Stereo Visual-Inertial Odometry With Online Initialization and Extrinsic Self-Calibration

Hongpei Yin, Peter Xiaoping Liu, *Fellow, IEEE*, and Minhua Zheng, *Member, IEEE*

Abstract— **Reliable extrinsic calibration is critical to fuse camera and inertial measurement unit (IMU), which are usually used in stereo visual-inertial odometry (VIO), however, it is difficult to obtain extrinsic parameters in practice. This article proposes a stereo VIO with the capability of calibrating the unknown extrinsic parameters online. The initial values of IMU-camera and camera–camera transformations are estimated during the initial process, which fully leverages commonly observed features between stereo cameras. The pose and velocity are estimated and the extrinsic parameters are jointly refined. In addition, the accelerometer and gyroscope biases, and gravity direction are taken into account. The proposed VIO scheme has been demonstrated to be effective without prior knowledge of extrinsic parameters.**

Index Terms— **Self-calibration, sensor fusion, visual-inertial odometry (VIO).**

I. Introduction

STEREO visual-inertial odometry (VIO) solves the problem of simultaneously state estimation for a mobile platform by fusing the measurements from two cameras and an inertial measurement unit (IMU). It has been increasingly used in various applications, such as AR/VR and drones. The fusion of IMU preintegration [1] measurement and vision feature tracking can boost the motion estimation performance [2]. The IMU measures linear acceleration and angular velocity in high frequency, ensuring robustness for aggressive motion and providing the true scale of translation. Although IMU has high measurement accuracy in a short time, it has a large cumulative error when used for a long time because of biases. The integration of vision measurement can compensate for this defect.

The extrinsic parameters of stereo VIO contain a coordinate transformation of IMU-camera and camera–camera pairs, they connect the IMU coordinate with every camera frame. Imprecise extrinsic calibration and slight change of extrinsic parameters during long-time operation would reduce the accuracy of stereo VIO [3], [4]. Reliable offline extrinsic calibration methods require an immobile visual marker and ideal movement of the instrumentation suite. It is difficult to guarantee the reliability of calibration in practical applications. In addition, the performance of VIO is highly dependent on a precise initialization, this process estimates accelerometer and gyroscope biases, velocity, and gravity direction [3], [5], [6].

Traditionally, VIO algorithms [7], [8], [9], [10] assume the extrinsic calibration is precise. In recent years, researchers have shown an increased interest in online initialization and extrinsic self-calibration for VIO [2], [3], [4], [5], [6], [11], [12]. However, most of the researches focused on monocular VIO, which are only available to estimate the transformation of one IMU-camera pair. To the best of our knowledge, there exist few studies on stereo VIO, which can estimate the transformation of two IMU-camera pairs. Fan et al. [4] focused on known but inaccurate calibration and formulated camera–camera, IMU-camera extrinsic parameters as state variables to be estimated. Huang et al. [3] presented a self-calibration method for stereo VIO, which performs two monocular visual front-end and constructs two maps with unknown scales. In practice, however, only one map needs to be retained. Besides, this method does not consider the co-visibility of cameras.

In general, when a device is equipped with instrumentation, including a stereo camera and IMU, there exist commonly observed features between cameras. For this common situation, we propose an optimization-based stereo VIO method that is capable of handling completely unknown extrinsic parameters. This method can complete the initialization procedure and estimate all the necessary extrinsic parameter pairs during the initial phase using the number of frames. The feature positions in both left images are right images tracked by Kanade–Lucas–Tomasi (KLT) optical flow algorithm [13]. We first perform monocular visual-inertial initialization and extrinsic calibration [5], [6], [11] with the left IMU-camera pair, providing the initial values of IMU biases, velocity, gravity, IMU-camera extrinsic parameters, and translation scale factor. And they are used as initial values for joint monocular visual-inertial bundle adjustment (VI-BA) to estimate keyframe poses and position of 3-D landmarks. Since the observations of 3-D landmarks in the right image are known, the pose of the right camera can be estimated by minimizing 3-D–2-D reprojection error. According to the above-mentioned steps, we achieve all initial values for stereo VIO, including IMU-camera and camera–camera extrinsic parameters. However, the initial guesses of extrinsic parameters are not accurate

Manuscript received 31 August 2022; revised 6 April 2023; accepted 16 May 2023. Date of publication 5 June 2023; date of current version 20 June 2023. The Associate Editor coordinating the review process was Dr. Sougata Kumar Kar. *(Corresponding authors: Peter Xiaoping Liu; Minhua Zheng.)*

Hongpei Yin is with the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China.

Peter Xiaoping Liu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: xpliu@sce.carleton.ca).

Minhua Zheng is with the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China, and also with the Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology (Beijing Jiaotong University), Ministry of Education, Beijing 100044, China (e-mail: mhzheng@bjtu.edu.cn).

Digital Object Identifier 10.1109/TIM.2023.3282674

enough to support the operation of stereo VIO. They are further refined during the stereo VI-BA.

The main contributions of this article are highlighted as follows.

1) We propose a stereo VIO, which utilizes a coarse-to-fine strategy to handle the unknown extrinsic parameters.
2) The initial guesses of IMU-camera and camera–camera transformations are estimated during the initial phase, which fully leverages commonly observed features between stereo cameras.
3) The extrinsic parameters are further optimized together with pose, velocity, and IMU biases.
4) Comprehensive evaluations indicate that the proposed stereo VIO works well even without prior extrinsic calibration.

The rest of this article is structured as follows. In Section II, we introduce the relevant works on VIO as well as their initialization and extrinsic calibration. Section III briefly introduces related concepts and the IMU measurement model. In Section IV, our methodology is described in detail. Section V presents the experiment results, including the evaluation for initialization, self-calibration, and the accuracy of the whole stereo VIO. Section VI concludes this article.

II. Related Work

A. Visual-Inertial Odometry

In recent years, a variety of VIO algorithms have been developed, they are used to estimate the motion of instrumentation by fusing the measurements from the camera and IMU. They can be categorized into filter-based methods and nonlinear optimization-based methods. Filter-based VIO methods mainly use multistate Kalman filter (MSCKF), IMU integration is used for state prediction and visual measurement is used for state correction. Li and Mourikis [14] analyzed the observability of MSCKF-based VIO, and used the first estimate Jacobian (FEJ) [15] method to solve the inconsistency problem. S-MSCKF [8] is a classical stereo VIO framework that can run in real time on a low-cost embedded platform. Geneva et al. [16] proposed OpenVINS based on MSCKF, providing a platform for researchers engaged in visual-inertial navigation. Huai and Huang [17] presented filter-based VIO in a robot-centered local coordinate system, avoiding the observability inconsistency problem in the world coordinate. Nonlinear optimization-based VIO uses VI-BA for state estimation by jointly minimizing the feature reprojection error and IMU preintegration [1] error. The nonlinear optimization is solved by Gauss–Newton's method or Levenburg–Marquardt's method. This method has higher accuracy than filter-based VIO but lower efficiency because of the iterative solution. The early optimization-based VIO is OKVIS [7], the IMU error is introduced in the cost function, and the form of the backend is the sliding window. Qin et al. [2], [5] proposed a robust monocular VIO with online biases correction and IMU-camera extrinsic calibration. Campos et al. proposed ORB-SLAM3 [10], a complete visual-inertial simultaneous localization and mapping (SLAM) system. Xia et al. proposed UniVIO [18], a VIO method for low-texture and varying illumination underwater sense which jointly optimizes projection errors and photometric errors. Xia et al. [19] improved the accuracy of VIO structural man-made environments by fusing point-line feature extraction. Because of the scale uncertainty for monocular motion estimation, monocular VIO needs to estimate the scale factor every time, which requires adequate movement excitation and ideal feature tracking during the initial phase. However, scale factor estimation is not required for stereo VIO because the positions of landmarks can be calculated by stereo vision when the extrinsic parameter of the stereo camera is known.

B. Online Initialization

Recently, researchers have noticed it is of great significance to estimate some parameters in the beginning phase, including IMU biases, velocity, and gravity direction. Besides, the monocular VIO needs to estimate the scale factor because the metric scale is unavailable for monocular structure from motion (SFM). This process is called initialization. The early initialization methods [20], [21] rely on short-term gyroscope integration, but they did not take IMU biases into account. Kaiser et al. [22] proposed a closed-form solution to estimate the initial values of IMU biases. Qin and Shen [5] proposed an online initialization method for monocular VIO on microaerial vehicle (MAV) which can run on-the-fly. It has been successfully integrated into a monocular VIO [2]. Campos et al. [6] formulated VIO initialization as a nonlinear optimization problem, which had shown to be more accurate than state-of-the-art. This method has been integrated into a versatile visual-inertial SLAM [10]. Wang and Cheng [23] proposed an IMU self-calibration method, the IMU intrinsic parameters are adaptively optimized according to the intensity of motion.

C. Extrinsic Self-Calibration

Imprecise extrinsic calibration would lead to poor performance for VIO. The common approach to achieving accurate extrinsic parameters is to employ offline methods, which require an immobile visual marker and ideal movement of the instrumentation suite. But this process is inaccessible in some cases. The solution is online extrinsic self-calibration. Recently, several online self-calibration approaches for monocular VIO have been developed. Li and Mourikis [14] proposed a modified MSCKF algorithm for VIO, the transformation between the camera and IMU is considered as a state parameter to be estimated. Yang and Shen [21] addressed the calibration problem of monocular VIO and a methodology that is available for calibrating IMU-camera extrinsic parameter is proposed. It had been integrated into a completed visual-inertial estimator in their later works [2], [24].

There have been extensive works focusing on handling the imprecise extrinsic parameters for multicamera VIO and stereo VIO methods. Jaekel et al. [25] accounted for the uncertainty of extrinsic parameters in their multistereo VIO framework, they modeled the uncertainty in outlier rejection and graph-based optimization. In order to limit the computational burden for the MSCKF-based multicamera method, Eckenhoff et al. [26] only clone the IMU pose to a single camera. Meanwhile, the spatial and temporal extrinsic parameters between all sensors are also estimated online. This work was extended to a multi-IMU multicamera system named MIMC-VINS [27]. Fan et al. [4] proposed an MSCKF-based stereo VIO, the extrinsic parameters are formulated into state variables. This method can successfully estimate precise extrinsic parameters when inaccurate calibration is given. A self-calibration method for stereo VIO with unknown extrinsic parameters was proposed by Huang et al. [3], they performed two monocular VIO initialization and calibration in parallel. Two sparse maps are created by two cameras, respectively, the extrinsic parameters are then estimated by aligning two maps. But only one map needs to be retained in practice. The main difficulty for self-calibration is to handling the unknown transformation of the cameras in the instrumentation suite. Besides, the existing research studies divided the instrumentation suite into several camera-IMU pairs, and the transformation between cameras is solved indirectly, which may bring in potential error. The camera–camera transformation is an important parameter for stereo VIO because the position of landmarks is computed by stereo vision. In our work, monocular visual-inertial initialization and extrinsic calibration are first performed using the left camera and IMU, keyframe poses, 3-D landmark positions, and IMU-camera transformation are estimated. The pose of the right camera is then directly solved by minimizing 3-D–2-D reprojection error using the observations of the landmark with the right camera, which will avoid potential error. The resultant extrinsic parameters are saved and reused for the next time we run the stereo VIO system.

III. Preliminaries

This section presents necessary mathematical notations and geometry concepts. In addition, IMU preintegration measurement model is also briefly reviewed.

A. Notations and Geometry Concepts

We start with defining mathematical symbols. In this article, vectors and matrixes are shown in bold. IMU frame is treated as the body frame, and b_i is the body frame taken from the ith camera. The observation of the camera is represented with the coordinate on the normalized image plane $\mathbf{x} = [x\ y\ 1]^T$, which can be directly computed with camera intrinsic parameters when the pixel location is known. $(\cdot)^w$ is considered as the world frame, $\mathbf{g}^w$ is the gravity vector in the world frame. We use $\mathbf{v}_i^w$ to denote the velocity of the ith frame. The measurement value of instrumentation is expressed as $(\hat{\cdot})$, which may be influenced by biases and noise. 3-D vector $\mathbf{p}$ denotes translation. Rotation is represented by quaternion $\mathbf{q}$ or rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$ [28]. $(\cdot)_{wb}$ is the transformation of the body frame with respect to the world frame, for example, $\mathbf{q}_{wb_i}$ denotes the rotation from the world frame to the ith body frame.

Quaternion is a 4-D complex number that represents 3-D rotation without singularity

$$\mathbf{q} = [s\ x\mathbf{i}\ y\mathbf{j}\ z\mathbf{k}] = [s\ \mathbf{v}]. \tag{1}$$

The multiplication of two quaternions is represented with $\otimes$, which is the superposition of two rotations. For example, given two quaternions $\mathbf{q}_{ab}$ and $\mathbf{q}_{bc}$, their superposition can be defined by

$$\begin{aligned}\mathbf{q}_{ac} &= \mathbf{q}_{ab} \otimes \mathbf{q}_{bc}\\ &= \mathcal{L}(\mathbf{q}_{ab}) \cdot \mathbf{q}_{bc} = \mathcal{R}(\mathbf{q}_{bc}) \cdot \mathbf{q}_{ab}\end{aligned} \tag{2}$$

where $\mathcal{L}(\mathbf{q}_{ab})$ and $\mathcal{R}(\mathbf{q}_{bc})$ represent left and right multiplication, respectively

$$\begin{aligned}\mathcal{L}(\mathbf{q}) &= \begin{bmatrix} s & -\mathbf{v}^T \\ \mathbf{v}_v & s\mathbf{I}_{3\times3} + [\mathbf{v}]_\times \end{bmatrix}\\ \mathcal{R}(\mathbf{q}) &= \begin{bmatrix} s & -\mathbf{v}^T \\ \mathbf{v}_v & s\mathbf{I}_{3\times3} - [\mathbf{v}]_\times \end{bmatrix}.\end{aligned} \tag{3}$$

Here, $[\mathbf{v}]_\times$ is the skew-symmetric matrix of a vector $\mathbf{v}$.

B. IMU Preintegration

An IMU provides a discrete-time sample of linear acceleration and angular velocity between two consecutive camera frames. Preintegration [1], [2] constrain the states of two camera frames using IMU measurement information, which can participate in the nonlinear optimization. Moreover, the measurement of IMU contains gravity, which makes the absolute pose of the system observable.

Given two consecutive camera frames i and j, the sampling interval of IMU is δt, and the transformation of position, velocity, and orientation can be written as

$$\begin{aligned}\mathbf{p}_{wb_j} &= \mathbf{p}_{wb_i} + \mathbf{v}_i^w \Delta t - \frac{1}{2}\mathbf{g}^w \Delta t^2 + \mathbf{q}_{wb_i} \iint_{t\in[i,j]} \left(\mathbf{q}_{b_i b_t} \mathbf{a}^{b_t}\right) \delta t^2\\ \mathbf{v}_j^w &= \mathbf{v}_i^w - \mathbf{g}^w \Delta t + \mathbf{q}_{wb_i} \int_{t\in[i,j]} \left(\mathbf{q}_{b_i b_t} \mathbf{a}^{b_t}\right) \delta t\\ \mathbf{q}_{wb_j} &= \mathbf{q}_{wb_i} \int_{t\in[i,j]} \mathbf{q}_{b_i b_t} \otimes \begin{bmatrix} 0 \\ \frac{1}{2}\boldsymbol{\omega}^{b_t} \end{bmatrix} \delta t\end{aligned} \tag{4}$$

in which

$$\begin{aligned}\boldsymbol{\alpha}_{b_i b_j} &= \iint_{t\in[i,j]} \left(\mathbf{q}_{b_i b_t} \mathbf{a}^{b_t}\right) \delta t^2\\ \boldsymbol{\beta}_{b_i b_j} &= \int_{t\in[i,j]} \left(\mathbf{q}_{b_i b_t} \mathbf{a}^{b_t}\right) \delta t\\ \mathbf{q}_{b_i b_j} &= \int_{t\in[i,j]} \mathbf{q}_{b_i b_t} \otimes \begin{bmatrix} 0 \\ \frac{1}{2}\boldsymbol{\omega}^{b_t} \end{bmatrix} \delta t.\end{aligned} \tag{5}$$

Equation (5) is known as preintegration terms, which are only related to IMU measurements between camera frames. Because the acceleration and angular velocity are affected by IMU biases, and biases are the states we need to estimate, we assume that the variation of preintegration is linear with biases. Therefore, $\boldsymbol{\alpha}_{b_i b_j}$, $\boldsymbol{\beta}_{b_i b_j}$, and $\mathbf{q}_{b_i b_j}$ are the function of IMU biases and they can be adjusted by first-order Taylor expansion approximations [1]

$$\begin{aligned}\boldsymbol{\alpha}_{b_i b_j} &\approx \hat{\boldsymbol{\alpha}}_{b_i b_j} + \mathbf{J}_{b_a}^{\alpha} \delta\mathbf{b}_{a_i} + \mathbf{J}_{b_g}^{\alpha} \delta\mathbf{b}_{g_i}\\ \boldsymbol{\beta}_{b_i b_j} &\approx \hat{\boldsymbol{\beta}}_{b_i b_j} + \mathbf{J}_{b_a}^{\beta} \delta\mathbf{b}_{a_i} + \mathbf{J}_{b_g}^{\beta} \delta\mathbf{b}_{g_i}\\ \mathbf{q}_{b_i b_j} &\approx \hat{\mathbf{q}}_{b_i b_j} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}\mathbf{J}_{b_g}^{\gamma} \delta\mathbf{b}_{g_i} \end{bmatrix}.\end{aligned} \tag{6}$$

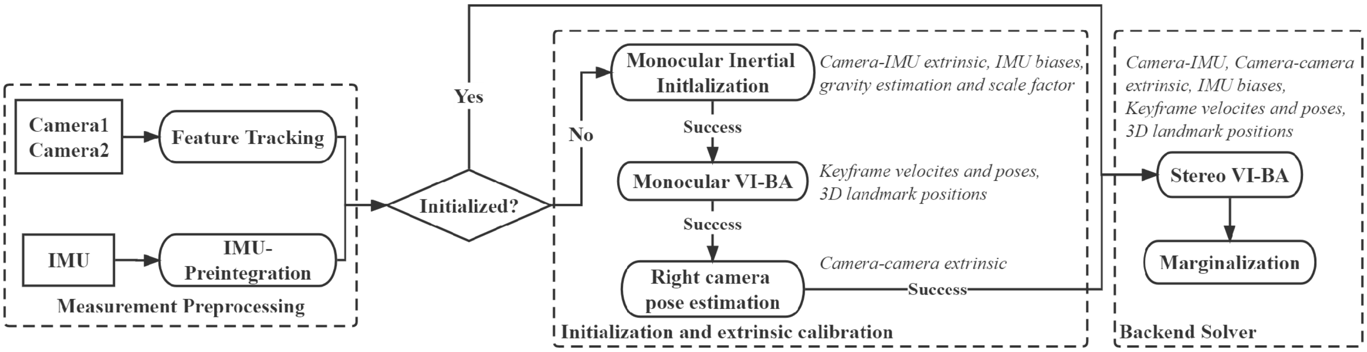

Camera1
Camera2
Feature Tracking
IMU
IMU-Preintegration
Measurement Preprocessing
Initialized?
Yes
No
Monocular Inertial Initlalization
Camera-IMU extrinsic, IMU biases, gravity estimation and scale factor
Success
Monocular VI-BA
Keyframe velocites and poses, 3D landmark positions
Success
Right camera pose estimation
Camera-camera extrinsic
Success
Initialization and extrinsic calibration
Camera-IMU, Camera-camera extrinsic, IMU biases, Keyframe velocites and poses, 3D landmark positions
Stereo VI-BA
Marginalization
Backend Solver

Fig. 1. Pipeline of the proposed stereo VIO method.

The residual of IMU measurement preintegration within two consecutive camera frames is defined as [1], [2]

$$\begin{bmatrix} \mathbf{r}_p \\ \mathbf{r}_q \\ \mathbf{r}_v \\ \mathbf{r}_{ba} \\ \mathbf{r}_{bg} \end{bmatrix} = \begin{bmatrix} \mathbf{q}_{b_i w}\left(\mathbf{p}_{b_i b_j} - \mathbf{v}_i^w \Delta t + \frac{1}{2}\mathbf{g}^w \Delta t^2\right) - \boldsymbol{\alpha}_{b_i b_j} \\ 2\left[\mathbf{q}_{b_j b_i} \otimes \left(\mathbf{q}_{b_i w} \otimes \mathbf{q}_{w b_j}\right)\right]_{xyz} \\ \mathbf{q}_{b_i w}\left(\mathbf{v}_{ij}^w + \mathbf{g}^w \Delta t\right) - \boldsymbol{\beta}_{b_i b_j} \\ \mathbf{b}_j^a - \mathbf{b}_i^a \\ \mathbf{b}_j^g - \mathbf{b}_i^g \end{bmatrix} \quad (7)$$

where $\mathbf{p}_{b_i b_j} = \mathbf{p}_{w b_j} - \mathbf{p}_{w b_i}$, $\mathbf{v}_{ij}^w = \mathbf{v}_j^w - \mathbf{v}_i^w$, and $[\cdot]_{xyz}$ is the real component of the quaternion.

IV. Methodology

This section presents the proposed stereo VIO in detail, especially, the initialization and extrinsic self-calibration process. Fig. 1 shows the pipeline of the proposed stereo VIO, which contains three main modules: 1) the measurement preprocessing module; 2) the initialization and extrinsic calibration module; and 3) the backend solver module. This system starts with preprocessing the measurements of instrumentation, which takes charge of feature tracking and IMU preintegration [1] calculation, providing constrain between two camera frames. The features are extracted with good features to track (GFTT) detector. The image is divided into several blocks according to its size, and new feature points are detected from the empty blocks with the highest Shi-Tomasi core [29]. The features are tracked with KLT optical flow [13], [30] algorithm. The initialization and extrinsic calibration module is the main contribution of our work, which will be presented in detail from Sections IV-A– IV-C. We focus on the uninitialized situation, in which the IMU biases and extrinsic parameters are unknown. The symbol of the parameters that need to be estimated in the initialization and extrinsic parameters calibration module are shown in Table I. This module provided initial guesses of IMU biases and extrinsic parameters for the backend solver (Section IV-D). The backend solver not only estimates the 3-D landmark positions, keyframe poses, and velocities but also further optimize the extrinsic parameters and IMU biases. To limit the computational complexity of the backend solver, we apply the same marginalization strategy as VINS-Mono [2].

TABLE I

Parameters That Need to be Estimated in the Initialization and Extrinsic Calibration Module

Symbol of parameter	Description
$\mathbf{q}_{bc^l}$ $(\mathbf{R}_{bc^l})$, $\mathbf{p}_{bc^l}$	Transformation from IMU to left camera.
$\mathbf{q}_{lr}$ $(\mathbf{R}_{lr})$, $\mathbf{p}_{lr}$	Transformation from left camera to right camera.
$\mathbf{b}_a, \mathbf{b}_g$	Accelerometer bias and gyroscope bias.
$\mathbf{g}^w$	Gravity vector in the world frame.
s	Scale factor of the camera poses.

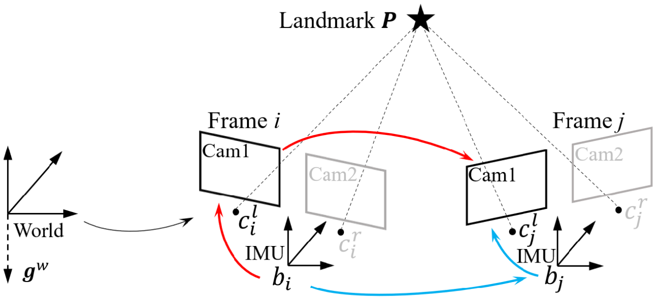


Fig. 2. Transformation from the world coordinate system to two consecutive frames i and j. The stereo camera places two cameras on the left (Cam1) and the right (Cam2). Because monocular inertial initialization only uses the IMU-Cam1 pair, the Cam2 is shown in gray.

A. Monocular Inertial Initialization

1) IMU-Camera Orientation Calibration and Gyroscope Bias Correction: Fig. 2 presents the transformation from the world coordinate system to two consecutive frames i and j. In the case of the left camera (Cam1 in Fig. 2), rotation from b_i to c_j^l can be calculated in two ways (red arrows and blue arrow in Fig. 2), and they are theoretically equal. The equivalent relations are expressed as

$$\mathbf{q}_{b_i b_j} \otimes \mathbf{q}_{bc^l} = \mathbf{q}_{bc^l} \otimes \mathbf{q}_{c_i^l c_j^l} \quad (8)$$

in which $\mathbf{q}_{b_i b_j}$ is the rotation from ith IMU to jth IMU, and $\mathbf{q}_{c_i^l c_j^l}$ is the rotation from ith left camera to jth left camera, $\mathbf{q}_{bc^l}$ is the IMU-camera rotation we need to estimate. According to (2) and (3), (8) can be written as the following linear equation:

$$\left(\mathcal{L}\left(\mathbf{q}_{b_i b_j}\right) - \mathcal{R}\left(\mathbf{q}_{c_i^l c_j^l}\right)\right)\mathbf{q}_{bc^l} = \mathbf{Q}_j^i \cdot \mathbf{q}_{bc^l} = \mathbf{0} \quad (9)$$

where $\mathbf{q}_{b_i b_j}$ can be computed by integrating the IMU measurement, and $\mathbf{q}_{c_i^l c_j^l}$ can be computed by the eight-point algorithm [31]. In practice, there exist N frames and corresponding geometry constraints, according to (9), they can be used to construct an overdetermined linear equation

$$\begin{bmatrix} w_1^0 \cdot \mathbf{Q}_1^0 \\ w_2^1 \cdot \mathbf{Q}_2^1 \\ \vdots \\ w_N^{N-1} \cdot \mathbf{Q}_N^{N-1} \end{bmatrix} \mathbf{q}_{bc^l} = \mathbf{Q}_N \cdot \mathbf{q}_{bc^l} = \mathbf{0}. \tag{10}$$

Because the calculation of $\mathbf{q}_{c_i^l c_j^l}$ highly relies on the quality of feature tracking, we use weight w to remove outliers

$$w_j^i = \begin{cases} 1, & r_j^i < \text{threshold} \\ \dfrac{\text{threshold}}{r_j^i}, & \text{otherwise.} \end{cases} \tag{11}$$

According to the relationship between the rotation matrix and axis angle, the residual r_j^i can be computed by the following equation:

$$r_j^i = \operatorname{acos}\left(\left(\operatorname{tr}\left(\hat{\mathbf{R}}_{bc^l}^{-1}\hat{\mathbf{R}}_{b_i b_j}^{-1}\hat{\mathbf{R}}_{bc^l}\hat{\mathbf{R}}_{c_i^l c_j^l}\right) - 1\right)\Big/2\right). \tag{12}$$

Theoretically, residual r_j^i tends to be zero. The weight in (10) is smaller when the residual becomes larger. After solving (10), the gyroscope bias can be estimated by the following equation:

$$\mathbf{b}_g^* = \arg\min_{\mathbf{b}^g} \sum_{i,j \in B} \left\| 2\left[\mathbf{q}_{c_0^l b_j}^{-1} \otimes \mathbf{q}_{c_0^l b_i} \otimes \mathbf{q}_{b_i b_j}\right]_{xyz} \right\|^2 \tag{13}$$

in which B is the set of camera frames, $\mathbf{q}_{c_0^l b_j}$ and $\mathbf{q}_{c_0^l b_i}$ can be computed with the accumulation of camera rotation and the estimated IMU-camera rotation $\mathbf{q}_{bc^l}$. According to (6), $\mathbf{q}_{b_i b_j}$ is the function of $\mathbf{b}_g$.

We only consider the transformation from IMU to the left camera in this step. The camera–camera extrinsic parameters will be estimated in the later steps, this will be presented in detail in Section IV-C.

2) Bias, Gravity, Scale Factor, and IMU-Camera Translation Estimation: In this step, we first perform monocular SFM [32] to compute the poses of the camera $\mathbf{R}_{c_0 c_k}$, $\hat{\mathbf{p}}_{c_0 c_k}$, in which k denotes kth frame and $\hat{\mathbf{p}}_{c_0 c_k}$ is the translation without metric scale. Then we estimate the necessary parameters by aligning visual measurement with IMU preintegration. They are defined as

$$\boldsymbol{\mathcal{X}}_I = [\mathbf{v}_{b_0}, \mathbf{v}_{b_1}, \ldots, \mathbf{v}_{b_N}, \mathbf{p}_{bc^l}, \mathbf{b}_a, \mathbf{b}_g, \mathbf{R}_{wg}, \mathbf{p}_{bc^l}, s] \tag{14}$$

where $\mathbf{v}_{b_k} (k \in N)$ is the velocity under kth frame, N is the number of camera frames, and $N = 15$ in our implementation. $\mathbf{b}_a$ and $\mathbf{b}_g$ denote IMU biases, which are assumed to be invariable in this step because it takes just 1–2 s to perform initialization. Although gyroscope bias $\mathbf{b}_g$ has been estimated in Section IV-A, it is further optimized in this step. $\mathbf{R}_{wg}$ is the gravity direction parameterized by two angles [6], [11], [33]. Gravity in the world coordinate is expressed as $\mathbf{g}^w = \mathbf{R}_{wg}\mathbf{g}_I$, where $\mathbf{g}_I = [0\ 0\ G]^T$ with $G = -9.81$. Let i and j be the consecutive camera frames, and the first camera c_0 frame is the world coordinate, the transformation between camera frame c_0 and the ith IMU body frame is

$$\begin{aligned} \mathbf{R}_{c_0 b_i} &= \mathbf{R}_{c_0 c_i}\mathbf{R}_{bc}^{-1} \\ s\hat{\mathbf{p}}_{c_0 b_i} &= s\hat{\mathbf{p}}_{c_0 c_i} - \mathbf{R}_{c_0 b_i}\mathbf{p}_{bc^l}. \end{aligned} \tag{15}$$

According to (6), (7), and (15), we can define the following IMU measurement residuals [6], [11]:

$$\begin{aligned} \mathbf{r}_p^{ij} &= \boldsymbol{\alpha}_{b_i b_j}(\mathbf{b}_a, \mathbf{b}_g) - s\mathbf{R}_{b_i c_0}\left(\hat{\mathbf{p}}_{c_0 c_j} - \hat{\mathbf{p}}_{c_0 c_i}\right) \\ &\quad - \mathbf{R}_{b_i b_j}\mathbf{p}_{bc^l} + \mathbf{p}_{bc^l} + \frac{1}{2}\mathbf{R}_{b_i c_0}\mathbf{R}_{c_0 g}\mathbf{g}_I \Delta t_k^2 - \mathbf{v}_{b_i}\Delta t_k \\ \mathbf{r}_v^{ij} &= \boldsymbol{\beta}_{b_i b_j}(\mathbf{b}_a, \mathbf{b}_g) \\ &\quad - \mathbf{R}_{b_i c_0}\left(\mathbf{R}_{c_0 b_j}\mathbf{v}_{b_j} + \mathbf{R}_{c_0 g}\mathbf{g}_I \Delta t_k - \mathbf{R}_{c_0 b_i}\mathbf{v}_{b_i}\right) \end{aligned} \tag{16}$$

where Δt_k is the time interval between frames. Given N cameras and IMU measurement residuals between them, the optimal estimation can be obtained by minimizing the sum of residuals

$$\boldsymbol{\mathcal{X}}_I^* = \arg\min_{\boldsymbol{\mathcal{X}}_I}\left\{\sum_{i,j \in N}\left\|\mathbf{r}_p^{ij} + \mathbf{r}_v^{ij}\right\|_2\right\}. \tag{17}$$

We optimize the gravity on-manifold [6], [33]

$$\mathbf{R}_{c_0 g} \leftarrow \mathbf{R}_{c_0 g}\operatorname{Exp}(\omega_1, \omega_2, 0) \tag{18}$$

where $\operatorname{Exp}(\cdot)$ is the exponential map from SO(3) to $\mathfrak{se}(3)$ [28], ω_1 and ω_2 are two angles used to parameterize the gravity direction.

B. Monocular VI-BA

After monocular inertial initialization, we go ahead with monocular VI-BA for high accuracy state estimation. The full state variables are defined as

$$\begin{aligned} \boldsymbol{\mathcal{X}} &= [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N, \rho_0, \rho_1, \ldots, \rho_M] \\ \boldsymbol{\xi}_k &= [\mathbf{p}_{wb_k}, \mathbf{v}_{wb_k}, \mathbf{q}_{wb_k}, \mathbf{b}_a, \mathbf{b}_g] \end{aligned} \tag{19}$$

in which $\boldsymbol{\xi}_k$ is the state of IMU at the same instant as the kth image frame, including position, velocity, and orientation in the world frame $\mathbf{p}_{wb_k}$, $\mathbf{v}_{wb_k}$, $\mathbf{q}_{wb_k}$ and IMU bias $\mathbf{b}_a, \mathbf{b}_g$. And ρ_i is the inverse depth [34] of the ith landmark from its first observation. The estimation of the scale factor provides initial values for $\mathbf{p}_{wb_k}$. We minimize the sum of residuals to obtain a maximum posterior estimation. The cost function is written as

$$\boldsymbol{\mathcal{X}}^* = \arg\min_{\mathcal{X}}\left\{\sum_{k \in N}\left\|\mathbf{r}_b^k(\boldsymbol{\mathcal{X}})\right\|_2 + \sum_{i,j \in N}\left\|\mathbf{r}_c^{i,j}(\boldsymbol{\mathcal{X}}, \mathbf{x}_{i,j})\right\|_2\right\} \tag{20}$$

where $\mathbf{r}_b^k(\mathcal{X})$ is the residual for IMU measurement [1], [2] (7). $\mathbf{r}_c^{i,j}(\boldsymbol{\mathcal{X}}, \mathbf{x}_{i,j})$ is the residual for monocular visual measurement, which is defined as the reprojection error on the normalized image plane in the current frame [2], [5]

$$\mathbf{r}_c^{i,j}(\boldsymbol{\mathcal{X}}, \mathbf{x}_{i,j}) = \left\|\left(\frac{\mathbf{P}_{c_j}}{Z_{c_j}} - \mathbf{x}_j\right)_{xy}\right\|_2 \tag{21}$$

in which $\mathbf{P}_{c_j} = [X_{c_j}\ Y_{c_j}\ Z_{c_j}]^T$ is the estimated position of the landmark in the jth frame, and $\mathbf{x}_k$ is the observation of the same landmark in the jth frame. And $(\cdot)_{xy}$ represents the first two dimensions of the vector.

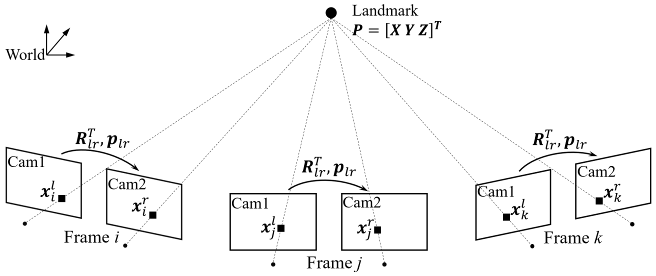

Landmark
$\boldsymbol{P} = [\boldsymbol{X}\ \boldsymbol{Y}\ \boldsymbol{Z}]^T$
World
$\boldsymbol{R}_{lr}^T, \boldsymbol{p}_{lr}$
$\boldsymbol{R}_{lr}^T, \boldsymbol{p}_{lr}$
$\boldsymbol{R}_{lr}^T, \boldsymbol{p}_{lr}$
Cam1
$\boldsymbol{x}_i^l$
Cam2
$\boldsymbol{x}_i^r$
Frame i
Cam1
$\boldsymbol{x}_j^l$
Cam2
$\boldsymbol{x}_j^r$
Frame j
Cam1
$\boldsymbol{x}_k^l$
Cam2
$\boldsymbol{x}_k^r$
Frame k

Fig. 3. One landmark and its projection in three consecutive frames. Cam1 and Cam2 represent left and right camera, respectively. The position of landmark $\mathbf{P}$ and the pose of the left camera are known after monocular inertial initialization. The observations of landmarks in both left and right cameras ($\mathbf{x}^l$ and $\mathbf{x}^r$) have been achieved from the measurement preprocessing module.

C. Right Camera Pose Estimation

The monocular VI-BA procedure has estimated the inverse depth for a set of landmarks and the pose of several body frames, therefore, we can compute the 3-D landmark positions in each left camera frame. Fig. 3 shows the position of one landmark and its projection in three consecutive frames. There exist hundreds of such landmarks. In addition, the observation of landmark in the right camera ($\mathbf{x}^r$ in Cam2, see Fig. 3) have been achieved from the measurement preprocessing module. Here, in each camera coordinate, we have a set of 3-D landmark positions in the left camera frame and corresponding observations in the right camera. Therefore, the camera–camera transformation can be estimated by minimizing 3-D–2-D reprojection error, let takes landmarks project to the kth frame as an example

$$\mathbf{R}_{lr}^*, \mathbf{p}_{lr}^* = \underset{\mathbf{R}_{lr},\mathbf{t}_{lr}}{\arg\min} \sum_{k\in N} \left\| \left(\boldsymbol{\pi}\left(\mathbf{R}_{lr}^T\mathbf{P}_{c_k}^l - \mathbf{R}_{lr}^T\mathbf{p}_{lr}\right) - \mathbf{x}_k^r\right)_{xy}\right\|_2 \tag{22}$$

in which $\boldsymbol{\pi}(\cdot)$ denotes the projection of a 3-D point to the normalized image plane.

D. Stereo VI-BA and Marginalization

The online initialization and extrinsic calibration module provides an initial guess of necessary parameters during the initial phase. But some of them need to be further optimized in the subsequent backend solver module. The whole state variables are defined as

$$\begin{aligned}\boldsymbol{\mathcal{X}} &= [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N, \boldsymbol{\xi}_{bc}, \rho_0, \rho_1, \ldots, \rho_M]\\ \boldsymbol{\xi}_k &= [\mathbf{p}_{wb_k}, \mathbf{v}_{wb_k}, \mathbf{q}_{wb_k}, \mathbf{b}_a, \mathbf{b}_g]\\ \boldsymbol{\xi}_{bc} &= \left[\mathbf{p}_{bc}^l, \mathbf{q}_{bc}^l, \mathbf{p}_{lr}, \mathbf{q}_{lr}\right].\end{aligned} \tag{23}$$

Different from monocular VI-BA, our stereo VI-BA method takes extrinsic parameters $\boldsymbol{\xi}_{bc}$ into account, including left IMU-camera and camera–camera transformation. The residual of visual measurement contains the reprojection error on both left and right camera. The projection of the landmark to the right camera is related to camera–camera transformation $\mathbf{p}_{lr}$ and $\mathbf{q}_{lr}$. Consider a landmark is first observed in the ith frame, the residual of visual measurement in the jth frame is defined as

$$\begin{aligned}\mathbf{r}_c^{i,j}(\boldsymbol{\mathcal{X}}, \mathbf{x}_k) &= f(\boldsymbol{\xi}_k, \boldsymbol{\xi}_{bc}, \rho_i)\\ &= \left\| \left(\frac{\mathbf{P}_{c_j}^l}{Z_{c_j}^l} - \mathbf{x}_k^l\right)_{xy} + \left(\frac{\mathbf{P}_{c_j}^r}{Z_{c_j}^r} - \mathbf{x}_k^r\right)_{xy} \right\|_2\end{aligned} \tag{24}$$

in which $\mathbf{P}_{c_j}^l$ can be computed by pose transformation from ith to jth frame

$$\begin{aligned}\mathbf{P}_{c_j}^l &= \mathbf{R}_{bc}^\top \mathbf{R}_{wb_j}^\top \mathbf{R}_{wb_i} \mathbf{R}_{bc} \mathbf{P}_{c_i}^l\\ &\quad + \mathbf{R}_{bc}^\top \left(\mathbf{R}_{wb_j}^\top \left(\left(\mathbf{R}_{wb_i}\mathbf{p}_{bc} + \mathbf{p}_{wb_i}\right) - \mathbf{p}_{wb_j}\right) - \mathbf{p}_{bc}\right)\end{aligned} \tag{25}$$

where $\mathbf{P}_{c_i}^l$ can be computed by the normalized image coordinate and the inverse depth under jth frame: $\mathbf{P}_{c_i} = \mathbf{x}_i/\rho_i$. $\mathbf{P}_{c_j}^r$ is computed by camera–camera transformation

$$\mathbf{P}_{c_j}^r = \mathbf{R}_{lr}^\top \mathbf{P}_{c_j}^l - \mathbf{R}_{lr}^\top \mathbf{t}_{lr}. \tag{26}$$

The whole stereo VI-BA is formulated as a cost function

$$\begin{aligned}\boldsymbol{\mathcal{X}}^* = \underset{\boldsymbol{\mathcal{X}}}{\arg\min} \Bigg\{ &\left\|\mathbf{r}_p - \mathbf{H}_p\boldsymbol{\mathcal{X}}\right\|_2 + \sum_{k\in N}\left\|\mathbf{r}_b^k(\boldsymbol{\mathcal{X}})\right\|_2\\ &+ \sum_{k\in N}\left\|\mathbf{r}_c^{i,j}(\boldsymbol{\mathcal{X}}, \mathbf{x}_k)\right\|_2 \Bigg\}\end{aligned} \tag{27}$$

including prior constrain $\|\mathbf{r}_p - \mathbf{H}_p\boldsymbol{\mathcal{X}}\|_2$, IMU measurement residual [1], [2] (7), and visual measurement residual (24). In order to bound the computational complexity, a sliding window optimization strategy is applied to selectively marginalize landmark and keyframe, in a manner inspired by [2]. If the average parallax between the incoming frame and the latest keyframe is larger than a certain threshold, or the amount of tracked features below a certain threshold, the incoming frame is treated as a new keyframe. If the incoming frame is keyframe, the oldest frame and the observed landmarks are marginalized. Otherwise, we marginalize the latest keyframe in the sliding window. The prior constrain is constructed by the marginalized state variables and Schur Complement [35], [36].

V. Experiments

We evaluate our stereo VIO system on the widely used benchmarking dataset EuRoC MAV [37]. It was collected with an on-board Micro Aerial Vehicle, containing synchronized stereo images with 20 Hz and IMU with 200 Hz. The accurate ground truth pose is achieved from the Leica MS50 laser tracker. The ground truth of extrinsic parameters and IMU biases are also provided. The dataset is collected from the Industrial Machine Hall and the Vicon Room without dynamic objects. The EuRoC MAV dataset has been a standard for VIO performance evaluation. All the experiments are carried out on a low-cost laptop with Ubuntu 16.04, Intel i5-4200 four-core 2.5 GHz CPU, and 8 GB of RAM. For each sequence in the EuRoC dataset, the average processing time of each step in Section IV are shown in Fig. 4. It takes less than 300 milliseconds to achieve initial guess of extrinsic parameters.

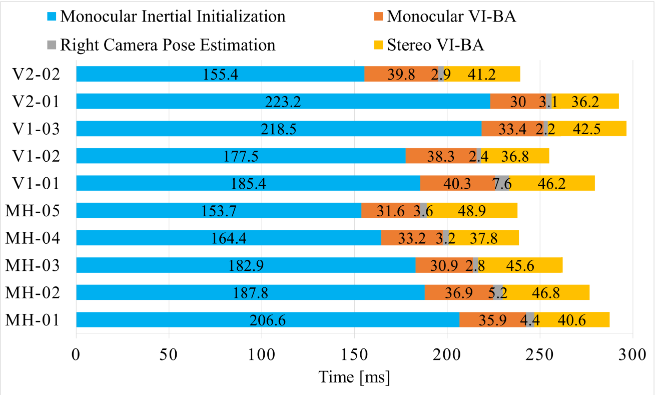


Fig. 4. Average processing time of each step in Section IV for each sequence in EuRoC dataset. Each sequence is performed 15 tests with different start time.

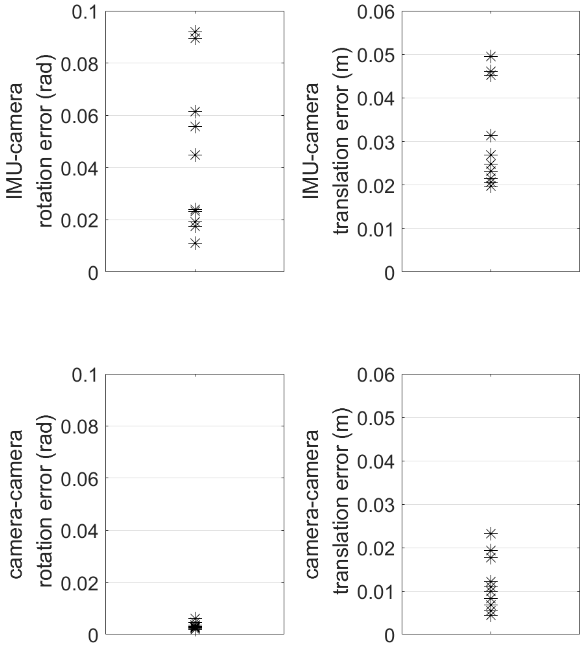


Fig. 5. IMU-camera and camera–camera calibration error using different start times in the V1-02 sequence.

A. Extrinsic Calibration Results

In this section, the performances of self-calibration are shown using the V1-02 sequence in the EuRoC dataset. We use rotation error and translation error to evaluate the extrinsic parameters' self-calibration result. The evaluation parameters are defined as the following formula:

$$\delta\theta_{\text{err}} = \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{estm}}\mathbf{R}_{\text{gt}}^{T}) - 1}{2}\right)$$
$$\delta t_{\text{err}} = \left\|\mathbf{t}_{\text{estm}} - \mathbf{t}_{\text{gt}}\right\|_2 \tag{28}$$

where the rotation error is $\delta\theta_{\text{err}}$ (radian) represents as axis angle of $\mathbf{R}_{\text{estm}}\mathbf{R}_{\text{gt}}^{T}$, and the translation error δt_{err} (m) is the norm of the difference between the two vectors. The subscript estm denotes our estimation result and gt denotes ground truth.

Fig. 5 shows the IMU-camera and camera–camera calibration error with ten different start times in the V1-02 sequence.

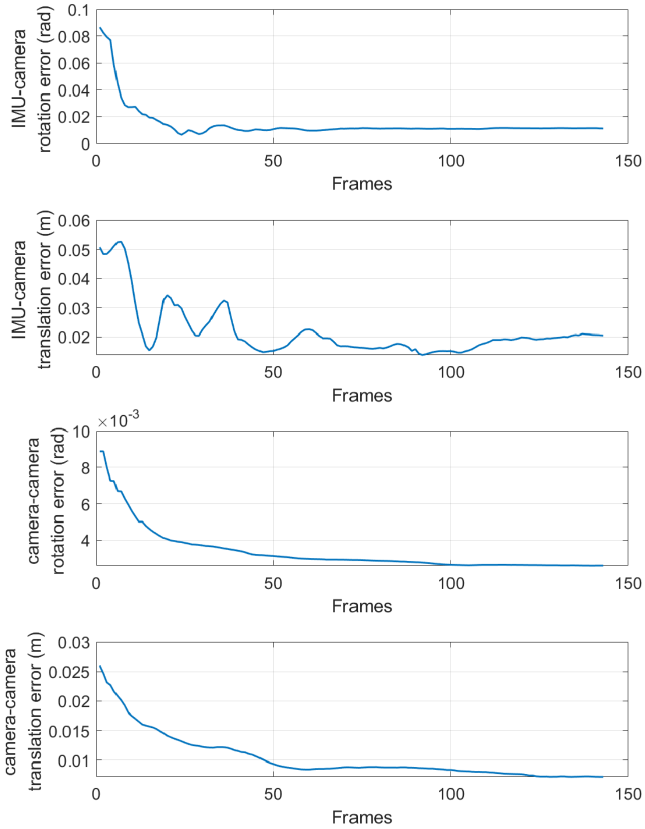


Fig. 6. Convergence performance of extrinsic parameters refinement in the backend solver. It takes less than 60 frames (3 s) for the errors to reach convergence.

The rotation errors of the IMU-camera are less than 0.1 radians (5.73 degrees) and the camera–camera rotation errors are even smaller. The norm of IMU-camera translation ground truth is $\|\mathbf{p}_{bc^l}\| = 0.069\ m$ and the norm of camera–camera translation ground truth is $\|\mathbf{p}_{lr}\| = 0.11\ m$. The camera–camera translation estimation has higher accuracy with a relative error of less than 20%, but the relative error of IMU-camera translation estimation is even larger than 50% in several start times. The precision of scale factor estimation is reflected in the camera–camera transformation result because the initial values of 3-D landmark positions in (22) are computed by the camera poses. Note that these are the calibration results during the initial phase. The extrinsic parameters are further refined in the subsequent nonlinear optimization, which is performed by the backend solver module. The extrinsic parameters estimation error change with the frame count can be seen in Fig. 6. Less than 60 frames (3 s) are enough for the errors to converge to a minimum. This experiment indicates that the initialization and extrinsic calibration module is available to provide reliable initial guesses of extrinsic parameters, and the backend solver can effectively further optimize the extrinsic parameters.

B. Bias Estimation Results

We use the V1-02 sequence to evaluate and analyze the performance of biases estimation results. In these experiments, the proposed online biases correlation method is compared with VINS-Fusion [24], which is the state-of-the-art stereo

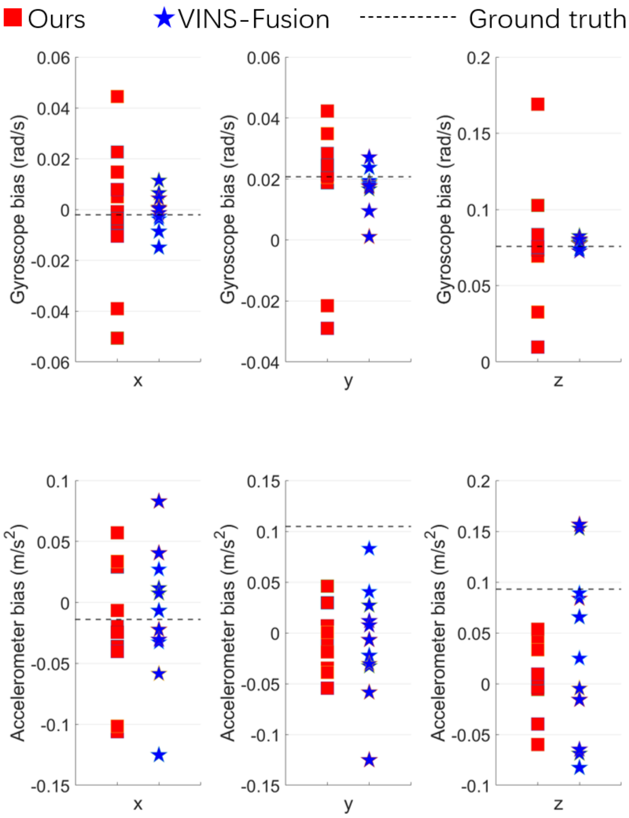


Fig. 7. IMU bias correction results during the initial procedure with different methods. The data contains the results of ten start times.

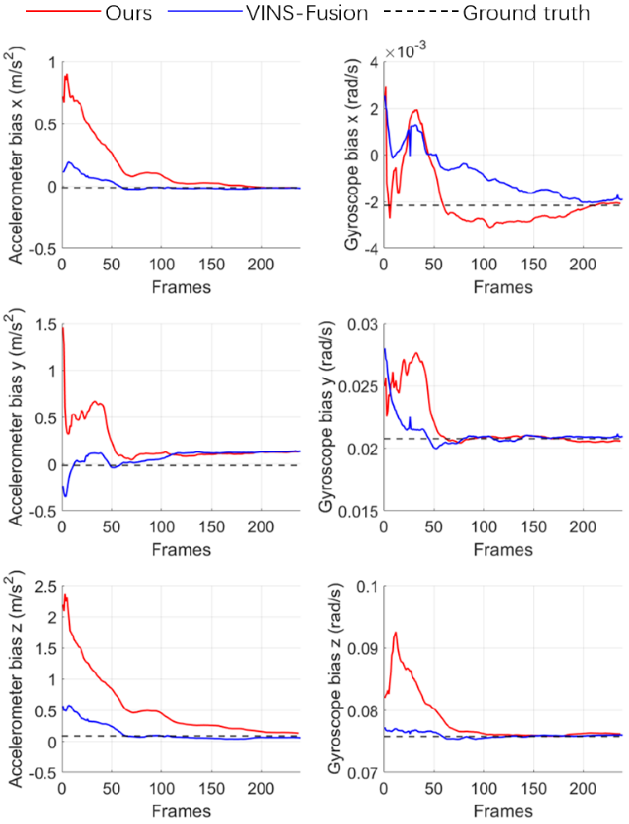


Fig. 8. IMU biases estimation results change with frame number using different methods.

VIO algorithm with online biases correlation. Fig. 7 compares the IMU biases correction performance between our method and VINS-Fusion during the initial procedure, which contains ten tests with different start times. If the points are closer to the dotted line, it indicates that the IMU biases correction results have higher accuracy. The result shows that the proposed method performs worse than VINS-Fusion on IMU biases estimation during the initial procedure. Compared with gyroscope bias, however, the accelerometer bias estimation results are closer to those of VINS-Fusion. It should be noted that the proposed biases correction method operates with completely unknown extrinsic parameters, but VINS-Fusion has accurate extrinsic parameters. The IMU biases estimation results are still acceptable under such hostile conditions.

Moreover, the IMU biases are further optimized in the backend solver module. The IMU biases estimation results change with frame number using our method and VINS-Fusion are shown in Fig. 8. In addition to motion estimation, the backend solver of the proposed stereo VIO not only optimizes the IMU biases but also further refines the extrinsic parameters. But VINS-Fusion only takes IMU biases into account. The gyroscope biases estimation during the initial procedure is easily affected by the imprecise extrinsic parameters. Therefore during the first 100 frames (5 s), the biases estimation of VINS-Fusion is much closer to the ground truth. Afterward, however, the proposed method can also make the estimated IMU biases close enough to the ground truth gradually along with the frame number. The aforementioned experimental results indicate that the proposed method is capable of estimating IMU biases with unknown extrinsic parameters. Although the convergence performance in the beginning is unsatisfactory, compared with VINS-Fusion which has precise extrinsic parameters, it can still converge to the ground truth in a similar time.

C. Trajectory Accuracy Comparison

The root mean square error (RMSE) of the absolute trajectory error (ATE) metric is used to evaluate the trajectory accuracy of the proposed stereo VIO. In this experiment, we compare our VIO method with the state-of-the-art VIO that work with a stereo camera, including VINS-Fusion [24] and S-MSCKF [8]. We turn off the loop closure mode of VINS-Fusion for fairness. It is worth noting that the proposed method operates with completely unknown extrinsic parameters, specifically, the rotation is set to 3 × 3 identity matrix, and the translation vector is set to zero. However, the state-of-the-art methods have precise extrinsic parameters provided by the EuRoC dataset. Table II shows the accuracy comparison results and trajectory length of the EuRoC dataset, S-MSCKF has the highest accuracy, and our method has a similar performance to VINS-Fusion. The accuracy performance indicates the proposed backend solver has a capacity for eliminating the impact of imprecise extrinsic parameters.

In addition, we make a qualitative comparison between VINS-Fusion with inaccurate initial extrinsic parameters, which are provided by the proposed initialization and extrinsic calibration module. The calibration error for each sequence

TABLE II

EXTRINSIC CALIBRATION ERROR AND TRAJECTORY ACCURACY COMPARISON IN EUROC DATASET. THE EXTRINSIC CALIBRATION RESULTS ARE PROVIDED BY INITIALIZATION AND EXTRINSIC CALIBRATION MODULE, WHICH ARE USED TO COMPUTE CALIBRATION ERROR. THE BEST RESULTS OF TRAJECTORY ACCURACY ARE SHOWN IN BOLD

Seq	Length	Extrinsic Calibration Error				Trajectory Accuracy Comparison: ATE RMSE(m)			
		IMU-Cam1 Trans (m)	IMU-Cam1 Rot (rad)	Cam1-Cam2 Trans (m)	Cam1-Cam2 Rot (rad)	S-MSCKF[1]	VINS-Fusion[2]	VINS-Fusion[3]	Proposed[4]
MH01	79.84	0.0693	0.1238	0.0272	0.0126	0.194	0.202	0.318	**0.174**
MH02	72.75	0.0516	0.1004	0.0294	0.0053	**0.143**	0.176	0.180	0.181
MH03	130.58	0.0799	0.0023	0.0118	0.0024	**0.272**	0.376	0.445	0.379
MH04	91.55	0.0309	0.1163	0.0047	0.0052	**0.189**	0.346	0.381	0.370
MH05	97.32	0.0512	0.1294	0.0181	0.0019	**0.307**	0.332	0.341	0.352
V101	58.51	0.0753	0.0968	0.0041	0.0015	**0.099**	0.146	0.159	0.151
V102	75.72	0.0277	0.0153	0.0270	0.0022	**0.167**	0.294	0.306	0.304
V103	78.77	0.0778	0.0245	0.0056	0.0024	**0.193**	0.257	0.294	0.260
V201	36.34	0.0424	0.1041	0.0095	0.0027	**0.074**	0.123	0.177	0.112
V202	83.01	0.0786	0.1129	0.0073	0.0025	**0.156**	0.240	0.256	0.245

[1] S-MSCKF operates with precise extrinsic parameters.

[2] VINS-Fusion operates with precise extrinsic parameters.

[3] VINS-Fusion operates with inaccurate extrinsic parameters. The initial values of extrinsic parameters are provided by the initialization and extrinsic calibration module, and the corresponding calibration errors are shown in the same row.

[4] The proposed method does not depend on any prior knowledge.

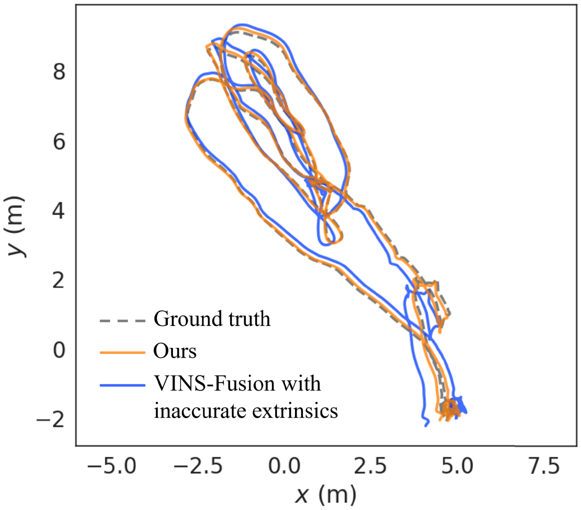


Fig. 9. Estimated trajectories comparison of MH-01 sequence. VINS-Fusion operates with inaccurate extrinsic parameters, but our method does not depend on any prior knowledge.

is shown in Table II. All stereo VIO methods suffer from the inaccurate extrinsic parameter condition. Regardless of the off-line extrinsic calibration method used, there may exist error. The possible sources of calibration errors include the inaccuracy of the sensors, the low quality of visual markers, and the imprecision of the calibration algorithm. It is difficult to determine the calibration error in practical applications. Therefore, it is possible for stereo VIO to suffer from such an inaccurate extrinsic parameter condition. In view of this situation, VINS-Fusion provides a function that estimates transformation between IMU and each camera online, which computes camera–camera transformation indirectly. It is easily influenced by one of the IMU-camera transformation estimation results, especially, when the initial value is inaccurate. The motion estimation of a stereo camera highly depends on camera–camera extrinsic parameter, because it is directly used to compute the landmark positions. But our method avoids this disadvantage by estimating camera–camera transformation directly. As can be seen in Table II, the trajectory accuracy of our method is slightly higher than VINS-Fusion in inaccurate extrinsic parameters situation. Fig. 9 shows the output trajectory comparison of the MH-01 sequence, in which VINS-Fusion with inaccurate extrinsic parameters drifts obviously, but our method can make the trajectory well-aligned with the ground truth. The above experiments show that the advantage of the proposed method is the capability of handling unknown extrinsic parameters, and the better performance in online extrinsic calibration in comparison with VINS-Fusion.

VI. CONCLUSION

In this article, we have presented a stereo VIO with online initialization and extrinsic self-calibration. This system contains three main modules: the measurement preprocessing module, the initialization and extrinsic calibration module, and the backend solver module. The initialization and extrinsic calibration module not only computes IMU biases, velocity, and gravity direction but also creatively estimates the IMU-camera and camera–camera transformation without prior knowledge. In addition, the proposed backend solver is able to further refine the IMU-camera and camera–camera extrinsic parameters. The experimental results indicate that the proposed method can accurately estimate the extrinsic parameters without initial value in seconds. In the meantime, IMU biases are also successfully estimated. The trajectory accuracy has a similar performance to VINS-Fusion and S-MSCKF with precise extrinsic parameters. In addition, the trajectory accuracy is slightly higher than VINS-Fusion in inaccurate extrinsic parameters situation. The proposed stereo VIO allows the device to realize "power-on-and-go" without tedious offline extrinsic calibration.

REFERENCES

[1] 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.

[2] 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.

[3] W. Huang, H. Liu, and W. Wan, "An online initialization and self-calibration method for stereo visual-inertial odometry," *IEEE Trans. Robot.*, vol. 36, no. 4, pp. 1153–1170, Aug. 2020.

[4] Y. Fan, R. Wang, and Y. Mao, "Stereo visual inertial odometry with online baseline calibration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 1084–1090.

[5] T. Qin and S. Shen, "Robust initialization of monocular visual-inertial estimation on aerial robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 4225–4232.

[6] C. Campos, J. M. M. Montiel, and J. D. Tardós, "Inertial-only optimization for visual-inertial initialization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 51–57.

[7] 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.

[8] K. Sun et al., "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 965–972, Apr. 2018.

[9] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 796–803, Apr. 2017.

[10] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[11] W. Huang and H. Liu, "Online initialization and automatic camera-IMU extrinsic calibration for monocular visual-inertial SLAM," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5182–5189.

[12] Y. Yang, P. Geneva, X. Zuo, and G. Huang, "Online self-calibration for visual-inertial navigation systems: Models, analysis and degeneracy," 2022, *arXiv:2201.09170*.

[13] 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.

[14] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *Int. J. Robot. Res.*, vol. 32, no. 6, pp. 690–711, May 2013, doi: 10.1177/0278364913481251.

[15] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "A first-estimates Jacobian EKF for improving SLAM consistency," in *Experimental Robotics*. Berlin, Germany: Springer, 2009, pp. 373–382.

[16] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4666–4672.

[17] Z. Huai and G. Huang, "Robocentric visual–inertial odometry," *Int. J. Robot. Res.*, vol. 41, no. 7, pp. 667–689, Jun. 2022, doi: 10.1177/0278364919853361.

[18] R. Miao, J. Qian, Y. Song, R. Ying, and P. Liu, "UniVIO: Unified direct and feature-based underwater stereo visual-inertial odometry," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[19] L. Xia, D. Meng, J. Zhang, D. Zhang, and Z. Hu, "Visual-inertial simultaneous localization and mapping: Dynamically fused point-line feature extraction and engineered robotic applications," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[20] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, *Initialization-Free Monocular Visual-Inertial State Estimation with Application to Autonomous MAVs*. Cham, Switzerland: Springer, 2016, pp. 211–227.

[21] 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.

[22] J. Kaiser, A. Martinelli, F. Fontana, and D. Scaramuzza, "Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation," *IEEE Robot. Autom. Lett.*, vol. 2, no. 1, pp. 18–25, Jan. 2017.

[23] Z. Wang and X. Cheng, "Adaptive optimization online IMU self-calibration method for visual-inertial navigation systems," *Measurement*, vol. 180, Aug. 2021, Art. no. 109478.

[24] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," 2019, *arXiv:1901.03642*.

[25] J. Jaekel, J. G. Mangelson, S. Scherer, and M. Kaess, "A robust multi-stereo visual-inertial odometry pipeline," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4623–4630.

[26] K. Eckenhoff, P. Geneva, J. Bloecker, and G. Huang, "Multi-camera visual-inertial navigation with online intrinsic and extrinsic calibration," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3158–3164.

[27] K. Eckenhoff, P. Geneva, and G. Huang, "MIMC-VINS: A versatile and resilient multi-IMU multi-camera visual-inertial navigation system," *IEEE Trans. Robot.*, vol. 37, no. 5, pp. 1360–1380, Oct. 2021.

[28] T. D. Barfoot, *State Estimation for Robotics*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[29] J. Shi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1994, pp. 593–600.

[30] J. Y. Bouguet, *Pyramidal Implementation of the Lucas–Kanade Feature Tracker Description of the Algorithm*. San Jose, CA, USA: Intel Corporation, Microprocessor Research Labs, 2000.

[31] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 6, pp. 580–593, Jun. 1997.

[32] J. L. Schönberger and J. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[33] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-Posteriori estimation," in *Proc. Robot., Sci. Syst.*, 2015.

[34] J. Civera, A. J. Davison, and J. Montiel, "Inverse depth parametrization for monocular SLAM," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 932–945, Oct. 2008.

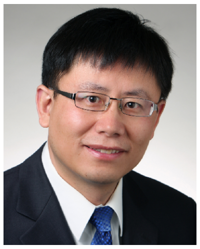
[35] K. Eckenhoff, L. Paull, and G. Huang, "Decoupled, consistent node removal and edge sparsification for graph-based SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 3275–3282.

[36] 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, Sep. 2010.

[37] M. Burri et al., "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, Sep. 2016.

Hongpei Yin received the B.S. degree in mechanical engineering and automation from the North China University of Technology, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in mechatronics with the Beijing Jiaotong University, Beijing.

His current research interests include visual SLAM and state estimation.



Peter Xiaoping Liu (Fellow, IEEE) received the B.Sc. degree in mechanical engineering and the M.Sc. degree in instrumentation and control engineering from Northern Jiaotong University, Beijing, China, in 1992 and 1995, respectively, and the Ph.D. degree in electrical and control engineering from the University of Alberta, Edmonton, AB, Canada, in 2002. He has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, since July 2002, where he is currently a Professor. His is also with Beijing Jiaotong University, Beijing, as an Adjunct Professor. His research interests include interactive networked systems and teleoperation, haptics, surgical simulation, and control and intelligent systems. Dr. Liu has served as an Associate Editor for several journals, including IEEE/ASME Transactions on Mechatronics, IEEE Transactions on Cybernetics, IEEE Transactions on Automation Science and Engineering, IEEE/CAA Journal of Automatica Sinica, and IEEE transactions on Instrumentation and Measurement. He is a licensed member of the Professional Engineers of Ontario (P.Eng), a Fellow of the Engineering Institute of Canada (FEIC), and a Fellow of Canadian Academy of Engineering (FCAE).

Minhua Zheng (Member, IEEE) received the B.Sc. degree in measurement control and information technology from Beihang University, Beijing, China, in 2010, and the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong, Hong Kong, in 2015. She is currently an Associate Professor with the School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing. Her research interests include robotics and intelligent systems, human–robot interaction, social robotics, and virtual surgery systems.