2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Online Initialization and Automatic Camera-IMU Extrinsic Calibration for Monocular Visual-Inertial SLAM

Weibo Huang, Hong Liu

Abstract— **Most of the existing monocular visual-inertial SLAM techniques assume that the camera-IMU extrinsic parameters are known, therefore these methods merely estimate the initial values of velocity, visual scale, gravity, biases of gyroscope and accelerometer in the initialization stage. However, it's usually a professional work to carefully calibrate the extrinsic parameters, and it is required to repeat this work once the mechanical configuration of the sensor suite changes slightly. To tackle this problem, we propose an online initialization method to automatically estimate the initial values and the extrinsic parameters without knowing the mechanical configuration. The biases of gyroscope and accelerometer are considered in our method, and a convergence criteria for both orientation and translation calibration is introduced to identify the convergence and to terminate the initialization procedure. In the three processes of our method, an iterative strategy is firstly introduced to iteratively estimate the gyroscope bias and the extrinsic orientation. Secondly, the scale factor, gravity, and extrinsic translation are approximately estimated without considering the accelerometer bias. Finally, these values are further optimized by a refinement algorithm in which the accelerometer bias and the gravitational magnitude are taken into account. Extensive experimental results show that our method achieves competitive accuracy compared with the state-of-the-art with less calculation.**

I. Introduction

Monocular visual-inertial SLAM is a technique which aims to track the incremental motion of a mobile platform and to simultaneously build a representative map for an environment using the measurements from a single on-board camera and an IMU sensor. Visual camera and inertial measurement unit (IMU) are ideal choice for SLAM techniques since the two sensor modalities are miniaturized size, cheap, low power consumption and acknowledged to complement each other. The rich representation of environments projected into an image helps to build a map and to estimate the trajectory of a camera up-to-scale, while the accurate short-term rigid body motion can be estimated by integrating the measurements of gyroscope and accelerometer contained in an IMU sensor. These properties make it possible for the visual-inertial setup to be applied to extensive practical applications, such as robot navigation [1], [2], autonomous or semi-autonomous driving [3], live metric 3D reconstruction [4], augmented and virtual reality [5], etc.

Several visual-inertial techniques have been presented in the literature, such as the recursive algorithms [6], [7] which commonly use the IMU measurements for state propagation, and the keyframe-based nonlinear optimization methods [8]–[10] which jointly minimize visual and inertial geometry error. However, the performances of these methods heavily depend on prior precise extrinsic calibration of the six-degree-of-freedom (6-DOF) transformation between the camera and the IMU. Extrinsic parameters are the bridge of the state transformation between camera reference frame and IMU reference frame. Incorrect calibration will introduce a systematic bias in motion estimation and degrade the overall navigation performance.

One alternative to obtain precise extrinsic parameters is to utilize offline methods [11]–[13]. However, these methods are complex and time-consuming since they usually require a professional user to carefully move the sensor suite in front of a stationary calibration target. Besides, it is usually required to repeat this process whenever the sensors are repositioned or significant mechanical stress is applied. Another alternative is to apply the online calibration approaches to jointly estimate the initial values and the extrinsic parameters. A self-calibration method based on unscented Kalman filter is proposed by Kelly *et al.* [14] to calibrate the extrinsic parameters. A closed-form solution is introduced by Martinelli [15] to estimate the initial values, and its revision is later proposed in [16] to automatically estimate the gyroscope bias. Yang and Shen [17] calibrate the extrinsic parameters and the initial values (except for IMU bias) with an optimization-based linear estimator. In their extended monocular visual-inertial system (VINS-Mono) [18], the IMU bias calibration is included in the sliding window nonlinear estimator.

Recently, an efficient IMU initialization method named VI ORB_SLAM was introduced by Mur-Artal *et al.* [19]. This method subdivides the initialization process into three simple sub-problems and achieves high accuracy in a short time. Unfortunately, the camera-IMU extrinsic parameters are still assumed as prerequisites in this method. Motivated by Mur-Artal's work, in this paper, three simple processes are introduced to automatically estimate the extrinsic parameters and the initial values with no need for any initial guess or any priori knowledge about the mechanical configuration of the sensor suite. Firstly, an iterative strategy is introduced to estimate the camera-IMU orientation and the gyroscope bias. Secondly, the extrinsic translation, the

This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046, 61673030, U1613209), Natural Science Foundation of Guangdong Province (No. 2015A030311034), Scientific Research Project of Guangdong Province (No. 2015B010919004), Specialized Research Fund for Strategic and Prospective Industrial Development of Shenzhen City (No. ZLZBCXLJZI20160729020003), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No. ZDSYS201703031405467).

The authors are with the Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School. {weibohuang, hongliu}@pku.edu.cn

visual scale and the gravity are roughly estimated. Finally, by considering the accelerometer bias and the magnitude of the gravitational acceleration, the scale, gravity and extrinsic translation are further refined. Although the proposed method is implemented based on the monocular ORB-SLAM [20], [21], our method is generic and could be applied to any keyframe-based visual odometry or SLAM system. The main contributions of this paper are identified as follows:

- Without knowing the mechanical configuration of the sensor suite, the scale, velocity, gravity, IMU biases, and the extrinsic parameters are jointly estimated.
- An iterative strategy is conducted to calculate the extrinsic orientation and gyroscope bias.
- A general convergence criteria is introduced to identify the convergence of the orientation and translation calibration and to terminate the initialization procedure.

II. Preliminaries

In this section, the necessary notations and some useful geometric concepts are briefly reviewed. Besides, the definition of the reference frames and the IMU preintegration on manifold are also presented as follows.

A. Notation

In this paper, vectors and matrixes are denoted in boldface, and a vector is expressed with respect to a specific reference frame by appending a left subscript, e.g. ${}_{\mathrm{A}}\mathbf{v}$ for the vector $\mathbf{v}$ expressed in frame $\{A\}$. If a vector describes the relative transformation from one reference frame to another frame, a right subscript is appended to indicate the transformed frame, e.g. ${}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}$ for the vector that defines the translation from camera frame $\{C\}$ to IMU body frame $\{B\}$. For the rotation matrix, the orientation from camera frame to IMU body frame is denoted as $\mathbf{R}_{\mathrm{CB}}$.

B. Geometric Concepts

1) Skew-symmetric matrix: The skew-symmetric matrix of a vector $\mathbf{a} = [a_x\ a_y\ a_z]^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$ is denoted as $[\mathbf{a}]_\times$:

$$[\mathbf{a}]_\times = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix}. \tag{1}$$

Given two vectors $\mathbf{a},\mathbf{b} \in \mathbb{R}^{3\times 1}$, the cross-product can be written as $\mathbf{a}\times\mathbf{b} = [\mathbf{a}]_\times\mathbf{b} = -[\mathbf{b}]_\times\mathbf{a}$.

2) Quaternion: A quaternion is a four-dimensional complex number that can be used to non-singularly parameterize the 3D rotation group $SO(3)$. It generally consists of a scalar part q_w and a vector part $\mathbf{q}_v$:

$$\mathbf{q} = q_w + q_x\mathbf{i} + q_y\mathbf{j} + q_z\mathbf{k} = \begin{bmatrix} q_w \\ \mathbf{q}_v \end{bmatrix} \tag{2}$$

where $q_w, q_x, q_y, q_z \in \mathbb{R}$, $\mathbf{q}_v = [q_x\ q_y\ q_z]^{\mathrm{T}} \in \mathbb{R}^{3\times 1}$. A quaternion can be normalised by dividing its norm $||\mathbf{q}||$, where $||\mathbf{q}|| = \sqrt{q_w^2 + q_x^2 + q_y^2 + q_z^2}$. In practice, quaternion arithmetic often requires that a quaternion describing an orientation needs to be normalised. As a result, the quaternions mentioned in the rest of this paper are normalised by default.

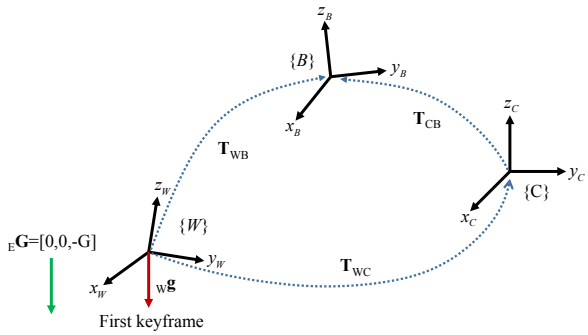

Fig. 1. Relationships among the world $\{W\}$, camera $\{C\}$, and IMU body $\{B\}$ reference frames. The camera and IMU are rigidly attached to a common bracket (e.g. printed circuit board (PCB)). The transformation $\mathbf{T}_{\mathrm{CB}} = [\mathbf{R}_{\mathrm{CB}}|{}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}]$ between the camera frame $\{C\}$ and the IMU body frame $\{B\}$ is usually unknown and has to be calibrated online. The ${}_{\mathrm{E}}\mathbf{G}$ and ${}_{\mathrm{W}}\mathbf{g}$ are respectively the vectors of gravitational acceleration in the earth's inertial frame $\{E\}$ and in the world frame $\{W\}$.

Given two unit quaternions $\mathbf{q}_{\mathrm{AB}}$ and $\mathbf{q}_{\mathrm{BC}}$, the compound orientation $\mathbf{q}_{\mathrm{AC}}$ can be defined by the quaternion multiplication:

$$\begin{aligned} \mathbf{q}_{\mathrm{AC}} &= \mathbf{q}_{\mathrm{AB}} \otimes \mathbf{q}_{\mathrm{BC}} \\ &= [\mathbf{q}_{\mathrm{AB}}]_L \cdot \mathbf{q}_{\mathrm{BC}} = [\mathbf{q}_{\mathrm{BC}}]_R \cdot \mathbf{q}_{\mathrm{AB}} \end{aligned} \tag{3}$$

where $\otimes$ is the quaternion multiplication operator, and the matrix representations for left and right quaternion multiplication are respectively defined as follows:

$$\begin{aligned} [\mathbf{q}]_L &= \begin{bmatrix} q_w & -\mathbf{q}_v^T \\ \mathbf{q}_v & q_w\mathbf{I}_{3\times 3} + [\mathbf{q}_v]_\times \end{bmatrix} \\ [\mathbf{q}]_R &= \begin{bmatrix} q_w & -\mathbf{q}_v^T \\ \mathbf{q}_v & q_w\mathbf{I}_{3\times 3} - [\mathbf{q}_v]_\times \end{bmatrix} \end{aligned} \tag{4}$$

where $\mathbf{I}_{3\times 3}$ is the 3×3 identity matrix, and $[\mathbf{q}_v]_\times$ represents the skew-symmetric matrix of the vector part of a quaternion.

An orientation can be expressed in an unit quaternion $\mathbf{q}$ or a rotation matrix $\mathbf{R}$. The transformation between these two formats is:

$$\begin{aligned} \mathbf{q} \mapsto \mathbf{R}:&\quad \mathbf{R} = (2q_w^2 - 1)\mathbf{I}_{3\times 3} + 2q_w[\mathbf{q}_v]_\times + 2\mathbf{q}_v\mathbf{q}_v^{\mathrm{T}} \\ \mathbf{R} \mapsto \mathbf{q}:&\quad q_w = \frac{\sqrt{\mathrm{tr}(\mathbf{R}) + 1}}{2} \\ &\quad \mathbf{q}_v = \begin{bmatrix} \frac{m_{32} - m_{23}}{4q_w} & \frac{m_{13} - m_{31}}{4q_w} & \frac{m_{21} - m_{12}}{4q_w} \end{bmatrix}^{\mathrm{T}} \end{aligned} \tag{5}$$

where the rotation matrix $\mathbf{R} \in SO(3)$, $\mathbf{R} = \{m_{ij}\}$, $m_{ij} \in \mathbb{R}$, $i, j \in [1,2,3]$, and the trace $\mathrm{tr}(\mathbf{R})$ is the sum of the elements on the main diagonal of $\mathbf{R}$.

C. Reference Frames

The relationships among the reference frames of our method are shown in Fig. 1. Since the visual odometry and the IMU integration measure the relative motion, the absolute attitude in the pre-fixed reference frame (e.g. the earth's inertial frame) can not be determined. Therefore, the reference frame (world frame $\{W\}$) of our method coincides with the coordinate system of the first keyframe which is determined by the monocular SLAM system. Our goal is to calibrate the rigid orientation $\mathbf{R}_{\mathrm{CB}} \in SO(3)$ and the translation

${}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}} \in \mathbb{R}^{3\times 1}$ between camera frame $\{C\}$ and IMU body frame $\{B\}$, meanwhile calculating the velocity ${}_{\mathrm{W}}\mathbf{v}_{\mathrm{B}}$, gyroscope bias $\mathbf{b}_{\mathrm{g}}$, accelerometer bias $\mathbf{b}_{\mathrm{a}}$, gravity ${}_{\mathrm{W}}\mathbf{g}$, and the visual scale s in the initialization stage.

Considering the scale factor s for the trajectory computed by monocular SLAM, the transformation between camera frame $\{C\}$ and IMU body frame $\{B\}$ is:

$$\mathbf{R}_{\mathrm{WB}} = \mathbf{R}_{\mathrm{WC}}\mathbf{R}_{\mathrm{CB}} \tag{6}$$

$$ {}_{\mathrm{W}}\mathbf{p}_{\mathrm{B}} = \mathbf{R}_{\mathrm{WC}} \cdot {}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}} + s \cdot {}_{\mathrm{W}}\mathbf{p}_{\mathrm{C}}. \tag{7}$$

D. IMU Model and Preintegration

An IMU commonly provides discrete time samples of the rotation rate and the acceleration of the sensor w.r.t. the body frame $\{B\}$. In principle, the change in pose of a strap-down IMU can be determined by integrating the gyroscope and accelerometer outputs with an initial linear speed. However, in addition to white sensor noises η_{g} and η_{a}, the measurement outputs ω_{B} and $\mathbf{a}_{\mathrm{B}}$ are subject to low-frequency drift biases $\mathbf{b}_{\mathrm{g}}$ and $\mathbf{b}_{\mathrm{a}}$ which limit the accuracy of inertial dead-reckoning over extended periods. In addition, the IMU accelerometer senses the force of gravitational attraction. The effect of the gravity ${}_{\mathrm{W}}\mathbf{g}$ in the world frame should be subtracted since the magnitude of the gravitational attraction is often large enough to dominate other measured accelerations.

The concept of *preintegrated IMU measurement* was first proposed in [22] and extended by Forster *et al.* [23] on the manifold space. Given two consecutive keyframes at time i and j, the corresponding IMU orientation $\mathbf{R}_{\mathrm{WB}}$, velocity ${}_{\mathrm{W}}\mathbf{v}_{\mathrm{B}}$ and position ${}_{\mathrm{W}}\mathbf{p}_{\mathrm{B}}$ can be computed by summarizing all measurements within this period:

$$\begin{aligned}
\mathbf{R}^{j}_{\mathrm{WB}} &= \mathbf{R}^{i}_{\mathrm{WB}} \prod_{k=i}^{j-1} \mathrm{Exp}\left((\omega^{k}_{\mathrm{B}} - \mathbf{b}^{k}_{\mathrm{g}} - \eta^{k}_{\mathrm{g}})\Delta t\right) \\
{}_{\mathrm{W}}\mathbf{v}^{j}_{\mathrm{B}} &= {}_{\mathrm{W}}\mathbf{v}^{i}_{\mathrm{B}} + {}_{\mathrm{W}}\mathbf{g}\Delta t_{ij} + \sum_{k=i}^{j-1} \mathbf{R}^{k}_{\mathrm{WB}}(\mathbf{a}^{k}_{\mathrm{B}} - \mathbf{b}^{k}_{\mathrm{a}} - \eta^{k}_{\mathrm{a}})\Delta t \\
{}_{\mathrm{W}}\mathbf{p}^{j}_{\mathrm{B}} &= {}_{\mathrm{W}}\mathbf{p}^{i}_{\mathrm{B}} + \sum_{k=i}^{j-1} \left({}_{\mathrm{W}}\mathbf{v}^{k}_{\mathrm{B}}\Delta t + \frac{1}{2}{}_{\mathrm{W}}\mathbf{g}\Delta t^{2} \right. \\
&\quad \left. + \frac{1}{2}\mathbf{R}^{k}_{\mathrm{WB}}(\mathbf{a}^{k}_{\mathrm{B}} - \mathbf{b}^{k}_{\mathrm{a}} - \eta^{k}_{\mathrm{a}})\Delta t^{2} \right)
\end{aligned} \tag{8}$$

where Δt denotes the IMU sampling interval, and $\Delta t_{ij} \doteq (j-i)\Delta t$. The $\mathrm{Exp}(\cdot)$ is the exponential map operator that maps Lie algebra $\mathfrak{so}(3)$ to Lie group. When the influence of IMU measurement noise is ignored, and the bias is assumed to remain constant during the preintegration period, a small bias correction $\delta\mathbf{b}^{i}_{(\cdot)}$ w.r.t. previously estimated $\bar{\mathbf{b}}^{i}_{(\cdot)}$ could be taken into account to rectify the preintegrated results. The expressions in (8) can be rewritten as:

$$\begin{aligned}
\mathbf{R}^{j}_{\mathrm{WB}} &= \mathbf{R}^{i}_{\mathrm{WB}}\Delta\bar{\mathbf{R}}_{ij}\mathrm{Exp}(\mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{R}}_{ij}}\delta\mathbf{b}^{i}_{\mathrm{g}}) \\
{}_{\mathrm{W}}\mathbf{v}^{j}_{\mathrm{B}} &= {}_{\mathrm{W}}\mathbf{v}^{i}_{\mathrm{B}} + {}_{\mathrm{W}}\mathbf{g}\Delta t_{ij} + \mathbf{R}^{i}_{\mathrm{WB}}\left(\Delta\bar{\mathbf{v}}_{ij} + \mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{v}}_{ij}}\delta\mathbf{b}^{i}_{\mathrm{g}} + \mathbf{J}^{\mathrm{a}}_{\Delta\bar{\mathbf{v}}_{ij}}\delta\mathbf{b}^{i}_{\mathrm{a}}\right) \\
{}_{\mathrm{W}}\mathbf{p}^{j}_{\mathrm{B}} &= {}_{\mathrm{W}}\mathbf{p}^{i}_{\mathrm{B}} + {}_{\mathrm{W}}\mathbf{v}^{i}_{\mathrm{B}}\Delta t_{ij} + \frac{1}{2}{}_{\mathrm{W}}\mathbf{g}\Delta t^{2}_{ij} \\
&\quad + \mathbf{R}^{i}_{\mathrm{WB}}\left(\Delta\bar{\mathbf{p}}_{ij} + \mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{p}}_{ij}}\delta\mathbf{b}^{i}_{\mathrm{g}} + \mathbf{J}^{\mathrm{a}}_{\Delta\bar{\mathbf{p}}_{ij}}\delta\mathbf{b}^{i}_{\mathrm{a}}\right)
\end{aligned} \tag{9}$$

where the Jacobians $\mathbf{J}^{\mathrm{g}}_{(\cdot)}$ and $\mathbf{J}^{\mathrm{a}}_{(\cdot)}$ describe how the measurements change due to a change in the bias estimation. The biases $\bar{\mathbf{b}}^{i}_{\mathrm{g}}$ and $\bar{\mathbf{b}}^{i}_{\mathrm{a}}$ remain constant during the preintegration and they can be precomputed at time i . The details of the Jacobians can be found in [23]. The terms of preintagration $\Delta\bar{\mathbf{R}}_{ij}$, $\Delta\bar{\mathbf{v}}_{ij}$ and $\Delta\bar{\mathbf{p}}_{ij}$ are independent of the states at time i and the gravity, in particular, they can be computed directly from the IMU outputs between two keyframes, as follows:

$$\begin{aligned}
\Delta\bar{\mathbf{R}}_{ij} &= \prod_{k=i}^{j-1} \mathrm{Exp}\left((\omega^{k}_{\mathrm{B}} - \bar{\mathbf{b}}^{i}_{\mathrm{g}})\Delta t\right) \\
\Delta\bar{\mathbf{v}}_{ij} &= \sum_{k=i}^{j-1} \Delta\bar{\mathbf{R}}_{ik}(\mathbf{a}^{k}_{\mathrm{B}} - \bar{\mathbf{b}}^{i}_{\mathrm{a}})\Delta t \\
\Delta\bar{\mathbf{p}}_{ij} &= \sum_{k=i}^{j-1} \left(\Delta\bar{\mathbf{v}}_{ik}\Delta t + \frac{1}{2}\Delta\bar{\mathbf{R}}_{ik}(\mathbf{a}^{k}_{\mathrm{B}} - \bar{\mathbf{b}}^{i}_{\mathrm{a}})\Delta t^{2}\right).
\end{aligned} \tag{10}$$

III. IMU Initialization and Extrinsic Calibration

This section details the proposed online initialization method to jointly calibrate the camera-IMU extrinsic parameters, as well as estimating the initial values, including velocity, scale factor, gravity, and the biases of gyroscope and accelerometer. In order to make all variables observable, our method requires that the pure monocular SLAM system has been performed for a few seconds and several keyframes have been determined.

A. Gyroscope Bias Estimation and Orientation Calibration

Assuming the gyroscope bias changes slowly and the change could be neglected in the initialization stage, a constant gyroscope bias can be estimated by minimizing the difference between camera rotations and IMU body rotations. As shown in Eq.(6), a prior calibrated extrinsic orientation $\mathbf{R}_{\mathrm{CB}}$ is needed for transforming the camera rotations into IMU body frame. However, this extrinsic orientation is usually unknown in most of practical applications. This problem can be solved by calculating the gyroscope bias and the extrinsic orientation using the same minimum function. However, the Jacobian calculation for extrinsic orientation on $SO(3)$ is complex due to the self-constraint.

In this paper, an iterative strategy is introduced to decouple the problem. At each iteration, the last calibrated extrinsic orientation is utilized to estimate a gyroscope bias. Then the integrated rotations are rectified by applying the estimated gyroscope bias. Finally a linear over-determined equation is developed to calculate a new extrinsic orientation.

1) Gyroscope Bias Estimation: The rotation relationship of two consecutive keyframes at time i and $i+1$ is described

as:

$$\mathbf{R}^{i}_{\mathrm{WB}} = \mathbf{R}^{i}_{\mathrm{WC}}\hat{\mathbf{R}}_{\mathrm{CB}}, \quad \mathbf{R}^{i+1}_{\mathrm{WB}} = \mathbf{R}^{i+1}_{\mathrm{WC}}\hat{\mathbf{R}}_{\mathrm{CB}} \tag{11}$$

where $\hat{\mathbf{R}}_{\mathrm{CB}}$ is the extrinsic orientation calibrated in last iteration.

Substituting Eq.(9) into Eq.(11), the relationship of preintegration rotation and camera rotation can be described as:

$$\Delta\bar{\mathbf{R}}_{i,i+1}\mathrm{Exp}(\mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{R}}}\delta\mathbf{b}_{\mathrm{g}}) = \hat{\mathbf{R}}_{\mathrm{BC}}\mathbf{R}^{i}_{\mathrm{CW}}\mathbf{R}^{i+1}_{\mathrm{WC}}\hat{\mathbf{R}}_{\mathrm{CB}} \tag{12}$$

where $\hat{\mathbf{R}}_{\mathrm{BC}}$, $\mathbf{R}^{i}_{\mathrm{CW}}$ are the inverses of $\hat{\mathbf{R}}_{\mathrm{CB}}$, $\mathbf{R}^{i}_{\mathrm{WC}}$ respectively. Using the logarithm map $\mathrm{Log}(\cdot)$, the minimum function for bias estimation is:

$$\begin{aligned}\delta\mathbf{b}^{*}_{\mathrm{g}} = \underset{\delta\mathbf{b}_{\mathrm{g}}}{\mathrm{argmin}} \sum_{i=1}^{N-1} \Big\| \mathrm{Log}\Big(\big(\Delta\bar{\mathbf{R}}_{i,i+1}\mathrm{Exp}(\mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{R}}}\delta\mathbf{b}_{\mathrm{g}})\big)^{\mathrm{T}} \\ \hat{\mathbf{R}}_{\mathrm{BC}}\mathbf{R}^{i}_{\mathrm{CW}}\mathbf{R}^{i+1}_{\mathrm{WC}}\hat{\mathbf{R}}_{\mathrm{CB}}\Big)\Big\|^{2}\end{aligned} \tag{13}$$

where $\|\cdot\|$ is the L2-norm. This equation can be solved with Gauss-Newton algorithm [24]. Since the gyroscope bias is assumed to remain constant in the initialization stage, and the gyroscope bias $\bar{\mathbf{b}}_{\mathrm{g}}$ is set to zero $\mathbf{0}_{3\times1}$ when integrating the $\Delta\bar{\mathbf{R}}_{i,i+1}$, the final estimated gyroscope bias is $\mathbf{b}^{*}_{\mathrm{g}} = \mathbf{0}_{3\times1} + \delta\mathbf{b}^{*}_{\mathrm{g}} = \delta\mathbf{b}^{*}_{\mathrm{g}}$. Once the gyroscope bias has been estimated, the preintegrations can be rectified by Eq.(10).

2) Extrinsic Orientation Calibration: The extrinsic orientation can be calibrated by aligning the camera rotations with the integrated IMU rotations. Here, the "delta" notations $\Delta_{\mathrm{C}}\mathbf{R}_{i,i+1} = \mathbf{R}^{i}_{\mathrm{CW}}\mathbf{R}^{i+1}_{\mathrm{WC}}$ and $\Delta_{\mathrm{B}}\mathbf{R}_{i,i+1} = \mathbf{R}^{i}_{\mathrm{BW}}\mathbf{R}^{i+1}_{\mathrm{WB}}$ are respectively defined to describe the relative rotation increments of the camera and the IMU body. The rotation relationship described in Eq.(11) becomes:

$$\mathbf{R}_{\mathrm{CB}}\Delta_{\mathrm{B}}\mathbf{R}_{i,i+1} = \Delta_{\mathrm{C}}\mathbf{R}_{i,i+1}\mathbf{R}_{\mathrm{CB}}. \tag{14}$$

With the quaternion representations $\mathbf{q}_{\mathrm{CB}}$, ${}_{\mathrm{B}}\mathbf{q}_{i,i+1}$ and ${}_{\mathrm{C}}\mathbf{q}_{i,i+1}$ for $\mathbf{R}_{\mathrm{CB}}$, $\Delta_{\mathrm{B}}\mathbf{R}_{i,i+1}$ and $\Delta_{\mathrm{C}}\mathbf{R}_{i,i+1}$ respectively, (14) can be described as:

$$\begin{aligned}&\mathbf{q}_{\mathrm{CB}} \otimes {}_{\mathrm{B}}\mathbf{q}_{i,i+1} = {}_{\mathrm{C}}\mathbf{q}_{i,i+1} \otimes \mathbf{q}_{\mathrm{CB}} \\ \Rightarrow\ &([{}_{\mathrm{B}}\mathbf{q}_{i,i+1}]_{R} - [{}_{\mathrm{C}}\mathbf{q}_{i,i+1}]_{L}) \cdot \mathbf{q}_{\mathrm{CB}} = \mathbf{Q}_{i,i+1} \cdot \mathbf{q}_{\mathrm{CB}} = \mathbf{0}_{4\times1}.\end{aligned} \tag{15}$$

Writing N keyframes i, $i+1$, ..., $i+N-1$ as 1, 2, ..., N for clarity of notation, all relations can be stacked into a linear over-determined equation:

$$\begin{bmatrix} w_{12}\cdot\mathbf{Q}_{12} \\ w_{23}\cdot\mathbf{Q}_{23} \\ \vdots \\ w_{N-1,N}\cdot\mathbf{Q}_{N-1,N} \end{bmatrix} \cdot \mathbf{q}_{\mathrm{CB}} = \mathbf{0}_{4(N-1)\times1} \tag{16}$$

where the $w_{i,i+1}$ is a weight for outlier handling. As the orientation calibration runs with incoming keyframes, the previously estimated result $\hat{\mathbf{q}}_{\mathrm{CB}}$ can be employed to calculate the residual. Specifically, the residual for two consecutive keyframes at time i and $i+1$ is defined as:

$$\mathbf{e}_{i,i+1} = \mathbf{Q}_{i,i+1} \cdot \hat{\mathbf{q}}_{\mathrm{CB}}. \tag{17}$$

The weight can be calculated as:

$$w_{i,i+1} = exp(-\|\mathbf{e}_{i,i+1}\| \cdot K_o) \tag{18}$$

where $exp(\cdot)$ indicates the exponential function, and K_o is a gain factor used to amplify the residual effect.

The orientation $\mathbf{q}_{\mathrm{CB}}$ in Eq.(16) can be obtained by solving the minimum function:

$$\mathbf{q}^{*}_{\mathrm{CB}} = \underset{\mathbf{q}_{\mathrm{CB}}:\|\mathbf{q}_{\mathrm{CB}}\|=1}{\mathrm{argmin}} \left\|\mathbf{A}_{4(N-1)\times4} \cdot \mathbf{q}_{\mathrm{CB}}\right\|. \tag{19}$$

In Eq.(19), there are $4(N-1)$ equations and 4 unknowns, therefore at least a pair of keyframes is required to calculate a solution. Once an orientation $\mathbf{q}^{*}_{\mathrm{CB}}$ has been obtained from Eq.(19), the corresponding matrix representation $\mathbf{R}^{*}_{\mathrm{CB}}$ can be calculated according to Eq.(5). The $\hat{\mathbf{R}}_{\mathrm{CB}}$ and $\hat{\mathbf{q}}_{\mathrm{CB}}$ which are respectively used for gyroscope bias estimation and residual computation can be updated for next iteration. The convergence criteria for orientation calibration is described in Section.III-D.

B. Scale, Gravity and Translation Approximation (no accelerometer bias)

Once the camera-IMU orientation has been calibrated, the scale s, gravity ${}_{\mathrm{W}}\mathbf{g}$ and extrinsic translation ${}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}$ can be approximately estimated without considering the accelerometer bias. Since the preintegrations have been rectified after the gyroscope bias estimation and the gyroscope bias is assumed constant, the $\mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{p}}}$ and $\mathbf{J}^{\mathrm{g}}_{\Delta\bar{\mathbf{v}}}$ can be set to zero. Substituting Eq.(6) and Eq.(7) into the third equation of Eq.(9), the position relationship of two consecutive keyframes can be obtained:

$$\begin{aligned}s\cdot{}_{\mathrm{W}}\mathbf{p}^{i+1}_{\mathrm{C}} =& s\cdot{}_{\mathrm{W}}\mathbf{p}^{i}_{\mathrm{C}} + {}_{\mathrm{W}}\mathbf{v}^{i}_{\mathrm{B}}\Delta t_{i,i+1} + \frac{1}{2}{}_{\mathrm{W}}\mathbf{g}\Delta t^{2}_{i,i+1} \\ &+ \mathbf{R}^{i}_{\mathrm{WC}}\mathbf{R}^{*}_{\mathrm{CB}}\Delta\bar{\mathbf{p}}_{i,i+1} + (\mathbf{R}^{i}_{\mathrm{WC}} - \mathbf{R}^{i+1}_{\mathrm{WC}}){}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}.\end{aligned} \tag{20}$$

Since the accelerometer bias is not considered at this stage, the IMU body rotation $\mathbf{R}^{i}_{\mathrm{WB}}$ is replaced by $\mathbf{R}^{i}_{\mathrm{WC}}\mathbf{R}^{*}_{\mathrm{CB}}$ to eliminate the drift caused by accelerometer bias. Eq.(20) aims to estimate s, ${}_{\mathrm{W}}\mathbf{g}$ and ${}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}$. Using the velocity relation in Eq.(9), two relations between three consecutive keyframes are calculated to reduce the burden of solving N velocities, which results in the following expression:

$$\begin{bmatrix} \lambda(i) & \beta(i) & \varphi(i) \end{bmatrix} \begin{bmatrix} s \\ {}_{\mathrm{W}}\mathbf{g} \\ {}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}} \end{bmatrix} = \gamma(i). \tag{21}$$

When writing keyframes i, $i+1$, $i+2$ as 1, 2, 3, we have:

$$\begin{aligned}\lambda(i) &= ({}_{\mathrm{W}}\mathbf{p}^{2}_{\mathrm{C}} - {}_{\mathrm{W}}\mathbf{p}^{1}_{\mathrm{C}})\Delta t_{23} - ({}_{\mathrm{W}}\mathbf{p}^{3}_{\mathrm{C}} - {}_{\mathrm{W}}\mathbf{p}^{2}_{\mathrm{C}})\Delta t_{12} \\ \beta(i) &= \frac{1}{2}(\Delta t_{12}\Delta t^{2}_{23} + \Delta t^{2}_{12}\Delta t_{23})\mathbf{I}_{3\times3} \\ \varphi(i) &= (\mathbf{R}^{2}_{\mathrm{WC}} - \mathbf{R}^{3}_{\mathrm{WC}})\Delta t_{12} - (\mathbf{R}^{1}_{\mathrm{WC}} - \mathbf{R}^{2}_{\mathrm{WC}})\Delta t_{23} \\ \gamma(i) &= \mathbf{R}^{1}_{\mathrm{WC}}\mathbf{R}^{*}_{\mathrm{CB}}\Delta\bar{\mathbf{p}}_{12}\Delta t_{23} - \mathbf{R}^{2}_{\mathrm{WC}}\mathbf{R}^{*}_{\mathrm{CB}}\Delta\bar{\mathbf{p}}_{23}\Delta t_{12} \\ &\quad - \mathbf{R}^{1}_{\mathrm{WC}}\mathbf{R}^{*}_{\mathrm{CB}}\Delta\bar{\mathbf{v}}_{12}\Delta t_{12}\Delta t_{23}.\end{aligned} \tag{22}$$

Similar to Eq.(16), a weight coefficient w_i is introduced to handle the outliers:

$$\begin{aligned}\mathbf{e}_{i} &= \hat{s}\cdot\lambda(i) + \beta(i){}_{\mathrm{W}}\hat{\mathbf{g}} + \varphi(i){}_{\mathrm{C}}\hat{\mathbf{p}}_{\mathrm{B}} - \gamma(i) \\ w_{i} &= exp(-\|\mathbf{e}_{i}\| \cdot K_t)\end{aligned} \tag{23}$$

where $\mathbf{e}_i$ represents the predicted residual, and K_t is a gain factor used to amplify the residual effect, and $\hat{s}$, ${}_{\mathrm{W}}\hat{\mathbf{g}}$ and ${}_{\mathrm{C}}\hat{\mathbf{p}}_{\mathrm{B}}$ indicate the last estimation results detailed in Section III-C.

All relations (21) for N keyframes can be stacked into a linear over-determined equation $\mathbf{B}_{3(N-2)\times 7}\cdot\mathbf{x}_{7\times 1}=\mathbf{C}_{3(N-2)\times 1}$ with the weights defined in (23) for outlier handling. The solution s^*, ${}_{\mathrm{W}}\mathbf{g}^*$, ${}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}^*$ can be obtained by solving the minimum function:

$$\mathbf{x}^*=\underset{\mathbf{x}}{\operatorname{argmin}}\left\|\mathbf{B}_{3(N-2)\times 7}\cdot\mathbf{x}_{7\times 1}-\mathbf{C}_{3(N-2)\times 1}\right\| \tag{24}$$

where $\mathbf{x}^*=[s^*\ \ {}_{\mathrm{W}}\mathbf{g}^{*\mathrm{T}}\ \ {}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}^{*\mathrm{T}}]^{\mathrm{T}}$. Since there are $3(N-2)$ equations and 7 unknowns, at least 5 keyframes is required to calculate a solution.

C. Accelerometer Bias Estimation, and Scale, Gravity and Translation Refinement

Note that the accelerometer bias has not been considered when computing the scale, gravity and camera-IMU translation. If the accelerometer bias is incorporated in Eq.(21), the chance of having an ill-conditioned system will increase since the gravity and accelerometer bias are hard to be distinguished. In our method, a rough gravity ${}_{\mathrm{W}}\mathbf{g}^*$ is approximated without considering the accelerometer bias in Section.III-B, and then the rough gravity is optimized by appending a perturbation. Using the already computed gravity ${}_{\mathrm{W}}\mathbf{g}^*$, the rotation between the earth's inertial frame and the world frame can be computed as follows:

$$\begin{aligned}&\mathbf{R}_{\mathrm{WI}}=\mathrm{Exp}(\tilde{\mathbf{v}}\theta)\\&\tilde{\mathbf{v}}=\frac{{}_{\mathrm{E}}\tilde{\mathbf{g}}\times{}_{\mathrm{W}}\tilde{\mathbf{g}}}{\|{}_{\mathrm{E}}\tilde{\mathbf{g}}\times{}_{\mathrm{W}}\tilde{\mathbf{g}}\|},\quad\theta=\mathrm{atan2}(\|{}_{\mathrm{E}}\tilde{\mathbf{g}}\times{}_{\mathrm{W}}\tilde{\mathbf{g}}\|,{}_{\mathrm{E}}\tilde{\mathbf{g}}\cdot{}_{\mathrm{W}}\tilde{\mathbf{g}})\\&{}_{\mathrm{E}}\tilde{\mathbf{g}}={}_{\mathrm{E}}\mathbf{G}/\|{}_{\mathrm{E}}\mathbf{G}\|,\quad{}_{\mathrm{W}}\tilde{\mathbf{g}}={}_{\mathrm{W}}\mathbf{g}^*/\|{}_{\mathrm{W}}\mathbf{g}^*\|.\end{aligned} \tag{25}$$

This rotation can be optimized by appending a perturbation $\delta\theta\in\mathbb{R}^{3\times 1}$:

$$\begin{aligned}&{}_{\mathrm{W}}\mathbf{g}=\mathbf{R}_{\mathrm{WI}}\mathrm{Exp}(\delta\theta)\cdot{}_{\mathrm{E}}\mathbf{G}\approx\mathbf{R}_{\mathrm{WI}}\cdot{}_{\mathrm{E}}\mathbf{G}-\mathbf{R}_{\mathrm{WI}}[{}_{\mathrm{E}}\mathbf{G}]_{\times}\delta\theta\\&\delta\theta=\begin{bmatrix}\delta\theta_{xy}^{\mathrm{T}}&0\end{bmatrix}^{\mathrm{T}},\quad\delta\theta_{xy}=[\delta\theta_x\ \ \delta\theta_y]^{\mathrm{T}},\\&{}_{\mathrm{E}}\mathbf{G}=[0\ \ 0\ \ -\mathrm{G}]^{\mathrm{T}}\end{aligned} \tag{26}$$

where the first-order approximation $\mathrm{Exp}(\delta\theta)\approx\mathbf{I}+[\delta\theta]_{\times}$ is applied, and G (normally $\mathrm{G}=9.81m\cdot s^{-2}$) represents the magnitude of the gravitational acceleration. Substituting Eq.(26) into Eq.(20) and including now the effect of accelerometer bias, we have:

$$\begin{aligned}s\cdot{}_{\mathrm{W}}\mathbf{p}_{\mathrm{C}}^{i+1}=&\,s\cdot{}_{\mathrm{W}}\mathbf{p}_{\mathrm{C}}^{i}+{}_{\mathrm{W}}\mathbf{v}_{\mathrm{B}}^{i}\Delta t_{i,i+1}-\frac{1}{2}\mathbf{R}_{\mathrm{WI}}[{}_{\mathrm{E}}\mathbf{G}]_{\times}\delta\theta\Delta t_{i,i+1}^2\\&+\mathbf{R}_{\mathrm{WC}}^{i}\mathbf{R}_{\mathrm{CB}}^{*}(\Delta\bar{\mathbf{p}}_{i,i+1}+\mathbf{J}_{\Delta\bar{\mathbf{p}}}^{\mathrm{a}}\delta\mathbf{b}_{\mathrm{a}})+(\mathbf{R}_{\mathrm{WC}}^{i}-\mathbf{R}_{\mathrm{WC}}^{i+1})\,{}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}\\&+\frac{1}{2}\mathbf{R}_{\mathrm{WI}}\cdot{}_{\mathrm{E}}\mathbf{G}\Delta t_{i,i+1}^2.\end{aligned} \tag{27}$$

Similar to the derivation of Eq.(21), we use the position relationships described in Eq.(9) to obtain the following expression:

$$\begin{bmatrix}\lambda(i)&\phi(i)&\zeta(i)&\xi(i)\end{bmatrix}\begin{bmatrix}s\\\delta\theta_{xy}\\\delta\mathbf{b}_{\mathrm{a}}\\{}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}\end{bmatrix}=\psi(i) \tag{28}$$

where $\lambda(i)$ remains the same as in (22), and $\phi(i)$, $\zeta(i)$ and $\xi(i)$ are computed as follows:

$$\begin{aligned}\phi(i)=&\left[-\frac{1}{2}\mathbf{R}_{\mathrm{WI}}[{}_{\mathrm{E}}\mathbf{G}]_{\times}(\Delta t_{12}\Delta t_{23}^2+\Delta t_{12}^2\Delta t_{23})\right]_{(:,1:2)}\\\zeta(i)=&\,\mathbf{R}_{\mathrm{WC}}^{2}\mathbf{R}_{\mathrm{CB}}^{*}\mathbf{J}_{\Delta\bar{\mathbf{p}}_{23}}^{\mathrm{a}}\Delta t_{12}-\mathbf{R}_{\mathrm{WC}}^{1}\mathbf{R}_{\mathrm{CB}}^{*}\mathbf{J}_{\Delta\bar{\mathbf{p}}_{12}}^{\mathrm{a}}\Delta t_{23}\\&+\mathbf{R}_{\mathrm{WC}}^{1}\mathbf{R}_{\mathrm{CB}}^{*}\mathbf{J}_{\Delta\bar{\mathbf{v}}_{12}}^{\mathrm{a}}\Delta t_{12}\Delta t_{23}\\\xi(i)=&\,(\mathbf{R}_{\mathrm{WC}}^{2}-\mathbf{R}_{\mathrm{WC}}^{3})\Delta t_{12}-(\mathbf{R}_{\mathrm{WC}}^{1}-\mathbf{R}_{\mathrm{WC}}^{2})\Delta t_{23}\\\psi(i)=&\,\mathbf{R}_{\mathrm{WC}}^{1}\mathbf{R}_{\mathrm{CB}}^{*}\Delta\bar{\mathbf{p}}_{12}\Delta t_{23}-\mathbf{R}_{\mathrm{WC}}^{2}\mathbf{R}_{\mathrm{CB}}^{*}\Delta\bar{\mathbf{p}}_{23}\Delta t_{12}\\&-\mathbf{R}_{\mathrm{WC}}^{1}\mathbf{R}_{\mathrm{CB}}^{*}\Delta\bar{\mathbf{v}}_{12}\Delta t_{12}\Delta t_{23}\\&-\frac{1}{2}\mathbf{R}_{\mathrm{WI}}\cdot{}_{\mathrm{E}}\mathbf{G}(\Delta t_{12}\Delta t_{23}^2+\Delta t_{12}^2\Delta t_{23})\end{aligned} \tag{29}$$

where $[\cdot]_{(:,1:2)}$ means the first two columns of the matrix.

A weight coefficient w_i for outlier rejection is introduced:

$$\begin{aligned}&\mathbf{e}_i=\hat{s}\cdot\lambda(i)+\phi(i)\delta\hat{\theta}_{xy}+\zeta(i)\delta\hat{\mathbf{b}}_{\mathrm{a}}+\xi(i)\,{}_{\mathrm{C}}\hat{\mathbf{p}}_{\mathrm{B}}-\psi(i)\\&w_i=exp(-\|\mathbf{e}_i\|\cdot K_r)\end{aligned} \tag{30}$$

where $\mathbf{e}_i$ and K_r are respectively the predicted residual and a gain factor to amplify the residual effect, and $\hat{s}$, $\delta\hat{\theta}_{xy}$, $\delta\hat{\mathbf{b}}_{\mathrm{a}}$ and ${}_{\mathrm{C}}\hat{\mathbf{p}}_{\mathrm{B}}$ are the last estimated results.

Similar to Eq.(24), a linear over-determined equation $\mathbf{D}_{3(N-2)\times 9}\cdot\mathbf{x}_{9\times 1}=\mathbf{E}_{3(N-2)\times 1}$ with the weights defined in (30) can be constructed to calculate the estimations s^*, $\delta\theta_{xy}^*$, $\delta\mathbf{b}_{\mathrm{a}}^*$ and ${}_{\mathrm{C}}\mathbf{p}_{\mathrm{B}}^*$. Since the accelerometer bias is set to zero $\mathbf{0}_{3\times 1}$ when integrating the $\Delta\bar{\mathbf{R}}_{i,i+1}$, $\Delta\bar{\mathbf{v}}_{i,i+1}$ and $\Delta\bar{\mathbf{p}}_{i,i+1}$, the final estimated accelerometer bias is $\mathbf{b}_{\mathrm{a}}^*=\mathbf{0}_{3\times 1}+\delta\mathbf{b}_{\mathrm{a}}^*=\delta\mathbf{b}_{\mathrm{a}}^*$. The gravity is refined by appending the perturbation, i.e. ${}_{\mathrm{W}}\mathbf{g}^*=\mathbf{R}_{\mathrm{WI}}\mathrm{Exp}(\delta\theta^*)\cdot{}_{\mathrm{E}}\mathbf{G}$.

D. Convergence Criteria

With the incoming new keyframe, the calibrated extrinsic orientation and translation will respectively converge to their stable values, which can be checked through calculating the standard deviation within a local period T_l. Specifically, we calculate the standard deviations σ_{oy}, σ_{op}, σ_{or} of the three axes (yaw, pitch, roll) of the historical calibrated orientations, and the σ_{tx}, σ_{ty}, σ_{tz} of the three axes of the historical calibrated translations within the local period. Besides, the quantity N_σ of the keyframes within the local period should be greater then a threshold th_N to exclude the case that a camera remains stable (e.g., new captured frames won't be determined as new keyframe if an aircraft hovers at one place). The extrinsic orientation calibration is convergent if $\sigma_{oy},\sigma_{op},\sigma_{or}<th_o$ and $N_\sigma<th_N$, and the extrinsic translation is convergent if $\sigma_{tx},\sigma_{ty},\sigma_{tz}<th_t$ and $N_\sigma<th_N$. Here th_o, th_t and th_N are three thresholds. If the calibrated orientation and translation are both convergent, the velocities of the historical keyframes could be computed by Eq.(9).

IV. Experiments And Discussions

In this section, performances of our online initialization method are evaluated on the EuRoC dataset [25] which provides accurate position ground-truth and camera-IMU extrinsic parameters. Eleven indoor sequences, ranging from slow flights under good visual conditions to dynamic flights

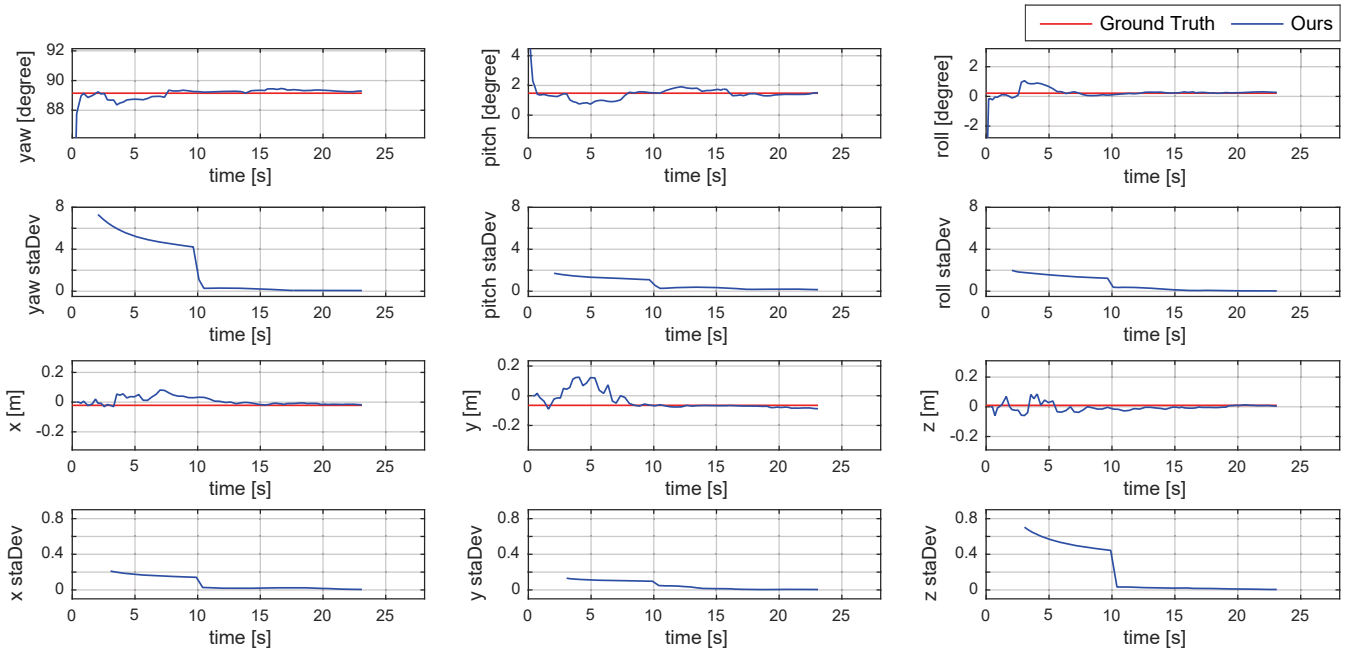

Fig. 2. Calibrated orientation, translation, as well as their corresponding standard deviations in *V2_01_easy*. The word "staDev" is the shorthand of "standard deviation". The ground-truth provided by the dataset is calibrated by *Kalibr*, with [89.147953 1.476930 0.215286] degrees in yaw, pitch, roll, and [-0.021640 -0.064677 0.009811] meters in x, y, z respectively.

with motion blur and poor illumination, were recorded with a Micro Aerial Vehicle (MAV). The two global-shutter, monochrome cameras and the IMU were hardware time-synchronized and were logged at a rate of 20 Hz and 200 Hz respectively. All the experiments are carried out with an Intel CPU i7-4720HQ (8 cores @2.60GHz) laptop computer with 8GB RAM.

A. Implementation Details

The proposed online initialization method is implemented in the *Local Mapping* thread of ORB_SLAM [20], specially, between the *Local BA* module and the *Local Keyframes Culling* module. Besides, the optimization theories introduced by Mur-Artal [19] is adopted to build our monocular visual-inertial SLAM system. The difference between Mur-Artal's work and ours is that our method automatically estimates the extrinsic parameters and the initial values, while the former assumes the extrinsic parameters are known. Some parameters mentioned above are set as follows: $K_o = 200.0$, $K_t = 100.0$, $K_r = 1.0$, $th_o = 0.1$, $th_t = 0.02$, $th_N = 10$, $T_l = 10$. In fact, these parameters can be set in a wide range and won't significantly impact the performance of our method.

B. Convergence Performance

In order to verify the convergence performance for the extrinsic calibration and the initial value estimation, the sequence *V2_01_easy* is expected to evaluate the proposed method. The time varied characteristic curves of the extrinsic parameters and the corresponding standard deviations are shown in Fig. 2. It can be seen that all the extrinsic parameters (yaw, pitch, roll for orientation and x, y, z for translation) are start to converge between 5 and 10 seconds. All the standard deviations satisfy the convergence criteria within 25 second. It is worth noting that the standard deviations of yaw, pitch and roll change dramatically at around 10 seconds. This phenomenon greatly dues to the quantity of keyframe is relatively small at the beginning, which results in the constraints of Eq.(19) is too weak to obtain a good solution. Because the standard deviation is calculated by the values within a local period T_l, the poor solutions in the first few seconds will have an effect on the standard deviation calculation within the local period. Fortunately, the extrinsic parameters tend to be stable with the increase of the keyframe quantity.

Fig. 3 shows the estimated results of the scale factor, gyroscope bias, accelerometer bias and the gravity. It can be seen that all the estimated values are almost convergent to stable values between 10 and 15 seconds. In particular, the yaw, pitch, roll curves shown in Fig. 2 and the gyroscope bias curves shown in Fig. 3 illustrate that the camera-IMU orientation calibration and the gyroscope bias estimation can achieve stable values in a very short time, which confirms that the iterative strategy described in Section III-A achieves good performance. It is worth noting that the curves of accelerometer bias suffer severe oscillation in the first ten seconds. It might be because the accelerometer bias and the gravity are hard to distinguished, therefore more keyframes are required for calculating stable values. Fig. 3 also shows that the processing time of our method is approximately linear to the quantity of keyframes.

C. Consistency Analysis

To analyze the consistency of the proposed initialization method, our method is compared with the state-of-the-art VINS-Mono [18]. In VINS-Mono, the extrinsic parameter-

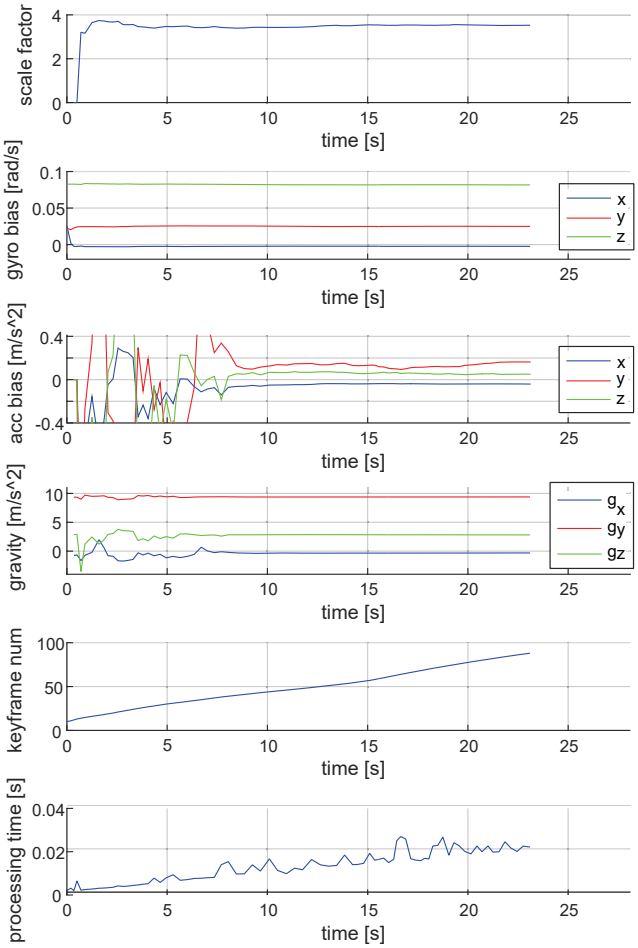

Fig. 3. IMU initialization results and processing time in *V2_01_easy*. The "acc" and "gyro" are the shorthands of "accelerometer" and "gyroscope" respectively.

s, velocity and gravity are roughly estimated by a linear initialization process, and then these values are refined by a nonlinear optimization process which takes accelerometer and gyroscope biases into account. Ten trials in the sequence *V2_01_easy* for our method, linear and nonlinear precesses of VINS-Mono are conducted in Fig. 4. It is observed that the proposed method achieves more accurate than the two processes of VINS-Mono on the extrinsic orientation calibration. As for translation calibration, it seems that the consistency of our method is slightly lower than the nonlinear results of VINS-Mono. This is probably due to the fact that there is no criteria to identify the convergence of these variables or to terminate the nonlinear optimization process in VINS-Mono, therefore the extrinsic parameters are persistently refined until the end of the sequence. In this way, all the measurements contained in the sequence are used to estimate parameters by VINS-Mono, which results in more consistent calibrated translations than ours. However, it spends more time to calculate accurate extrinsic parameters and there is no guarantee that the intermediate calibration is precise enough to obtain good localization performance. On the contrary, our method provides reasonable and comparable calibration results in a few seconds, which are accurate enough to support the further localization and mapping work.

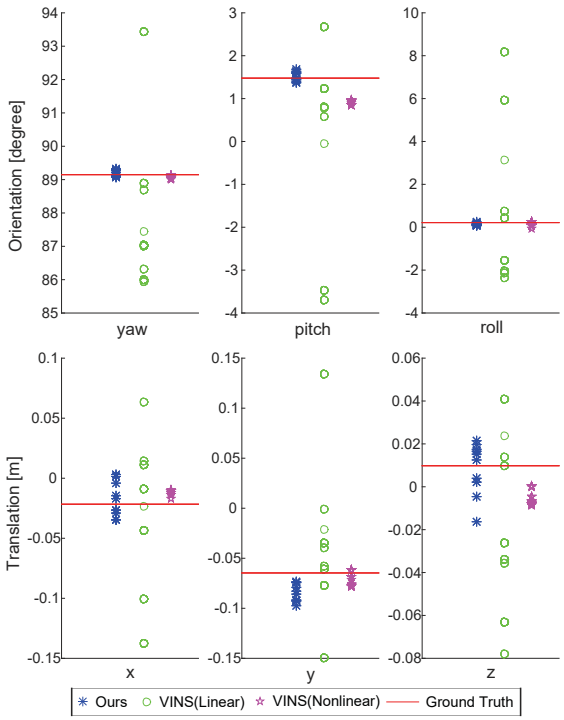

Fig. 4. Consistency analysis of different methods in *V2_01_easy*.

D. Motion Estimation Performance

Table I shows the errors of the calibrated orientation and translation, and the absolute translational root-mean square error (RMSE) of the keyframe trajectory for each sequence. The listed results of VI ORB-SLAM is obtained from [19]. The VINS-Mono [18] is performed without knowing the prior extrinsic parameters. All the results of our method and VINS-Mono are the median over five executions in each sequence. The ideal scale factor $\hat{s}$ is measured by aligning the estimated keyframe trajectory to the best fit with the ground-truth. The scale error can be calculated as $|s^* - \hat{s}|/|\hat{s}| \times 100\%$.

It can be seen that in spite of the extrinsic parameters are unknown in our method, the RMSE error of our method is not much worse than VI ORB-SLAM. Conversely, our method exhibits more accurate performance on the *MH_01_easy*, *MH_02_easy* and *MH_04_difficult* sequences. Comparing with VINS-Mono, we can conclude that the most of the trajectories estimated by our method are more accurate than VINS-Mono, especially on the machine hall sequences. This demonstrates that the calibration results provided by our method are more reasonable than VINS-Mono to support the motion estimation, although the extrinsic parameters calibrated by our method are less consistent than the nonlinear process of VINS-Mono. The results on the *V1_03_difficult* sequence show that VINS-Mono is more robust against the motion blur and illumination change, while our method and VI ORB-SLAM both fail tracking. However, the extrinsic parameters can still be accurately estimated by our method on this sequence. The emphasized data in the Table I also shows that our method is able to calibrate accurate extrinsic parameters, achieving a typical precision of 0.6 degree for orientation and 0.05 meter for translation.

TABLE I

EXTRINSIC CALIBRATION ERROR AND KEYFRAME TRAJECTORY ACCURACY IN EUROC DATASET.

	Ours (No Full BA)									VINS-Mono	VI ORB-SLAM
	Orientation Error (degree)			Translation Error (m)			RMSE	Scale	RMSE(m)	RMSE	RMSE
	e_yaw	e_pitch	e_roll	e_x	e_y	e_z	(m)	Error(%)	GT scale[1]	(m)	(m)
V1_01_easy	-0.122	*-0.484*	0.108	-0.003	-0.014	0.010	**0.056**	1.1	0.049	0.069	0.027
V1_02_medium	-0.275	-0.039	0.161	-0.006	0.007	0.018	0.044	1.1	0.042	**0.007**	0.028
V1_03_difficult	-0.053	-0.147	0.151	-0.015	-0.025	*0.047*	—[2]	—	—	**0.158**	—
V2_01_easy	-0.148	-0.030	-0.056	-0.004	0.023	0.005	**0.048**	1.9	0.019	0.058	0.032
V2_02_medium	*-0.521*	-0.160	0.128	-0.039	-0.003	0.011	0.071	2.1	0.053	**0.062**	0.041
V2_03_difficult	0.043	0.193	0.009	*-0.043*	0.005	0.017	**0.100**	2.1	0.095	0.253	0.074
MH_01_easy	-0.122	0.145	0.231	-0.014	0.009	0.010	**0.050**	0.9	0.045	0.135	0.075
MH_02_easy	0.090	0.427	0.293	-0.010	0.014	0.024	**0.031**	0.0	0.031	0.112	0.084
MH_03_medium	0.116	0.233	0.277	-0.016	0.012	0.038	**0.093**	1.1	0.073	0.125	0.087
MH_04_difficult	0.084	0.074	*0.373*	-0.007	-0.023	0.035	**0.081**	1.1	0.072	0.169	0.217
MH_05_difficult	-0.071	0.366	0.130	-0.014	*-0.035*	0.046	**0.133**	0.0	0.133	0.260	0.082

[1] GT scale: the estimated keyframe trajectory is scaled to the best fit with the ground-truth trajectory.

[2] "—" means that the tracking is lost at some point and a significant portion of the sequence is not processed by the system. The max absolute errors for each axis of orientation and translation are emphasized with underlines.

V. CONCLUSIONS

In this paper, we propose a novel online initialization method for monocular visual-inertial SLAM without knowing the mechanical configuration of the sensor suite. Specifically, our method automatically estimates the visual scale, velocity, gravity, biases of gyroscope and accelerometer, and calibrates the camera-IMU extrinsic parameters while the system is performing free motion in environments. To simplify the initialization problem, three simple processes are included in our method. Our method is able to automatically identify the convergence of the calibration parameters so that the initialization stage can be terminated. Experiments demonstrate that our method achieves competitive accuracy and consistency compared with the state-of-the-art methods. A limitation of our method is the assumption that the camera and IMU measurements are hardware time-synchronized. To overcome this, we plan to correct the time offset in our future work.

REFERENCES

[1] A. Stelzer, H. Hirschmüller, and M. Görner, "Stereo-vision-based navigation of a six-legged walking robot in unknown rough terrain," *Int. J. Robot. Res.*, vol. 31, no. 4, pp. 381–402, 2012.

[2] H. Liu, Z. Wang, and P. Chen, "Feature points selection with flocks of features constraint for visual simultaneous localization and mapping," *Int. J. Adv. Robot. Syst.*, vol. 14, no. 1, pp. 1–11, 2016.

[3] S. Weiss, D. Scaramuzza, and R. Siegwart, "Monocular-SLAM–based navigation for autonomous micro helicopters in GPS-denied environments," *J. Field Robot.*, vol. 28, no. 6, pp. 854–874, 2011.

[4] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys, "Live metric 3d reconstruction on mobile phones," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 65–72.

[5] T. Oskiper, S. Samarasekera, and R. Kumar, "Multi-sensor navigation algorithm using monocular camera, IMU and GPS for large scale augmented reality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2012, pp. 71–80.

[6] M. Li and A. I. Mourikis, "Improving the accuracy of EKF-based visual-inertial odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 828–835.

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

[8] 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.

[9] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 1885–1892.

[10] 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.

[11] J. Rehder and R. Siegwart, "Camera/IMU calibration revisited," *IEEE Sensors J.*, vol. 17, no. 11, pp. 3257–3268, 2017.

[12] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2013, pp. 1280–1286.

[13] P. Furgale, T. D. Barfoot, and G. Sibley, "Continuous-time batch estimation using temporal basis functions," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 2088–2095.

[14] J. Kelly and G. S. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *Int. J. Robot. Res.*, vol. 30, no. 1, pp. 56–79, 2011.

[15] A. Martinelli, "Closed-form solution of visual-inertial structure from motion," *Int. J. Comput. Vision*, vol. 106, no. 2, pp. 138–152, 2014.

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

[17] 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, 2017.

[18] Y. Lin, F. Gao, T. Qin, W. Gao, T. Liu, W. Wu, Z. Yang, and S. Shen, "Autonomous aerial navigation using monocular visual-inertial fusion," *J. Field Robot.*, 2017.

[19] 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, 2017.

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

[21] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, 2017.

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

[23] 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, 2017.

[24] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3607–3613.

[25] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.