tri-focal tensor geometry. Results of simulation experiments have demonstrated the effectiveness of the proposed method.

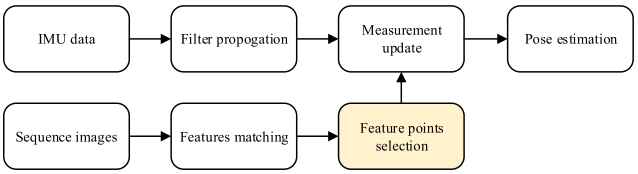


FIGURE 1. Process of the feature-based visual inertial navigation combined with the proposed feature selection method.

The remainder of this paper is organized as follows: Section II introduces the VIO framework this paper employed. The proposed method is presented in Section III. A number of comparisons are performed by experiments in Section IV. Finally, Section V summarizes the key conclusions of this work.

II. VISUAL EGO-MOTION ESTIMATION

Visual navigation could be classified into direct methods and feature-based methods. The direct methods could estimate a completely dense reconstruction by minimizing the photometric error. However, large scale movements and rotations will affect the estimated results. The feature-based methods could only estimate a sparse reconstruction based on salient points matching and epipolar geometry. Compared with the direct methods, the feature-based methods have better robustness in large scale motion estimation. Therefore, a feature-based method which uses MSCKF based on tri-focal tensor geometry is employed as our ego-motion estimation framework [24], [25]. Fig. 1 is the process of the feature-based visual inertial navigation combined with the proposed feature selection method.

The system state of the selected VIO algorithm is divided into two parts: nominal state and error state. The nominal state, which is described in Equation (1), includes the pose information and bias of the IMU.

$$\hat{x}_k = \begin{bmatrix} \hat{x}_{IMU}^T & (\hat{p}_{I1}^G)^T & (\hat{q}_{I1}^G)^T & (\hat{p}_{I2}^G)^T & (\hat{q}_{I2}^G)^T \end{bmatrix}^T \quad (1)$$

where $\hat{x}_{IMU}$ is the current IMU state which is shown in Equation (2).

$$\hat{x}_{IMU} = \begin{bmatrix} (\hat{p}_I^G)^T & (\hat{q}_I^G)^T & (\hat{v}_I^G)^T & \hat{b}_a^T & \hat{b}_g^T \end{bmatrix}^T \quad (2)$$

$\hat{p}_I^G$ is the IMU position in global frame, $\hat{q}_I^G$ the quaternion from IMU frame to global, $\hat{v}_I^G$ the IMU velocity. $\hat{b}_a$ and $\hat{b}_a$ are the bias of accelerometer and gyroscope. In the nominal state, $\hat{p}_{I1}^G$ and $\hat{q}_{I1}^G$ are IMU position and quaternion of last but one, $\hat{p}_{I2}^G$ and $\hat{q}_{I2}^G$ are the last.

Since the error state of MSCKF is used in Kalman iteration, the nominal and error state should be kept corresponding. The error state is expressed in Equation (3).

$$\tilde{x}_k = \begin{bmatrix} \tilde{x}_{IMU}^T & (\delta\tilde{p}_{I1}^G)^T & (\delta\theta_{I1}^G)^T & (\delta\tilde{p}_{I2}^G)^T & (\delta\theta_{I2}^G)^T \end{bmatrix}^T \quad (3)$$

where $\tilde{x}_{IMU}$ is the current error of IMU state which is shown in Equation (4).

$$\tilde{x}_{IMU} = [\,(\delta\tilde{p}_I^G)^T \quad (\delta\theta_I^G)^T \quad (\delta\tilde{v}_I^G)^T \quad \tilde{b}_a^T \quad \tilde{b}_g^T\,]^T \tag{4}$$

$\delta\tilde{p}_I^G$ is the position error, $\delta\theta_I^G$ the attitude error, $\delta\tilde{v}_I^G$ the velocity error, $\tilde{b}_a$ and $\tilde{b}_g$ are the bias error of accelerometer and gyroscope. In the error state, $\delta\tilde{p}_{I1}^G$ and $\delta\theta_{I1}^G$ are the position and attitude errors of last but one, $\delta\tilde{p}_{I2}^G$ and $\delta\theta_{I2}^G$ are the last. The framework of MSCKF has two processes: filter propagation and measurement update.

A. FILTER PROPAGATION

In the MSCKF, the nominal state could be easily predicted with kinematic equation by Runge Kutta. The propagation of error state is given as Equation (5).

$$\dot{\tilde{x}}_k = F\tilde{x}_k + \Gamma n_i \tag{5}$$

where F and Γ are the state transition matrix and noise transition matrix which are expressed in Equation (6) and (7), and n_i is the noise vector of IMU.

$$F = \begin{bmatrix} 0_{3*3} & 0_{3*3} & I_3 & 0_{3*3} & 0_{3*3} & 0_{3*12} \\ 0_{3*3} & F_{22} & 0_{3*3} & 0_{3*3} & -I_3 & 0_{3*12} \\ 0_{3*3} & F_{32} & 0_{3*3} & F_{34} & 0_{3*3} & 0_{3*12} \\ 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*12} \\ 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*12} \\ 0_{12*3} & 0_{12*3} & 0_{12*3} & 0_{12*3} & 0_{12*3} & 0_{3*12} \end{bmatrix} \tag{6}$$

$$\Gamma = \begin{bmatrix} 0_{3*3} & 0_{3*3} & 0_{3*3} & 0_{3*3} \\ -I_3 & 0_{3*3} & 0_{3*3} & 0_{3*3} \\ 0_{3*3} & -R\left(\hat{q}_I^G\right) & 0_{3*3} & 0_{3*3} \\ 0_{3*3} & 0_{3*3} & I_3 & 0_{3*3} \\ 0_{3*3} & 0_{3*3} & 0_{3*3} & I_3 \\ 0_{12*3} & 0_{12*3} & 0_{12*3} & 0_{12*3} \end{bmatrix} \tag{7}$$

The elements in Equation (6) are as follows:

$$\begin{cases} F_{22} = -\lfloor\hat{\omega}\times\rfloor \\ F_{32} = -R\left(\hat{q}_I^G\right)\lfloor\hat{a}\times\rfloor \\ F_{34} = -R\left(\hat{q}_I^G\right) \end{cases}$$

In the above equations, $\lfloor\hat{\omega}\times\rfloor$ and $\lfloor\hat{a}\times\rfloor$ are the antisymmetric matrix of angular speed and acceleration, and $R(\hat{q}_I^G)$ is the rotation matrix obtained from the quaternion.

The error prediction matrix ϕ can be obtained by Taylor series. The process of Taylor series is described in Equation (8).

$$\phi = \exp(F\Delta t) = I + F\Delta t + 1/2!F^2\Delta t^2 + \cdots \tag{8}$$

The definition of noise covariance matrix Q is as following:

$$Q = n_i n_i^T \tag{9}$$

Subsequently, the error system process noise W will be calculated by using the noise covariance:

$$W = \int \phi(\tau)\,\Gamma Q \Gamma^T \phi(\tau)^T d\tau \tag{10}$$

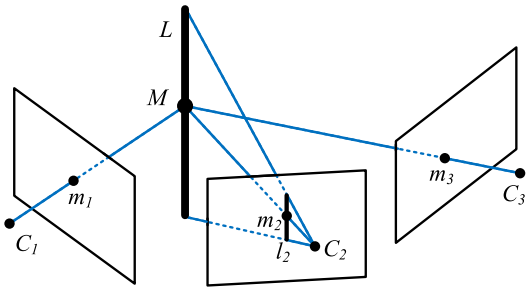


FIGURE 2. Point-Line-Point correspondence of tri-focal tensor.

Lastly, the error state covariance matrix P can be calculated as Equation (11).

$$P_k = \phi P_{k-1|k-1}\phi^T + W \tag{11}$$

B. MEASUREMENT UPDATE

The observation model is based on tri-focal tensor which describes the geometric relations among three camera images. The tri-focal tensor could recover the camera motion without the real position of features. Fig. 2 is the point-line-point correspondence among three views.

Where m_1, m_2 and m_3 are the normalized plane coordinates of one feature point in three views respectively, which can be obtained by the intrinsic matrix of camera and pixel location of feature point. l_2 is the line perpendicular to the epipolar line in second view. It is assumed that $P1[I|0]$, $P2[A|a_4]$ and $P3[B|b_4]$ are the projection matrices at three camera viewpoints, where A and B are rotation matrixes from the first view to the second and the third one, and a_4 and b_4 are the relevant translation vector. The tri-focal tensor is shown in Equation (12).

$$T_i = b_4 a_i^T - a_4 b_i^T \tag{12}$$

where a_i and b_i are the i-column elements of $P2$ and $P3$. Then the correspondence of point-line-point can be expressed in Equation (13).

$$m_3 = \left(\sum_i m_{1i} T_i^T\right) l_2 \tag{13}$$

where m_{1i} represents the i-column of m_1.

The main consideration of MSCKF is the pose of the consecutive camera. As the error states directly affect the projection location of feature points, the measurement model in the visual navigation framework is the location error of matching points. Since the filter state includes three consecutive states of the camera, the measurement model is represented by the epipolar geometry of adjacent images and the tri-focal tensor of three consecutive images. The corresponding measurement model of the i-th feature point is given by Equation (14).

$$z_i = \begin{bmatrix} m_2^T R_{1,2}^T \lfloor t_{12}\times\rfloor m_1 \\ m_3^T R_{2,3}^T \lfloor t_{23}\times\rfloor m_2 \\ K\left(\sum_i m_{1i} T_i^T\right) l_2 \end{bmatrix} \tag{14}$$

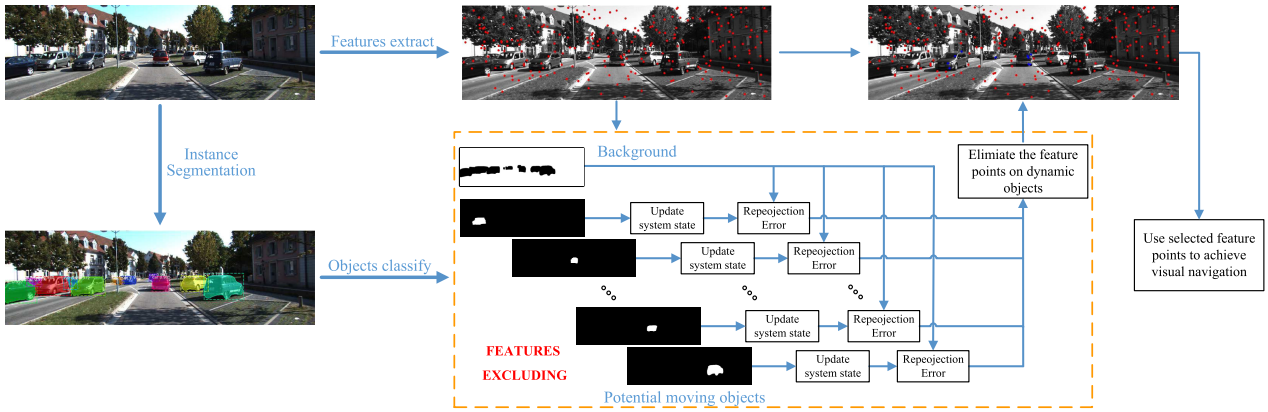


FIGURE 3. Procedure of the proposed features selection algorithm performs in visual navigation system.

$R_{i,j}$ means the rotation matrix from i to j, $\lfloor t_{ij}\times\rfloor$ is the anti-symmetric matrix of relevant translation vector, and K is the intrinsic matrix of the camera. Since the measurement model is nonlinear, the Sigma-Point approach is used to update the system error state. The error state will be used to correct the nominal state after the measurement update, which is shown in Equation (15).

$$X_k = \hat{X}_k + \tilde{X}_k \tag{15}$$

Additionally, the RANSAC algorithm is also applied to select inliers during the measurement update process. The method used here is similar to 1-point RANSAC. Equation (16) is the inliers decision formula.

$$\{m_1, m_2, m_3\}^{Inliers} = \left\{ \{m_1, m_2, m_3\} \left\| m_3 - \left(\sum_i m_{1i} T_i^T \right) l_2 \right\| < t \right\} \tag{16}$$

The RANSAC in visual navigation could detect the outliers of feature point, and it enables the VIO to work on a certain dynamic environment. However, this method still has a number of limitations. For example, the RANSAC may fail with multiple objects moving simultaneously or several large dynamic objects occupying an ample space in the image. A practical and robust features selection algorithm is necessary for visual navigation.

III. FEATURE POINTS SELECTION ALGORITHM BASED ON INSTANCE SEGMENTATION

To reduce the adverse influence caused by moving objects on ego-estimation, the features on dynamic targets should be eliminated. A naive method is to remove all the feature points that lie on the potential moving targets, resulting in a certain number of valid feature points are also lost. This paper proposed a features selection algorithm based on instance segmentation. The procedure of the proposed method is depicted in Fig. 3. This work prefers Mask R-CNN [26] to obtain the object instance information. Additionally, the feature points which are located on the potential dynamic objects will be selected. The Kalman filter, in turn, updated by the features on each potential moving object. Subsequently, according to the reprojection error to eliminate the improper feature points. The remaining feature points which are trustworthy can be used for visual navigation.

FIGURE 4. Instance segmentation results of Mask R-CNN.

A. INSTANCE SEGMENTATION ALGORITHM

Computer vision has been improved rapidly in recent years, especially in terms of object detection and segmentation. The instance segmentation, as a combination of object detection and semantic segmentation, requires finding and accurately segmenting the object. Since the proposed method needs to get the precise segmentation information of the potential moving objects, the instance segmentation technology is adopted as the first step to implement the features selection algorithm. The Mask R-CNN is used to implement the function in this paper.

Mask R-CNN is an extension of the Faster R-CNN [27], a segmentation mask branch that executes in parallel with other branches is introduced in the head section of the model. Moreover, the Mask R-CNN prefers RoIAlign instead of RoIPool to improve the accuracy of the mask. Through the above improvements to the past network, Mask R-CNN has a proper execution. The segmentation results of image are depicted in Fig. 4.

In the proposed method, potential dynamic and static categories are segmented by Mask R-CNN. The potential dynamic objects mainly include person, animal, vehicle, etc. The static category is considered as the background, which includes construction, landmark, etc. The features on the

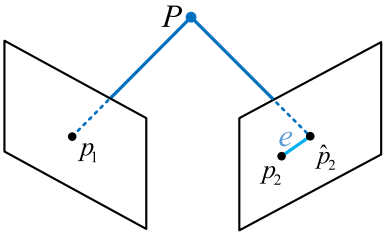


FIGURE 5. Reprojection of the feature points.

background are generally considered to be trustworthy for visual navigation. Due to the Mask R-CNN has no ability to distinguish whether the target is moving or not, the proposed method still need a features excluding strategy.

B. FEATURES EXCLUDING STRATEGY

In this part, the matching points on each potential moving target are used for state update, respectively. The reprojection error [28] of the features on the background is calculated by updated state and adopted to evaluate whether the potential moving objects are in motion or not. The specific implementation is as follows.

Firstly, an object index map should be constructed by encoder the pixel on each potential moving object. According to the location of each matching point, the number of matching points on each target is obtained. Feature points on each target are used as measurements to update the system state separately. The reprojection error is presented in Fig. 5, which refers to the distance between observed and estimated locations of feature points.

Where p_2 and $\hat{p}_2$ are measurement and reprojection of feature point respectively. e means reprojection error. In this paper, the features on each potential moving target are respectively used to calculate the reprojection error of matching points on the background. The reprojection of features in our work could be calculated by tri-focal tensor:

$$\hat{m}_3 = \left(\sum_i m_{1i} \hat{T}_i^T \right) l_2 \tag{17}$$

where $\hat{m}_3$ is the reprojection location of the feature point in the third normal coordinate, and $\hat{T}$ is the tri-focal tensor based on the system update. To calculate the reprojection error in image, the normal coordinate should be convert to pixel coordinate by the intrinsic matrix:

$$\begin{pmatrix} \hat{u} & \hat{v} \end{pmatrix} = K^{-1} \hat{m}_3 \tag{18}$$

where the $\hat{u}$ and $\hat{v}$ are the reprojection location in pixel coordinate. The average of reprojection error reflects the state of the potential moving target, and the moving possibility of each object is depicted in Equation (19).

$$PM_i = \frac{1}{n} \sum_{j=1}^{n} \left(\sqrt{\left(\hat{u}_{ij} - u_j\right)^2 + \left(\hat{v}_{ij} - v_j\right)^2} \right) \tag{19}$$

where PM_i reflects the moving possibility of the i-th potential dynamic object, $\hat{u}_{ij}$ and $\hat{v}_{ij}$ are the predict position of j-th matching feature by the system update using the features on i-th potential dynamic object, u_j and v_j the measurement position, n is the number of feature points on the background.

(a) Naive approach

(b) Mask approach

(c) Proposed approach

FIGURE 6. Effect of proposed features selection algorithm.

The feature points on targets with high moving probability should be eliminated. According to the experiments of the selected dataset, the threshold of the probability chooses 1.8. Fig. 6 describes the effect of the proposed method. In this example, there are eight potential moving objects have feature points, the quantitative results of calculation based on the proposed method is shown in Table 1. In Fig .6, the points are the features extracted from the image. Red points are used for visual navigation, and the blue points are improper features which calculated by naive mask and our method. In Fig. 6(b), although some cars were parked on the roadside, all of them were supposed to be the moving objects, and the features on these cars are eliminated. Fig. 6(c) shows that our method could distinguish the dynamic targets effectively, and the feature points are utilized maximized.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed method was evaluated using a public dataset KITTI [29], [30]. The KITTI is one of the largest datasets for computer vision algorithm, which is used to evaluate multiple fields such as stereographic images, optical flow, visual odometry (VO), object detecting and tracking.

The data acquisition platform of KITTI was equipped with multiple sensors, including two grey and two color cameras, a Velodyne laser scanner, and an OXTS GPS/IMU. The dataset also provides corrected and synchronized data. In the experiments, the pose information of the GPS/IMU are introduced as ground truth, and the left color camera is the image

TABLE 1. Quantitative results of calculation based on the proposed method.

No.	Moving possibility	Number of features	Status
1	0.9245	5	Not moving
2	1.4432	1	Not moving
3	0.6609	8	Not moving
4	1.8863	4	Moving
5	2.6690	1	Moving
6	1.8889	5	Moving
7	0.9245	4	Not moving
8	0.6419	9	Not moving

TABLE 2. Details of the four routes in this study.

No.	Sequence No. in KITTI	Start Image	End Image
Test_1	2011_09_26_drive_0005_sync	0000000000	0000000154
Test_2	2011_09_26_drive_0022_sync	0000000500	0000000799
Test_3	2011_09_26_drive_0059_sync	0000000000	0000000373
Test_4	2011_09_26_drive_0096_sync	0000000000	0000000475

sensor. The resolution of the images is 1242 × 375 pixels. Since the proposed method is applied to dynamic scenarios, four groups of raw data are selected as experimental data. The details of these data are shown in Table 2.

The performance of the proposed method in VIO is also evaluated by using the root mean square error (RMSE) and Hausdorff distance [31], [32].

RMSE shows the differences between the estimated data and the ground-truth. Since trajectory is displayed in 2D coordinates, the calculation formula of RMSE is Equation (20).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\left(\hat{x}_i - x_i\right)^2 + \left(\hat{y}_i - y_i\right)^2\right)}{n}} \tag{20}$$

where $\hat{x}_i$ and $\hat{y}_i$ are the estimated position, x_i and y_i are the ground truth, n means the number of position points. Due to it computes the error between two points with the same index, the measurement of RMSE is symmetric.

The Hausdorff distance is the maximum distances between a point in one set and its nearest point in another set, which has been widely used in shape matching. In this issue, the formula is modified as Equation (21).

$$hd\left(G, E\right) = \frac{1}{N_G}\sum_{g\in G}\left\{\min_{e\in E}\left\{d\left(g, e\right)\right\}\right\} \tag{21}$$

where G and E are the trajectories of ground truth and estimated. g and e are the corresponding position point. N_G means the point number of G. d represents the Euclidean distance. Different from RMSE, the Hausdorff distance is asymmetric, the value from G to R is different from G to E. In this paper, we used the bigger one as the measurement value. The formula is described in Equation (22):

$$HD\left(G, E\right) = \max\left\{hd\left(G, E\right), hd\left(G, E\right)\right\} \tag{22}$$

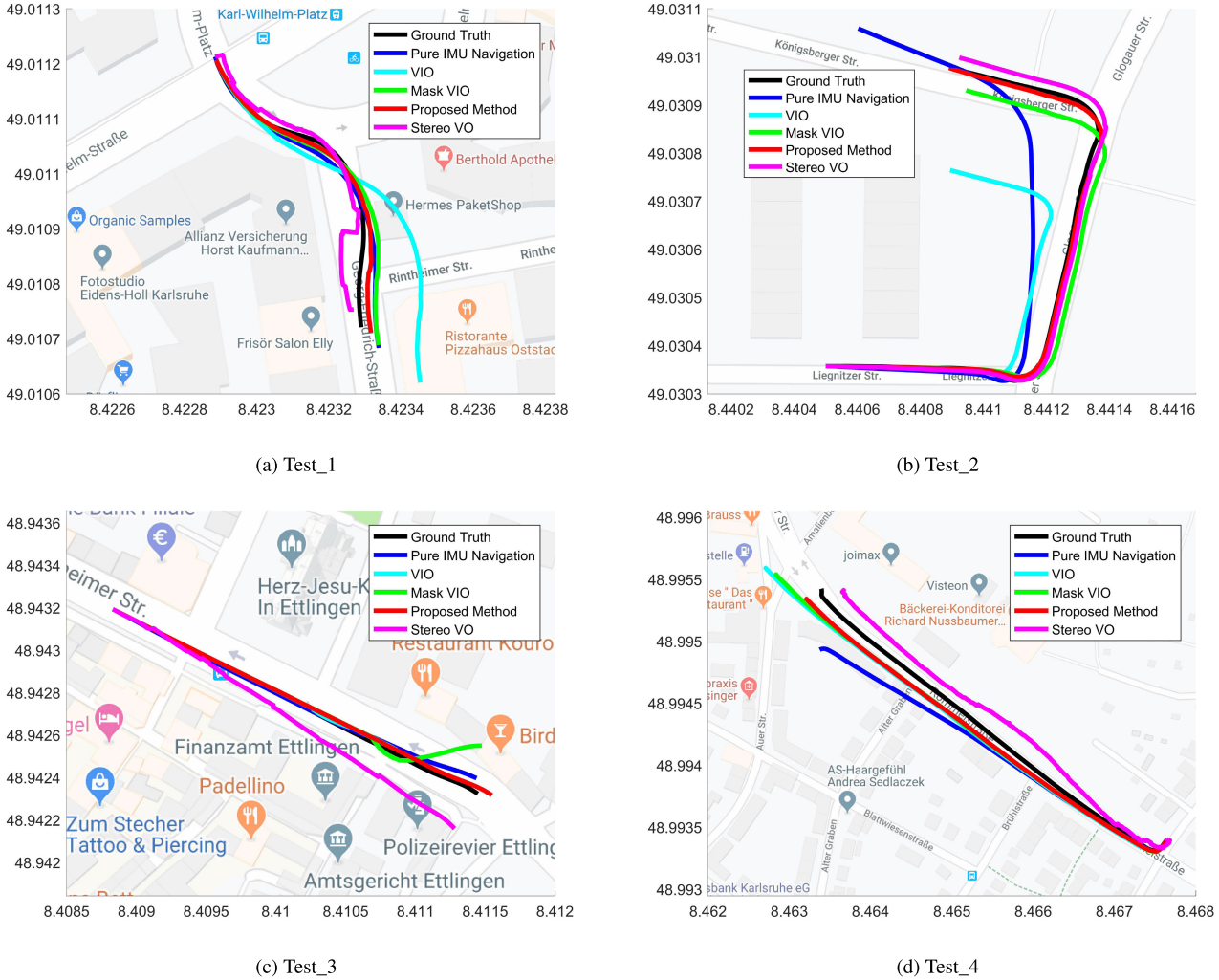
To verify the effectiveness of the proposed method in dynamic environments, pure IMU navigation, original VIO, naive mask VIO, the proposed method, and stereo VO were used to generate the trajectories. Fig. 7 is the experimental results that show the comparisons between the proposed algorithm and other approaches.

TABLE 3. The position RMSE and Hausdorff distance values computed between the ground truth trajectory and the estimated.

No.	Algorithm	RMSE(m)	Hausdorff Distance(m)
Test_1	IMU	2.8517	1.2921
	VIO	7.7141	1.9355
	Mask VIO	2.7210	1.3967
	Proposed method	2.0816	0.9826
	Stereo VO	2.3415	1.3971
Test_2	IMU	12.0162	5.4551
	VIO	14.6351	4.5468
	Mask VIO	7.1478	2.2353
	Proposed method	5.2301	1.6046
	Stereo VO	9.6545	4.4696
Test_3	IMU	5.6305	1.7693
	VIO	31.5043	1.5589
	Mask VIO	7.9709	2.8410
	Proposed method	4.4327	1.7855
	Stereo VO	12.5365	10.1206
Test_4	IMU	28.6721	13.1284
	VIO	21.1688	6.2501
	Mask VIO	18.2445	5.1425
	Proposed method	9.2548	4.9490
	Stereo VO	14.0194	10.5652

In Fig. 7, four groups of vehicle paths were estimated by above methods. Black lines stand for the ground truth which were produced by GPS/IMU. Blue lines represent pure IMU navigation, cyan lines are original VIO trajectories, green lines are the naive mask VIO trajectories, the paths obtained by the proposed method are shown as red lines, and the magenta lines are stereo VO trajectories. The pure IMU navigation trajectories usually have a bias, which will increase over time. In dynamic environments, due to the disturbance of the moving targets, the estimated results of stereo VO and VIO deviated from the ground truth. The Mask VIO could improve navigation accuracy to some extent. The proposed algorithm performs better than other methods.

The RMSE and Hausdorff distance values computed between ground truth and each trajectory estimation are reported in Table 3. Since the feature-based methods have a contingency, the RMSE and Hausdorff distance of above visual navigation methods are obtained by conduct experiments repeatedly five times and calculate the average. The error accumulation in pure IMU navigation leads to the trajectories differ from the ground truth, especially in Test_2 and Test_4. Both the RMSE and Hausdorff distance in these two tests show the IMU navigation has a large error. The selected data all have moving objects which introduce adverse interference for visual navigation. Therefore, almost all the results of RMSE and Hausdorff distance of original VIO are worse than other methods. It is worth noting that, even the Hausdorff distance of VIO in Test_3 are better than others, it does not mean the VIO performs good similarity. The trajectory of VIO in Test_3 has an inversion at 26s the trajectory after rotation basically coincides with that before. As a consequence,



(a) Test_1 (b) Test_2

(c) Test_3 (d) Test_4

FIGURE 7. The ground truth trajectory and the trajectories obtained using different methods.

TABLE 4. Process duration of VIO, Mask VIO, and the proposed method(CPU and GPU could compute in parallel).

	CPU time(ms)	GPU time(ms)
VIO	232.56	None
Mask VIO	259.92	≈350
Proposed method	312.44	≈350

the Hausdorff distance can not be used as an evaluative criterion alone. The results of stereo VO are not stable in a dynamic environment, as pure visual navigation is more susceptive to dynamic targets. The Mask VIO removed the features on all potential moving object, and the performance is better than VIO. However, excessive elimination of feature points leads to the matching points are not used adequately. The proposed method uses the reprojection error to eliminate the feature points located on moving objects, so the valid feature points are utilized to the maximum. From the above test results, almost all the RMSE and Hausdorff distance of the proposed method are better than other methods.

The above algorithms have been solved on a 2.40 GHz E5-2640 workstation that has 32 GB of ram, and the GPU is Nvidia Quadro M5000. Table 4 shows the average process duration for one calculation cycle of VIO, Mask VIO, and the proposed method. We could conclude that the proposed method added a few additional computation time in the CPU process. Compared to the CPU process time, the process duration in GPU is relative long. Therefore, a lightweight network that could satisfy the requirement of real-time should be adopted in future work. Additionally, because the sequence images contain a large number of repeated targets, using the combination of deep-learning and consistency check method could prompt the proposed method applied in the real-time visual navigation system.

V. CONCLUSION

In this paper, we proposed a features selection algorithm based on instance segmentation. Compared with the existing methods, the proposed method does not need to calculate the 3D position information of each feature point, so it has low computational complexity. The deep neural networks are used to realize accurate instance segmentation. Therefore, the object segmentation has good stability. According to the

reprojection error, this algorithm removes the features on dynamic targets to achieve maximum utilization of the valid features. The proposed method combines with the MSCKF based on tri-focal tensor and has been evaluated by the KITTI dataset. The RMSE metric is used to obtain the error of the estimated trajectory, and Hausdorff distance is utilized to measure the inconsistency between the estimated path and ground-truth. The algorithm performs well in both the RMSE and Hausdorff.

The proposed algorithm has high efficiency and robustness in the features selection and can be applied to visual navigation for ego-motion estimation. We will study the application of this algorithm to other scenarios and improve its performance in future.

REFERENCES

[1] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual navigation for mobile robots: A survey," *J. Intell. robotic Syst.*, vol. 53, no. 3, pp. 263–269, Nov. 2008, doi: 10.1007/s10846-008-9235-4.
[2] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, Dec. 2011, doi: 10.1109/MRA.2011.943233.
[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007, doi: 10.1109/TPAMI.2007.1049.
[4] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.
[5] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Rover navigation using stereo EGO-motion," *Robot. Auto. Syst.*, vol. 43, no. 4, pp. 215–229, 2003, doi: 10.1016/S0921-8890(03)00004-6.
[6] Y. Liu, D. Yang, J. Li, Y. Gu, J. Pi, and X. Zhang, "Stereo visual-inertial SLAM with points and lines," *IEEE Access*, vol. 6, pp. 69381–69392, 2018, doi: 10.1109/ACCESS.2018.2880689.
[7] G. Yang, L. Zhao, and J. Mao, "Optimization-based, simplified stereo visual-inertial odometry with high-accuracy initialization," *IEEE Access*, vol. 7, pp. 39054–39068, 2019, doi: 10.1109/ACCESS.2019.2902295.
[8] A. Martinelli, "Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 44–60, Feb. 2012, doi: 10.1109/TRO.2011.2160468.
[9] P. Corke, J. Lobo, and J. Dias, "An introduction to inertial and visual sensing," *Int. J. Robot. Res.*, vol. 26, no. 6, pp. 519–535, Jun. 2017, doi: 10.1177/0278364907079279.
[10] E. S. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *IEEE Trans. Robot.*, vol. 30, no. 4, pp. 407–430, Jan. 2011, doi: 10.1177/0278364910388963.
[11] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 690–711, Aug. 2018, doi: 10.1109/TRO.2018.2853729.
[12] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *Int. J. Robot. Res.*, vol. 32, no. 6, pp. 690–711, Jun. 2013, doi: 10.1177/0278364913481251.
[13] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. ICRA*, Roma, Italy, Apr. 2007, pp. 3565–3572.
[14] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme," in *Proc. IV*, San Diego, CA, USA, Jun. 2010, pp. 486–492.
[15] J. Civera, O. G. Grasa, A. J. Davison, and J. M. M. Montiel "1-Point RANSAC for extended Kalman filtering: Application to real-time structure from motion and visual odometry," *J. Field Robot.*, vol. 27, no. 5, pp. 609–631, Aug. 2010, doi: 10.1002/rob.20345.
[16] Y. Fang and B. Dai, "An improved moving target detecting and tracking based on optical flow technique and Kalman filter," in *Proc. ICCSE*, Hubei, China, Jan. 2009, pp. 1197–1202.
[17] A. I. Mourikis and S. I. Roumeliotis, "Semantic monocular SLAM for highly dynamic environments," in *Proc. IROS*, Madrid, Spain, Oct. 2018, pp. 393–400.
[18] B. Kitt, F. Moosmann, and C. Stiller, "Moving on to dynamic environments: Visual odometry using feature classification," in *Proc. IROS*, Taiwan, China, Oct. 2010, pp. 5551–5556.
[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[20] B. Romera-Paredes and P. H. S. Torr, "Recurrent instance segmentation," in *Proc. ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 312–329.
[21] L. Xiao, J. Wang, and X. Qiu, "Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment," *Robot. Auto. Syst.*, vol. 117, pp. 1–16, Jul. 2019, doi: 10.1016/j.robot.2019.03.012.
[22] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "DS-SLAM: A semantic visual SLAM towards dynamic environments," in *Proc. IROS*, Madrid, Spain, Oct. 2018, pp. 1168–1174.
[23] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018, doi: 10.1109/LRA.2018.2860039.
[24] J. S. Hu and M. Y. Chen, "A sliding-window visual-IMU odometer based on Tri-focal tensor geometry," in *Proc. ICRA*, Hong Kong, Jun. 2014, pp. 3963–3968.
[25] X. Dong, B. He, and X. Dong, "Monocular visual-IMU odometry using multi-channel image patch exemplars," *Multimedia Tools Appl.*, vol. 76, no. 9, pp. 11975–12003, Oct. 2016, doi: 10.1007/s11042-016-3927-8.
[26] K. He, G. Georgia, D. Piotr, and G. Ross, "Mask R-CNN," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 2961–2969.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. CVPR*, Sep. 2015, pp. 91–99.
[28] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003, pp. 363–406.
[29] A. Geiger, P. Lenz, and C. Stiller, "Vision meets robotics: The KITTI datase," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Aug. 2013, doi: 10.1177/0278364913491297.
[30] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. Int. Conf. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
[31] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. 12th Int. Conf. Pattern Recognit.*, Oct. 1994, pp. 566–568.
[32] X. Dong, X. Dong, J. Dong, and H. Zhou, "Monocular visual-IMU odometry: A comparative evaluation of detector-descriptor-based methods," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2019.2919003.

XIAOKAI MU received the M.S. degree from the School of Physics and Electronic-Electrical Engineering, Ningxia University, in 2016. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Ocean University of China. His research interests include integrated navigation and SLAM.

BO HE (M'01) was born in Qingdao, Shandong, China, in 1971. He received the M.S. and Ph.D. degrees from the Harbin Institute of Technology, China, in 1996 and 1999, respectively. From 2000 to 2003, he was a Postdoctoral Fellow with Nanyang Technological University, Singapore, where he worked on mobile robots, unmanned vehicles, research works included precise navigation, and control and communication. In 2004, he joined the Ocean University of China (OUC), where he is currently a Full Professor and the Deputy Head of the Department of Electronics Engineering, College of Information Science and Engineering. His current research interests include AUV design and applications, AUV SLAM, AUV control, and machine learning.

XIN ZHANG received the B.S. degree from the Department of Electronic Engineering, Ocean University of China, in 2016, where she is currently pursuing the Ph.D. degree. Her research interests include AUV SLAM and vision-aided inertial navigation.

TIANHONG YAN received the B.S. degree in mechanical engineering from Liaoning Technical University, Liaoning, China, in 1993, the M.S. degree in mechanical engineering from the Xian University of Technology, Shanxi, China, in 1996, and the Ph.D. degree in aerospace engineering from the Harbin Institute of Technology, Harbin, China, in 1999.

In March 1999, he joined the Micro-Satellite Research and Development Department, Shanghai Institute of Technological Physics, Chinese Academy of Sciences, as a Member of Research Staff, where he is working on integrated system design technology for Micro-Satellite. In November 2000, he began working on servo-mechanical aspects of hard disk drives with the Center for Mechanics of Micro-Systems, Nanyang Technological University, Singapore. Since September 2002, he has been with ASM Technologies, Singapore, focusing on motion control and electronics packaging equipments. From 2004 to 2008, he has been with the National Optical Lithography Tool Research and Development Center of China, Shanghai Micro-Electronics Equipment Company, Ltd., as a Senior System Engineer and an Assistant Manager. He joined China Jiliang University, as a Full Professor, in 2009, where he is currently focusing on autonomous underwater vehicle, robotics and high-speed and high-precision mechatronic systems.

XU CHEN received the B.S. degree from the School of Physical Engineering, Qufu Normal University, in 2018. She is currently pursuing the master's degree with the Department of Electronic Engineering, Ocean University of China. Her research interests include visual navigation and AUV SLAM.

RUI DONG received the B.S. degree from the School of Physical and Optoelectronic Engineering, Weifang University, in 2018. She is currently pursuing the master's degree with the Department of Electronic Engineering, Ocean University of China. Her research interests include visual navigation and AUV SLAM.