Paper:

A Robust Visual Inertial Navigation System Based on Low-cost Inertial Measurement Unit

Zelong Zhang*, Kaoru Hirota, Yaping Dai, Zhiyang Jia

School of Automatic, Beijing Institute of Technology
No.5 Zhongguancun South Street, Haidian District, Beijing 100081, P.R.China
E-mail: zzl178207829@163.com

Abstract. A visual-inertial system (VINS) based on vision sensor is vulnerable to environment illumination and texture, the problem of initial scale ambiguity still exists in a monocular VINS system. The fusion of a monocular camera and an inertial measurement unit (IMU) can effectively solve the scale blur problem, improve the robustness of the system, and obtain higher positioning accuracy. Based on a monocular visual inertial navigation system (VINS-mono), a state-of-the-art fusion performance of monocular vision and IMU, an initialization scheme is designed which can calculate the acceleration bias as a variable during the initialization process so that it can be applied to low-cost IMU sensors. The experimental result on the EuRoc dataset shows that the initial values obtained through the initialization process can be efficiently used for launching nonlinear visual-inertial state estimator and positioning accuracy of the improved VINS-mono has been improved by about 8% than VINS-mono.

Keywords: SLAM, VIO, VINS-mono, sensors fusion

1. Introduction

With the innovations of sensors and algorithms [1], mobile robots are getting smaller and smarter and are addressing new applications in medicine, agriculture, and security applications [2,3]. Simultaneous Localization and Mapping (SLAM) has always been a research hotspot of the robot navigation for many decades.

However, in the navigation process, only relying on the monocular camera information to locate the mobile robot, there is a scale ambiguity problem, which leads to the failure to obtain the real trajectory length. This is the scale ambiguity problem in monocular SLAM, which limits its wide application. RGB-D camera can obtain color image and depth image at the same time, but its measurement distance is limited and contains too much noise [4]. A two-dimensional laser scanner is widely used in indoor positioning, but it contains too little information to perform complex tasks. Three-dimensional laser scanners are not widely used because of its high price. In order to solve this problem, more and more solutions tend to use the sensor fusion method, making use of the different characteristics of data acquired by sensors to complement each sensor's advantages and achieve better results [5,6]. In different sensor modes, the combination of a monocular camera and an IMU has good robustness and low-cost characteristic, so this combination is a potential solution.

The main advantage of this monocular visual-inertial navigation system (VINS) is to have the metric scale, as well as roll and pitch angles, all observable. This enables navigation tasks that require metric state estimates. In addition, the integration of IMU measurements can dramatically improve motion tracking performance by bridging the gap between losses of visual tracks due to illumination change, texture-less area, or motion blur. In fact, the monocular VINS is not only widely used in mobile robots, drones, and mobile devices, it is also the minimum sensor setup for sufficient self and environmental perception. The representative work of VINS scheme include ROVIO[7], KOVIS[8], VINS-mono[9] and VINS-fusion[10].

Based on a monocular visual-inertial navigation system (VINS-mono), a state-of-the-art fusion performance of monocular vision and IMU, this paper designs a new initialization scheme that can calculate the acceleration bias as a variable during the initialization process so that it can be applied to low-cost IMU sensors. Through experiment and analysis in the EuRoc dataset, the result shows that the initial values obtained through this process can efficiently be used to launch nonlinear visual-inertial state estimator and positioning accuracy of the improved VINS-mono has been improved by about 8% than VINS-mono.

The rest of this study is organized as follows. The improved VINS-mono system overall framework is given in Section 2. Then, the initialization process of the improved VINS-mono is described in Section 3. The experiment result and analysis are shown in Section 4. Finally, a conclusion is drawn in Section 5.

2. System Overall Framework

Monocular VINS and visual SLAM are essentially state estimation problems. Based on the VINS-mono project, the IMU and camera data are tightly coupled by nonlinear optimization method. The function module of the improved VINS-mono includes five parts: data prepro-

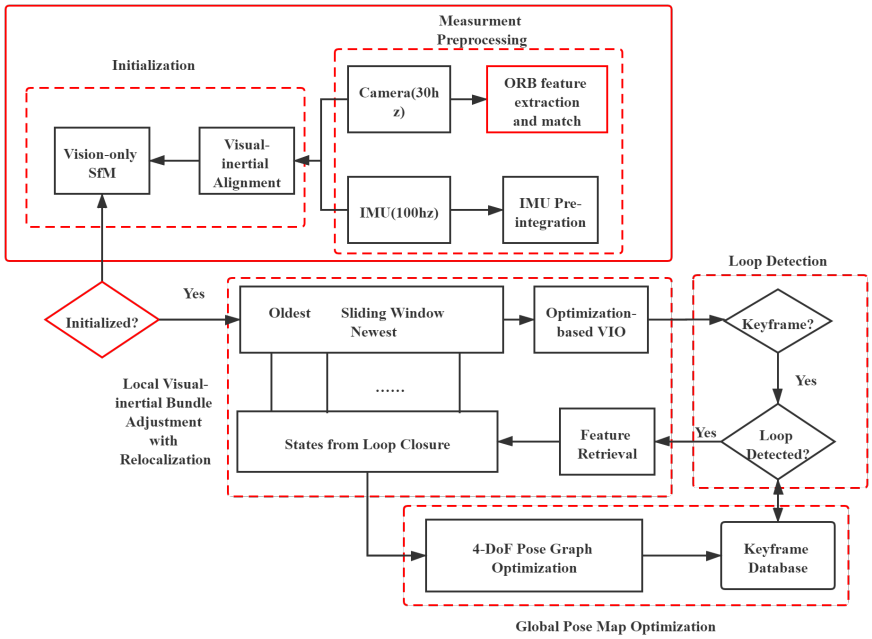

Fig. 1. The improved visual-inertial navigation system architecture framework.

cessing, initialization, back-end nonlinear optimization, Closed-loop Detection and closed-loop optimization. The code mainly opens four threads, including front-end image matching, back-end nonlinear optimization, Closed-loop Detection and closed-loop optimization. The overall framework of the improved VINS-mono system is shown in Figure 1, in which the red solid line represents the improved part compared with VINS-mono. The main functions of each functional module are as follows.

(1) Image and IMU preprocessing. Pyramid representation is used to process the image. The feature points of the sphere are extracted from each layer of the image. The feature point method is used to match the adjacent frames. Outliers were removed by random sample consistency (RANSAC) [11]. Finally, the tracking feature points are pushed to the image queue, and a notification is sent to the back end for processing. The position, velocity and rotation (PVQ) of the current time are obtained by integrating IMU data, and the pre-integration increment, Jacobian matrix of adjacent frames and the covariance term of pre-integration error for back-end optimization are calculated.

(2) Initialization. Structure from Motion (SfM) [12] is used for pure visual estimation of pose and 3D position of all keyframes in sliding window. Then, the initial parameters are calculated with IMU pre integration results.

(3) Local visual-inertial bundle adjustment with relocalization. The visual constraint, IMU constraint and closed-loop constraint are transformed into a large objective function for nonlinear optimization to solve the velocity, position, attitude and deviation of all frames in the sliding window.

(4) Closed-loop detection and optimization. Open source dictionary package (DBoW3) is used for closed-loop detection. When the detection is successful, the whole camera trajectory is closed-loop optimized.

3. The Initialization Process of the Improved VINS-mono

VINS-mono does not initialize acceleration bias b_a which simply sets its initial value to zero, which is not applicable to low-cost IMU. The initialization result directly affects the robustness and positioning accuracy of the entire tightly coupled system.

In this paper, a new initialization scheme is designed, which can calculate the acceleration bias b_a during the initialization process so that it can be applied to low-cost IMU sensors. Besides, the ORB feature point method is used instead of optical flow method to make the initialization model more accurate and robust during the initialization process.

The VINS-mono visual processing uses the optical flow tracking method. The accuracy of the pose solved by the optical flow tracking is not as good as the feature point matching, which has a great influence on the accuracy of the initialization and is directly related to the accuracy of the subsequent motion estimation. In order to improve this situation, in this paper, the ORB feature point method is used for pose estimation in the initialization phase. The procedure of the initialization is shown in Figure 2.

3.1. Visual SfM

The visual initialization uses the key frames image sequences in the initial time about 10 s to perform the pose

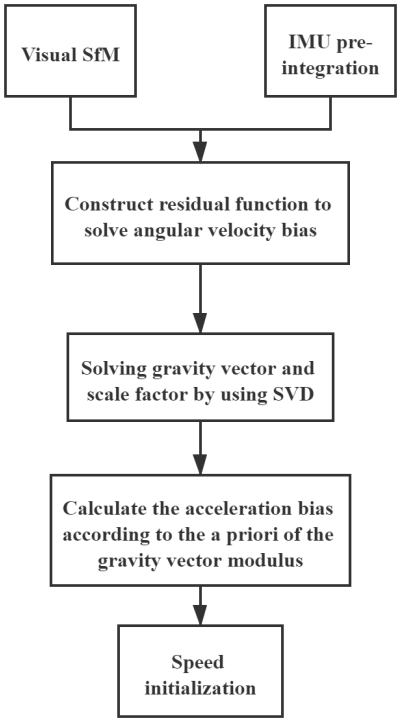

Fig. 2. Procedure of initializing of the improved VINS-mono.

calculation and triangulation as well as further global optimization. The selection of the image key frame is mainly based on the distance of the parallax, and when the parallax distance is greater than a certain threshold, it is selected as a key frame. The vision based SFM technique is used to obtain the more accurate pose and image point coordinates of the key frame sequence. This provides more accurate pose parameters for IMU initialization. In order to make the visual initialization independent of the scene, that is, to determine whether the initial scene is flat or non-planar, a relatively accurate initial value can be obtained. The two initial key frames images of the system adopt a parallel computing fundamental matrix and a homography matrix method and choose the right model according to a specific mechanism. The scene scale is fixed, and the triangle points are initialized according to the initial two key frames, then the Perspective-n-Point (PnP) algorithm is used to restore motion and continuously triangulate to restore the map point coordinates. After tracking a sequence, a Bundle Adjustment (BA) is constructed based on the projection error of the image coordinates for global optimization, and the optimized map points and poses are obtained, as shown in Figure 3.

3.2. Visual-inertial Alignment

The purpose of visual-inertial alignment is to use the results of the visual SfM to decouple the IMU and calculate its initial values separately. The initialization process can be decomposed into four small problems in order to solve:

(1) Estimation of the gyroscope bias.

(2) Estimation of the scale and the gravitational acceleration.

(3) Estimation of the acceleration bias and the optimization of the scale and gravity.

(4) Speed estimation.

In order to describe the movement of the rigid body in three-dimensional space and the positional relationship between the camera and the IMU sensor mounted on the rigid body, the positional transformation relationship is defined as shown in Figure 4. The IMU coordinate system and the rigid body coordinate system (body) are defined as coinciding. T_{BC} represents the transformation of the coordinates in the camera coordinates to the IMU coordinate system, and it is composed of R_{BC}. T_{WB} denotes the transformation relationship between the rigid body coordinate system and the world coordinate system (R_{WB} denotes the rotating part, and W_P denotes the translation part). P_l and Z_l represent world coordinates and image plane coordinates, respectively.

3.3. Gyro Bias Estimation

The bias of the gyroscope can be decoupled from the result of the rotation calculated by the visual SfM and the result of the IMU pre-integration. During the initialization process, it can be assumed that b_g is a constant and does not change over time. Throughout the initialization process, the rotation of the adjacent key frames can be solved by the visual SfM. The rotation between adjacent frames can also be obtained by the pre-integration of the IMU. Assuming that the rotation matrix obtained by visual SfM is accurate, the value of b_g can be calculated indirectly using the difference between the two rotation matrices corresponding to Lie algebra. The exponential map (at the identity) Exp: $so(3)\rightarrow SO(3)$ associates an element of the Lie Algebra to a rotation and coincides with the standard matrix exponential (Rodrigues' formula).

The calculation formula is as follows:

$$\underset{b_g}{argmin}\sum_{i=0}^{n-1}\left\|Log\left(\left(R_{BW}^{i+1}R_{WB}^{i}\right)^{T}\triangle R_{i,i+1}Exp\left(J_{\triangle R}^{g}b_g\right)\right)\right\| \tag{1}$$

where the Jacobians $J_{\triangle R}^{g}$ account for a first-order approximation of the effect of changing the gyroscope biases without explicitly recomputing the pre-integrations. Both pre-integrations and Jacobians can be efficiently computed iteratively as IMU measurements arrive [13]. The above formula $R_{WB}^{(\cdot)}=R_{WC}^{(\cdot)}R_{CB}$, n represents the number of key frames, and $\triangle R_{i,i+1}$ represents the integral value of the gyroscope between two adjacent key frames. The superscript i represents the time of the key frame. $R_{WC}^{(\cdot)}$can be obtained with visual SfM, and R_{CB} is the rotation matrix of the IMU coordinate system in the camera coordinate system. Formula (1) can be solved with the Levenberg–Marquard algorithm based on nonlinear optimization, which is more robust than the Gauss–Newton method, and the value of b_g can be decoupled.

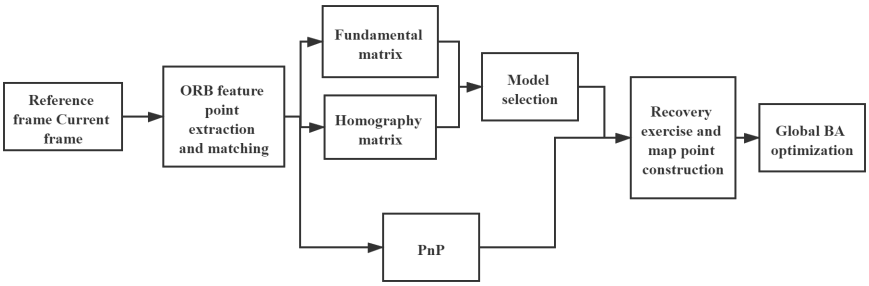

Fig. 3. Visual Structure from Motion (SfM) flowchart.

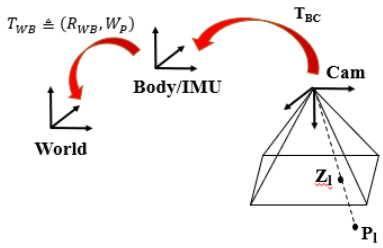

Fig. 4. Conversion relations of different coordinate systems.

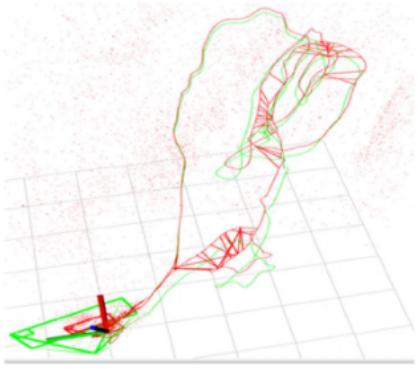

Fig. 5. The improved VINS-mono system running on the EuRoc dataset

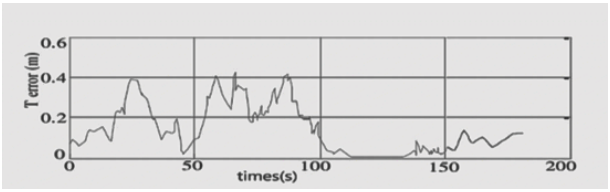

Fig. 6. The graph describes the trajectory error change of the improved system over time in dataset testing

4. Experimental Result

Running our improved VINS-mono system on the EuRoc dataset, the result are shown in the figure 5. The purpose of a test is to test the convergence of system variables during initialization. The green line represents the running track and the red line represents the result of loop detection. In about 10 seconds, the deviation of accelerometer and angular velocity sensor with IMU converge to a stable value, while the scale estimation is close to the optimal value. Through the similarity transformation between the attitude estimation and the real attitude, the optimal value is obtained. After 10s, the condition number decreases significantly and converges, which indicates that the system converges faster. Since the dataset contained the real track coordinates, the accuracy of the track in the modified system could be worked out by the calculation of an error between the estimated trajectory and the real trajectory. According to Figure 6, the trajectory result error was small, and the cumulative error was properly eliminated when dataset were running in the system. This was because the speed of collecting data with the drone was slow enough for the system to detect in a closed loop and hence to make a holistic optimization.

The overall error of the improved VINS-mono system is 0.068m, and the calibration error is 0.079%, which is about 8% higher than that of VINS-mono. The results show that the new initialization method can effectively solve the variable initialization problem of VINS system composed of low-cost sensors.

5. Conclusions

A new initialization scheme is designed which can calculate the acceleration bias as a variable during the initialization process so that it can be applied to low-cost IMU sensors. The results show that the improved VINS-mono scheme completes the entire initialization process within approximately 10 seconds can efficiently facilitate initialization with low-cost sensors. Due to the stricter initialization scheme to avoid the result from falling into the local minimum, the positioning accuracy is also improved.

The improved VINS-mono scheme still uses bag-of-word for loopback detection, but it can easily cause false results for loopback detection especially in an indoor environment that has many similar scenes. Therefore, fur-

ther improvement of the robustness of the system loop detection is needed. Besides, this scheme can generate sparse point clouds information and it is necessary to generate dense 3D point clouds information of environment based on the video stream captured by camera real-time in the next work. In addition, it is necessary to fuse more sensor information to improve the positioning accuracy and robustness further in the next step.

References:

[1] Bloss, R. Sensor innovations helping unmanned vehicles rapidly growing smarter, smaller, more autonomous and more powerful for navigation, mapping, and target sensing. Sens. Rev. 2015, 35, 6–9.

[2] Bloss, R. Unmanned vehicles while becoming smaller and smarter are addressing new applications in medical, agriculture, in addition to military and security. Ind. Robot 2014, 41, 82–86.

[3] Birk, A.; Schwertfeger, S.; Pathak, K. A networking framework for teleoperation in safety, security, and rescue robotics. IEEE Wirel. Commun. 2009, 16, 6–13.

[4] Wen, C.; Qin, L. Three-Dimensional Indoor Mobile Mapping with Fusion of Two-Dimensional Laser Scanner and RGB-D Camera Data. IEEE Geosci. Remote. Sens. Lett 2013, 11, 843–847.

[5] Hsu, L. T.; Wen, W. New Integrated Navigation Scheme for the Level 4 Autonomous Vehicles in Dense Urban Areas; IEEE Symposium on Position Location and Navigation (PLANS): Portland, OR, USA, 2020.

[6] Chiang, K.-W.; Le, D.T.; Duong, T.T.; Sun, R. The Performance Analysis of INS/GNSS/V-SLAM Integration Scheme Using Smartphone Sensors for Land Vehicle Navigation Applications in GNSS-Challenging Environments. Remote Sens. 2020, 12, 1732, doi:10.3390/rs12111732.

[7] Lynen, S. M.; Achtelik, W.S.; Weiss, M. A robust and modular multi-sensor fusion approach applied to mav navigation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–8 November 2013.

[8] Bloesch, M.S.; Omari, M.; Siegwart, R. Robust visual inertial odometry using a direct ekf-based approach. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots & Systems, Hamburg, Germany, 28 September–2 October 2015; pp. 298–304.

[9] Qin, T.; Pei, L. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. IEEE Trans. Robot. 2018, 34, 98–116.

[10] Qin, T.; Pan, J. A General Optimization-based Framework for Local Odometry Estimation with Multiple Sensors. arXiv 2019, arXiv:1901.03638.

[11] Fischler, M.A.; Bolles,R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 1981, 24, 381–395.

[12] Hartley, R.; Zisserman, A. Multiple View Geometry in Computer Vision; Cambridge University Press: Cambridge, UK, 2004; pp.24–29.

[13] Engel, J.; Koltun, V. Direct Sparse Odometry. IEEE Pattern. Anal. 2017, 40, 611–625.