Research on Visual-Inertial SLAM Technology with GNSS Assistance

Lin Zhao, Xiaohan Wang, Xiaoze Zheng, and Chun Jia(✉)

Harbin Engineering University, Harbin, China
jiachun@hrbeu.edu.cn

Abstract. In the new era of robustness and perception, the Visual-Inertial Odometry (VIO), which is tightly coupled by the camera and the Inertial Measurement Unit (IMU), can obtain high-precision local pose results in unknown environment. Its low cost and miniaturization have received widespread attention. However, due to the limitation of the measurement principle, in the long-term runs, error will still accumulate. In addition, the outdoor large-scale environment is also a major challenge facing VIO. The Global Navigation Satellite System (GNSS) can provide accurate global estimates for VIO in an open outdoor environment and correct drift caused by long-term operation. Similarly, VIO can still perform in environments where GNSS is denied, which makes it possible for seamless indoor and outdoor navigation. Therefore, this paper proposes a visual-inertial SLAM algorithm assisted by GNSS. Taking the optimized tightly coupled VIO as the main body, and the pose information obtained by GNSS is combined with the VIO solution result to enhance the global positioning while ensuring the accuracy of the local pose accuracy. To this end, a simulation experiment based on the KITTI data set was carried out. The results show that the VIO system with the aid of GNSS can achieve the accuracy of 1.687 m error average, 1.176 m standard deviation, and 2.056 m root mean square error, which is nearly 80% higher than that without assistance. And it can also play a role in the environment where GNSS is denied, and the robustness of the system is also enhanced.

Keywords: SLAM · VIO · GNSS · Sensor fusion · Non-linear optimization

1 Introduction

In recent 30 years, with the development of computer vision, machine learning and other technologies and the improvement of hardware manufacturing level, visual SLAM, which selects camera as the main sensor, with its advantages of rich information collection and relatively low cost, has gradually been valued by the industry, and has played a great application potential in many fields such as robot, UAV and AR/VR [1, 2]. According to the era division of the research process of SLAM by Cadena, after the classic era and algorithm analysis era, the current SLAM research is in the era of robustness and perception [3]. At this stage, the main characteristics and research keystone of SLAM focus on system robustness performance, high-level understanding, resource awareness and task driven perception.

C. Yang and J. Xie (Eds.): *China Satellite Navigation Conference (CSNC 2022) Proceedings*, LNEE 909, pp. 425–434, 2022.
https://doi.org/10.1007/978-981-19-2580-1_36

Obviously, SLAM using a single sensor is difficult to meet the development needs of the times, and does not meet the needs of current theoretical research. It can be said that SLAM problem promotes the idea of multi-sensor information fusion to a certain extent. Among them, through the joint optimization of sparse visual feature point observation and IMU measurement, the visual-inertial SLAM (or Visual Inertial System, VINS) can effectively deal with the blur of fast-moving camera and correct the accumulated error of IMU, so as to further improve the robustness of the system. It provides an effective solution for the miniaturization and low cost of SLAM, and is the research hotspot in multi-sensor information fusion SLAM [4].

However, a large number of studies show that although the camera can correct the cumulative error of IMU, due to the limitation of odometer measurement principle and the lack of explicit loop closure, VIO still has cumulative error in long-term operation. In the outdoor large-scale environment, GNSS is still a more easy-to-use and reliable global positioning method [5]. GNSS can provide accurate global pose estimation for VIO, which is conducive to correcting the drift of VIO and improving the accuracy and continuity of navigation and positioning. When GNSS refuses, VIO can still play a role in maintaining the need for navigation.

Based on this, this paper proposes a visual-inertial SLAM scheme assisted by GNSS. Taking the tightly coupled VIO based on optimization as the main body, the IMU angular velocity and acceleration information are modeled and optimized together with the visual features, and the pose information obtained by the combination of GNSS original observations is fused with the VIO solution results, so as to ensure the local pose accuracy and enhance the global positioning accuracy at the same time. It can achieve long-term and high-precision positioning results in indoor and outdoor environments.

2 Visual-Inertial Odometry

2.1 Visual Front End Based on Optical Flow Tracking

There are two main implementation methods for the front end of traditional visual SLAM. First, based on the method of feature points, the key points and descriptors are calculated to realize feature extraction and matching, and the camera motion is optimized by minimizing the reprojection error. This method is stable and not easy to be affected by illumination changes, but it takes a long time to calculate the descriptor and performs poorly in the case of missing features. The second is the direct method, which only requires light and shade changes in the scene, and estimates the camera motion by minimizing the photometric error. The direct method eliminates the calculation of key points and descriptors and avoids the lack of features. However, it completely depends on gradient search, has the problem of non-convexity, and the gray invariant assumption is difficult to meet in the real environment.

Considering the shortcomings of the above two methods, we choose the compromise scheme of key point extraction and optical flow tracking. This method retains the feature points, only extracts the key points, and uses the optical flow tracking method to replace the descriptor matching, which can take into account the accuracy and robustness to a certain extent.

The key point extraction algorithm here is Shi-Tomasi corner extraction algorithm [6], which is an improved method based on Harris corner extraction [7]. The basic principle of Harris corner extraction algorithm is to take the target pixel as the center, calculate the change of gray curvature in its window, and select the point with the largest curvature change as the feature point.

On this basis, KLT optical flow tracking is used to track the feature points. KLT algorithm is an improved method based on LK optical flow tracking, which tracks the same feature points of two consecutive images [8]. KLT algorithm has three assumptions: First, the image should keep the brightness constant; Second, movement is continuous or small; Third, the space is consistent, and the motion changes of adjacent points are similar. These three points can be met under normal circumstances.

2.2 IMU Preintegration

Because the sampling frequencies of IMU and camera are inconsistent, in order to use them for data fusion, the problem of data synchronization must be considered. The processing method of IMU preintegration [9] solves this problem well.

The so-called IMU preintegration is to take the IMU relative measurement information between two key frames at different times as a constraint to estimate the IMU state at the next time (Fig. 1).

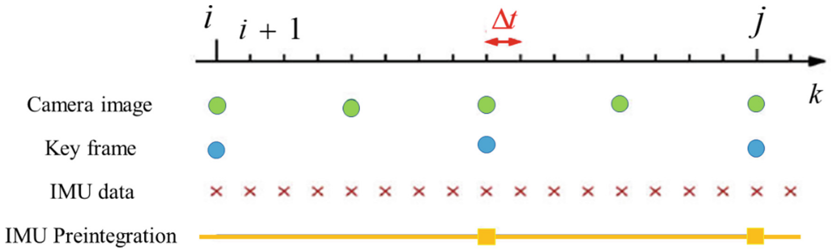

Fig. 1. IMU preintegration

Generally, six-axis IMU can output three-dimensional acceleration and three-dimensional angular velocity information, and its measurement model is:

$$\begin{aligned} \hat{\boldsymbol{a}}_{\mathrm{B}} &= \boldsymbol{R}_{\mathrm{B}}^{\mathrm{W}}\left(\boldsymbol{a}_{\mathrm{W}} - \boldsymbol{g}_{\mathrm{W}}\right) + \boldsymbol{b}_{\mathrm{a}} + \boldsymbol{\eta}_{\mathrm{a}} \\ \hat{\omega}_{\mathrm{B}} &= \boldsymbol{\omega}_{\mathrm{B}} + \boldsymbol{b}_{\mathrm{g}} + \boldsymbol{\eta}_{\mathrm{g}} \end{aligned} \tag{1}$$

Among them, the subscripts B and W represent the IMU coordinate system and the world coordinate system respectively. The subscripts a and g represent accelerometer and gyroscope, respectively. $\boldsymbol{a}_{\mathrm{W}}$ and $\boldsymbol{g}_{\mathrm{W}}$ represents the real acceleration of motion and gravitational acceleration in the world coordinate system, and $\boldsymbol{R}_{\mathrm{B}}^{\mathrm{W}}$ represents the rotation matrix between the IMU coordinate system and the world coordinate system; $\boldsymbol{\omega}_{\mathrm{B}}$ represents the angular velocity in the IMU coordinate system; $\boldsymbol{b}_{\mathrm{a}}$ and $\boldsymbol{b}_{\mathrm{g}}$ represent the zero offset of the accelerometer and the gyroscope respectively; $\boldsymbol{\eta}_{\mathrm{a}}$ and $\boldsymbol{\eta}_{\mathrm{g}}$ represent respectively Measurement noise of two sensors.

In order to calculate the motion information from the measured values of IMU, the following kinematic model is introduced:

$$\begin{aligned} \dot{\boldsymbol{p}}_{\mathrm{W}}^{\mathrm{B}} &= \boldsymbol{v}_{\mathrm{W}}^{\mathrm{B}} \\ \dot{\boldsymbol{v}}_{\mathrm{W}}^{\mathrm{B}} &= \boldsymbol{a}_{\mathrm{W}} \\ \dot{\boldsymbol{R}}_{\mathrm{W}}^{\mathrm{B}} &= \boldsymbol{R}_{\mathrm{W}}^{\mathrm{B}} \boldsymbol{\omega}_{\mathrm{B}}^{\wedge} \end{aligned} \tag{2}$$

Among them, $\boldsymbol{p}_{\mathrm{W}}$ and $\boldsymbol{v}_{\mathrm{W}}$ respectively represent the three-dimensional position and velocity in the world coordinate system, and $(\cdot)^{\wedge}$ represents the antisymmetric matrix operator.

Based on IMU measurement model and kinematics model, the preintegration formula of IMU can be further obtained:

$$\begin{aligned} \boldsymbol{p}_{\mathrm{W}}^{\mathrm{B}}(t+\Delta t) &= \boldsymbol{p}_{\mathrm{W}}^{\mathrm{B}}(t) + \boldsymbol{v}_{\mathrm{W}}^{\mathrm{B}}(t)\Delta t + \frac{1}{2}\boldsymbol{g}_{\mathrm{W}}\Delta t^{2} \\ &\quad + \frac{1}{2}\boldsymbol{R}_{\mathrm{W}}^{\mathrm{B}}(t)\cdot\left(\hat{\boldsymbol{a}}_{\mathrm{B}}(t) - \boldsymbol{b}_{\mathrm{a}}(t) - \boldsymbol{\eta}_{\mathrm{ad}}(t)\right)\Delta t^{2} \\ \boldsymbol{v}_{\mathrm{W}}^{\mathrm{B}}(t+\Delta t) &= \boldsymbol{v}_{\mathrm{W}}^{\mathrm{B}}(t) + \boldsymbol{g}_{\mathrm{W}}\Delta t + \boldsymbol{R}_{\mathrm{W}}^{\mathrm{B}}(t)\cdot\left(\hat{\boldsymbol{a}}_{\mathrm{B}}(t) - \boldsymbol{b}_{\mathrm{a}}(t) - \boldsymbol{\eta}_{\mathrm{ad}}(t)\right)\Delta t \\ \boldsymbol{R}_{\mathrm{W}}^{\mathrm{B}}(t+\Delta t) &= \boldsymbol{R}_{\mathrm{W}}^{\mathrm{B}}(t)\cdot\operatorname{Exp}\left(\left(\hat{\omega}_{\mathrm{B}}(t) - \boldsymbol{b}_{\mathrm{g}}(t) - \boldsymbol{\eta}_{\mathrm{gd}}(t)\right)\Delta t\right) \end{aligned} \tag{3}$$

The above three formulas are the preintegration calculation formula of IMU. The so-called preintegration is to integrate the original output data of IMU to obtain the integration of measured values. Therefore, this integration can be carried out immediately when the IMU measurement data are obtained, and it can be "pre" integrated before the optimization calculation. At the same time, in the process of optimization calculation, even if the position $\boldsymbol{p}$, velocity $\boldsymbol{v}$ or rotation matrix $\boldsymbol{R}$ change, it will not affect the preintegrated value, so there is no need to re integrate the IMU data.

2.3 Nonlinear Optimization Based on Sliding Window

The essence of SLAM problem is state estimation. In order to describe the generality of, motion equation and observation equation are commonly used to describe:

$$\begin{cases} \boldsymbol{x}_{k} = f\left(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k}, \boldsymbol{w}_{k}\right) \\ z_{k,j} = h\left(\boldsymbol{y}_{j}, \boldsymbol{x}_{k}, \boldsymbol{v}_{k,j}\right) \end{cases} \tag{4}$$

Among them, the upper formula represents the motion equation, $\boldsymbol{x}_k$ represents the camera pose at the moment k, $\boldsymbol{u}_k$ represents the input or reading of the motion sensor (here is IMU), $\boldsymbol{w}_k$ represents the noise of the process; the lower formula represents the observation equation, $z_{k,j}$ represents the observation that the camera observes the landmark $\boldsymbol{y}_k$ at the location $\boldsymbol{x}_k$, $\mathbf{v}_{k,j}$ represents the observed noise of the process.

The so-called SLAM problem is how to estimate the position of the camera and the position of the road sign after obtaining the motion sensor input and the observation of the road sign, that is, the problem of "positioning" and "map construction". Mathematically, it is a process of solving the maximum a posterior (MAP). Assuming that all measurements are independent of each other and that the noise of each measurement conforms to the zero mean Gaussian distribution, the MAP problem can be further transformed into the minimum estimation of the sum of cost functions:

$$\begin{aligned}\mathcal{X}^* &= \arg\max_{\mathcal{X}} p(\mathcal{X}|z) \\ &= \arg\max_{\mathcal{X}} p(\mathcal{X})p(z|\mathcal{X}) \\ &= \arg\max_{\mathcal{X}} p(\mathcal{X}) \prod_{i=1}^{n} p\left(z_i|\mathcal{X}\right) \\ &= \arg\min_{\mathcal{X}} \left\{ \left\| \boldsymbol{r}_p - \boldsymbol{H}_p \mathcal{X} \right\|^2 + \sum_{i=1}^{n} \left\| \boldsymbol{r}\left(z_i, \mathcal{X}\right) \right\|_{\mathrm{P_i}}^2 \right\}\end{aligned} \tag{5}$$

where, z represents the set of independent sensors and $\left\{\boldsymbol{r}_p, \boldsymbol{H}_p\right\}$ represents the a priori information of the system.

In the VIO back end mentioned in this paper, in order to ensure the optimization efficiency and save computing resources, a sliding window is set up. The whole back end uses the tight coupling method to integrate the measurement information of vision and IMU, and uses the Bundle Adjustment (BA) method in computer vision to minimize the feature re projection position error of all points in the sliding window and the observation error of IMU. On this basis, the nonlinear optimization method is used to iteratively optimize the pose of key frames, IMU state and the three-dimensional coordinates of road markings.

After the nonlinear optimization method is used to optimize all variables in the sliding window, in order to maintain a constant number of frames in the sliding window, the key frames need to be removed, and the corresponding key frame pose variables, IMU state variables and some road marking variables related to the key frames will be removed. Directly discarding the variables will cause the system to lose information and reduce the accuracy of the system, even the system is not observable. Therefore, in order to solve this problem, the method of marginalization is often used when moving out the variables to convert the variable information into the constraints between the remaining variables, so as to retain the variable information as much as possible.

3 VIO/GNSS Fusion Algorithm

As for the fusion of VIO and GNSS, our processing strategy here is to take the tightly coupled VIO as the main body, and loosely couple the GNSS global pose estimation with the VIO pose results, so as to achieve the effect of GNSS assisting VIO.

3.1 Spatiotemporal Datum Unification

In order to fuse the sensor data, we need to transform the data of different sensors in time and space, so that the simultaneous interpreting of time and coordinates is achieved.

Time unification is realized by PPS signal of GNSS receiver. For the unification of space, the main work is the transformation of coordinate system. After collecting data, GNSS receiver obtains the current longitude, latitude and elevation information through software settlement. Then, the position in the geodetic coordinate system (LLA) needs to be converted to the earth centered earth fixed (ECEF) coordinate system.

3.2 Fusion Method Based on Kalman Filter

Kalman Filter (KF) is an algorithm for optimal estimation of system state by using linear system state equation and system input and output observation data. In the loose combination system, GNSS system and VIO system work independently. VIO outputs position and attitude information, GNSS outputs position information, uses KF algorithm to optimally estimate the system error, and finally the system outputs the corrected VIO parameters (Fig. 2).

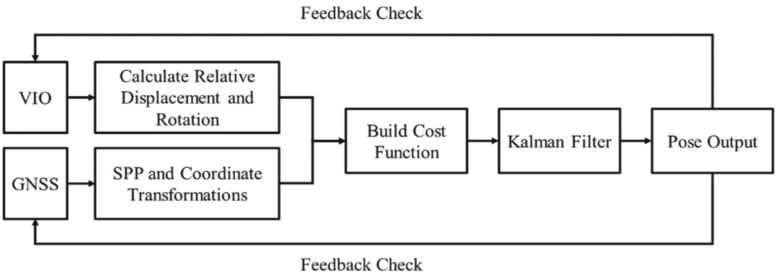
Feedback Check
VIO
Calculate Relative Displacement and Rotation
GNSS
SPP and Coordinate Transformations
Build Cost Function
Kalman Filter
Pose Output
Feedback Check

Fig. 2. Flow chart of VIO and GNSS fusion algorithm

The error state vector is selected as:

$$\tilde{x} = \left[\tilde{P}_x, \tilde{P}_y, \tilde{P}_z, \tilde{V}_x, \tilde{V}_y, \tilde{V}_z\right]^{\mathrm{T}} \tag{6}$$

The error state transfer equation of the combined system is:

$$\dot{\tilde{x}}(t) = \boldsymbol{F}(t)\cdot\tilde{\boldsymbol{x}}(t) + \boldsymbol{G}(t)\cdot\boldsymbol{w}(t) \tag{7}$$

where, the state transition matrix $\boldsymbol{F}(t)=\begin{bmatrix} 0_{3\times3} & \boldsymbol{I}_{3\times3} \\ 0_{3\times3} & 0_{3\times3} \end{bmatrix}$ and the noise driving matrix $\boldsymbol{G}(t)=\begin{bmatrix} \boldsymbol{I}_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & \boldsymbol{I}_{3\times3} \end{bmatrix}$, the system noise is $\boldsymbol{w}(t)$.

The system measurement model is:

$$\boldsymbol{Z}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{V}_k \tag{8}$$

Among them, the quantity measurement $\boldsymbol{Z}_k = \begin{bmatrix} P_x - \widehat{P}_x \\ P_y - \widehat{P}_y \\ P_z - \widehat{P}_z \\ V_x - \widehat{V}_x \\ V_y - \widehat{V}_y \\ V_z - \widehat{V}_z \end{bmatrix}$, and the measurement matrix $\boldsymbol{H}_k = \boldsymbol{I}_{6\times6}$, $\boldsymbol{V}_k$ is the measurement noise.

According to the above model, the optimal unbiased estimation of the combined system can be obtained by using the two steps of "prediction" and "update" of Kalman filter.

4 Simulation and Discussion

4.1 Simulation Dataset

KITTI Dataset [10] is the world's largest computer vision algorithm evaluation dataset released by Karlsruhe Institute of technology in Germany. This dataset can be used to evaluate stereo images, visual ranging, 3D object detection, 3D tracking and so on. KITTI Dataset includes real image data of urban, rural and highway scenes. The sampling platform of KITTI Dataset includes two gray cameras, two color cameras and a GPS navigation system.

Select 2011_10_03_drive_0027_Synced data as test data. The data set has a total length of 3700 m and lasts 430 s. The running track of the proposed algorithm is shown in the figure below (Fig. 3).

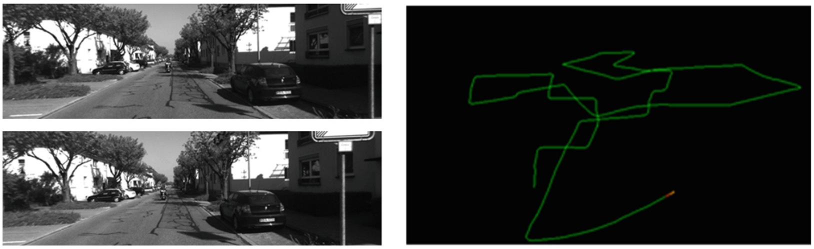

Fig. 3. The camera's left and right eye real-time images and dataset running trajectory

Compared with the official real trajectory map, the early trajectory is more accurate, while the later trajectory drifts, which shows that the long-distance accuracy of VIO is indeed poor without explicit loop closure.

4.2 Analysis of Simulation Results

In order to quantitatively analyze the data, EVO is selected for analysis. EVO is a python tool used to evaluate the measurement data and output estimation of SLAM system. Its core function is to draw the camera motion trajectory. In this paper, EVO is used to draw and compare the motion trajectory, calculate the Absolute Pose Error (APE) and align and compare the trajectory. In order to obviously compare the performance of the algorithms before and after GNSS assistance, we directly draw the error mapping of the two algorithms, as shown in the figure below (Figs. 4, 5).

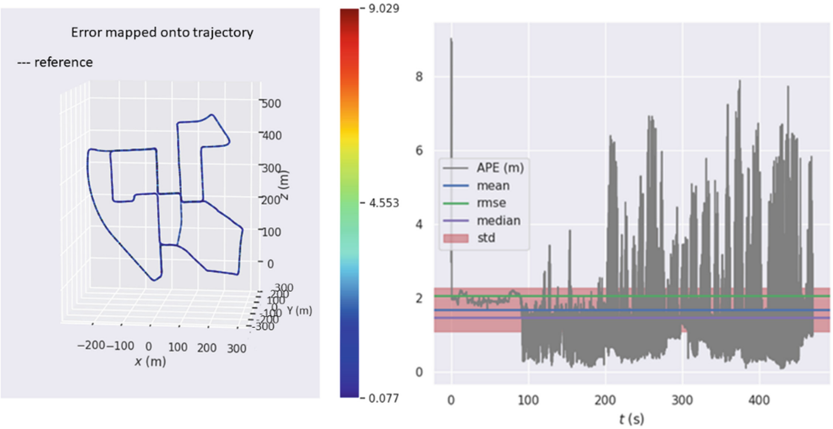

Fig. 4. VIO trajectory error map and APE without GNSS assistance

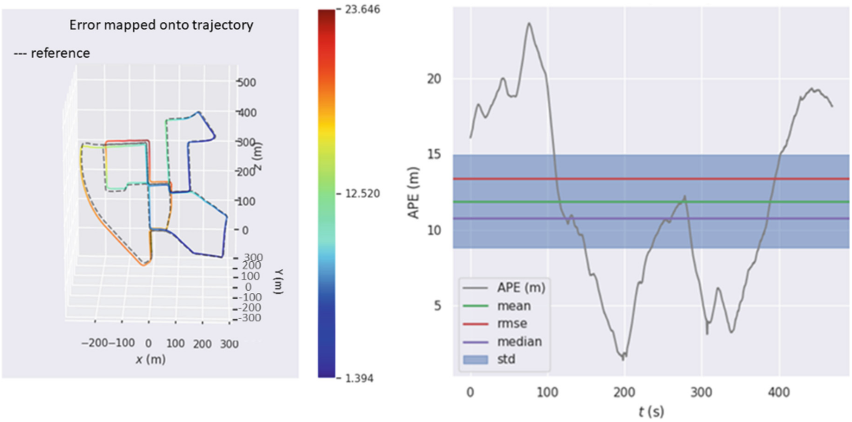

Fig. 5. VIO trajectory error map and APE with GNSS assistance

Relevant data are shown in the table below (Table 1).

Table 1. Results of Metric Absolute Pose Error

Method	RMSE(m)	Mean(m)	Median(m)	Std(m)
Only VIO	13.385	11.896	10.763	6.136
VIO with GNSS assistance	2.056	1.687	1.461	1.176

According to the comparison of APE data, in an outdoor environment with GNSS assistance, the VIO trajectory error is further reduced, the root mean square error is increased from 13.385 m to 2.056 m, the average error is increased from 11.896 m to 1.687 m, and the standard deviation is increased from 6.136 m. The increase is 1.176 m. After the GNSS correction is applied to the long-running VIO system, the navigation accuracy and robustness of the system have been improved to a certain extent. In trajectories with poor GNSS performance, pure VIO can still provide accurate pose estimation, which shows that the GNSS-assisted VIO algorithm in this article is effective.

5 Conclusion

This paper proposes a GNSS-assisted visual-inertial SLAM technology solution, which takes the optimized tightly coupled VIO as the main body, integrates the global positioning results of GNSS, and carries out a simulation experiment based on the KITTI data set. The results show that in terms of long-term positioning, the VIO system assisted by GNSS can achieve an accuracy of 1.687 m, a standard deviation of 1.176 m, and a root mean square error of 2.056 m, which is nearly 80% higher than that without assistance. Even when GNSS cannot be used indoors, the system described in this article degenerates to VIO, which can also provide stable 3D navigation and map construction. The proposed algorithm effectively solves the long-term drift problem of VIO, and it is rejected in GNSS.

Due to the limitations of actual conditions, only simulation experiments have been done, and the effect of VIO operation in the event of GNSS failure has not been evaluated. Since only Single-Point Positioning (SPP) based on pseudorange and KF loose coupling are used, this algorithm does not further explore the use of GNSS. Subsequently, we will conduct research on the fusion of Real-Time Kinematic (RTK) and Precise Point Positioning (PPP) with VIO. Furthermore, GVIO, which tightly coupled with the raw observations of GNSS, will be our key research work in future.

Acknowledgements. This research was jointly funded by the National Key Research and Development Program (No. 2021YFB3901300), the National Natural Science Foundation of China (Nos. 61773132, 61633008, 61803115, 62003109), the 145 High-tech Ship Innovation Project sponsored by the Chinese Ministry of Industry and Information Technology, the Heilongjiang Province Research Science Fund for Excellent Young Scholars (No. YQ2020F009), and the Fundamental Research Funds for Central Universities (Nos. 3072019CF0401, 3072020CFT0403).

References

1. Davison, A.J., Reid, I.D., Molton, N.D., et al.: MonoSLAM: real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1052–1067.9 (2007)
2. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artif. Intell. Rev. **43**(1), 55–81 (2012). https://doi.org/10.1007/s10462-012-9365-8
3. Cadena, C., Carlone, L., Carrillo, H., et al.: Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. IEEE Trans. Rob. **32**(6), 1309–1332 (2016)
4. Leutenegger, S., Lynen, S., Bosse, M., et al.: Keyframe-based visual-inertial odometry using nonlinear optimization. Int. J. Robot. Res. **34**(3), 314–334 (2015)
5. Li, L., Li, Z., Yuan, H., Wang, L., Hou, Y.: Integrity monitoring-based ratio test for GNSS integer ambiguity validation. GPS Solut. **20**(3), 573–585 (2015). https://doi.org/10.1007/s10291-015-0468-y
6. Shi, J., Tomasi: Good features to track. In: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600. IEEE, Seattle, WA, USA (1994)
7. Harris, C.J., Stephens, M.: A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference, Manchester, pp. 147–151 (1988)
8. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence, pp. 674–679. Moran Kaufmann Publishers Inc., San Francisco, CA, USA (1981)
9. Forster, C., Carlone, L., Dellaert, F., et al.: On-Manifold preintegration for real-time visual-inertial odometry. IEEE Trans. Rob. **33**(1), 1–21 (2017)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE, Providence, RI, USA (2012)