IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 10, 2020, accepted March 1, 2020, date of publication March 3, 2020, date of current version March 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978154

Integrated Autonomous Relative Navigation Method Based on Vision and IMU Data Fusion

WENLEI LIU[1], SENTANG WU[1], YONGMING WEN[2], AND XIAOLONG WU[3]

[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
[2]Science and Technology on Inforamtion Systems Engineering Laboratory, Beijing Institute of Control and Electronics Technology, Beijing 100038, China
[3]Navigation and Control Technology Institute, NORINCO Group, Beijing 100089, China

Corresponding author: Wenlei Liu (liuwenlei@buaa.edu.cn)

This work was supported by the Industrial Technology Development Program under Grant B1120131046.

ABSTRACT An integrated autonomous relative navigation method based on vision and IMU data fusion was proposed in this paper, which can improve the position accuracy effectively and has strong adaptability to environmental changes. Firstly, IMU pre-integration formula based on Runge Kutta method was derived, which can improve the pre-integration position accuracy and reduce the accumulated error effectively. Secondly, an inverse depth estimation method based on the mixed probability model was proposed during the system initialization process, which can improve the accuracy of camera depth estimation and provide better initial conditions for back-end optimization. Thirdly, a sliding window filtering method based on the probability graph was proposed, which can avoid repeated calculations and improve the sliding window filtering efficiency. Forthly, combined with the advantages of the direct method and the feature point method, a mixed re-projection optimization method was proposed, which can expand the application scope of the method and improve the optimization accuracy effectively. Finally, in the closed-loop optimization, a closed-loop optimization method based on similar transformation is proposed to eliminate the accumulated error. In order to verify the environmental adaptability of the method and the impact of closed-loop detection on the relative navigation system, indoor and outdoor experiments were carried out with a hand-held camera and an IMU. EuRoC dataset was used in the experiments and the proposed method was compared with some classical methods. The experimental results showed that this method has high accuracy and robustness.

INDEX TERMS Data fusion, relative navigation, pre-integration, probability graph, sliding window filtering, mixed re-projection, similar transformation.

I. INTRODUCTION

With the continuous development of driverless technology, in order to adapt to various complex environmental conditions and resist random interferences, the autonomous navigation is more and more valued by many researchers. The inertial navigation technology and the vision-based SLAM (simultaneous localization and mapping) technology are two important autonomous navigation methods. The inertial navigation technology has a relatively high position accuracy and can output stable navigation data in a short period of time, but its accumulated error is increased gradually and navigation accuracy becomes more and more poor with time, thus leading to divergence. Therefore, the inertial navigation technology has a poor operation stability in long-term operation. On the other hand, the vision-based navigation technology has the advantages of simple equipment, low cost and relatively high position accuracy, but it can be greatly affected by external environmental conditions and there are problems such as scale drift. Therefore, a single navigation technology can hardly be applied to all kinds of environments. An effective way to improve the overall performance of the navigation system is to adopt the method of integrated relative navigation, which can take the advantages and avoid the disadvantages of both navigation technologies to integrate the data from all navigation equipment, so as to improve the accuracy of the navigation system greatly. For example, when the camera is only rotated, its rotation value cannot be calculated with the triangulation method and its translation vector cannot be calculated with the beam adjustment method, but its rotation can be compensated with the observation value on the gyroscope, so as to calculate its average parallax, which may

The associate editor coordinating the review of this manuscript and approving it for publication was Md Asaduzzaman.

not affect the real rotation result. In this paper, the integrated autonomous relative navigation method based on vision and IMU data fusion was studied. The pre-integration precision was improved with Runge Kutta method, the sliding window filtering precision was improved with the sliding window principle based on the probability graph, and the global optimization precision was improved with the combination of the direct method and the feature point method.

The direct method is used to estimate the motion of the camera directly with the gray-scale information of the image and can also operate normally when the number of feature points is small and the texture is fuzzy. In reference [1], the working principle and application of optical flow method are introduced in detail, and an inversely integrated algorithm is proposed based on the traditional optical flow method, which effectively reduces the loss of image information. Reference [2] focuses on the sparse direct method, which neglects the smoothness of the direct method and does not depend on the descriptors of the feature points, and proposes a photometric calibration method based on exposure time, lens halo and nonlinear response function. In references [3], [4], a semi direct method is proposed, in which the feature points are tracked with the direct method, the feature points are processed with the triangulation method and the corner points and edge pixels under the conditions with fuzzy textures and fast motions are tracked with the inverse depth estimation method. In reference [5], a semi direct visual localization (SDVL) method is proposed to improve the efficiency of feature matching. The three-dimensional point parametric tracking thread of inverse depth includes motion model, direct image alignment and feature matching optimization. In order to keep the luminosity unchanged during the measurement process, the direct method based on histogram equalization was adopted in this paper.

The vision locating method based on image features is used stably and is insensitive to the light, so it has strong robustness. In reference [6], a vision locating method for large scenes (LSD-SLAM) is proposed, to reconstruct 3D environment into the location map of key frames in real time with the high-precision pose estimation based on the direct alignment of images and the relevant semi dense depth map. In references [7], [8], a monocular vision SLAM based on ORB features (ORB-SLAM), is proposed. The system has strong robustness and can be initialized and relocate automatically, to generate a compact and traceable map. When the scene is changed, the map could be expended automatically, and the multi-threaded processing method can be used also effectively to improve the operation speed. In reference [9], an incremental pose optimization ORB-SLAM based on similarity transformation is proposed, to effectively solve the scale drift problem of ORB-SLAM and eliminate the accumulated error through global optimization. With the mixed inverse depth estimation method based on the probability graph, the uncertainty of depth estimation can be effectively solved and the robustness of the depth estimation can be improved. In order to adapt to a variety of complex environmental conditions, a mixed optimization method of direct method and feature point method was used in the back-end optimization in this paper.

The sliding window filtering technology is used widely in vision SLAMs, which can avoid repeated calculations and improve the operation speed effectively. In reference [10], a sliding window filter for SLAM based on feature-based 6-DOF (degree of freedom) batch processing with constant on-line time approximation least square SLAM is proposed, to achieve constant time complexity and linear space complexity. In reference [11], an observability constrained sliding window filter (OC-SWF) method is proposed to calculate the linearization points of Hessian to ensure the correct dimensions of Hessian zero space, minimize the linear error and avoid the influx of these non-existent information. In references [12], [13], a sliding window filtering method based on delay status marginalization is proposed, to extend the off-line batch least squares solution to the fast on-line incremental solution. In reference [14], a new mixed sliding window optimizer is proposed to realize the information fusion of the close coupled vision aided inertial navigation system. Based on the multi-status constraint method, a new distributed edge method is designed. In this paper, a sliding window filtering method based on probability graph was proposed, to improve the sliding window filtering efficiency greatly.

The above parts have described the methods commonly used in visual SLAM, and the following parts will describe the relative navigation method based on the fusion of vision and IMU, which includes loose coupling and tight coupling mainly. The loose coupling refers to the fusion after the pose is calculated with the vision and IMU data and the tight coupling refers to the pose calculation after the features of the image are added to the status vectors and optimized.

The loose coupling navigation method has been researched earlier, and some achievements have been made. In references [15], [16], a new algorithm of monocular vision inertial measurement range is proposed, to integrate the advantages of EKF (extended kalman filter) method and direct photometric error minimization method and incorporate the photometric error minimization into EKF measurement model directly. In reference [17], a new semi direct monocular vision synchronous locating and mapping (SLAM) system is proposed, to maintain the fast performance of the direct method and the high precision and closed-loop capability of the feature method. In reference [18], a probability cost function combined with the re-projection error of landmarks with the inertia terms is proposed, to limit the optimization to a bounded key frame window with marginalization processing, so as to ensure the real-time processability of the problem. In reference [19], a similar image synchronous position and mapping method is proposed to integrate the observation results from IMU and vision sensors so as to ensure a constant time output. The research results of the loose coupling fusion method, which

lays a foundation for the research on tight coupling fusion method.

With the development of computer vision technology and the improvement of computing power, the advantages of tight coupling fusion method is attracting more and more attentions and many solutions have been proposed by scholars. In references [20]–[22], a real-time vision aided inertial navigation algorithm based on extended Kalman filter (EKF) is proposed and a measurement model is derived, to express the geometric constraints generated when observing static features from multiple camera poses. In references [23]–[25], a robust and universal monocular vision inertial status estimator (VINS-Mono) is proposed, to obtain high-precision vision inertial measurement range with the tight coupling nonlinear optimization and pre-integrated IMU measurements and feature observations and enhance the global consistency with 4-DOF pose optimization. In reference [26], a pre-integration theory was proposed, to process the structure of rotation group, and prove that the pre-integration inertial measurement unit model can be seamlessly integrated into the vision inertial measurements under the unified frame of factor graph. In reference [27], a new analytic pre-integration theory of sensor fusion based on graph is proposed, and the closed form solution of pre-integration equation is derived, so as to improve the accuracy of status estimation. In this paper, the tight coupling fusion method of IMU pre-integration based on Runge Kutta method was used to achieve accurate autonomous navigation.

Fig.1 is the schematic diagram of vision and IMU data fusion system. In the front-end processing, the IMU pre-integration formula was derived and Runge Kutta method was used for calculation to improve the accuracy. Then the system was initialized, the initial pose of the camera was estimated with the sparse direct method, the inverse depth of the camera was estimated with mixed probability distribution model based on the probability graph, the rotation matrix between the camera and IMU was calibrated, and the gyroscope offset was corrected. Finally, the velocity, gravity and scale factors were initialized. In the back-end optimization, the sliding window filtering principle based on the probability graph was proposed to improve the filtering efficiency greatly, and the back-end optimization was carried out with the re-projection error function calculated with the fusion of the direct method and the feature point method to reduce the overall error of the system. In the global optimization of the closed-loop detection, the word bag model was used for detection and optimization. When a closed-loop was detected, the global optimization would be relocated to reduce the scale drift and cumulative error of the system.

This paper is organized as below: IMU pre-integration based on Runge Kutta method and system initialization will be introduced in Section II - Front-end Processing; the sliding window filtering principle based on the probability graph and the re-projection error optimization method mixed with the direct method and the feature point method will be introduced in Section III - Back-end Optimization; the closed-loop

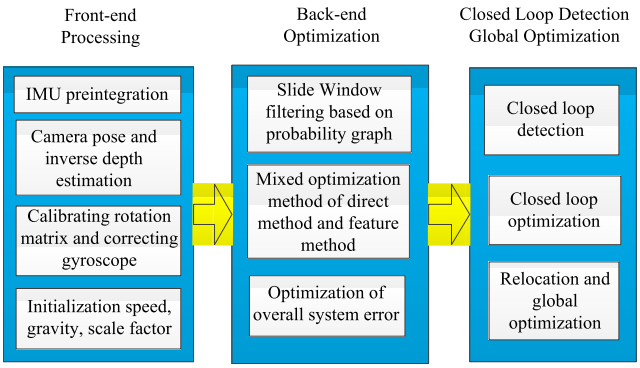


FIGURE 1. Schematic diagram of vision and IMU data fusion system.

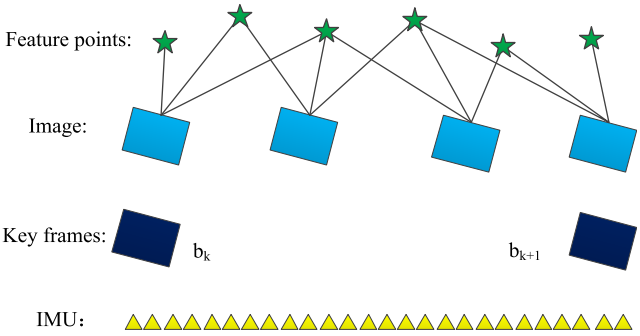


FIGURE 2. Relationship between IMU data and image data.

detection based on the word bag model and relocation global optimization method will be introduced in Section IV; the experiments with EuRoC dataset and the indoor and outdoor experiments with hand-held camera and IMU will be introduced in Section V, and the methods proposed in this paper will be summarized in Section VI.

II. FRONT-END PROCESSING

A. IMU PRE-INTEGRATION

In this paper, $(\cdot)^w$, $(\cdot)^b$ and $(\cdot)^c$ are defined as the world coordinates system, the ontology coordinates system and the camera coordinates system separately, q_b^w or R_b^w is the rotation matrix from the ontology coordinates system to the world coordinates system, p_b^w is the translation vector of the carrier's position in the key frame relative to the world coordinates system, b_k is the body coordinates system corresponding to the k^{th} image, and c_k is the camera coordinates system corresponding to the k^{th} image. $\otimes$ is the multiplication operator between quaternions, and $\hat{(\cdot)}$ is the actual measurement. The relationship between IMU data and image data is shown in Fig.2.

$\hat{a}_t$, $\hat{w}_t$ are the measurements with the accelerometer and the gyroscope in IMU. The relationship between them and the actual values, offsets, noises and accelerations of gravity can be expressed as below [28], [29]:

$$\begin{cases} \hat{a}_t = a_t + a_{bt} + R_w^{b_t} g^w + n_a \\ \hat{w}_t = w_t + w_{bt} + n_w \end{cases} \tag{1}$$

where, a_t and w_t are the actual acceleration and the angular velocity respectively, $R_w^{b_t}$ is the rotation matrix of the

ontology coordinates system transformed from the world coordinates system at the time of t, $g^w = \begin{bmatrix} 0 & 0 & g \end{bmatrix}^T$ is the direction of gravity acceleration, n_a and n_w are the measurement noises following a zero-mean Gauss distribution of $n_a \sim N\left(0, \sigma_{n_a}^2\right)$, $n_w \sim N\left(0, \sigma_{n_w}^2\right)$, a_{bt} and w_{bt} are the accelerometer deviation and the angular velocimeter deviation, respectively, which are random walk deviation, and whose derivatives follow a zero-mean Gauss distribution as follows:

$$\dot{a}_{bt} = n_{ba} \sim N\left(0, \sigma_{ba}^2\right), \dot{w}_{bt} = n_{bw} \sim N\left(0, \sigma_{bw}^2\right) \quad (2)$$

Assume that Δt_k is the time interval from t_k to t_{k+1}. In this time period, the recurrence relationship of position, speed and direction from b_k image to b_{k+1} image is as follows:

$$\begin{cases} p_{b_{k+1}}^w = p_{b_k}^w + v_{b_k}^w \Delta t_k \\ \qquad + \iint_{t\in[t_k,t_{k+1}]} \left(R_{b_t}^w \left(\hat{a}_t - a_{bt} - n_a\right) - g^w\right) dt^2 \\ v_{b_{k+1}}^w = v_{b_k}^w + \int_{t\in[t_k,t_{k+1}]} \left(R_{b_t}^w \left(\hat{a}_t - a_{bt} - n_a\right) - g^w\right) dt \\ q_{b_{k+1}}^w = q_{b_k}^w \otimes \int_{t\in[t_k,t_{k+1}]} \frac{1}{2} q_t^{b_k} \otimes \left(\hat{w}_t - w_{bt} - n_w\right) dt \end{cases} \quad (3)$$

The reference coordinates system of the above formula is transformed from the world coordinates system w to the ontology coordinates system b_k at the k^{th} key frame time, and the following formulas are obtained:

$$\begin{cases} R_w^{b_k} p_{b_{k+1}}^w = R_w^{b_k} \left(p_{b_k}^w + v_{b_k}^w \Delta t_k - \frac{1}{2} g^w \Delta t_k^2\right) \\ \qquad + \iint_{t\in[t_k,t_{k+1}]} \left(R_t^{b_k} \left(\hat{a}_t - a_{bt} - n_a\right)\right) dt^2 \\ R_w^{b_k} v_{b_{k+1}}^w = R_w^{b_k} \left(v_{b_k}^w - g^w \Delta t_k\right) \\ \qquad + \int_{t\in[t_k,t_{k+1}]} \left(R_t^{b_k} \left(\hat{a}_t - a_{bt} - n_a\right)\right) dt \\ q_w^{b_k} \otimes q_{b_{k+1}}^w = \int_{t\in[t_k,t_{k+1}]} \frac{1}{2} q_t^{b_k} \otimes \left(\hat{w}_t - w_{bt} - n_w\right) dt \end{cases} \quad (4)$$

Let:

$$\begin{cases} \alpha_{b_{k+1}}^{b_k} = \iint_{t\in[t_k,t_{k+1}]} \left(R_t^{b_k} \left(\hat{a}_t - a_{bt} - n_a\right)\right) dt^2 \\ \beta_{b_{k+1}}^{b_k} = \int_{t\in[t_k,t_{k+1}]} \left(R_t^{b_k} \left(\hat{a}_t - a_{bt} - n_a\right)\right) dt \\ q_{b_{k+1}}^{b_k} = \int_{t\in[t_k,t_{k+1}]} \frac{1}{2} q_t^{b_k} \otimes \left(\hat{w}_t - w_{bt} - n_w\right) dt \end{cases} \quad (5)$$

It can be seen from the above formulas that the pre-integration sub-items $\alpha_{b_{k+1}}^{b_k}, \beta_{b_{k+1}}^{b_k}, q_{b_{k+1}}^{b_k}$ takes the key frame b_k as the reference coordinate system, and the result is the relative motion of the key frame b_{k+1} with respect to the key frame b_k, which is only related to the measurement values with IMU, and is not affected by the position, speed and rotation of the key frame. $R_t^{b_k}, q_t^{b_k}$ is the rotation matrix of key frame b_k at time t.

In the case of small disturbances, the rotation quaternion $q_t^{b_k}$ is over parameterized, which can be simplified as a three-dimensional angle vector $\theta_i^{b_k}$ expressed as $\delta q_t^{b_k} = \begin{bmatrix} 1 & \frac{1}{2}\delta\theta_t^{b_k} \end{bmatrix}^T$. With the above formulas, the linear dynamic error equation within a continuous time period can be derived as follows [23]:

$$\begin{bmatrix} \delta\dot{\alpha}_t^{b_k} \\ \delta\dot{\beta}_t^{b_k} \\ \delta\dot{\theta}_t^{b_k} \\ \delta\dot{a}_{bt} \\ \delta\dot{w}_{bt} \end{bmatrix} = \begin{bmatrix} 0 & I & 0 & 0 & 0 \\ 0 & 0 & -R_t^{b_k}\left(\hat{a}_t - a_{bt}\right)\times & -R_t^{b_k} & 0 \\ 0 & 0 & -\left(\hat{w}_t - w_{bt}\right)\times & 0 & -I \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta\alpha_t^{b_k} \\ \delta\beta_t^{b_k} \\ \delta\theta_t^{b_k} \\ \delta a_{bt} \\ \delta w_{bt} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ -R_t^{b_k} & 0 & 0 & 0 \\ 0 & -I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} n_a \\ n_w \\ n_{ba} \\ n_{bw} \end{bmatrix} = F_t \delta z_t^{b_k} + G_t n_t \quad (6)$$

where, $[\cdot]_\times$ is the antisymmetric matrix derived with the vector; assume $q_v = \begin{bmatrix} q_x & q_y & q_z \end{bmatrix}^T$, then $[q_v]_\times = \begin{bmatrix} 0 & -q_z & q_y \\ q_z & 0 & -q_x \\ -q_y & q_x & 0 \end{bmatrix}$.

Based on the definition of the derivative, the following recurrence formula can be obtained:

$$\delta z_{t+\delta t}^{b_k} = \delta z_t^{b_k} + \delta \dot{z}_t^{b_k} \delta t = (I + F_t \delta t)\, \delta z_t^{b_k} + (G_t \delta t)\, n_t \quad (7)$$

Assume that the measurement time interval between two adjacent frames b_k and b_{k+1} of IMU is δt, the measurement is carried out with IMU between two key frames, i corresponds to the measured discrete time point, and the value measured with IMU is $\hat{a}_i^{b_k}, \hat{w}_i^{b_k}, \hat{q}_i^{b_k}$ respectively. In order to improve the accuracy, the pre-integration formula is discretized with the fourth-order Runge Kutta method as follows (8), as shown at the bottom of the next page, where

$$\begin{cases} f_{22} = I - \left[\frac{7}{8}\hat{w}_i + \frac{7}{24}\hat{w}_{i+1} - \frac{7}{6}w_{b_i}\right]_\times \delta t \\ f_{12} = -\frac{7}{8} R\left(\hat{q}_i^{b_k}\right) \left[\hat{a}_i - a_{b_i}\right]_\times \delta t \\ \qquad - \frac{7}{24} R\left(\hat{q}_{i+1}^{b_k}\right) \left[\hat{a}_{i+1} - a_{b_{i+1}}\right]_\times f_{22} \delta t \\ f_{14} = \frac{49}{144} R\left(\hat{q}_{i+1}^{b_k}\right) \left[\hat{a}_{i+1} - a_{b_{i+1}}\right]_\times \delta t^2 \\ v_{11} = -\frac{49}{192} R\left(\hat{q}_{i+1}^{b_k}\right) \left[\hat{a}_{i+1} - a_{b_{i+1}}\right]_\times \delta t^2 \\ v_{13} = -\frac{49}{576} R\left(\hat{q}_{i+1}^{b_k}\right) \left[\hat{a}_{i+1} - a_{b_{i+1}}\right]_\times \delta t^2 \end{cases} \quad (9)$$

Assume that Jacobian matrix of the system status at the initial time (i.e. the time corresponding to the key frame b_k) is $J_{b_k} = I$, and the covariance matrix is $P_{b_k} = 0$, then the recurrence formula of Jacobian matrix and covariance matrix is as follows:

$$\begin{cases} J_{i+1} = FJ_i \\ P_{i+1} = FP_iF^T + VQV^T \end{cases} \tag{10}$$

where, Q is the covariance matrix of noise signal n, and all noises are independent of each other. Q is a 18×18 diagonal matrix expressed as follows:

$$Q_{18\times 18} = diag\left(\sigma_{n_a}^2, \sigma_{n_w}^2, \sigma_{n_a}^2, \sigma_{n_w}^2, \sigma_{ba}^2, \sigma_{bw}^2\right) \tag{11}$$

Assuming that the variation of the pre-integration is linear with the deviation, the first order approximation can be expressed as follows:

$$\begin{aligned} \alpha_{b_{k+1}}^{b_k} &\approx \hat{\alpha}_{b_{k+1}}^{b_k} + J_{a_b}^{\alpha_{b_{k+1}}^{b_k}} \delta a_{b_k} + J_{a_w}^{\alpha_{b_{k+1}}^{b_k}} \delta w_{b_k} \\ \beta_{b_{k+1}}^{b_k} &\approx \hat{\beta}_{b_{k+1}}^{b_k} + J_{a_b}^{\beta_{b_{k+1}}^{b_k}} \delta a_{b_k} + J_{a_w}^{\beta_{b_{k+1}}^{b_k}} \delta w_{b_k} \\ q_{b_{k+1}}^{b_k} &\approx q_{b_{k+1}}^{b_k} \otimes \begin{bmatrix} 1 \\ \frac{1}{2} J_{a_w}^{\theta_{b_{k+1}}^{b_k}} \delta w_{b_k} \end{bmatrix} \end{aligned} \tag{12}$$

where, $J_{a_b}^{\alpha_{b_{k+1}}^{b_k}}, J_{a_w}^{\alpha_{b_{k+1}}^{b_k}}, J_{a_b}^{\beta_{b_{k+1}}^{b_k}}, J_{a_w}^{\beta_{b_{k+1}}^{b_k}}, J_{a_w}^{\theta_{b_{k+1}}^{b_k}}$ are the sub-blocks corresponding to Jacobian matrix J_{i+1}. For example, the meaning of $J_{a_b}^{\alpha_{b_{k+1}}^{b_k}}$ is $\frac{\delta \alpha_{b_{k+1}}^{b_k}}{\delta a_b}$.

With the covariance matrix, the measurement model of IMU can be expressed as follows:

$$\begin{bmatrix} \hat{\alpha}_{b_{k+1}}^{b_k} \\ \hat{\beta}_{b_{k+1}}^{b_k} \\ \hat{q}_{b_{k+1}}^{b_k} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} R_w^{b_k}\left(p_{b_{k+1}}^w - p_{b_k}^w - v_{b_k}^w \Delta t_k + \frac{1}{2} g^w \Delta t_k^2\right) \\ R_w^{b_k}\left(v_{b_{k+1}}^w - v_{b_k}^w + g^w \Delta t_k\right) \\ \left(q_{b_k}^w\right)^{-1} \otimes q_{b_{k+1}}^w \\ a_{b_{k+1}} - a_{b_k} \\ w_{b_{k+1}} - w_{b_k} \end{bmatrix} \tag{13}$$

In this paper, the pre-integration of IMU is taken as the initial measurement values for processing the key frames b_k and b_{k+1}, the measurement error function is obtained as follows:

$$\begin{aligned} &\begin{bmatrix} \delta \alpha_{b_{k+1}}^{b_k} \\ \delta \beta_{b_{k+1}}^{b_k} \\ \delta q_{b_{k+1}}^{b_k} \\ \delta a_b \\ \delta w_b \end{bmatrix} \\ &= \begin{bmatrix} R_w^{b_k}\left(p_{b_{k+1}}^w - p_{b_k}^w - v_{b_k}^w \Delta t_k + \frac{1}{2} g^w \Delta t_k^2\right) - \alpha_{b_{k+1}}^{b_k} \\ R_w^{b_k}\left(v_{b_{k+1}}^w - v_{b_k}^w + g^w \Delta t_k\right) - \beta_{b_{k+1}}^{b_k} \\ 2\left(q_{b_k}^w\right)^{-1} \otimes q_{b_{k+1}}^w \otimes \left(q_{b_{k+1}}^{b_k}\right)^{-1} \\ a_{b_{k+1}} - a_{b_k} \\ w_{b_{k+1}} - w_{b_k} \end{bmatrix} \end{aligned} \tag{14}$$

Because the pre-integration is based on the key frame and integrates the IMU measurement value between two adjacent key frames, it provides the initial value for fusion with image data, so the pre-integration will not cause cumulative error. In the process of back-end optimization, because the relative position of the pre-integration relative to the reference frame is unchanged, it can avoid repeated integration in the

$$\begin{aligned} \begin{bmatrix} \delta \alpha_{i+1}^{b_k} \\ \delta \beta_{i+1}^{b_k} \\ \delta \theta_{i+1}^{b_k} \\ \delta a_{b_{i+1}} \\ \delta a_{w_{i+1}} \end{bmatrix} &= \begin{bmatrix} I & \delta t I & \frac{1}{2} f_{12} \delta t & -\left(\frac{7}{16} R\left(\hat{q}_i^{b_k}\right) + \frac{7}{48} R\left(\hat{q}_{i+1}^{b_k}\right)\right) \delta t^2 & \frac{1}{2} f_{14} \delta t \\ 0 & I & f_{12} & -\left(\frac{7}{8} R\left(\hat{q}_i^{b_k}\right) + \frac{7}{24} R\left(\hat{q}_{i+1}^{b_k}\right)\right) \delta t & f_{14} \\ 0 & 0 & f_{22} & 0 & -\frac{7}{6} \delta t I \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} \delta \alpha_i^{b_k} \\ \delta \beta_i^{b_k} \\ \delta \theta_i^{b_k} \\ \delta a_{b_i} \\ \delta a_{w_i} \end{bmatrix} \\ &+ \begin{bmatrix} \frac{7}{16} R\left(\hat{q}_i^{b_k}\right) \delta t^2 & \frac{1}{2} v_{11} \delta t & \frac{7}{48} R\left(\hat{q}_{i+1}^{b_k}\right) \delta t^2 & \frac{1}{2} v_{13} \delta t & 0 & 0 \\ \frac{7}{8} R\left(\hat{q}_i^{b_k}\right) \delta t & v_{11} & \frac{7}{24} R\left(\hat{q}_{i+1}^{b_k}\right) \delta t & v_{13} & 0 & 0 \\ 0 & \frac{7}{8} \delta t I & 0 & \frac{7}{24} \delta t I & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta t I & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta t I \end{bmatrix} \begin{bmatrix} n_{a_i} \\ n_{w_i} \\ n_{a_{i+1}} \\ n_{w_{i+1}} \\ n_{ba} \\ n_{bw} \end{bmatrix} \\ &= F \delta z_i + V n \end{aligned} \tag{8}$$

optimization process, reduce the calculation amount and improve the calculation speed.

B. SYSTEM INITIALIZATION

Before further optimization, the system needs to be initialized with the loose coupling method. In this paper, the pose and inverse depth of the camera were estimated with the monocular camera, the rotation matrix between the camera and IMU was calibrated, the gyroscope offset was corrected, and finally, the velocity, gravity and scale factors were initialized, to align the camera estimation results with IMU results.

1) ESTIMATION OF POSE AND INVERSE DEPTH OF MONOCULAR CAMERA

The camera pose and inverse depth are the basis of system initialization. In this paper, the sparse direct method was used for pose estimation, and the inverse depth estimation method based on probability map was used.

Before using the sparse direct method, the histogram equalization method was used to the pre-processing in order to ensure the unchanged gray scale. The histogram equalization method had a faster calculation speed, could highlight the details of the image, and reduce the impact of the changes of light intensity on the gray scale, to ensure the consistency of the gray scale measurement.

The basic principle of the sparse direct method is to estimate the position of corresponding matching point with the position and pose of the current camera and optimize the camera pose by minimizing the photometric error. The optimized error function is follows:

$$\min_{\xi} \Delta(\xi) = \sum_{i=1}^{N} \|e_i\|^2 = \sum_{i=1}^{N} \|I_1(P_i) - I_2(h(P_i))\|^2 \quad (15)$$

where, $I_1(P_i)$ is the gray value of the i^{th} feature point in the first image, $I_2(h(P_i))$ is the gray value of the i^{th} feature point in the second image, N is the number of feature points, $h(P_i)$ is obtained with the re-projection method, R is the rotation matrix, t is the translation vector, ξ is the Lie algebra of R and t, ρ_i is the corresponding inverse depth information, and K is the internal parameter matrix of the camera. The conversion formula is as follows:

$$h(P_i) = \rho_i K(RP_i + t) = \rho_i K\left(\exp\left(\xi^{\wedge}\right) P_i\right) \quad (16)$$

The Jacobian matrix of optimization function can be obtained with Lie algebra, and the pose can be obtained by using the damped least square method, which is Levenberg Marquardt (LM) method.

In this paper, the inverse depth was estimated with Gaussian uniform mixture probability distribution based on the probability graph. The mixed probability distribution model has strong robustness and external interference signals have little influence on it, so the accuracy of depth estimation can be improved effectively.

With the actual inverse depth ρ, the Gaussian distribution precision λ and the scale coefficient of the correct data π, the probability distribution of the inverse depth measurement value x of the mixed model is as follows [30]:

$$p(x|\rho, \lambda, \pi) = \pi N(x|\rho, \lambda^{-1}) + (1-\pi)U(x) \quad (17)$$

where, $N(x|\rho, \lambda^{-1})$ is a Gaussian distribution with a mean value of ρ and a variance of λ^{-1}, and $U(x)$ is an interference signal following an uniform distribution.

According to Bayesian theorem: *posterior* $\propto$ *likelihood* $\times$ *prior*, the joint probability distribution of all random variables can be obtained as follows:

$$P(X, Z, \pi, \rho, \lambda) = p(X|Z, \rho, \lambda)p(Z|\pi) \cdot p(U|Z, \pi)p(\rho|\lambda)p(\lambda)p(\pi) \quad (18)$$

Finally, the parameters of the inverse depth were estimated with the variational inference method.

2) CALIBRATION OF THE ROTATION MATRIX BETWEEN CAMERA AND IMU

The calibration accuracy of the relative rotation between the camera and IMU is very important to the system fusion result. In this paper, the rotation matrix was solved with the nonlinear optimization method.

Assuming that the relative rotation of the camera between two adjacent key frames $R_{c_{k+1}}^{c_k}$ or $q_{c_{k+1}}^{c_k}$, the relative rotation matrix between two adjacent key frames derived with the pre-integration of IMU is $R_{b_{k+1}}^{b_k}$ or $q_{b_{k+1}}^{b_k}$, and the relative rotation matrix between camera and IMU is R_c^b or q_c^b, the relationship can be obtained as follows:

$$R_{b_{k+1}}^{b_k} R_c^b = R_c^b R_{c_{k+1}}^{c_k} \quad (19)$$

Assuming that $q = q_w + q_x i + q_y j + q_z k = q_w + q_v$, the above formula can be transformed into the quaternion form as follows:

$$\begin{aligned} &q_{b_{k+1}}^{b_k} \otimes q_c^b = q_c^b \otimes q_{c_{k+1}}^{c_k} \\ &\Rightarrow \left(Q_l\left(q_{b_{k+1}}^{b_k}\right) - Q_r\left(q_{c_{k+1}}^{c_k}\right)\right) q_c^b = Q_{b_{k+1}}^{b_k} q_c^b = 0 \end{aligned} \quad (20)$$

where, $Q_l(q) = \begin{bmatrix} q_w I + [q_v]_\times & q_v \\ q_v & q_w \end{bmatrix}$, $Q_r(q) = \begin{bmatrix} q_w I - [q_v]_\times & q_v \\ q_v & q_w \end{bmatrix}$.

After n consecutive key frames are processed, the relationship is as follows:

$$\begin{bmatrix} w_{b_1}^{b_0} Q_{b_1}^{b_0} \\ w_{b_2}^{b_1} Q_{b_2}^{b_1} \\ \vdots \\ w_{b_n}^{b_{n-1}} Q_{b_n}^{b_{n-1}} \end{bmatrix} q_c^b = Q_n q_c^b = 0 \quad (21)$$

where, $w_{b_i}^{b_{i-1}}$ is the weighting coefficient, which is obtained from the relative rotation matrix previously measured with IMU.

The relative rotation matrix q_c^b between the camera and IMU can be derived by solving the above formula.

3) CORRECTION OF GYROSCOPE BIAS

Some deviation of the gyroscope may be generated during the integration process, so it needs to be corrected in the initialization stage. Assume that the reference frame is c_0, the rotations of two consecutive key frames b_k and b_{k+1} with respect to the reference frame is $q_{b_k}^{c_0}$, $q_{b_{k+1}}^{c_0}$, respectively, and the relative rotation matrix $q_{b_{k+1}}^{b_k}$ between the two key frames can be derived with the pre-integration of IMU, then the target function of the gyro bias correction is:

$$\begin{cases} \min\limits_{w_{b_k}} \sum\limits_{k\in C} \left\| {q_{b_{k+1}}^{c_0}}^{-1} \otimes q_{b_k}^{c_0} \otimes q_{b_{k+1}}^{b_k} \right\| \\ q_{b_{k+1}}^{b_k} \approx \hat{q}_{b_{k+1}}^{b_k} \otimes \begin{bmatrix} 1 \\ \frac{1}{2} J_{a_w}^{\theta_{b_{k+1}}^{b_k}} \delta w_{b_k} \end{bmatrix} \end{cases} \tag{22}$$

where, C is the set of all key frames relative to the reference frame c_0.

The ideal result of the above objective function is an unit quaternion $\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^T$ with the real part of 1 and the virtual parts of 0. Let $(q)_{vec}$ be the virtual parts of the quaternion, then the objective function can be simplified as the form with virtual parts only:

$$J_{a_w}^{\theta_{b_{k+1}}^{b_k}} \delta w_{b_k} = 2 \left({q_{b_{k+1}}^{b_k}}^{-1} \otimes {q_{b_k}^{c_0}}^{-1} \otimes q_{b_{k+1}}^{c_0} \right)_{vec} \tag{23}$$

The above formula is transformed into the form of positive definite matrix, then the nonlinear optimization LM algorithm is used to solve the above formula, and the optimal solution of δw_{b_k} is obtained, and thus the gyro bias is corrected.

C. INITIALIZATION OF SPEED, GRAVITY AND SCALE FACTORS

Assuming that $(\cdot)^{c_0}$ is the camera coordinates system relative to the reference frame c_0, the translation vector and rotation matrix of the key frame measured only with the camera information relative to the reference frame is $\left(\tilde{p}_{c_k}^{c_0}, q_{c_k}^{c_0}\right)$, and the translation vector and rotation matrix of the camera relative to IMU body is $\left(p_c^b, q_c^b\right)$, the following relationship can be derived:

$$\begin{cases} q_{b_k}^{c_0} = q_{c_k}^{c_0} \otimes q_c^b \\ s\tilde{p}_{b_k}^{c_0} = s\tilde{p}_{c_k}^{c_0} - R_{b_k}^{c_0} p_c^b \end{cases} \tag{24}$$

where, s was the scale factor, which can be derived with the inverse depth estimation method based on the probability graph, namely $s = 1/\rho$.

Assuming that the variables of the speed, gravity and scale factors to be optimized were expressed as $\Omega = \left[v_{b_0}^{b_0}, v_{b_1}^{b_1}, \cdots, v_{b_n}^{b_n}, g^{c_0}, s\right]^T$, where $v_{b_k}^{b_k}$ is IMU measurement speed corresponding to the k^{th} key frame, and g^{c_0} is the coordinates representation of the gravity acceleration in the reference frame c_0, with the pre-integration formula, the following relations can be obtained:

$$\begin{cases} \alpha_{b_{k+1}}^{b_k} = R_{c_0}^{b_k} \left(s\left(p_{b_{k+1}}^{c_0} - p_{b_k}^{c_0}\right) - R_{b_k}^{c_0} v_{b_k}^{b_k} \Delta t_k + \frac{1}{2} g^{c_0} \Delta t_k^2 \right) \\ \beta_{b_{k+1}}^{b_k} = R_{c_0}^{b_k} \left(R_{b_{k+1}}^{c_0} v_{b_{k+1}}^{b_{k+1}} - R_{b_k}^{c_0} v_{b_k}^{b_k} + g^{c_0} \Delta t_k \right) \end{cases} \tag{25}$$

The formula (24) can be substituted into (25), to obtain the following formula:

$$\begin{bmatrix} -I\Delta t_k & 0 & \frac{1}{2} R_{c_0}^{b_k} \Delta t_k^2 & R_{c_0}^{b_k} \left(\tilde{p}_{b_{k+1}}^{c_0} - \tilde{p}_{b_k}^{c_0}\right) \\ -I & R_{c_0}^{b_k} R_{b_{k+1}}^{c_0} & R_{c_0}^{b_k} \Delta t_k & 0 \end{bmatrix} \times \begin{bmatrix} v_{b_k}^{b_k} \\ v_{b_{k+1}}^{b_{k+1}} \\ g^{c_0} \\ s \end{bmatrix} = \begin{bmatrix} \alpha_{b_{k+1}}^{b_k} + R_{c_0}^{b_k} R_{b_{k+1}}^{c_0} p_c^b - p_c^b \\ \beta_{b_{k+1}}^{b_k} \end{bmatrix} \tag{26}$$

With the linear least square method, the optimal values of velocity, gravity and scale factors can be obtained.

III. BACK-END OPTIMIZATION

After the front-end processing, the status error is usually large. In order to obtain a higher navigation and position accuracy, the key frames were selected from the sequence images for optimization with the front-end processing results as the initial values in this paper.

A. SLIDING WINDOW FILTERING PRINCIPLE BASED ON THE PROBABILITY GRAPH

The sliding window filtering is very important in the back-end optimization. It can be used to ensure the processing precision, avoid the repeated calculations of camera status, reduce the amount of calculations and improve the speed of operation. In this paper, a sliding window filtering method based on the probability graph was proposed, to apply the relevant knowledge of probability graph into the sliding window filtering, and its basic principle is shown in the figure below.

In Fig.3, the pentagram indicates the landmarks, the square indicates the probability factor function with the normal distribution, the circle indicates the data observed with the camera, and the triangle indicates the data measured with IMU. Assume that the status discarded by sliding window filtering is the old one, i.e. $x_o : x_{o1}, x_{o2}, \cdots, x_{om}$, the remaining status by sliding window filtering is the reserved one, i.e. $x_r : x_{r1}, x_{r2}, \cdots, x_{rm}$., and the newly added key frames in window is the new one, i.e. $x_n : x_{n1}, x_{n2}, \cdots, x_{nm}$. The yellow dotted line box indicates the filtered old status, the blue dotted line box indicates the latest observation status, and the green dotted line box indicates the key frame to be filtered out in the latest observation, that is, it is classified into the old status. During the edge processing with the sliding window filtering, in order to ensure the continuity of the pre-integration

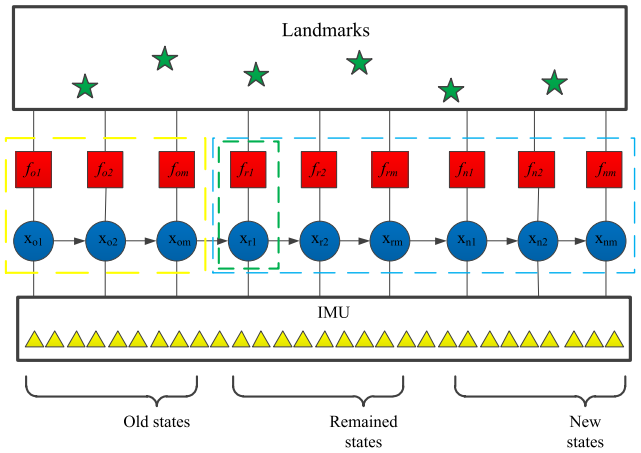


FIGURE 3. Schematic diagram of sliding window filtering based on the probability graph.

of IMU, only the key frames are updated, and no IMU measurement value is discarded.

As can be seen from Fig.3, the key frames are divided into discarded ones, reserved key ones and newly added key ones. Based on the probability graph model, the following formula can be obtained:

$$\begin{aligned} p(x|z) &\propto p(x_o)\, p(x_r, x_n|z) \\ &= \frac{1}{\sqrt{2\pi \hat{P}_0}} \exp\left(-\frac{1}{2}\left\|x_o - \hat{x}_o\right\|^2_{\hat{P}_0}\right) \\ &\quad \cdot \frac{1}{\sqrt{2\pi Q}} \exp\left(-\frac{1}{2}\left\|z - h(x_r, x_n)\right\|^2_Q\right) \end{aligned} \tag{27}$$

where, z is all available observations, i.e. observation landmark information, $p(x_o) = \prod_{i=1}^{m} f_{oi}$ is the joint probability density of the discarded key frames. $p(x_r, x_n|z) = \prod_{i=1}^{m} (f_{ri} f_{oi})$ is the joint probability density of the reserved key frames and the newly added key frames, the index term $\left\|x_o - \hat{x}_o\right\|^2_{\hat{P}_0} = \left(x_o - \hat{x}_o\right)^T \hat{P}_0^{-1} \left(x_o - \hat{x}_o\right)$ is Mahalanobis distance, where $\hat{x}_o, \hat{P}_0$ are the mean and covariance of the prior distribution of the discarded key frames, respectively, $h(x_r, x_n)$ is the re-projection function, and Q is the covariance matrix of error.

By transforming the above formula into the form of negative logarithm function, the error function Δ can be derived as follows:

$$\Delta = \left\|x_o - \hat{x}_o\right\|^2_{\hat{P}_0} + \left\|z - h(x_r, x_n)\right\|^2_Q \tag{28}$$

Assuming that $x_{o'}$ is the key frame filtered out from the latest measurement, and $x_{r'}$ is the remaining key frame from the latest measurement, by using the least square method, the error function $\left\|z - h(x_r, x_n)\right\|^2_Q$ can be transformed into the following form:

$$\begin{bmatrix} Q_{o'o'} & Q_{o'r'} \\ Q^T_{o'r'} & Q_{r'r'} \end{bmatrix} \begin{bmatrix} \delta x_{o'} \\ \delta x_{r'} \end{bmatrix} = \begin{bmatrix} g_{o'} \\ g_{r'} \end{bmatrix} \tag{29}$$

where, $Q_{o'o'}, Q_{o'r'}, Q_{r'r'}$ are the covariances between the filtered key frames and the remaining key frames, and $g_{o'}, g_{r'}$ are the constant terms obtained after the least square processing.

With Shure transformation, the error function can be derived as follows:

$$\begin{cases} \delta x_{o'} = (Q_{o'r'})^{-1} (g_{o'} - Q_{o'r'} \delta x_{r'}) \\ \delta x_{r'} = \left(Q_{r'r'} - Q^T_{o'r'} (Q_{o'o'})^{-1} Q_{o'r'}\right)^{-1} \\ \qquad \left(g_{r'} - Q^T_{o'r'} (Q_{o'o'})^{-1} g_{o'}\right) \end{cases} \tag{30}$$

Because the discarded key frames are independent of each other, the error function of the key frames filtered after the latest observation can be added to the error function of the old status as the prior distribution of the error of the old status of the next measurement for further optimization.

B. IMAGE RE-PROJECTION ERROR BASED ON THE MIXTURE OF DIRECT METHOD AND FEATURE POINT METHOD

During the optimization process with the direct method, the feature matching is not required, and there is less dependence on matching feature points, therefore, it has stronger adaptability and robustness. The feature point method is not sensitive to light intensity and has high positioning accuracy. In this paper, a hybrid method of direct method and feature point method is used to optimize the image re projection.

Assuming that P_i^w is the global coordinates point and $p_i^c = \left[u_i^c \; v_i^c\right]^T$ is the projection coordinates in the camera coordinates system corresponding to the global point, $T_w^b = \left[R_w^b \; p_w^b\right]$ is the transformation matrix from the world coordinates system to the ontology coordinates system, $T_b^c = \left[R_b^c \; p_b^c\right]$ is the transformation matrix from the ontology coordinates system to the camera coordinates system, $h(\cdot)$ is the projection function, and ρ_i is the inverse depth in the camera coordinates system, the re-projection function of the direct method can be derived as follows:

$$\begin{aligned} &I\left(h\left(P_i^w\right)\right) \\ &= I\left(T_b^c T_w^{b_j} P_i^w\right) = I\left(T_b^c T_w^{b_j} T_{b_i}^w T_c^b \left(\frac{1}{\rho_i} h^{-1}\left(p_i^w\right)\right)\right) \\ &= I\left(R_b^c \left(R_w^{b_j} \left(R_{b_i}^w \left(R_c^b \left(\frac{1}{\rho_i} h^{-1}\left(p_i^w\right) + p_c^b\right) + p_{b_i}^w\right) - p_w^{b_j}\right) - p_b^c\right)\right) \end{aligned} \tag{31}$$

where $I(\cdot)$ is the gray function corresponding to the space point.

The optimal function of the direct method is $r\left(\hat{z}^{c_k}\right) = I\left(P_j^w\right) - I\left(h\left(P_i^w\right)\right)$.

The position and pose of ORB feature points were optimized mainly with BA algorithm (the minimum re-projection error method). The optimized function is as follows:

$$\min_{\xi} \Delta(\xi) = \sum_{i=1}^{N} \|e_i\|^2 = \sum_{i=1}^{N} \left\|h\left(P_i^w\right) - \mu_i\right\|^2 \tag{32}$$

where $\mu_i = [u_i, v_i]^T$ is the pixel coordinates of the projection.

The expression of image re-projection error based on the mixture of the direct method and the feature point method is as follows:

$$\begin{aligned}&\sum_{k\in C}\left\|r_C\left(\hat{z}^{c_k},\aleph\right)\right\|^2\\&=\xi\sum_{i,j}H\left(I\left(P_j^w\right)-I\left(h\left(P_i^w\right)\right)\right)\\&\quad+(1-\xi)\sum_{i}H\left(h\left(P_i^w\right)-\mu_i\right)\end{aligned}\tag{33}$$

where, $H(\cdot)$ is Huber kernel function, with which the problem of error growth caused by mismatching can be solved effectively, and ξ is the proportion coefficient of the direct method in the re-projection error, within a range of $0\sim1$, in which 0 indicates the re-projection error method based on the feature points, and 1 indicates the re-projection error method based on the direct method.

With the error minimization method, the optimal solution of the above formula can be derived.

The value of the scale coefficient ξ depends on different application scenarios and environmental conditions. When the number of feature points is relatively small and the gray level changes between the two adjacent frames is not obvious, the proportion coefficient of the direct method based on the histogram equalization should be increased; when the number of feature points is relatively large and the tracking effect is good, the method based on ORB feature points should account for a larger proportion. Therefore, the value of ξ can be determined based on either the number of feature points, or the change of light intensity. The flexibility in determination of ξ enables a wider application scope and a stronger robustness of the method.

C. OVERALL OPTIMIZATION OF SYSTEM ERROR

The status variables for the back-end optimization of the whole system mainly include the position $p_{b_i}^w$, the speed $v_{b_i}^w$, the pose $q_{b_i}^w$, the accelerometer offset a_{bt}, the gyroscope offset w_{bt} of IMU at the corresponding time of the i^{th} key frame, the external parameter from the camera to IMU x_c^b and the measured inverse depth ρ_i of the i^{th} feature point. The expression of the status variables is as follows:

$$\begin{cases}\aleph=\left[x_0,x_1,\cdots,x_n,x_c^b,\rho_0,\rho_1,\cdots,\rho_m\right]\\x_i=\left[p_{b_i}^w,v_{b_i}^w,q_{b_i}^w,a_{bt},w_{bt}\right],\quad i\in[0,n]\\x_c^b=\left[p_c^b,q_c^b\right]\end{cases}\tag{34}$$

The overall optimization function of the system error is as follows:

$$\begin{aligned}\min_{\aleph}\Bigg\{&\left\|r_p\left(x_o,\aleph\right)\right\|^2+\sum_{k\in B}\left\|r_B\left(\hat{z}_{b_{k+1}}^{b_k},\aleph\right)\right\|_{P_{b_{k+1}}^{b_k}}^2\\&+\sum_{k\in C}\left\|r_C\left(\hat{z}^{c_k},\aleph\right)\right\|_{P^{c_k}}^2\Bigg\}\end{aligned}\tag{35}$$

where, $\left\|r_p\left(x_o,\aleph\right)\right\|^2$ is Mahalanobis distance error of prior information composed of key frames filtered out by sliding window filtering, $\sum_{k\in B}\left\|r_B\left(\hat{z}_{b_{k+1}}^{b_k},\aleph\right)\right\|_{P_{b_{k+1}}^{b_k}}^2$ is Mahalanobis distance error sum of IMU pre-integration derived with Formula (14), B is the set of all pre-integration items, $P_{b_{k+1}}^{b_k}$ is the covariance matrix of IMU pre-integration noises, $\sum_{k\in C}\left\|r_C\left(\hat{z}^{c_k},\aleph\right)\right\|_{P^{c_k}}^2$ is the sum of image re-projection errors, C is the set of all key frames, P^{c_k} is the covariance matrix of re-projection noises.

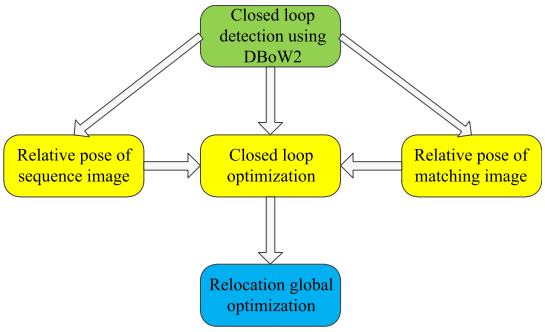


FIGURE 4. Flow chart of closed-loop detection and global optimization.

Each of the above optimization functions is expanded with the first-order Taylor formula, to derive the incremental equation as follows:

$$\begin{aligned}&\left(J_p+\sum\left(J_{b_{k+1}}^{b_k}\right)^T\left(P_{b_{k+1}}^{b_k}\right)^{-1}J_{b_{k+1}}^{b_k}\right.\\&\qquad\left.+\sum\left(J^{c_k}\right)^T\left(P^{c_k}\right)^{-1}J^{c_k}\right)\delta\aleph\\&=\Delta_p+\Delta_B+\Delta_C\end{aligned}\tag{36}$$

where, J_p is Jacobian matrix of the prior information, $J_{b_{k+1}}^{b_k}$ is the Jacobian matrix of IMU pre-integration, J^{c_k} is the Jacobian matrix of re-projection error, and Δ_p, Δ_B, Δ_C are the constant terms of the prior information, IMU pre-integration and re-projection in incremental equation respectively.

The optimal solution of the incremental equation can be derived with LM algorithm.

IV. CLOSED LOOP DETECTION AND GLOBAL OPTIMIZATION

Fig.4 is the flow chart of the closed-loop detection and global optimization. Firstly, DBoW2 is used for closed-loop detection. Secondly, when a closed-loop is detected, it will be optimized. The relative position and pose of the closed-loop optimization consists of two parts, i.e., the relative position and pose of the sequence image and the relative position and pose of the matching image. Finally, the global optimization of relocation is carried out.

A. CLOSED LOOP DETECTION

In this paper, the classic binary word bag model (DBoW2) was used in closed-loop detection. In order to ensure that each image has a sufficient number of feature points, the binary vector brief descriptor was used to describe an image [31].

A dictionary for all feature descriptors was prepared to transform the matching of the images into the matching of the feature descriptors corresponding to each image. With this method, the storage of key frame information can be reduced and the feature matching speed can be increased.

B. CLOSED LOOP OPTIMIZATION

It was assumed that the j^{th} key frame in the window is matched with the i^{th} key frame in the database and the corresponding ontology coordinates are b_j and b_i, respectively, when a closed-loop is detected.

The relative position and pose of closed-loop optimization are mainly composed of the relative positions and poses of the sequence image and the matching image. The relative position and pose of the sequence image are those of the i^{th} key frame corresponding to the j^{th} key frame calculated with IMU and monocular vision integrated navigation, and the measurement expression is as follows:

$$\begin{cases} \hat{p}_{ij}^{b_i} = \hat{R}_w^{b_i}\left(\hat{p}_{b_j}^w - \hat{p}_{b_i}^w\right) \\ \hat{q}_{ij}^{b_i} = \left(\hat{q}_{b_i}^w\right)^{-1} \otimes \left(\hat{q}_{b_j}^w\right) \end{cases} \tag{37}$$

where, $\hat{p}_{b_i}^w$, $\hat{p}_{b_j}^w$ are the translation vectors of the i^{th} key frame and j^{th} key frame measured with IMU and monocular vision integrated navigation, $\hat{q}_{b_i}^w$, $\hat{q}_{b_j}^w$ are the rotation quaternions of the i^{th} key frame and j^{th} key frame measured, $\hat{p}_{ij}^{b_i}$ is the translation vector of the i^{th} key frame in the database relative to the j^{th} key frame in the ontology coordinates system b_i as the reference coordinates system, and $\hat{q}_{ij}^{b_i}$ is the rotation quaternion of the i^{th} key frame in the database relative to the j^{th} key frame in the ontology coordinates system b_i as the reference coordinates system.

With the estimation values of the i^{th} and j^{th} frames, the error optimization function can be derived as follows:

$$\begin{cases} \delta p_{ij}^{b_i} = R_w^{b_i}\left(p_{b_j}^w - p_{b_i}^w\right) - \hat{p}_{ij}^{b_i} \\ \delta q_{ij}^{b_i} = \left(q_{b_i}^w\right)^{-1} \otimes \left(q_{b_j}^w\right) \otimes \left(\hat{q}_{ij}^{b_i}\right)^{-1} \end{cases} \tag{38}$$

The relative position and pose of the matching image can be solved by similarity transformation of the inverse depth information. Because the accumulated error may be generated when the position and pose are calculated with integrated navigation of IMU and vision, the world coordinates system will change from w_i to w_j. The main purpose of relocation is to eliminate the error and change the world coordinates system from w_j to w_i. The error function $\delta S_{w_j}^{w_i}$ of the similarity transformation is as follows:

$$\delta S_{w_j}^{w_i} = H\left(S_{b_i}^{w_i} \cdot \left(S_{b_j}^{w_j} \cdot S_{b_i}^{b_j}\right)^{-1}\right) \tag{39}$$

where, $S_{b_i}^{b_j}$ is the similarity transformation of 3D point of the j^{th} frame and 2D point of the i^{th} frame, $S_{b_j}^{w_j}$ is the similarity transformation of the j^{th} frame in the world coordinates system w_j, $S_{b_i}^{w_i}$ is the similarity transformation of the i^{th} frame in the world coordinates system w_i. $H(\cdot)$ is a Hube kernel function to limit the growth rate of the error and reduce the impact of error matching on the closed-loop optimization results.

The error function of closed-loop optimization consists of two parts: the sequence error expressed as Formula (38) and the matching error expressed as Formula (39). The optimization objective function can be expressed as follows:

$$r_L\left(\delta p_{ij}^{b_i}, \delta q_{ij}^{b_i}, \delta S_{w_j}^{w_i}, \aleph\right) = r_L\left(\delta p_{ij}^{b_i}, \delta q_{ij}^{b_i}, \aleph\right) + r_L\left(\delta S_{w_j}^{w_i}, \aleph\right) \tag{40}$$

C. GLOBAL OPTIMIZATION OF RELOCATION

When a closed-loop is detected, due to the accumulated error in the sequence locating process, in addition to the error between the matched i^{th} key frame and the j^{th} key frame, the rest of the key frames are also affected by the accumulated error. Therefore, the relocation and global optimization should be carried out for all key frames to reduce the position error of all the key frames. During the relocating process, the closed-loop optimization error should be taken into account, so as to calculate the relative position and pose relationships of all key frames in the closed-loop detection. The objective function of relocation global optimization is as follows:

$$\min_{\aleph}\left\{\begin{array}{l} \left\|r_p\left(x_o, \aleph\right)\right\|^2 + \sum_{k\in B}\left\|r_B\left(\hat{z}_{b_{k+1}}^{b_k}, \aleph\right)\right\|_{P_{b_{k+1}}^{b_k}}^2 \\ + \sum_{k\in C}\left\|r_C\left(\hat{z}^{c_k}, \aleph\right)\right\|_{P^{c_k}}^2 \\ + \sum_{i,j\in L}\left\|r_L\left(\delta p_{ij}^{b_i}, \delta q_{ij}^{b_i}, \delta S_{w_j}^{w_i}, \aleph\right)\right\|_{P^{c_l}}^2 \end{array}\right\} \tag{41}$$

where, L is the set of all key frames in closed-loop detection, and P^{c_l} is the covariance matrix of closed-loop optimization.

The optimized relative positions and poses of all key frames in the closed-loop can be obtained by solving the above formula. If there is a tracking loss during the relocating process, a high relocating precision can also be obtained by the relocation and global optimization of the non key frames.

V. EXPERIMENT

In this paper, EuRoC MAV dataset was used for the experiments, and compared with the common methods to verify the effectiveness of the proposed method. The indoor and outdoor environmental data were collected with hand-held camera and IMU, and processed with the method proposed, to verify the environmental adaptability of this method.

A. EXPERIMENTS BASED ON EUROC DATASET

EuRoC dataset contains the image information collected with the binocular camera (with a camera frequency of 20Hz) and corresponding inertial navigation information (with an IMU frequency of 100Hz) and is suitable for the verification of the fusion navigation method based on the vision and IMU data. Four methods were compared in this paper, including

TABLE 1. Experimental comparison results.

	Root mean square error (m)			
Dataset	Stereo-SLAM	VINS-SLAM	VIORB-SLAM	Proposed method
MH_02_easy	0.935	0.345	0.198	0.096
MH_03_medium	0.891	0.567	0.221	0.109
V1_01_easy	0.452	0.227	0.185	0.087
V1_03_difficult	0.519	0.179	0.194	0.094
V2_01_easy	0.601	0.476	0.162	0.089
V2_03_difficult	0.532	0.452	0.157	0.262

Stereo-SLAM method based on binocular camera fusion, VINS-SLAM method based on monocular camera and IMU fusion, VIORB-SLAM method based on feature point camera and IMU fusion [32], and SLAM method based on mixed feature camera and IMU fusion proposed in this paper. The experimental results obtained with the EuRoC data set are shown in TABLE 1.

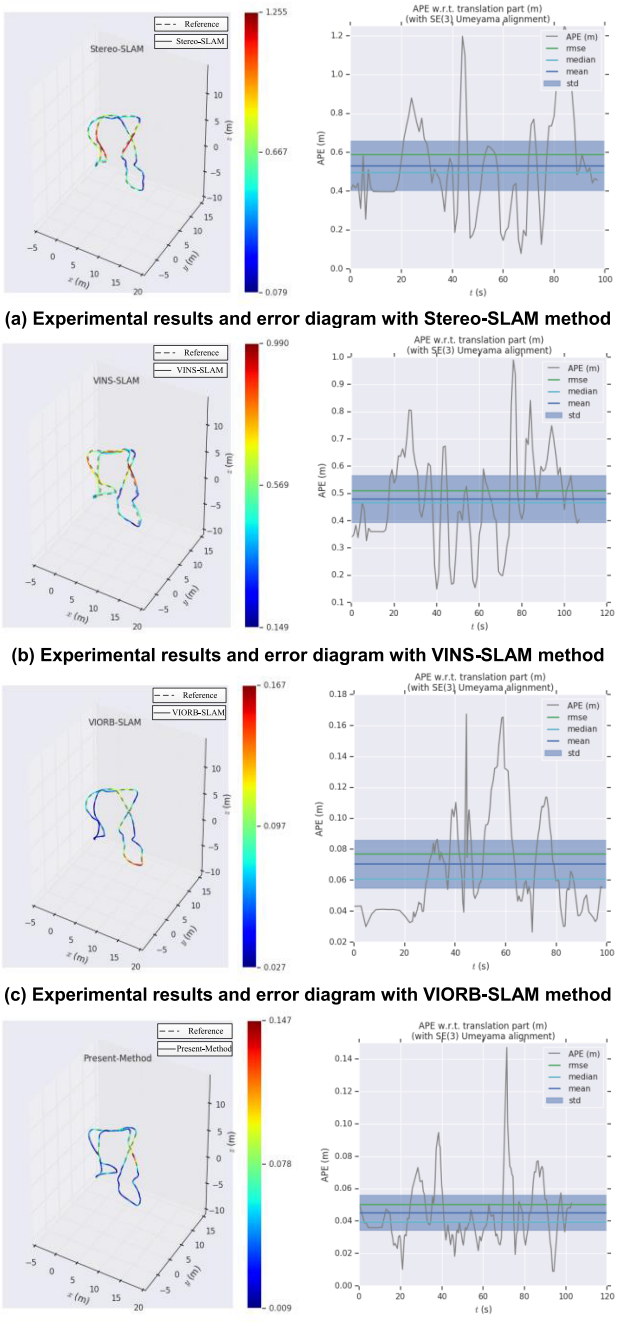
It can be seen from TABLE 1 that the method proposed in this paper has relatively small root mean square error for most of the EuRoC data sets, and the method has good adaptability and robustness.

In this paper, the representative MH_04_difficult dataset was selected for the experiment. This dataset contains clear textures, fuzzy images generated by rapid movement, backgrounds with obvious light and dark changes, etc. These features may increase the difficulty during the dynamic navigation position process, and may also effectively verify the robustness and stability of the algorithm. The experimental results and error diagrams with the four methods are shown in the figures below.

In Fig.5, the fitting diagrams of the experimental data and the real track are shown in the left, and the corresponding error diagrams are shown in the right. In the error diagrams, APE refers to absolute percentage error, rmse refers to root mean square error, median refers to median error, mean refers to average error, and std refers to standard deviation. It can be seen that the errors of results obtained with these methods are in the sequence: Stereo-SLAM > VINS-SLAM > VIORB-SLAM > Present-Method.

With the pure vision SLAM method used to fuse the binocular camera image, the scale drift problem with the monocular SLAM can be solved, but when the distance is relatively long, the position accuracy will be seriously reduced due to the limitation of the binocular camera baseline, and the adaptability to environmental conditions is relatively poor because it only uses the feature information of the image for position. Therefore, the position accuracy of the pure vision SLAM method is the worse and the error is larger than those of the other three methods with the monocular camera and IMU.

In the fusion locating methods with monocular vision and IMU data, the pre-integration of IMU can reduce the calculations and improve the calculation speed effectively. The measurement value of IMU can compensate the depth uncertainty of the monocular camera effectively. The front-end of



(a) Experimental results and error diagram with Stereo-SLAM method

(b) Experimental results and error diagram with VINS-SLAM method

(c) Experimental results and error diagram with VIORB-SLAM method

(d) Experimental results and error diagram with the method proposed in this paper

FIGURE 5. Experimental results and error diagrams based on EuRoC dataset.

VINS-SLAM was processed with the optical flow method, to obtain a good adaptability to environmental conditions with obvious light changes and achieve good position accuracy in the case of rapid movement. However, due to the lack of processing of image feature points, any possible mismatching, which may affect the position accuracy. The front-end of VIORB-SLAM was processed with ORB features, to obtain a relatively high position accuracy and the locating method is suitable for cases with relatively slow moving speed.

TABLE 2. Camera calibration and IMU initialization results.

Camera calibration results	f_x	f_y	c_x	c_y
	503.4387	503.2271	628.6168	433.6075
Transformation matrix from camera to IMU	$\begin{bmatrix} 0.9997 & 0.0046 & 0.0056 \\ -0.0046 & 0.9999 & -0.0007 \\ -0.0056 & 0.0007 & 0.9999 \end{bmatrix}$			
IMU parameter	n_w	n_a	w_{bt}	a_{bt}
	0.12	0.009	0.00005	0.00005
Acceleration of gravity $\boldsymbol{g}$	9.8019			

The matching method based on ORB features can improve the matching accuracy, thus improving the position accuracy.

In this paper, Runge Kutta method was used for IMU pre-integration, which can improve the pre-integration accuracy effectively, so as to provide more accurate position and pose information for the fusion algorithm. In the front-end processing, the sparse direct method was used, to ensure the position accuracy and meet the speed requirements, so as to improve the robustness and adaptability of front-end processing. The inverse depth was estimated with Gaussian uniform mixture probability distribution method, to improve the position accuracy effectively. In the back-end optimization with VINS-SLAM and VIORB-SLAM, the sliding window principle was applied and the image relocating error was optimized in the global optimization. The sliding window processing method based on the probability graph was used in the back-end optimization and the mixed relocation optimization method based on the direct method and the feature point method was used in the global optimization, to improve the accuracy of the back-end optimization effectively.

B. EXPERIMENTS WITH HANDHELD CAMERA AND IMU

The experimental equipment used in this paper were shown in Fig.6. The processor of the laptop is Intel i7-6500U with a frequency of 2.6GHz, and a memory of 16G, which can provide powerful computing power for real-time data processing. The binocular camera has a image acquisition frequency of 50Hz. The image size is 1280 × 800 pixels. The baseline length of the binocular camera is 12 cm. The data acquisition frequency of IMU is 1000Hz. In the experiment, the left camera of binocular camera and IMU data are used for fusion. Before further experiments, the camera and IMU need to be calibrated and initialized. The experimental results are shown in TABLE 2.

After initialization, the accuracy of data fusion between camera and IMU can be improved. In order to verify the applicability and robustness of the proposed method, this experiment consisted of two parts: indoor part and outdoor part.

C. INDOOR ENVIRONMENT EXPERIMENT

The indoor environment experiment was carried out in the laboratory. The hand-held camera and IMU were moved around the laboratory to collect and process the real-time information and obtain the experimental trajectory. In this paper, the operation trajectories with different methods were compared, and the importance of the closed-loop detection to the global optimization was verified. The experimental results are as follows:

The total length of indoor moving trajectory was 20 m. As can be seen from Fig.7: when no closed-loop was detected, based on the operation results with VIORB-SLAM

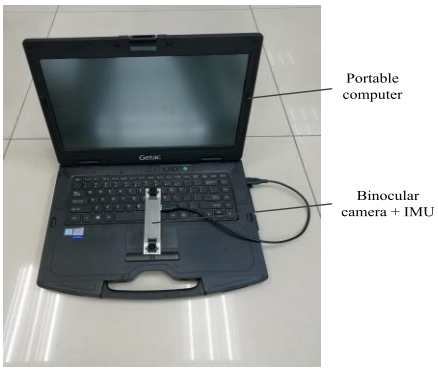
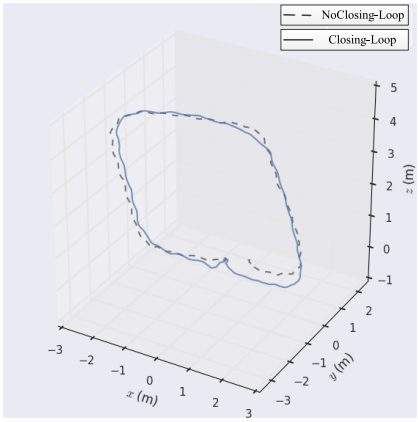
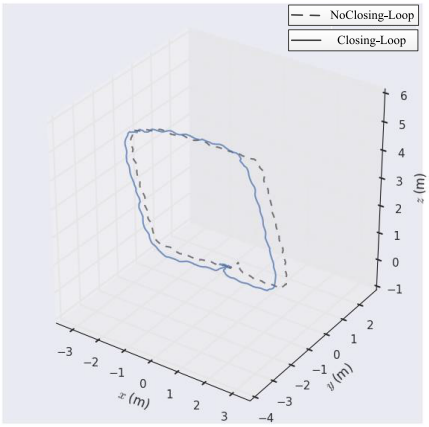


FIGURE 6. Handheld camera and IMU experimental devices.



(a) Experimental results with VIORB-SLAM method



(b) Experimental results with the proposed method

FIGURE 7. Comparison of results of the indoor environment experiments.

method, the final drift errors in x-axis, y-axis and z-axis were $[0.392\ 0.03\ 0.264]$, accounting for 2.37% of that in the total length; based on the operation results with the method proposed in this paper, the final drift errors in x-axis, y-axis and z-axis were $[-0.237\ 0.008\ -0.012]$, accounting for 1.18% of that in the total length.

If the starting point of the indoor experiment is taken as the origin, the hand-held camera and IMU circle the laboratory and return to the starting point again, and maintain the same position and posture as the beginning. It can be seen from the operation results of VIORB-SLAM that the drift errors of pitch, yaw and roll were $[0.231^\circ\ 0.201^\circ\ 1.53^\circ]$. According to the operation results of the proposed method, the drift errors of pitch, yaw and roll were $[0.083^\circ\ 0.05^\circ\ 0.733^\circ]$.

In this paper, Runge Kutta method was used to improve the pre-integration precision of IMU, the mixed inverse depth estimation method based on the probability graph was used to improve the precision of camera depth estimation, and the mixed optimization method of the direct method and the feature point method was used to improve the precision of the global optimization. Therefore, when no closed loop is detected, the proposed method can reduce the drift error and improve the robustness of the whole system effectively.

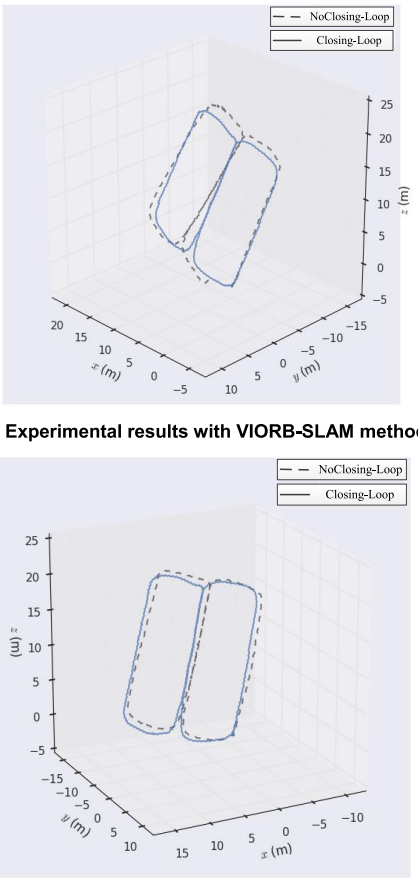
When the closed-loop detection was added, the two methods could eliminate the drift error effectively. In this paper, the closed-loop optimization method based on similar transformation is used to eliminate the accumulated error. Therefore, the closed-loop detection is very important in the whole relative navigation process and could eliminate the scale drift error of the camera and the inertial accumulation error of IMU, so as to reduce the global error and improve the position accuracy.

D. OUTDOOR ENVIRONMENT EXPERIMENT

The outdoor environment experiment was carried out in the campus. The hand-held camera and IMU were moved in 8-shape track on a playground to collect the real-time data and obtain the moving trajectory for the closed-loop detection experiment and the two methods were compared. The experimental results were as follows:

The total length of operation trajectory in outdoor environment was 120m. As can be seen from Fig.8: when no closed-loop was detected, based on the operation results with VIORB-SLAM method, the final drift errors in x-axis, y-axis and z-axis were $[3.152\ 1.364\ 0.620]$, accounting for 2.92% of that in the total length; based on the operation results of the method proposed in this paper, the final drift errors in x-axis, y-axis and z-axis were $[-0.698\ -0.311\ 1.741]$, accounting for 1.58% of that in the total length.

If the starting point of the outdoor experiment is taken as the origin, the hand-held camera and IMU circle around the campus and return to the starting point again, and maintain the same position and posture as the beginning. It can be seen from the operation results of VIORB-SLAM that the drift errors of pitch, yaw and roll were $[1.616^\circ\ 6.61^\circ\ 7.033^\circ]$. According to the operation results of the proposed method, the drift errors of pitch, yaw and roll were $[0.967^\circ\ 1.033^\circ\ 3.133^\circ]$.



(a) Experimental results with VIORB-SLAM method

(b) Experimental results with the proposed method

FIGURE 8. Comparison results of outdoor environment experiments.

The experimental results showed that the proposed method can reduce the outdoor locating error and improve the stability of system operation effectively when no closed-loop is detected. When the closed-loop detection was added, the two methods could also reduce the drift error of the system and improve the accuracy of global position effectively in outdoor environment.

The above experiments showed that the method proposed in this paper can be used in large-scale environments, and is relatively stable in the scenes with relatively fuzzy textures and with relatively obvious light changes. In cases of fast movement, it can be used for tracking and locating. Even in cases of tracking loss, it can be also used for quick relocating.

It can also be seen from the experiments that the locating error in the outdoor environment is larger than that in the indoor environment. In the indoor environment, there is relatively less interferences and stable light intensity, which is easy to extract the feature points and use the direct method for locating. In the outdoor environment, there is uncertain and unstable light intensity, the scale of the external environment is usually large, and the depth estimation error of the camera is relatively large, which increases the locating error in the

outdoor environment. The method proposed in this paper can effectively reduce locating error and provide strong stability in large-scale environments with strong interferences.

VI. CONCLUSION

During the fusion locating process with vision and IMU data, the pre-integration of IMU is very important in the whole optimization process, which can avoid the problem of repeated calculations effectively in the optimization process. In this paper, the pre-integration method based on Runge Kutta method was used to improve the pre-integration efficiency.

In the initialization of the system, the sparse direct method based on the histogram equalization was used in this paper to calculate the position and pose of the camera and the positions and poses of large number of key frames could be calculated quickly, to provide the initial values for the back-end optimization. A mixed probability model was used to estimate the inverse depth of the camera, which was robust and less affected by external interferences.

In the back-end optimization, the sliding window filtering method based on the probability graph model was proposed in this paper, which was relatively intuitive and convenient, and was easy to update the old and new status, so as to avoid the repeated calculations of the camera status, reduce the calculations, and improve the calculation speed.

In the calculation of the re-projection error, a mixed re-projection method integrated with the high speed of the direct method and the high precision and closed-loop ability of the feature point method was proposed in this paper, which could not only increase the robustness of the re-projection method, but also improve the calculation speed. In the global optimization of relocation, the mixed re-projection error was also added to improve the accuracy and speed of the global optimization.

In the closed-loop optimization, a closed-loop optimization method based on similar transformation is proposed to eliminate the accumulated error.

In conclusion, the integrated autonomous relative navigation method based on the vision and IMU data fusion proposed in this paper has stronger robustness and higher calculation accuracy.

REFERENCES

[1] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.

[2] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[3] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, Apr. 2017.

[4] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 15–22.

[5] E. Perdices and J. Cañas, "SDVL: Efficient and accurate semi-direct visual localization," *Sensors*, vol. 19, no. 2, p. 302, 2019.

[6] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 834–849.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[8] R. Mur-Artal and J. D. Tardos, "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.

[9] W. Liu, S. Wu, Z. Wu, and X. Wu, "Incremental pose map optimization for monocular vision SLAM based on similarity transformation," *Sensors*, vol. 19, no. 22, p. 4945, 2019.

[10] G. P. Huang, A. I. Mourikis, S. I. Roumeliotis, "An observability-constrained sliding window filter for SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 65–72.

[11] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "An observability-constrained sliding window filter for SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 65–72.

[12] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *J. Field Robot.*, vol. 27, no. 5, pp. 587–608, 2010.

[13] K. Eckenhoff, L. Paull, and G. Huang, "Decoupled, consistent node removal and edge sparsification for graph-based SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 3275–3282.

[14] J. Jiang, X. Niu, R. Guo, and J. Liu, "A hybrid sliding window optimizer for tightly-coupled vision-aided inertial navigation system," *Sensors*, vol. 19, no. 15, p. 3418, 2019.

[15] P. Tanskanen, T. Naegeli, M. Pollefeys, and O. Hilliges, "Semi-direct EKF-based monocular visual-inertial odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 6073–6078.

[16] J. Deng, S. Wu, H. Zhao, and D. Cai, "Measurement model and observability analysis for optical flow-aided inertial navigation," *Opt. Eng.*, vol. 58, no. 08, p. 1, Aug. 2019.

[17] S.-P. Li, T. Zhang, X. Gao, D. Wang, and Y. Xian, "Semi-direct monocular visual and visual-inertial SLAM with loop closure detection," *Robot. Auto. Syst.*, vol. 112, pp. 201–210, Feb. 2019.

[18] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[19] T. Lupton and S. Sukkarieh, "Visual-Inertial-Aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 61–76, Feb. 2012.

[20] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3565–3572.

[21] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 5303–5310.

[22] M. Quan, S. Piao, M. Tan, and S.-S. Huang, "Accurate monocular visual-inertial SLAM using a map-assisted EKF approach," *IEEE Access*, vol. 7, pp. 34289–34300, 2019.

[23] 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. 1004–1020, Aug. 2018.

[24] Y. Lin, F. Gao, T. Qin, W. Gao, T. Liu, W. Wu, Z. Yang, and S. Shen, "Autonomous aerial navigation using monocular visual-inertial fusion," *J. Field Robot.*, vol. 35, no. 1, pp. 23–51, Jan. 2018.

[25] Z. Yang and S. Shen, "Monocular visual–inertial state estimation with online initialization and camera–IMU extrinsic calibration," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 39–51, Jan. 2017.

[26] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, Feb. 2017.

[27] K. Eckenhoff, P. Geneva, and G. Huang, "Closed-form preintegration methods for graph-based visual–inertial navigation," *Int. J. Robot. Res.*, vol. 38, no. 5, pp. 563–586, Apr. 2019.

[28] J. Solà, "Quaternion kinematics for the error-state Kalman filter," 2017, *arXiv:1711.02508*. [Online]. Available: http://arxiv.org/abs/1711.02508

[29] S. Shen, Y. Mulgaonkar, and N. Michael, "Initialization-free monocular visual-inertial state estimation with application to autonomous MAVs," in *Experimental Robotics*. Cham, Switzerland: Springer, 2016, pp. 211–227.

[30] W. Liu, S. Wu, X. Wu, and K. Li, "Mixed probability inverse depth estimation based on probabilistic graph model," *IEEE Access*, vol. 7, pp. 72591–72603, 2019.

[31] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.
[32] R. Mur-Artal and J. D. Tardos, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 796–803, Apr. 2017.

WENLEI LIU was born in Yantai, China. He received the B.S. degree from the Shandong University of Science and Technology, Qingdao, in 2012, and the M.S. degree from Beihang University, Beijing, in 2016, where he is currently pursuing the Ph.D. degree.

His main research interests include high precision visual relative navigation and the data fusion of multidata sensors to achieve the purpose of cooperative detection and guidance.

SENTANG WU received the Ph.D. degree in dynamics, ballistics, and aircraft motion control systems from National Aviation University, Ukraine, in 1992.

He is currently a Professor of automation science and electrical engineering and a Ph.D. Tutor with Beihang University, Beijing, China. He is also the Navy Missile Expert with the National Defense Basic Research Institute and a member of the Academic Committee. His research interests include the theory and application of nonlinear stochastic systems, computer information processing and control, and aircraft cooperative control, precision, and guidance.

YONGMING WEN was born in 1988. He received the Ph.D. degree from Beihang University. His main research interests include aircraft autonomous formation, flight control systems, and intelligent cooperative guidance and control systems.

XIAOLONG WU was born in Chengdu, China, in 1988. He received the B.S. degree from Sichuan University, Chengdu, in 2010, and the Ph.D. degree from Beihang University, Beijing, in 2017.

He is currently a Research Associate. His research interests include overall design of the guidance and control system for unmanned aerial vehicles, autonomous decision, online planning, and application of computer vision.