Robust Features and Accurate Inliers Detection Framework: Application to Stereo Ego-motion Estimation

Haigen MIN, Xiangmo ZHAO, Zhigang XU and Licheng ZHANG
College of Information Engineering, University of Chang'an
Xi'an, Middle of the South Second Ring Road 710064 - CHINA
[e-mail: xmzhao@chd.edu.cn]
*Corresponding author: Xiangmo ZHAO

Received March 15, 2016; revised June 7, 2016; accepted November 28, 2016; published January 31, 2017

Abstract

In this paper, an innovative robust feature detection and matching strategy for visual odometry based on stereo image sequence is proposed. First, a sparse multiscale 2D local invariant feature detection and description algorithm AKAZE is adopted to extract the interest points. A robust feature matching strategy is introduced to match AKAZE descriptors. In order to remove the outliers which are mismatched features or on dynamic objects, an improved random sample consensus outlier rejection scheme is presented. Thus the proposed method can be applied to dynamic environment. Then, geometric constraints are incorporated into the motion estimation without time-consuming 3-dimensional scene reconstruction. Last, an iterated sigma point Kalman Filter is adopted to refine the motion results. The presented ego-motion scheme is applied to benchmark datasets and compared with state-of-the-art approaches with data captured on campus in a considerably cluttered environment, where the superiorities are proved.

Keywords: local invariant feature, AKAZE, Ego-motion estimation, RANSAC, Iterated sigma point Kalman Filter

https://doi.org/10.3837/tiis.2017.01.016 **ISSN : 1976-7277**

1. Introduction

An accurate self-localization module is the core component of autonomous navigation system. The technology has received widespread attention from researchers in the development of intelligent mobile robot. The traditional positioning methods include compass [1], inertial measurement unit [2], wheel odometer [3], GPS [4] and their combinations [5-7], etc. The positioning system has rigid requirements as the application environment of intelligent mobile robot is becoming complex. Accuracy, real-time, robustness, portability, and energy consumption jointly restrict the performance of the positioning system. No matter which one fails, this one will be the shortest board as barrel effect shows. However, there are some disadvantages or limitations of traditional localization methods, e.g., the drift error increases quickly when wheels slip in uneven terrain or other adverse conditions [8]. The degree of GPS accuracy for civilian use may be close to meter while the cost of high precision GPS is too much. What's more, in GPS-denied environment (e.g., in cities with skyscrapers, forests, tunnels, outer space, etc.), GPS is ineffective because of invalid or absent signals [9]. In 1960s, NASA designed the mobile robot prototype for lunar exploration [10]. Then, the Soviet Union launched the remote lunar exploration robot Lunokhod No.1 and No.2. The rigorous working environment (e.g., Odometer misjudgment caused by wheel slip or lock, no signal of GPS) of planetary exploration robot puts forward unprecedented challenges for localization algorithms.

In order to solve the problems above, methods based on visual odometry (VO) have received great attentions. In 1983, Moravec [11] introduced the stereo visual odometry in the intelligent planetary exploration robot. Then, Matthies and Shafer [12] implemented the visual localization algorithm in 1987. American Jet Propulsion Laboratory first designed an autonomous robot under unstructured environment based on visual odometry [13]. Visual localization algorithms improved the positioning precision of the planetary exploration robot, also greatly expanded the scope of action of the robot, and ensured the robot completing task more safe and efficient. In 2004, planetary exploration robots Spirit and Opportunity successfully landed on Mars. They extracted 3D information from stereo vision system to estimate the camera pose, which couldn't be accurately estimated by traditional wheel odometer [14]. Mars Science Laboratory launched Mars exploration robot, which was equipped with several high-definition cameras. They further improved the consistency of the vision system. European Space Agency (ESA) and China National Space Administration (CNSA) also carried out lunar exploration programs, in which visual odometer module was very critical to self-localization.

Unmanned autonomous vehicle is a hot research topic in the field of intelligent mobile robot. In order to promote the development of related technologies, DARPA organized the DARPA Ground Challenge. In 2004 and 2005, the Challenge was held in outdoor environment while the third Challenge was held in urban environment in 2007. The positioning and navigation system based on vision played an important role in the whole process [15]. From 2012 to 2015, DARPA requested the participating teams to use the humanoid robot to complete some tasks, e.g., driving cars, climbing stairs, moving across clutter terrains. The provided humanoid robot Atlas perceived the environment information mainly with stereo camera rig. Google Company was not satisfied with the modified car used for autonomous driving, and he launched a new generation car with no steering wheel, brake and accelerator pedal. Up till the end of 2015, the prototype car had run about 2 million km and got the permission from several states in the

United States. Except the wheeled mobile robots mentioned above, there are some mobile robots (e.g., walking robot, flying robot, etc.) who don't support wheel odometer. Boston Dynamics' Big Dog can load 150Kg and walk on rough road at the rate of 8Km/h, aiming at material transportation in the war or the disaster. JPL installed a stereo vision system on the Big Dog to reconstruct the 3D terrain and look for passable road [16]. Also it contained a visual odometer to determine the pose of the robot.

In this study, we build on the main ideas of the Andreas Geiger's LIBVISO2 [31] and the robust feature AKAZE [41], to design an ego-motion estimation method for mobile vehicles, especially ones involving stereo camera rig and robust invariant feature scheme whose main contributions are as follows.

1) The introduction of robust feature AKAZE into visual odometry. We compare several 2D invariant feature algorithms from stability, accuracy and efficiency, and then AKAZE is adopted to implement feature detection and description.

2) Improvement on the RANSAC. The conventional RANSAC algorithm requires numerous iterations, which increases computational complexity. We explore the relationship between feature points on the consecutive image to accelerate the convergence procedure.

3) Highlight the iteration process in filtering work. The iterated sigma point Kalman Filter is suitable for vehicle motion refinement, a non-linear problem. Especially, we apply the iteration for convergence, which is faster than non-iterative process.

4) Experiments are conducted on public data set and data captured with our mobile platform. We analyze the results from quantitative and qualitative aspects to demonstrate the superiority of our proposed method.

The reminder of this manuscript is organized as follows: Section II is dedicated to review of related work. Section III briefly explains the system model of our platform. An outline of the overall algorithm, which includes the robust invariant feature AKAZE , the outlier removal procedure based on the improved RANSAC algorithm, motion estimation based on the geometric constraint, and the iterated sigma point Kalman Filter based refinement, are detailed in Section IV. Finally, our experiments and results are presented in Section V, and our conclusion is in Section VI.

2. Related Work

In general, there are two ways to estimate ego-motion in map, which are visual odometry and visual simultaneous localization and mapping (SLAM) [17]. Also, ego-motion estimation based on visual information can be roughly divided into two classes, namely monocular visual odometry (single camera architecture) and multiocular visual odometry (two or more cameras architecture) [18]. In other aspect, these approaches can be further separated into feature based methods, appearance based methods and hybrid methods [19].

In computer vision, techniques of camera pose estimation and 3-dimensional scene reconstruction based on image sequences belong to SFM (Structure from motion) [20-21]. VO can be thought of as a special case of SFM. SFM focuses on the 3-dimensional reconstruction as well as camera pose estimation and usually refines with bundle adjustment, while VO devotes to the real-time and accurate estimation of camera movement. The research on VO began with Moravec [22], who designed a planetary rover equipped with what he termed a slider stereo [8]. His innovative work also include Moravec corner detector [23]. Compared with sparse feature method, optical flow always requires tracking a set of frames [24]. The accuracy decreases dramatically if the movement between the adjacent images is too large.

Makadia [25] provided a dense matching method based on harmonic Fourier transform to calculate the relative motion. This method performs well with little texture, but calculation cost is high. Generally, the optical flow method is applied in monocular VO, while the binocular VO mainly uses the method based on features. Monocular and stereo VO are the two pipelines of visual research. Stereo VO can eliminate scale ambiguity and measure the movement of 6DOF(degree of freedom). Shafer [12] and Matthies [26] employed stereo VO to demonstrate its superiority of accuracy and robustness. Olson [27] extended this work by incorporating the global attitude sensors (e.g., compass and panoramic camera) , which reduced the accumulative error to a degree.

The concept of VO was formally put forward by Nister [28] and realized with a real-time VO system. The later researches were mostly based on his VO framework. The most successful VO applications were NASA's Mars Exploration Rover, Spirit and Opportunity [14]. Image pyramid was introduced to help feature tracking in 2011 on Curiosity. Howard [29] implemented a stereo VO with adopted Harris and Fast feature to ensure the real-time performance, and employed feature matching method [30] to find the corresponding feature points. Geiger [31-32] used a simple Sobel template operator to detect feature. They tested the VO algorithm on the KITTI benchmark dataset and obtained the positioning result with high efficiency and accuracy. KITTI vision benchmark suite was established by Andreas (MPI Tubingen), Philip Lenz (KIT), Christoph Stiller (KIT) and Raquel Urtasun (University of Toronto), etc. Bellavia [33] proposed the key frame matching and loop matching strategy to build stereo camera SLAM system. Badino [34] integrated features from multiple frames to improve the accuracy of the motion estimation. Jwu-Sheng [17] combined a monocular and an inertial measurement unit(IMU) to form an visual odometer. They employed trifocal tensor geometry information and multi-state constraint Kalman filter in the algorithm architecture to reduce the time consuming and enhance the accuracy of the algorithm. Besides the fusion applications of visual information and IMU, visual odometry is usually used as supplement to global positioning system (GPS) [35-37]. Fusion approaches are becoming the mainstream of research about the practical positioning and navigation system.

We have so far discussed some of the pioneering works in VO area. The mentioned similar works lay a good foundation on the development of VO, but they suffer from a low accuracy and high computational complexity. All of the aforementioned works devote to balance these two aspects, but they haven’t achieved unprecedented performance. Our work not only focuses on accuracy, but also tries to reduce algorithm complexity. Then we balance these two aspects to achieve the optimal performance.

3. Stereo Based System Model

In this section, the system model and motion parameterisation are briefly introduced as the base of next section. **Fig. 1** shows our experimental platform equipped with a high-resolution stereo camera rig (Basler Ace1600 GigE, image size 1600×1200 pixels, 30 Hz) and differential GPS (NovAtel OEM6TM GNSS). Data are captured in a considerably cluttered environment on campus.

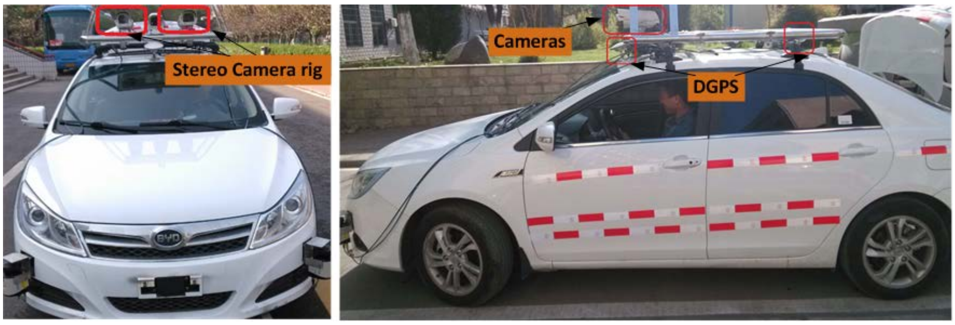


Fig. 1. Experimental platform with vehicular stereo camera rig and DGPS.

3.1 System Model

In general, the stereo camera rig we use in our method can be viewed as a linear pinhole model camera that conforms to the central projection [38-39]. The geometrical relationships of images involve four coordinate systems: the world coordinate system, the camera coordinate system, the physical coordinate system of the image, and the pixel coordinate system. According to the definition of the right-hand spiral rule, the model of the pinhole camera is shown in **Fig. 2.**

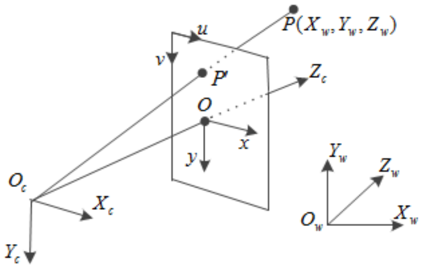


Fig. 2. The pinhole camera model.

The transformation between point P in the world coordinate system and its projection point P' in the pixel coordinate (u,v) is

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & \gamma & u_0 & 0 \\ 0 & \beta & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R(r) & \lfloor t \rfloor_\times \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_1 M_2 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1}$$

with

$$M_1 = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix},\ M_2 = \begin{bmatrix} R(r) & t \\ 0^T & 1 \end{bmatrix}, \tag{2}$$

$$R(r) = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{bmatrix}, \lfloor t \rfloor_\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \tag{3}$$

where α、β are the effective focal lengths along the u-axis and the v-axis, respectively. A correction parameter γ is needed for the highly accurate camera model. $\gamma = \alpha \tan\theta$, where θ is the deviation in degrees of the axis of the CCD array. M_1 is the intrinsic parameter and M_2

is the extrinsic parameter. $R(r)$ and t are respectively the rotation matrix and translation vector , with $\lfloor t \rfloor_{\times}$ the skew symmetric matrix of t .

3.2 Motion Parameterization

Motion parameterization, i.e., determining the spatial position of the camera coordinate system relatives to the world reference system, is expressed by a rotation matrix $\lfloor t \rfloor_{\times}$ and a translation vector $R(r)$.

The rotation matrix is defined in (4) and the rotation angle is parameterized by the Euler angle:

$$R(\theta,\Phi,\Psi)=\mathrm{R}_Z(\theta)\cdot\mathrm{R}_X(\Phi)\cdot\mathrm{R}_Y(\Psi) \tag{4}$$

In spatial motion, when the ego-motion vector $(V_X,V_Y,V_Z,w_X,w_Y,w_Z)$ of a wheeled car platform and the time difference ΔT between consecutive frames are known, the values of t and $R(r)$ in each time step can be obtained using (5) and (6). Here, V_i and w_i represent the speed of translation and rotation, respectively:

$$t=(V_X\cdot\Delta T,V_Y\cdot\Delta T,V_Z\cdot\Delta T)^T \tag{5}$$

$$R(r)=(w_X\cdot\Delta T,w_Y\cdot\Delta T,w_Z\cdot\Delta T) \tag{6}$$

4. Algorithm Overview

This section reveals the algorithm, mainly including robust feature detection and matching, the improved RANSAC and iterated sigma point Kalman Filter-based refinement. The overall algorithm is depicted in detail with the flow chart (as seen in Fig. 3).

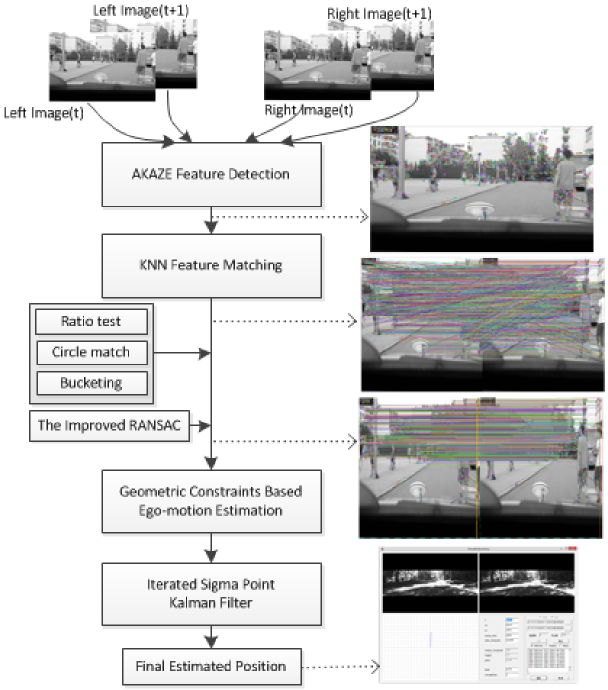


Fig. 3. Sketch of the algorithm architecture. It is explained in detail in the flowing sections.

4.1 Robust Feature

SIFT and SURF algorithms are multiscale 2D invariant features in nonlinear space. However, Gaussian blurring obtains the invariance and robustness, spoiling accuracy and distinctiveness. Alcantarilla [40] presented a 2D feature in a nonlinear scale space by means of nonlinear diffusion filtering. Even though Alcantarilla enhanced the accuracy and robustness of 2D invariant features, the efficiency was not significantly improved compared with SURF. Next year, the author proposed a novel and fast multiple invariants feature AKAZE [41] based on KAZE [40].

The paper of KAZE described three conductivity functions. Here we choose conductivity function g_2 which reserves wider region:

$$g_2 = \frac{1}{1+\frac{\left|\nabla L_\sigma\right|^2}{\lambda^2}} \tag{7}$$

where λ is the contrast factor that determines the level of diffusion. ∇ and div are respectively the gradient and divergence, and L is the image luminance. ∇L_σ is the gradient of a Gaussian smoothed L. σ indicates the Gaussian transformation of L.

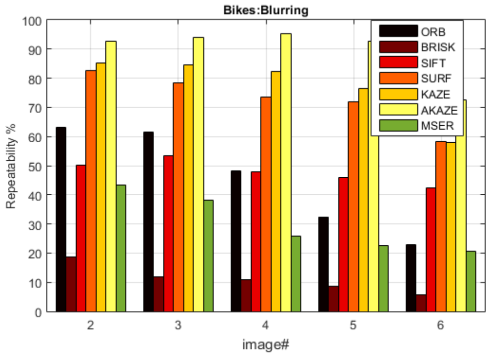
Since the Additive Operator Splitting (AOS) is computationally intense, a numerical scheme called Fast Explicit Diffusion (FED) is introduced into the feature detection procedure in nonlinear space. What's more, a Modified-Local Difference Binary descriptor is used as scale and rotation invariant with low storage requirements. From these two innovation, AKAZE shows outstanding performance compared to the state-of-the-arts. To demonstrate the excellent performance of AKAZE, we present comparison of experiment results obtained on the evaluation set of Mikolajczyk [42]. et al.

4.1.1 Repeatability

The detector repeatability [42] reveals a significant evaluation criterion of local invariant features. Repeatability is defined as the ratio between the corresponding points of two images and the number of the less features of one image. In our case, when the overlap error is smaller than 40%, we consider a correspondence between two regions, defined by:

$$1-\frac{A\bigcap H^t BH}{A\bigcup H^t BH}<\lambda \tag{8}$$

where A and B are the two regions. H is the corresponding homography between the two images. λ determines the overlap error and $\lambda = 0.4$



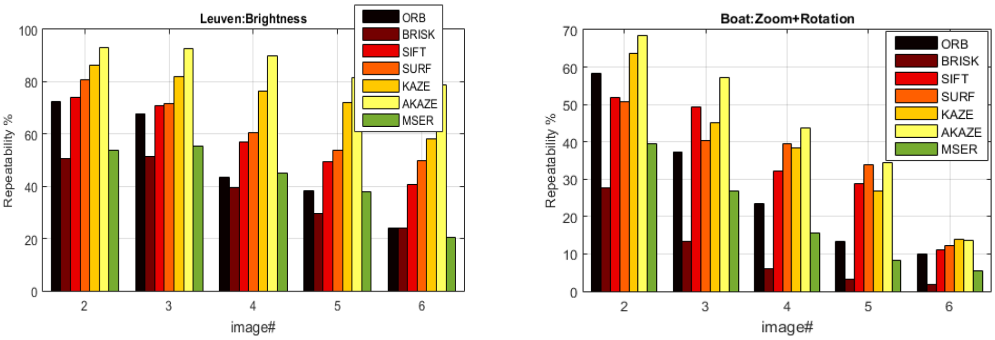


Fig. 4. Repeatability score under different transformation.

Fig. 4 depicts the repeatability scores in the Oxford dataset [42-43] with different photometric and geometric transformations, likes Gaussian blur, lighting variation, scale and rotation changes. Each sequence contains 6 images and the first image is regarded as the reference, so we can get 5 group histograms in each graph.

No matter with what kind of transformation, the repeatability of AKAZE scores top compared with other local invariant features, e.g., ORB [44], BRISK [45], SIFT [46], SURF [47], MSER [48]. SIFT and SURF have a good anti-noise ability and perform similar in lighting variation, scale and rotation changes, while the repeatability scores of ORB, BRISK and MSER decline rapidly in different conditions.

4.1.2 Precision-Recall

Precision-Recall [43] curve represents the whole performance of the local invariant feature algorithm. Eq.(9) shows the definition of recall and precision:

$$recall = \frac{T}{P},\ precision = \frac{T}{F} \qquad (9)$$

where T is the correct matching points. P and F are the all correspondences and matches, respectively.

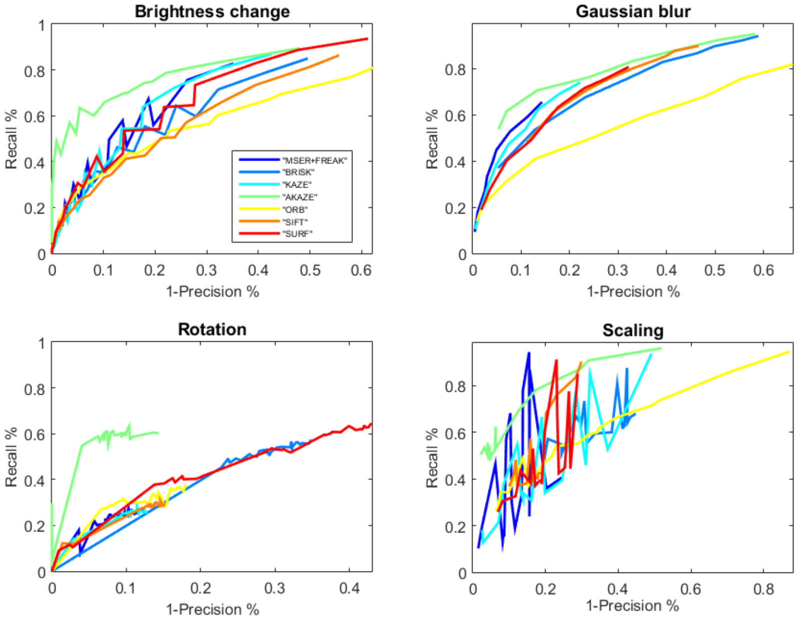


Fig. 5. Precision-Recall curve.

Fig. 5 shows the recall versus 1-precision relationship curve combined detection and description performance. So besides the detection process, feature descriptors and feature matching procedures are also adopted to finish this evaluation. With FREAK [49], we descript the MSER feature detector. From this curve, we can see that AKAZE performs best with different transformations, and SURF and KAZE have similar performance.

4.1.3 Efficiency

Considering the requirement of some real-time systems for local invariant features algorithms, high execution efficiency is an essential matter that we have to consider. We test the whole algorithm execution efficiency, including feature detection, feature description and feature matching three main steps, on a i7-4790 CPU 3.6 GHz computer, as shown in **Fig. 6.**

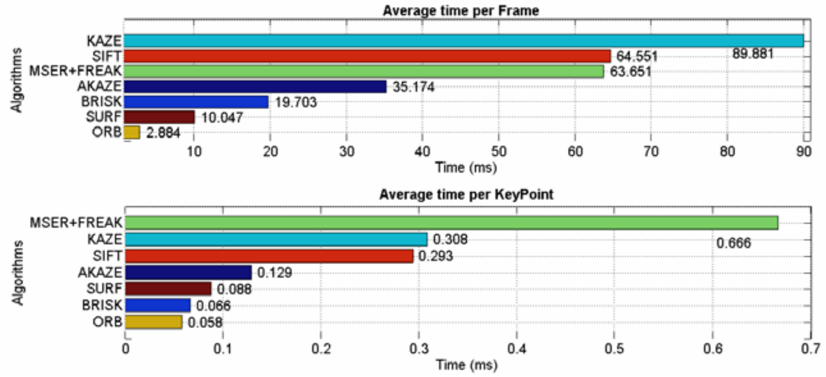


Fig. 6. Execution efficiency of different local invariant features.

Only considering the efficiency, KAZE is not appropriate for rigorous real-time system. ORB and BRISK score high in average time per frame and average time per keypoint, however they are not robust in view of repeatability and precision-recall curve. MSER and SIFT perform similar in average time per frame, while they are time-consuming because of fewer feature points. SURF and AKAZE perform well both in average time per frame and average time per keypoint, which can satisfy the needs of the real-time system.

4.2 Circle Match and Bucketing

We employ the Euclidean Distance to measure the similarity between AKAZE descriptors. The K-Nearest-Neighbor (KNN) is adopted to perform the matching procedure between consecutive frames. The ratio test is introduced to remove the mismatches to get a robust feature matching set. When the ratio between the best candidate matching point distance and the second best candidate matching point distance is smaller than the threshold, we take it as a safe match, otherwise we remove the match. In our case, we set the ratio threshold $\lambda = 0.6$. A bucket concept [18] is adopted to choose a subset of the matching feature points. A small number and uniform distribution of feature points reduce the computational complexity of the overall algorithm.

Fig. 7 shows the strategy of circle matching. $I_{L,k}$ and $I_{R,k}$ are the current left and right frame, $I_{L,k-1}$ and $I_{R,k-1}$ are the last left and right frame and $I_{L,k-2}$ and $I_{R,k-2}$ are the left and right frame before the last. $P^{(\prime)}_{L(R),k(k-1,k-2)}$ are the feature point in corresponding frame images as

Fig. 7 shows. Only when start point $P_{L,\ k}$ and end point $P'_{L,\ k}$ is the same feature point $(P_{L,\ k} = P'_{L,\ k})$, we declare a successful match, otherwise we delete the current feature point.

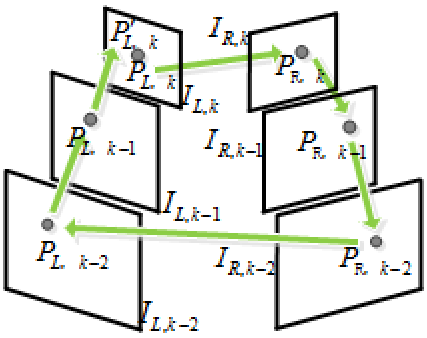


Fig. 7. Feature matching between image triples.

4.3 Outlier Removal Based on the Improved RANSAC

The result of feature matching in last section contains features of static as well as dynamic objects. In order to maintain the accuracy of subsequent calculations of vehicle position, further elimination of mismatches is needed. The traditional method for the elimination of erroneous matching is the RANSAC algorithm [50]. However, this method requires numerous iterations, which increases computational complexity. Moreover, it often fails to eliminate mismatching. To avoid these shortcomings of traditional RANSAC, we propose an improved RANSAC algorithm based on geometrical constraints.

There are relationships between two matching points set: (1) the slope of each matching pair is equal or close to each of the others, and (2) the length of each matching pair is equal or close to each of the others. We obtain the set of matching points of the frame $P_1 = \{ p_{1i}(x, y) \mid p_{1i} \in I_1 \} (i = 0,1,...,n-1)$ and the corresponding set of matching points of the given frame $P_2 = \left\{ p_{2j}(x, y) \mid p_{2j} \in I_2 \right\} (j = 0,1,...,n-1)$, where n is the sum of matching points. A tuple $M\left(E, C\right)$ represents a geometric constraints model, i.e., there is a constraint $C = \left(c_j \mid j = 0,1,...m\right)$ with regard to a finite set of elements $E = \left(e_i \mid i = 0,1,...n\right)$. To calculate the geometrical relationship between P_1 and P_2, we assume a point set P :

$$P = \left\{(P_1[i], P_2[j]) \mid i = j = 0,1,...,.n-1\right\} \tag{10}$$

Each element of P is a matching point pair. Based on the geometrical constraint model M , we search the matching point set $(P_1[i], P_2[j])$ that satisfies the geometrical constraints C and reject the matching point pair that does not satisfy it.

Experiments show that our proposed algorithm eliminates either mismatching points or points on dynamic objects, reducing the number of iterations and improves computational efficiency compared with the traditional RANSAC algorithm. Thus it improves the efficiency of the image matching algorithm to a greater degree, as shown in **Table 1**.

Table 1. Comparison between conventional RANSAC and the improved RANSAC

Group No.	N_{ori}	$P_{ori}^{correct}(\%)$	$T_{ori}(ms)$	N_{our}	$P_{our}^{correct}(\%)$	$T_{our}(ms)$
1	245	80.0	29.0	162	100.0	18.3
2	178	84.6	20.8	112	100.0	12.8
3	156	84.4	18.9	104	99.9	12.3
4	99	87.7	11.6	82	100	10.1

In the above table, N_{ori} is the number of matching points and T_{ori} is the average computation time for the conventional RANSAC. N_{our} is the number of matching points and T_{ori} is the average computation time for the improved RANSAC. The definition of $P_{ori}^{correct}$ and $P_{our}^{correct}$ is as follows: $N_{ori}^{correct}$ is the number of correct matching points for conventional RANSAC, and $N_{our}^{correct}$ is the number of correct matching points for the improved RANSAC.

$$P_{ori}^{correct} = \frac{N_{ori}^{correct}}{N_{ori}} \quad , \quad P_{our}^{correct} = \frac{N_{our}^{correct}}{N_{our}} \tag{11}$$

The improved RANSAC reduces the overall running time, meanwhile enhance the algorithm accuracy.

After the procedure of outlier removal, we get a robust inlier matching set which is the prerequisite of ego-motion estimation. Performing minimization using equation (12) obtains the estimation of $[R(r),t]$.

$$c^2(P_k, P_{k-1}) = \arg min \sum_{i}^{n^{Inliers}} \left\| P_k - (R(r)P_{k-1} + t) \right\|^2 \tag{12}$$

where i is a feature point. k is the time instant. P_k and P_{k-1} are triangulated 3D points at instants k and $k-1$, respectively.

Combining motion parameterization in Section 3.2 and Equation (12), we can calculate the initial estimation of 6DOF motion parameters.

4.4 The Iterated Sigma Point Kalman Filter Refinement

The theory of Kalman Filter is widely applied into dynamic system to estimate the instantaneous state. It provides a solution that may directly reduce the effects of disturbance noises including system and measurement noises. The errors in the parameters can also normally be handled as noise [51]. In traditional case, the relation between the instantaneous state $(V_X, V_Y, V_Z, w_X, w_Y, w_Z)^T$ and disturbed measurement is linear, namely, the Kalman filter is a linear filter that provides a prediction and an update step [52]. The Extended Kalman Filter [EKF] [53] is an extension of Kalman Filter, which provides the non-linear solutions using a first order Taylor expansion [54]. However, this reduces the accuracy of estimation result. A better choice is the usage of Kalman Filter based on the Unscented Transform (UT) [55]. Such filters propagate mean and covariance based on sigma points, like Unscented Kalman Filtef (UKF) [56] or Sigma Point Kalman Filter (SPKF) [57]. We highlight the importance of iteration in update step, which leads to a faster convergence.

Comparing the statistically linear error propagation and first order Taylor expansion error propagation, statistically linear error propagation is more accurate [57]. In the Sigma Points methods, regression points are selected by:

$$\begin{cases} x_0 = s \\ x_i = s + \gamma \cdot \sqrt{P_i} \qquad i = 1, \ldots, L \\ x_i = s - \gamma \cdot \sqrt{P_i} \qquad i = L+1, \ldots, 2L \end{cases} \tag{13}$$

where γ is parameter that determines the state space. s is the mean matrix and P is the state error covariance. i is the index of sigma point and we set L=5 in our case.

The system state mean and covariance are calculated by linear weighted regression from the Sigma points X and the transformed regression points Z :

$$\begin{cases} \bar{Z} = \sum_{i=0}^{2L} \omega_i^{(a)} Z_i \\ \operatorname{cov}(Z,Z) = \sum_{i=0}^{2L} \omega_i^{(c)} (Z_i - \bar{Z})(Z_i - \bar{Z})^T \\ \operatorname{cov}(X,Z) = \sum_{i=0}^{2L} \omega_i^{(c)} (X_i - s)(Z_i - \bar{Z})^T \end{cases} \tag{14}$$

where $\omega_i^{(a)}$ and $\omega_i^{(c)}$ are used to compute the state mean and covariance [58].

$$\begin{cases} K = \operatorname{cov}(X,Z)[\operatorname{cov}(Z,Z) + R]^{-1} \\ x_{i+1} = s + K[z - Z_i - (\operatorname{cov}(X,Z))^T P^{-1}(s_k - x_i)] \\ P_{k+1} = P_k - K \operatorname{cov}(Z,Z) K^T \end{cases} \tag{15}$$

where K is the Kalman gain. P and R are respectively state and measurement error covariance matrices.

The ISPKF measurement update is defined by Eq.(13) to Eq.(15), which is benefited from iteration process and statistically linearized error propagation.

5. Experimental Classification Results and Analysis

To illustrate the advantage of our proposed method (Proposed method-AKAZE), we evaluate our stereo motion estimation method on the KITTI dataset [32]. This benchmark consists of 22 stereo sequences and we randomly take the sequence 05 and sequence 09 as comparison data sets. The result of GPS/IMU from OXTS RT 3003 is regarded as the ground truth and the trajectory of VISO2 [31] and Proposed Method-SURF are used as references. The Proposed Method-SURF is the same pipeline with the Proposed Method-AKAZE and the only change is replacing the feature AKAZE with SURF. We test our proposed method on this dataset and draw the trajectories of ours with the references and ground truth in the same graph.

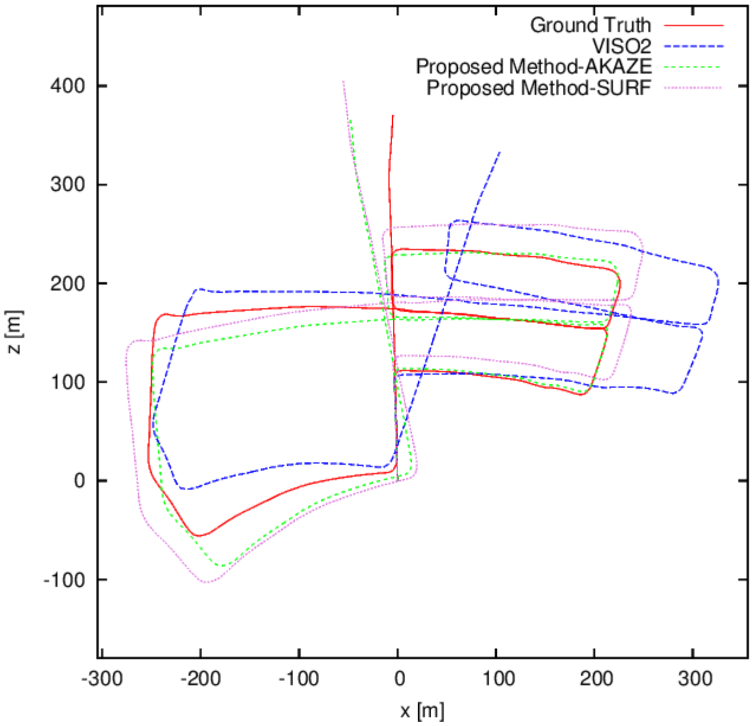


Fig. 8. Motion estimation on sequence 05 with ground truth, VISO2 stereo and our Proposed Method-AKAZE/-SURF.

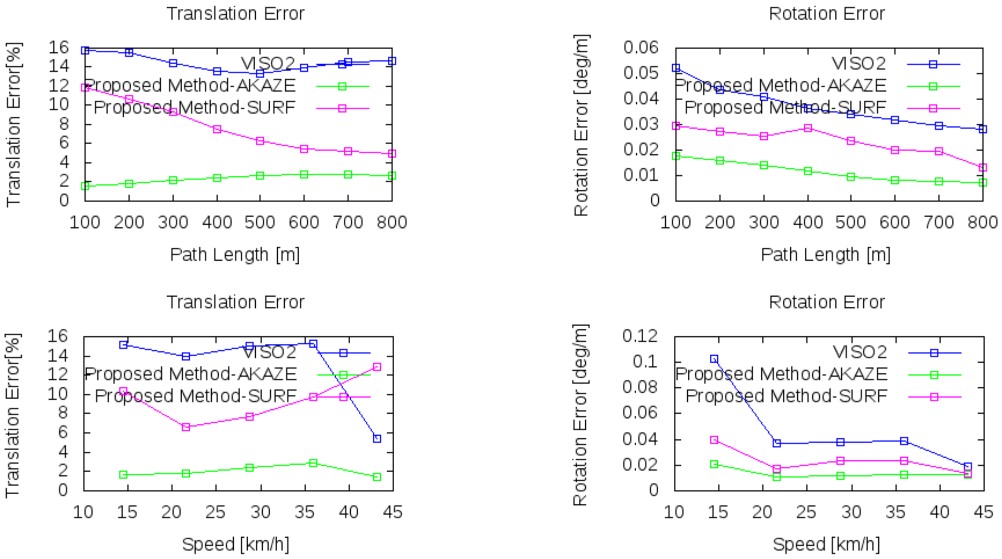


Fig. 9. Translation and rotation error of VISO2 stereo VO and our Proposed Method-AKAZE/-SURF on sequence 05.

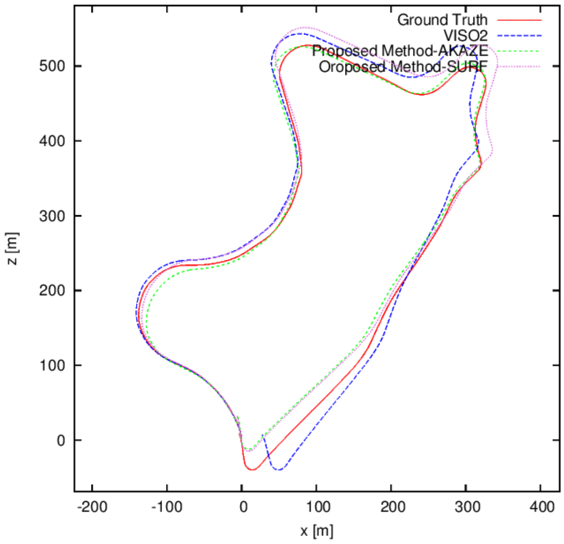


Fig. 10. Motion estimation on sequence 09 with ground truth, VISO2 stereo and our Proposed Method-AKAZE/-SURF.

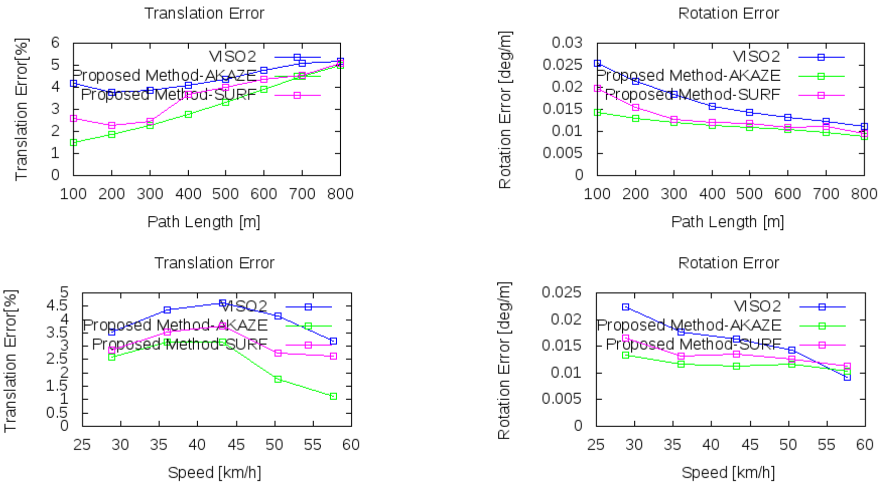
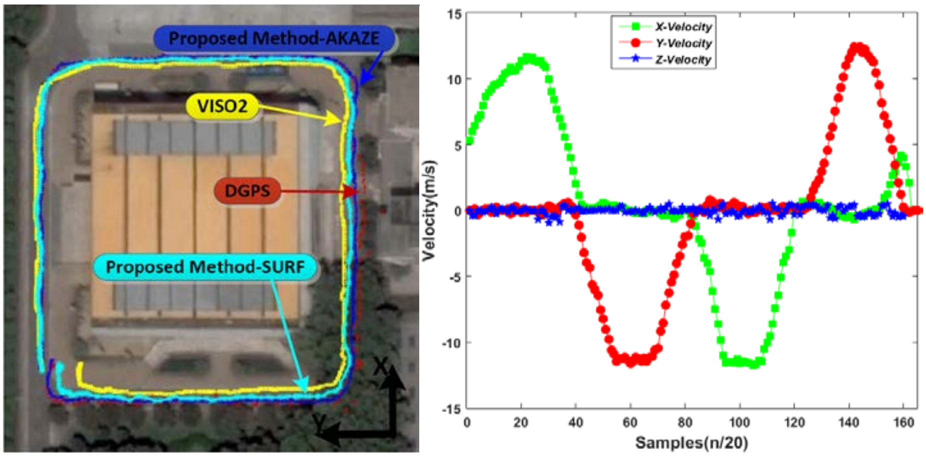


Fig. 11. Translation and rotation error of VISO2 stereo VO and our Proposed Method-AKAZE/-SURF on sequence 09.

From **Fig. 8** and **Fig. 10**, it can be seen that our method is closer to the ground truth than VISO2 stereo method. No matter with regard to rotation error or translation error, our method performs better, as shown in **Fig. 9** and **Fig. 11**. In Section 4.1, taking into consideration of accuracy, stability and efficiency, the feature SURF and AKAZE perform more similar. So we compare these two features' performance in application of VO. The trajectory of Proposed Method-SURF is close to the Proposed Method-AKAZE and they almost have the same varying trend. Clearly, the latter is more accurate in terms of rotation and translation.

To further demonstrate that our method was suitable for practical application, we used the vehicle platform presented in Section II to capture image data. Our approach was tested with one high-resolution stereo camera rig mounted on top of the vehicle. The distance from the center of the camera lens to the ground was 1.56m. The speed of the vehicle ranged from 0 km/h to 50 km/h.



(a) Trajectory estimated using different methods (b) Estimated velocity along x-, y-, and z-axes

Fig. 12. Comparison of the Proposed Method-AKAZE with other ego-motion estimation methods (DGPS, VISO2 stereo and Proposed Method-SURF).

The trajectories recovered using the proposed method and other methods are shown in **Fig.12**. The trajectory was approximately 1.3 km and lasted 203s. Further, we compared the velocities along the x-, y-, and z-axes, and calculated the Euler angles relative to each axis. We regard the trajectory of DGPS as the ground truth. The Proposed Method-AKAZE is closer to the ground truth than the stereo method of VISO2 and the Proposed Method-SURF. Moreover, our method has less drift than DGPS in a GPS-resistant environment, which ensures a more reliable performance. As shown in **Table 2**, the proposed method outperforms other methods in terms of overall RMSE and end-point error.

Table 2. Overall RMSE and end-point error results of our experiment

Algorithm	Overall position RMSE (m)	Overall orientation RMSE (deg)	End-point position RMSE (m)	End-point orientation RMSE (deg)
Proposed Method-AKAZE	3.5371	2.0632	1.6548	2.4521
Proposed Method-SURF	5.3167	3.8189	6.1054	3.8367
VISO2 stereo	6.4820	4.5281	11.6634	6.4937

5. Conclusion

In this work, we present a novel approach for 6DoF ego-motion of stereo visual odometry based on robust features. A new method for vehicle positioning based on stereo vision is proposed and compared with traditional visual odometry techniques. With our key enhancements to the adopted approach and algorithm, overall the proposed method can generate highly accurate ego-motion estimation results in a manner suited to real-time applications. To further improve our research, we are working on a better model of the system by combining it with the DGPS method. In our future research, the local positioning method and global localization method will be invariably combined in a practical navigation system.

References

[1] S. Funke, R. Schirrmeister, S. Skilevic et al., "Compass-Based Navigation in Street Networks," *Web and Wireless Geographical Information Systems*, pp. 71-88, 2015. Article (CrossRef Link).

[2] M. Luna, G. Meifeng, Z. Xinxi et al., "An indoor pedestrian positioning system based on inertial measurement unit and wireless local area network," in *Proc. of Control Conference (CCC), 2015 34th Chinese. IEEE*, pp. 5419-5424, 2015. Article (CrossRef Link).

[3] Z. Wang, J. Tan, and Z. Sun, "Error Factor and Mathematical Model of Positioning with Odometer Wheel," *Advances in Mechanical Engineering*, vol. 7, no. 1, pp. 305981-305981, 2015. Article (CrossRef Link).

[4] N. M. Drawil, H. M. Amar, and O. A. Basir, "GPS localization accuracy classification: A context-based approach," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 1, pp. 262-273, 2013. Article (CrossRef Link).

[5] K. Saadeddin, M. F. Abdel-Hafez, and M. A. Jarrah, "Estimating Vehicle State by GPS/IMU Fusion with Vehicle Dynamics," *Journal of Intelligent & Robotic Systems*, vol. 74, no. 1-2, pp. 147-172, 2014. Article (CrossRef Link).

[6] J. Georgy, T. Karamat, U. Iqbal et al., "Enhanced MEMS-IMU/odometer/GPS integration using mixture particle filter," *Gps Solutions*, vol. 15, no. 3, pp. 239-252, 2011. Article (CrossRef Link).

[7] Soares dos Santos, Douglas, Cairo L. Nascimento, and Wagner Chiepa Cunha. "Autonomous navigation of a small boat using IMU/GPS/digital compass integration," in *Proc. of Systems Conference (SysCon), 2013 IEEE International*, pp. 468-474, 2013. Article (CrossRef Link).

[8] S. Davide, and F. Friedrich, "Visual Odometry: Part I: The First 30 Years and Fundamentals," *IEEE Robotics & Automation Magazine*, 2011. Article (CrossRef Link).

[9] Ahrens, S., Levine, D., Andrews, G., & How, J. P., "Vision-based guidance and control of a hovering vehicle in unknown, gps-denied environments," in *Proc. of Robotics & Automation, ICRA. IEEE International Conference on*, pp. 3155-3160, 2009. Article (CrossRef Link).

[10] M. G. Bekker, "Mechanics of locomotion and lunar surface vehicle concepts," *Sae Transactions*, vol. 72, no. 12, pp. 0148-7191, 1964. Article (CrossRef Link).

[11] Moravec, Hans P, "The Stanford cart and the CMU rover," in *Proc. of the IEEE*, vol. 71, no. 7, pp. 872-884, 1983. Article (CrossRef Link).

[12] Matthies, Larry, and S. A. Shafer. "Error Modeling in Stereo Navigation," *Autonomous Robot Vehicles*. Springer New York, pp. 239-248, 1990. Article (CrossRef Link).

[13] Matthies, L. H., "Stereo Vision for Planetary Rovers," *Intern Journal Computer Vision*, vol. 8, pp. 71-91, 1992. Article (CrossRef Link).

[14] Maimone, Mark, Y. Cheng, and L. Matthies, "Two years of Visual Odometry on the Mars Exploration Rovers," *Journal of Field Robotics*, vol. 24, no.3, pp. 169-186, 2007. Article (CrossRef Link).

[15] S. Thrun, M. Montemerlo, H. Dahlkamp et al., "Stanley: The robot that won the DARPA Grand Challenge," *Journal of field Robotics*, vol. 23, no. 9, pp. 661-692, 2006. Article (CrossRef Link).

[16] M. Raibert, K. Blankespoor, G. Nelson et al., "BigDog, the Rough-Terrain Quaduped Robot," in *Proc. of IFAC Proceedings*, pp. 10822-10825, 2011. Article (CrossRef Link).

[17] Hu, Jwu Sheng, and M. Y. Chen, "A sliding-window visual-IMU odometer based on tri-focal tensor geometry," *Robotics and Automation (ICRA)*, pp. 3963-3968, 2014. Article (CrossRef Link).

[18] Kitt, B., Geiger, A., & Lategahn, H., "Visual Odometry based on Stereo Image Sequences with RANSAC-based Outlier Rejection Scheme," in *Proc. of IEEE Intelligent Vehicles Symposium*, vol. 43, no. 6, pp. 486 – 492, 2010. Article (CrossRef Link).

[19] Scaramuzza, D., F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC," in *Proc. of IEEE International Conference on Robotics & Automation*, pp. 4293-4299, 2009. Article (CrossRef Link).

[20] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 5828, pp. 133-135, 1981. Article (CrossRef Link).

[21] C. G. Harris, and J. M. Pike, “3D positional integration from image sequences,” *Image & Vision Computing*, vol. 6, no. 2, pp. 87-90, 1988. Article (CrossRef Link).

[22] Moravec, Hans Peter, "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," *DTIC Document*, 1980. Article (CrossRef Link).

[23] Morevec, Hans P., "Towards Automatic Visual Obstacle Avoidance," in *Proc. of Int. Joint Conf. Artificial Intelligence*, Cambridge, USA, pp. 584-584, August 1977. Article (CrossRef Link).

[24] Horn, Berthold K. P., and B. G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 7, no. 81, pp. 185–203, 1980. Article (CrossRef Link).

[25] Makadia, Ameesh, C. Geyer, and K. Daniilidis, "Correspondence-free Structure from Motion," *International Journal of Computer Vision*, vol. 5, no. 3, pp. 311-327, 2007. Article (CrossRef Link).

[26] Matthies L., "Dynamic Stereo Vison," *Carnegie Mellon University Computer Science Department*, 1989. Article (CrossRef Link).

[27] Olson, C. F., et al., "Robust Stereo Ego-motion for Long Distance Navigation," in *Proc. of CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, pp. 453-458, 2000. Article (CrossRef Link).

[28] Nistér, David, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. of CVPR 2004,Proceedings of the 2004 IEEE*, vol.1, pp. I-652-I-659, 2004. Article (CrossRef Link).

[29] Howard, A., "Real-Time Stereo Visual Odometry for Autonomous Ground Vehicles," in *Proc. of IROS 2008. IEEE/RSJ International Conference on IEEE*, pp. 3946-3952, 2008. Article (CrossRef Link).

[30] Hirschmuller, H., P. R. Innocent, and J. M. Garibaldi, "Fast, unconstrained camera motion estimation from stereo without tracking and robust statistics," in *Proc. of Control, Automation, Robotics and Vision, 2002, 7th International Conference on IEEE*, vol.2, pp. 1099-1104, 2002. Article (CrossRef Link).

[31] Geiger, Andreas, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," in *Proc. of IEEE Intelligent Vehicles Symposium*, vol. 32, no. 14, pp. 963-968, 2011. Article (CrossRef Link).

[32] Geiger, Andreas, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. of Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on IEEE*, pp. 3354-3361, 2012. Article (CrossRef Link).

[33] Bellavia, Fabio, et al., "Robust Selective Stereo SLAM without Loop Closure and Bundle Adjustment," in *Proc. of Image Analysis and Processing – ICIAP 2013*. Springer Berlin Heidelberg, pp. 462-471, 2013. Article (CrossRef Link).

[34] Badino, H., A. Yamamoto, and T. Kanade, "Visual Odometry by Multi-frame Feature Integration," in *Proc. of 2013 IEEE International Conference on Computer Vision Workshops (ICCVW) IEEE Computer Society*, pp. 222-229, 2013. Article (CrossRef Link).

[35] Wei, Lijun, et al., "GPS and Stereovision-Based Visual Odometry: Application to Urban Scene Mapping and Intelligent Vehicle Localization," *International Journal of Vehicular Technology*, vol. 2011, pp. 5-6, 2011. Article (CrossRef Link).

[36] J Rehder，K Gupta，S Nuske，"S Singh, “Global Pose Estimation with Limited GPS and Long Range Visual Odometry,” in *Proc. of 2012 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 627-633, 2012. Article (CrossRef Link).

[37] Schneider, Johannes, and W. Förstner, "Real-Time Accurate Geo-Localization of a MAV with Omnidirectional Visual Odometry and GPS," in *Proc. of Computer Vision - ECCV 2014 Workshops. Springer International Publishing*, pp. 483-501, 2014. Article (CrossRef Link).

[38] Corke, Peter I., "Visual Control of Robots: High-performance Visual Servoing," *Number Isbn Research Studies Press Ltd*, 1996. Article (CrossRef Link).

[39] Huang, Po Chia, et al., "A Voxel-Driven System Matrix Design for Multipinhole SPECT with Overlapping Projection," in *Proc. of IEEE Nuclear Science Symposium Conference Record*, pp. 3924-3927, 2009. Article (CrossRef Link).

[40] Alcantarilla, Pablo Fernández, A. Bartoli, and A. J. Davison, "KAZE Features," *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, pp. 214-227, 2012. Article (CrossRef Link).

[41] PF Alcantarilla，A Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," in *Proc. of British Machine Vision Conference (BMVC)*, Bristol, UK, September 2013. Article (CrossRef Link).

[42] K. Mikolajczyk, T. Tuytelaars, C. Schmid et al., “A Comparison of Affine Region Detectors,” *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43-72, 2005. Article (CrossRef Link).

[43] Krystian, Mikolajczyk, and S. Cordelia, "A performance evaluation of local descriptors." *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 27, no. 10, pp. 1615-1630, 2005. Article (CrossRef Link).

[44] E. Rublee, V. Rabaud, K. Konolige et al., “ORB: An efficient alternative to SIFT or SURF,” in *Proc. of Computer Vision (ICCV), 2011 IEEE International Conference on*, vol. 58, no. 11, pp. 2564-2571, 2011. Article (CrossRef Link).

[45] Leutenegger, Stefan, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," in *Proc. of Computer Vision (ICCV), 2011 IEEE International Conference on IEEE*, pp. 2548-2555, 2011. Article (CrossRef Link).

[46] Lowe, David G., "Object recognition from local scale-invariant features," in *Proc. of the International Conference on Computer Vision*, vol. 2, pp. 1150-1157, 2001. Article (CrossRef Link).

[47] H. Bay, T. Tuytelaars, and L. V. Gool, “SURF: Speeded Up Robust Features,” *Computer Vision & Image Understanding*, vol. 110, no. 3, pp. 404-417, 2006. Article (CrossRef Link).

[48] J. Matas, O. Chum, M. Urban et al., “Robust wide-baseline stereo from maximally stable extremal regions,” *Image & Vision Computing*, vol. 22, no. 10, pp. 761-767, 2004. Article (CrossRef Link).

[49] Ortiz, Raphael, "FREAK: Fast Retina Keypoint," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, pp. 510-517, 2012. Article (CrossRef Link).

[50] Das, A., and S. L. Waslander, "Outlier rejection for visual odometry using parity space methods," in *Proc. of Robotics and Automation (ICRA), 2014 IEEE International Conference on IEEE*, pp. 3613-3618, 2014. Article (CrossRef Link).

[51] Janiszewski D., "Extended Kalman Filter Based Speed Sensorless PMSM Control with Load Reconstruction," in *Proc. of Conference of the IEEE Industrial Electronics Society IEEE*, pp. 1465-1468, 2006. Article (CrossRef Link).

[52] Bertozzi, M., et al., "Pedestrian localization and tracking system with Kalman filtering," in *Proc. of Intelligent Vehicles Symposium, 2004 IEEE*, pp. 584-589, 2004. Article (CrossRef Link).

[53] Cruz, Sérgio, et al., "FPGA implementation of a sequential Extended Kalman Filter algorithm applied to mobile robotics localization problem," in *Proc. of Circuits and Systems (LASCAS), 2013 IEEE Fourth Latin American Symposium on. IEEE*, pp. 1-4, 2013. Article (CrossRef Link).

[54] A. H. Haddad, “Applied optimal estimation,” in *Proc. of the IEEE*, vol. 64, no. 4, pp. 574-575, 1976. Article (CrossRef Link).

[55] Angrisani, L., P. D'Apuzzo, and L. M. R. Schiano, "Unscented transform: A powerful tool for measurement uncertainty evaluation," *IEEE Transactions on Instrumentation & Measurement*, vol. 55, no. 3, pp. 737-743, 2006. Article (CrossRef Link).

[56] Huang, Guoquan P., Anastasios I. Mourikis, and Stergios I. Roumeliotis, "A quadratic-complexity observability-constrained unscented Kalman filter for SLAM," *Robotics, IEEE Transactions on*, vol. 29, no. 5, pp. 1226-1243, 2013. Article (CrossRef Link).

[57] Tang, Youmin, et al., "A practical scheme of the sigma-point Kalman filter for high-dimensional systems," *Journal of Advances in Modeling Earth Systems*, vol. 6, no. 1, pp. 21-37, 2014. Article (CrossRef Link).

[58] R. Van Der Merwe, A. Doucet, N. De Freitas et al., "The unscented particle filter," in *Proc. of NIPS*, pp. 584-590, 2000. Article (CrossRef Link).

Hai-Gen Min received the B.S. and M.S. degrees in the Department of computer science and he is currently pursuing the Ph.D. degree in the Department of Traffic Information Engineering & Control from Chang’an University, China. His research interests include computer vision, intelligent vehicle and internet of vehicles.

Xiang-Mo Zhao received the B.S. degree in Chongqing University and then received his M.S. and Ph.D. degrees in the Department of computer science from Chang’an University, China. He is currently a professor at Chang’an University in the Department of Traffic Information Engineering & Control. His research interests include distributed network measurement and control technology and intelligent transportation system.

Zhi-Gang Xu received the B.S., M.S. and Ph.D. degree in the Department of Traffic Information Engineering & Control from Chang’an University, China. He is currently an associate professor in the Department of the computer science. His research interests include the traffic image and video processing technology and intelligent detection technology.

Li-Cheng Zhang received the B.S. and M.S. degrees in Traffic Information Engineering & Control from Chang’an University,China. He is currently pursuing the Ph.D. degree of the same professional. His research interests include wireless sensor networks, vehicle detection technology.