Monocular Visual–Inertial State Estimation With Online Initialization and Camera–IMU Extrinsic Calibration

Zhenfei Yang, *Student Member, IEEE*, and Shaojie Shen, *Member, IEEE*

***Abstract*— There have been increasing demands for developing microaerial vehicles with vision-based autonomy for search and rescue missions in complex environments. In particular, the monocular visual–inertial system (VINS), which consists of only an inertial measurement unit (IMU) and a camera, forms a great lightweight sensor suite due to its low weight and small footprint. In this paper, we address two challenges for rapid deployment of monocular VINS: 1) the initialization problem and 2) the calibration problem. We propose a methodology that is able to initialize velocity, gravity, visual scale, and camera–IMU extrinsic calibration on the fly. Our approach operates in natural environments and does not use any artificial markers. It also does not require any prior knowledge about the mechanical configuration of the system. It is a significant step toward plug-and-play and highly customizable visual navigation for mobile robots. We show through online experiments that our method leads to accurate calibration of camera–IMU transformation, with errors less than 0.02 m in translation and 1° in rotation. We compare out method with a state-of-the-art marker-based offline calibration method and show superior results. We also demonstrate the performance of the proposed approach in large-scale indoor and outdoor experiments.**

***Note to Practitioners*—This paper presents a methodology for online state estimation in natural environments using only a camera and a low-cost micro-electro-mechanical systems (MEMS) IMU. It focuses on addressing the problems of online estimator initialization, sensor extrinsic calibration, and nonlinear optimization with online refinement of calibration parameters. This paper is particularly useful for applications that have superior size, weight, and power constraints. It aims for rapid deployment of robot platforms with robust state estimation capabilities with almost no setup, calibration, or initialization overhead. The proposed method can be used in platforms including handheld devices, aerial robots, and other small-scale mobile platforms, with applications in monitoring, inspection, and search and rescue.**

***Index Terms*— Calibration, estimator initialization, sensor fusion, state estimation, visual navigation, visual–inertial systems (VINSs).**

Manuscript received January 3, 2016; accepted March 10, 2016. This paper was recommended for publication by Associate Editor A. M. Hsieh and Editor Y. Sun upon evaluation of the reviewer's comments. This work was supported in part by the Hong Kong University of Science and Technology under Grant R9341. This paper is a revised and extended version of "Monocular Visual-Inertial Fusion with Online Initialization and Camera-IMU Calibration" presented at the IEEE International Symposium on Safety, Security, and Rescue Robotics, West Lafayette, IN, USA, October 2015.

The authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: zyangag@connect.ust.hk; eeshaojie@ust.hk).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. The Supplementary Material contains the following. Three experiments are presented in the video to demonstrate the performance of our self-calibrating monocular visual-inertial state estimation method. The first experiment details the camera-IMU extrinsic calibration process in a small indoor experiment. The second experiment evaluates the performance of the overall system in a large-scale indoor environment with highlights to the online calibration process. The third experiment presents the state estimation results in a large-scale outdoor environment using different camera configurations. This material is 52.6 MB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2016.2550621

I. Introduction

THERE have been increasing demands for developing high maneuverability robots, such as microaerial vehicles (MAVs) with vision-based autonomy for search and rescue missions in confined environments. Such robots and sensor suites should be miniaturized, rapidly deployable, and require minimum maintenance even in hazardous environments. The monocular visual–inertial system (VINS), which consists of only a low-cost MEMS inertial measurement unit (IMU) and a camera, has become a very attractive sensor choice due to its superior size, weight, and power characteristics. In fact, the monocular VINS is the minimum sensor suite that allows both accurate state estimation and sufficient environment awareness.

However, the algorithmic challenges for processing information from monocular VINS are significantly more involved than stereo- [2] or RGB-D-based [3] configurations due to the lack of direct observation of visual scale. The performance of state-of-the-art nonlinear monocular VINS estimators [4]–[7] relies heavily on the accuracy of initial values (velocity, attitude, and visual scale) and the calibration of camera–IMU transformation. In time-critical search and rescue missions, careful initialization (launch) of the robot platform or explicit calibration by professional users is often infeasible. In fact, it is desirable to simply plug sensors onto the MAV, throw it into the air, and have everything operational. This implies that the initialization and the camera–IMU extrinsic calibration procedure should be performed with no prior knowledge about either the dynamical motion or mechanical configuration of the system.

Our earlier works [7]–[9] focused on initialization and tightly coupled fusion of monocular VINSs, but with the assumption of known camera–IMU transformation. In this paper, we relax this assumption and propose a method for joint initialization and extrinsic calibration without any prior knowledge about the mechanical configuration of the system. This paper is a revised version of [10]. Several signifi-

Fig. 1. Our monocular VINS setup with unknown camera–IMU extrinsic calibration. Colored coordinate axes are plotted (red: x-axis, green: y-axis, and blue: z-axis). Note that there are multiple 90° offsets between the camera frame and the IMU frame. The rotation offsets are unknown to the estimator and have to be calibrated online.

cant extensions have been made. First, we provide a more thorough discussion of the full monocular VINS pipeline, with additional details on the formulation of IMU preintegration (Section IV) and nonlinear VINS with calibration refinement (Section VI). Second, we compare our approach against Kalibr, a state-of-the-art marker-based offline calibration method [1] (Section VIII-B) and show superior results. We show through online experiments that our method leads to accurate calibration of camera–IMU transformation with errors of 0.02 m in translation and 1° in rotation. Finally, we present large-scale outdoor experiments performed onboard a quadrotor aerial vehicle with different sensor placements and analyze the results (Section VIII-D). Our approach is a substantial step toward minimum sensing for plug-and-play robotic systems. We identify the contributions of this paper as fourfold.

- A probabilistic optimization-based initialization procedure that is able to recover all the essential navigation quantities (initial velocity and attitude, visual scale, and camera–IMU calibration) without any prior system knowledge or artificial calibration objects.
- Principal criteria to identify convergence and termination for the initialization procedure.
- A tightly coupled monocular visual–inertial fusion methodology for accurate state estimation with online calibration refinement.
- Online experiments in complex indoor and outdoor environments.

The rest of this paper is organized as follows. In Section II, we discuss the relevant literature. We give an overview of the complete system pipeline in Section III. In Section IV, we present IMU preintegration, a core technique for summarizing and utilizing high-rate IMU measurements without knowing the initial attitude or velocity. We detail our linear initialization and camera–IMU calibration procedure in Section V. A tightly coupled, self-calibrating, nonlinear optimization-based monocular VINS estimator, which is built on top of [7] and [8], is presented in Section VI. We present a two-way marginalization scheme for handling of degenerate motions in Section VII and discuss implementation details and present experimental results in Section VIII. Finally, this paper is concluded with a discussion of possible future research in Section IX.

II. Related Work

There is a rich body of scholarly work on VINS state estimation with either monocular [4], [5], [11], stereo [6], or RGB-D cameras [3]. We can categorize solutions to VINS as filtering based [4], [5], [11]–[14] or graph optimization/bundle adjustment based [6], [7], [15]. Filtering-based approaches may require fewer computational resources due to the marginalization of past states, but the early fix of linearization points may lead to suboptimal results. On the other hand, graph optimization-based approaches may improve performance via iterative relinearization at the expense of higher computational demands. In real-world applications, marginalization [16] is usually employed for both filtering- and optimization-based approaches to achieve constant computational complexity. Conditioning is also a popular method within the computation vision community to enforce constant computation [17]. A comparison between filtering- and graph optimization-based approaches [18] demonstrates nearly identical results. However, the platform for verification is equipped only with an optical flow sensor that is insufficient for extended feature tracking. This limits the power of graph-based approaches, as a well-connected graph is never constructed.

Another way to categorize VINS solutions is to consider them as loosely coupled [11], [19] or tightly coupled [4]–[6], [12]–[15], [20]. Loosely coupled approaches usually consist of a standalone vision-only state estimation module such as PTAM [17] or SVO [21], and a separate IMU fusion module [19]. These approaches do not consider the visual and inertial information coupling, making them incapable of correcting drifts in the vision-only estimator. Tightly coupled approaches perform systematic fusion of visual and IMU measurements and usually lead to better results [6]. In particular, for the monocular VINS, tightly coupled methods are able to implicitly incorporate the metric scale information from IMU into scene depth estimation, thus removing the need for explicit visual scale modeling.

However, all the aforementioned VINS solutions rely on accurate initialization of system motions and accurate camera–IMU calibration. This is particularly critical for monocular VINS due to the lack of direct observation of visual scale. There is a wide body of work trying to deal with the velocity, attitude, and visual scale initialization problems for monocular VINS. Pioneering work in this area was proposed by Lupton and Sukkarieh [22], who performed the estimation in the body frame of the first pose in the sliding window. An IMU preintegration technique is proposed to handle multirate sensor measurements. They showed that the nonlinearity of the system mainly arises only from rotation drift. Recent results suggest that by assuming the orientation is known, VINS may be solved in a linear closed form [23]–[27]. It has been shown that both the initial gravity vector and the body frame velocity can be estimated linearly. These results have the significant implication that a good initialization of the VINS problem may actually be unnecessary. In particular, [25] and [26] analytically show conditions from which initial values are solvable. However, [23] is limited to using a fixed small number of IMU measurements, which makes it very sensitive to IMU noise. Approaches that utilize multiple IMU

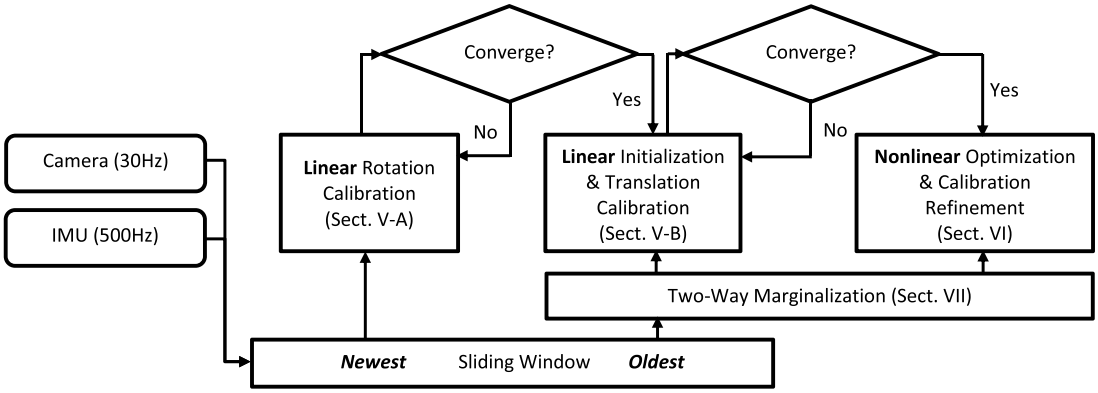

Fig. 2. Block diagram illustrating the full pipeline of the proposed approach. Each module is discussed in the corresponding sections marked in the diagram.

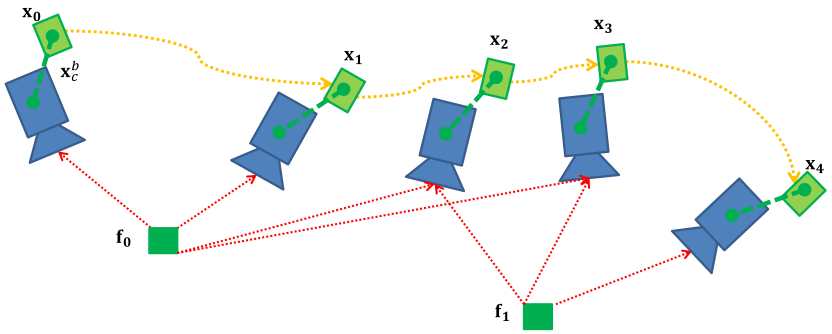

Fig. 3. Example of sliding window with five IMU states $\mathbf{x}_k$ and two features $\mathbf{f}_l$. Dotted lines represent the pre-integrated IMU measurements and visual measurements. Note that there is a constant but unknown camera–IMU extrinsic calibration $\mathbf{x}_c^b$. All aforementioned quantities are jointly estimated in our framework. This sliding window model applies to both linear initialization (Section V) and nonlinear optimization (Section VI).

measurements in a sliding window [24]–[27] do not scale well to a large number of IMU measurements since they rely on double integration of accelerometer output over an extended period of time. Moreover, these closed-form approaches do not take the noise characteristic of the system into account, producing suboptimal results.

For camera–IMU calibration, Li and Mourikis [5], Weiss [19], and Heng [28] consider incorporating the camera–IMU transformation into the state vector for the nonlinear estimator. However, the convergence of the calibration parameters still depends on the accuracy of the initial values, and the calibration performance is not systematically analyzed in any of these papers. Our work is related to [24], as both aim to jointly initialize the motion of the system as well as the camera–IMU calibration. However, [24] is a geometric method without consideration of sensor noise. In particular, the formulation of IMU measurements in [24] results in unbounded IMU error over time, which leads to downgraded performance as more IMU measurements are incorporated. On the other hand, our probabilistic formulation explicitly bounds the error for each measurement using a sliding window approach and is thus able to fuse an extensive number of sensor measurements until good initial values are obtained. Also, [24] only shows results with simulated data, while we present extensive experimental results with real sensor data.

III. Overview

Our proposed monocular VINS estimator consists of three phases, as illustrated in Fig. 2. The first phase initializes the rotation between the camera and the IMU in a linear fashion (Section V-A). The second phase handles on-the-fly initialization of velocity, attitude, visual scale, and camera–IMU translation with a probabilistic linear sliding window approach (Section V-B), as illustrated in Fig. 3. This phase is an extension of [8] and [9] by relaxing the known camera–IMU calibration assumption. Finally, the third phase that focuses on high-accuracy nonlinear graph optimization (Fig. 3) for both state estimation and calibration refinement will be detailed in Section VI. Note that the three phases run sequentially and continuously with automatic switching. This suggests that all the user needs to do is to move the monocular VINS sensor suite freely with sufficient motion in natural environments. Our estimator is able to automatically identify convergence and switch to the next phase (Sections V-A2 and V-B5). Both linear and nonlinear estimators utilize a two-way marginalization scheme (Section VII), which was first proposed in [9], for handling of degenerate motion.

We begin by defining notations. We consider $(\cdot)^w$ as the earth's inertial frame, $(\cdot)^b$ as the current IMU body frame, and $(\cdot)^{c_k}$ as the camera body frame while taking the kth image. We further note $(\cdot)^{b_k}$ as the IMU body frame while the camera is taking the kth image. Note that IMU usually runs at a higher rate than the camera, and that multiple IMU measurements may exist in the interval $[k, k+1]$. $\mathbf{p}_Y^X$, $\mathbf{v}_Y^X$, and $\mathbf{R}_Y^X$ are the 3-D position, velocity, and rotation of frame Y with respect to frame X. Specially, $\mathbf{p}_t^X$ represents the position of the IMU body frame at time t with respect to frame X. A similar convention is followed for other parameters. The

camera–IMU transformation is an unknown constant that we denote as $\mathbf{p}_c^b$ and $\mathbf{R}_c^b$. Besides rotation matrices, we also use quaternions ($\mathbf{q} = [q_x, q_y, q_z, q_w]$) to represent rotation. The Hamilton notation is used for quaternions. $\mathbf{g}^w = [0, 0, g]^T$ is the gravity vector in the world frame, and $\mathbf{g}^{b_k}$ is the earth's gravity vector expressed in the IMU body frame during the kth image capture. In addition, we use $\hat{(\cdot)}$ for noisy sensor measurements and preintegrated IMU measurements.

IV. IMU Preintegration

The IMU preintegration technique was first proposed in [22] and is advanced to consider on-manifold uncertainty propagation in [7] and [8]. Further improvements to incorporate IMU biases and integrate with full SLAM framework are proposed in [20]. Here, we give an overview of its motivation and usage within our monocular VINS framework.

Given two time instants that correspond to two images, we can write the IMU propagation model for position and velocity in the earth's inertial frame as follows:

$$\begin{aligned}\mathbf{p}_{b_{k+1}}^w &= \mathbf{p}_{b_k}^w + \mathbf{v}_{b_k}^w \Delta t + \iint_{t\in[k,k+1]} \left(\mathbf{R}_t^w \mathbf{a}_t^b - \mathbf{g}^w\right) dt^2 \\ \mathbf{v}_{b_{k+1}}^w &= \mathbf{v}_{b_k}^w + \int_{t\in[k,k+1]} \left(\mathbf{R}_t^w \mathbf{a}_t^b - \mathbf{g}^w\right) dt \end{aligned} \tag{1}$$

where $\mathbf{a}_t^b$ is the instantaneous linear acceleration in the IMU body frame and Δt is the time difference $[k, k+1]$ between two image captures. It can be seen that the rotation between the world frame and the body frame is required for IMU state propagation. This global rotation can be determined only with known initial attitude, which is hard to obtain in many applications. However, as introduced in [7], if the reference frame of the whole system is changed to the frame of the first image capture b_0, and the frame for IMU propagation is changed to b_k, we can preintegrate the parts in (1) that are related to linear acceleration $\mathbf{a}$ and angular velocity $\boldsymbol{\omega}$ as follows:

$$\begin{aligned}\boldsymbol{\alpha}_{b_{k+1}}^{b_k} &= \iint_{t\in[k,k+1]} \mathbf{R}_t^{b_k} \mathbf{a}_t^b dt^2 \\ \boldsymbol{\beta}_{b_{k+1}}^{b_k} &= \int_{t\in[k,k+1]} \mathbf{R}_t^{b_k} \mathbf{a}_t^b dt \\ \mathbf{R}_{b_{k+1}}^{b_k} &= \int_{t\in[k,k+1]} \mathbf{R}_t^{b_k} \lfloor \boldsymbol{\omega}_t^b \times \rfloor dt \end{aligned} \tag{2}$$

where $\lfloor \boldsymbol{\omega}_t^b \times \rfloor$ is the skew-symmetric matrix from $\boldsymbol{\omega}_t^b$. $\mathbf{R}_t^{b_k}$ is the incremental rotation from b_k to the current time t, which is available through short-term integration of gyroscope measurements. Therefore, (1) can be rewritten as

$$\begin{aligned}\mathbf{p}_{b_{k+1}}^{b_0} &= \mathbf{p}_{b_k}^{b_0} + \mathbf{R}_{b_k}^{b_0} \mathbf{v}_{b_k}^{b_k} \Delta t - \mathbf{g}^{b_0} \Delta t^2/2 + \mathbf{R}_{b_k}^{b_0} \boldsymbol{\alpha}_{b_{k+1}}^{b_k} \\ \mathbf{v}_{b_{k+1}}^{b_{k+1}} &= \mathbf{R}_{b_k}^{b_{k+1}} \mathbf{v}_{b_k}^{b_k} - \mathbf{g}^{b_0} \Delta t + \mathbf{R}_{b_k}^{b_{k+1}} \boldsymbol{\beta}_{b_{k+1}}^{b_k} \\ \mathbf{R}_{b_{k+1}}^{b_0} &= \mathbf{R}_{b_k}^{b_0} \mathbf{R}_{b_{k+1}}^{b_k}. \end{aligned} \tag{3}$$

It can be seen that the preintegration parts (2) can be obtained solely with IMU measurements within $[k, k+1]$. With this formulation, the dependency on global orientation is removed. Therefore, using the camera–IMU rotation calibration obtained in Section V-A, we are able to formulate the joint problem of monocular VINS initialization and camera–IMU translation calibration as a linear problem that can be solved without any prior knowledge of the system (Section V-B). This preintegration technique also enables summarizing multiple IMU measurements into a standalone measurement term. This new measurement term can be relinearized during the iterative nonlinear optimization (Section VI) to achieve better accuracy. This is in contrast to the existing graph-based VINS approach [6] that performs IMU integration in the global frame. Detailed treatments of uncertainty propagation during IMU preintegration for both linear (Section V-B) and nonlinear (Section VI) estimators are presented in the corresponding sections.

V. Estimator Initialization and Camera–IMU Extrinsic Calibration

We now detail our online estimator initialization approach to recover all critical states, including velocity, attitude (gravity vector), depth of features, and camera–IMU extrinsic calibration. Our initialization procedure does not require any prior knowledge about the mechanical configuration of the sensor suite. It also does not require the estimator to be started from stationary, making it particularly useful for dynamically launching aerial robots in a search and rescue setting. The initialization and calibration process can be formulated as solving two sets of linear systems, which will be discussed in Sections V-A and V-B, respectively.

A. Linear Initialization of Camera–IMU Rotation

The constant camera–IMU rotation offset can be obtained by aligning two rotation sequences from the IMU and the camera.

1) Linear Rotation Calibration: We assume that sufficient features can be tracked, and the incremental rotation between two images $\mathbf{R}_{c_{k+1}}^{c_k}$ can be estimated using the classic five-point algorithm [29] with RANdom Sample Consensus (RANSAC)-based outlier rejection. We further denote the corresponding rotation obtained by integrating gyroscope measurements represented in the IMU frame as $\mathbf{R}_{b_{k+1}}^{b_k}$. The following equation holds for any k:

$$\mathbf{R}_{b_{k+1}}^{b_k} \cdot \mathbf{R}_c^b = \mathbf{R}_c^b \cdot \mathbf{R}_{c_{k+1}}^{c_k}. \tag{4}$$

With a quaternion representation for rotation, we can write (4) as

$$\begin{aligned}&\mathbf{q}_{b_{k+1}}^{b_k} \otimes \mathbf{q}_c^b = \mathbf{q}_c^b \otimes \mathbf{q}_{c_{k+1}}^{c_k} \\ &\quad \Rightarrow \left[\mathcal{Q}_1\left(\mathbf{q}_{b_{k+1}}^{b_k}\right) - \mathcal{Q}_2\left(\mathbf{q}_{c_{k+1}}^{c_k}\right)\right] \cdot \mathbf{q}_c^b = \mathbf{Q}_{k+1}^k \cdot \mathbf{q}_c^b = \mathbf{0} \end{aligned} \tag{5}$$

where

$$\begin{aligned}\mathcal{Q}_1(\mathbf{q}) &= \begin{bmatrix} q_w \mathbf{I}_3 + \lfloor \mathbf{q}_{xyz} \times \rfloor & \mathbf{q}_{xyz} \\ -\mathbf{q}_{xyz} & q_w \end{bmatrix} \\ \mathcal{Q}_2(\mathbf{q}) &= \begin{bmatrix} q_w \mathbf{I}_3 - \lfloor \mathbf{q}_{xyz} \times \rfloor & \mathbf{q}_{xyz} \\ -\mathbf{q}_{xyz} & q_w \end{bmatrix} \end{aligned} \tag{6}$$

are matrix representations for left and right quaternion multiplication, $\lfloor \mathbf{q}_{xyz} \times \rfloor$ is the skew-symmetric matrix from the first three elements $\mathbf{q}_{xyz}$ of a quaternion, and $\otimes$ is the quaternion multiplication operator.

With multiple incremental rotations between pairs of consecutive images, we are able to construct the overconstrained linear system as

$$\begin{bmatrix} w_1^0 \cdot \mathbf{Q}_1^0 \\ w_2^1 \cdot \mathbf{Q}_2^1 \\ \vdots \\ w_N^{N-1} \cdot \mathbf{Q}_N^{N-1} \end{bmatrix} \cdot \mathbf{q}_c^b = \mathbf{Q}_N \cdot \mathbf{q}_c^b = \mathbf{0} \tag{7}$$

where N is the index of the latest frame that keeps growing until the rotation calibration is completed. Note that the incremental rotation measurements obtained from the five-point algorithm contain outliers due to wrong correspondences or numerical errors under degenerated motions. We use weight w_{k+1}^k derived from the robust norm for better outlier handling. As the rotation calibration runs with incoming measurements, we are able to use the previously estimated camera–IMU rotation $\hat{\mathbf{R}}_c^b$ as the initial value to weight the residual in a similar fashion to the Huber norm [30]. Specifically, the residual is defined as the angle norm in the angle-axis representation of the residual rotation matrix

$$r_{k+1}^k = \operatorname{acos}\left(\left(\operatorname{tr}\left(\hat{\mathbf{R}}_c^{b^{-1}} \mathbf{R}_{b_{k+1}}^{b_k{}^{-1}} \hat{\mathbf{R}}_c^b \mathbf{R}_{c_{k+1}}^{c_k}\right) - 1\right)/2\right). \tag{8}$$

The weight is a function of the residual

$$w_{k+1}^k = \begin{cases} 1, & r_{k+1}^k < \text{threshold} \\ \dfrac{\text{threshold}}{r_{k+1}^k}, & \text{otherwise.} \end{cases} \tag{9}$$

If there are no sufficient features for estimating the camera rotation, w_{k+1}^k is set to zero. The solution to the above linear system can be found as the right unit singular vector corresponding to the smallest singular value of $\mathbf{Q}_N$.

2) Termination Criteria: Successful calibration of the camera–IMU rotation $\mathbf{R}_c^b$ relies on sufficient rotation excitation. Under sufficient rotation, the null space of $\mathbf{Q}_N$ should be rank one. However, under degenerate motions in one or more axes, the null space of $\mathbf{Q}_N$ may be larger than one. Therefore, by checking whether the second smallest singular value of $\mathbf{Q}_N$, $\sigma_{\mathbf{R}}^{\min 2}$, is large enough, we have a good indicator of whether sufficient rotation excitation is achieved. We set a singular value threshold $\sigma_{\mathbf{R}}$. The camera–IMU rotation calibration process terminates if $\sigma_{\mathbf{R}}^{\min 2} > \sigma_{\mathbf{R}}$. A convergence plot can be found in Fig. 6.

B. Linear Initialization of Velocity, Attitude, Feature Depth, and Camera–IMU Translation

Once the camera–IMU rotation is fixed, we can estimate the camera–IMU translation together with an initialization of velocity, attitude, and feature depth, as well as the IMU poses with respect to the initial reference frame, as in (3).

1) Linear Sliding Window Estimator: We use a tightly coupled sliding window formulation (Fig. 3) for incorporating a large number of IMU and camera measurements with constant computational complexity. The initialization is done in the IMU frame, with the full state vector defined as (the transpose is ignored for simplicity of presentation)

$$\begin{aligned} \mathcal{X} &= \left[\mathbf{x}_n, \mathbf{x}_{n+1}, \ldots \mathbf{x}_{n+N}, \mathbf{p}_c^b, \lambda_m, \lambda_{m+1}, \ldots \lambda_{m+M}\right] \\ \mathbf{x}_k &= \left[\mathbf{p}_{b_k}^{b_0}, \mathbf{v}_{b_k}^{b_k}, \mathbf{g}^{b_k}\right] \end{aligned} \tag{10}$$

where $\mathbf{x}_k$ is the kth IMU state, the gravity vector $\mathbf{g}^{b_k}$ determines the attitude (roll and pitch angles), N is the number of IMU states in the sliding window, M is the number of features that have sufficient parallax within the sliding window, n and m are starting indexes of the sliding window, and λ_l is the depth of the lth feature from its first observation. $\mathbf{p}_{b_0}^{b_0} = [0, 0, 0]$ is preset. Note that we reuse the sensor measurements that we used for the camera–IMU rotation calibration in this linear initialization phase, but with $\mathbf{R}_c^b$ fixed as a constant. We also directly use the incremental ($\mathbf{R}_{b_{k+1}}^{b_k}$) and relative ($\mathbf{R}_{b_{k+1}}^{b_0}$) rotations obtained from short-term integration of gyroscope measurements. As this linear initialization can usually be done in only a few seconds, using the IMU rotation directly will not cause significant drifts.

The linear initialization is done with maximum likelihood estimation by minimizing the sum of the Mahalanobis norm of all measurement errors from the IMU and the monocular camera within the sliding window

$$\min_{\mathcal{X}} \left\{ \left\| \mathbf{r}_p - \mathbf{H}_p \mathcal{X} \right\|^2 + \sum_{k \in \mathcal{B}} \left\| \hat{\mathbf{z}}_{b_{k+1}}^{b_k} - \mathbf{H}_{b_{k+1}}^{b_k} \mathcal{X} \right\|_{\mathbf{P}_{b_{k+1}}^{b_k}}^2 + \sum_{(l,j) \in \mathcal{C}} \left\| \hat{\mathbf{z}}_l^{c_j} - \mathbf{H}_l^{c_j} \mathcal{X} \right\|_{\lambda \mathbf{P}_l^{c_j}}^2 \right\} \tag{11}$$

where $\mathcal{B}$ is the set of all IMU measurements, $\mathcal{C}$ is the set of all observations between any features and any camera poses, and $\mathbf{H}_{b_{k+1}}^{b_k}, \mathbf{H}_l^{c_j}$ are corresponding measurement matrices. $\{\mathbf{r}_p, \mathbf{H}_p\}$ is the (optional) prior, which will be discussed in Section V-B4. Since incremental and relative rotations are known, (11) can be solved in a noniterative linear fashion.

2) Linear IMU Measurement Model: The linear IMU measurement $\{\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \mathbf{H}_{b_{k+1}}^{b_k}\}$ between consecutive frames k and $k+1$ is derived from (3), with the exception that all rotations ($\mathbf{R}_{b_k}^{b_0}$ and $\mathbf{R}_{b_{k+1}}^{b_k}$) are considered as known. We also introduce the propagation of body frame gravity vectors

$$\begin{aligned} \hat{\mathbf{z}}_{k+1}^k &= \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k} \\ \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} \\ \hat{\mathbf{0}} \end{bmatrix} = \mathbf{H}_{b_{k+1}}^{b_k} \mathcal{X} + \mathbf{n}_{b_{k+1}}^{b_k} \\ &= \begin{bmatrix} \mathbf{R}_{b_0}^{b_k} \left(\mathbf{p}_{b_{k+1}}^{b_0} - \mathbf{p}_{b_k}^{b_0}\right) - \mathbf{v}_{b_k}^{b_k} \Delta t + \mathbf{g}^{b_k} \dfrac{\Delta t^2}{2} \\ \mathbf{R}_{b_{k+1}}^{b_k} \mathbf{v}_{b_{k+1}}^{b_{k+1}} - \mathbf{v}_{b_k}^{b_k} + \mathbf{g}^{b_k} \Delta t \\ \mathbf{R}_{b_{k+1}}^{b_k} \mathbf{g}^{b_{k+1}} - \mathbf{g}^{b_k} \end{bmatrix}. \end{aligned} \tag{12}$$

The covariance $\mathbf{P}_{b_{k+1}}^{b_k}$ of the linear IMU measurement has the following form:

$$\mathbf{P}_{b_{k+1}}^{b_k} = \begin{bmatrix} {}^{\boldsymbol{\alpha}\boldsymbol{\beta}}\mathbf{P}_{b_{k+1}}^{b_k} & \mathbf{0} \\ \mathbf{0} & {}^{\mathbf{g}}\mathbf{P}_{b_{k+1}}^{b_k} \end{bmatrix}. \tag{13}$$

Given known rotations, the joint covariance matrix ${}^{\boldsymbol{\alpha\beta}}\mathbf{P}^{k}_{k+1}$ is independent of the gravity covariance ${}^{\mathbf{g}}\mathbf{P}^{b_k}_{bk+1}$. Consider the continuous-time state-space model of the preintegrated IMU measurements derived from the first two equations in (2)

$$\begin{aligned}\begin{bmatrix}\dot{\boldsymbol{\alpha}}^{b_k}_t\\ \dot{\boldsymbol{\beta}}^{b_k}_t\end{bmatrix} &= \begin{bmatrix}\mathbf{0} & \mathbb{I}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}^{b_k}_t\\ \boldsymbol{\beta}^{b_k}_t\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \mathbf{R}^{b_k}_t\end{bmatrix}\mathbf{a}^b_t\\ &= \mathbf{F}_t\cdot\begin{bmatrix}\boldsymbol{\alpha}^{b_k}_t\\ \boldsymbol{\beta}^{b_k}_t\end{bmatrix} + \mathbf{G}_t\cdot\mathbf{a}^b_t. \end{aligned} \tag{14}$$

${}^{\boldsymbol{\alpha\beta}}\mathbf{P}^{b_k}_{b_{k+1}}$ can be calculated recursively using first-order discrete-time propagation within the time interval $[k, k+1]$ with initial covariance $\mathbf{P}^{b_k}_{b_k} = \mathbf{0}$

$$\begin{aligned}{}^{\boldsymbol{\alpha\beta}}\mathbf{P}^{b_k}_{t+\delta t} &= (\mathbb{I}+\mathbf{F}_t\delta t)\cdot{}^{\boldsymbol{\alpha\beta}}\mathbf{P}^{b_k}_t\cdot(\mathbb{I}+\mathbf{F}_t\delta t)^T\\ &\quad+(\mathbf{G}_t\delta t)\cdot\mathbf{Q}_t\cdot(\mathbf{G}_t\delta t)^T\end{aligned} \tag{15}$$

where $\mathbf{Q}_t$ is the covariance of the additive Gaussian noise of the accelerometer measurements

$$\hat{\mathbf{a}}^b_t = \mathbf{a}^b_t + {}^a\mathbf{n}_t,\ {}^a\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t). \tag{16}$$

3) Linear Camera Measurement Model: The linear camera measurement model $\{\hat{\mathbf{z}}^{c_j}_l, \mathbf{H}^{c_j}_l\}$ for the observation of the lth feature in the jth image is defined as

$$\begin{aligned}\hat{\mathbf{z}}^{c_j}_l &= \hat{\mathbf{0}} = \mathbf{H}^{c_j}_l\mathcal{X} + {}^{\lambda}\mathbf{n}^{c_j}_l\\ &= \mathbf{M}\cdot f^{b^{-1}}_c\left(f^{b_0^{-1}}_{b_j}\left(f^{b_0}_{b_i}\left(f^{b}_{c}\left(\lambda_l\begin{bmatrix}u^{c_i}_l\\ v^{c_i}_l\\ 1\end{bmatrix}\right)\right)\right)\right)\end{aligned} \tag{17}$$

where $\mathbf{M}$ is defined as

$$\mathbf{M} = \begin{bmatrix}-1 & 0 & \hat{u}^{c_j}_l\\ 0 & -1 & \hat{v}^{c_j}_l\end{bmatrix} \tag{18}$$

and $[u^{c_i}_l, v^{c_i}_l]$ is the noiseless first observation of the lth feature that happened in the ith image. $[\hat{u}^{c_j}_l, \hat{v}^{c_j}_l]$ is the observation of the same feature in the jth image. The function $f^m_n(.)$ that transforms a 3-D point $\mathbf{r}$ from frame n to frame m and its inverse function $f^{m^{-1}}_n(.)$ are defined as

$$\begin{aligned}f^m_n(\mathbf{r}) &= \mathbf{R}^m_n\cdot\mathbf{r} + \mathbf{p}^m_n\\ f^{m^{-1}}_n(\mathbf{r}) &= \mathbf{R}^n_m\cdot\left(\mathbf{r}-\mathbf{p}^m_n\right)\end{aligned} \tag{19}$$

where all rotation matrices $\mathbf{R}^X_Y$ are known quantities.

${}^{\lambda}\mathbf{n}^{c_j}_l$ is the additive Gaussian measurement noise for the linear camera model, with the covariance matrix in the following form:

$${}^{\lambda}\mathbf{P}^{c_j}_l = {\lambda^{c_j}_l}^2\mathbf{P}^{c_j}_l \tag{20}$$

where $\mathbf{P}^{c_j}_l$ is the feature observation noise in the normalized image plane. Although we can only initialize the unknown $\lambda^{c_j}_l$ as the average scene depth, we find in practice that the solution is insensitive to the initial value of $\lambda^{c_j}_l$ as long as it is set to be *larger* than the actual depth.

4) Solution to the Linear Estimator: The linear cost function (11) can be rearranged into the following form:

$$(\boldsymbol{\Lambda}_p + \boldsymbol{\Lambda}_{\mathcal{B}} + \boldsymbol{\Lambda}_{\mathcal{C}})\mathcal{X} = (\mathbf{b}_p + \mathbf{b}_{\mathcal{B}} + \mathbf{b}_{\mathcal{C}}) \tag{21}$$

where $\{\boldsymbol{\Lambda}_{\mathcal{B}}, \mathbf{b}_{\mathcal{B}}\}$ and $\{\boldsymbol{\Lambda}_{\mathcal{C}}, \mathbf{b}_{\mathcal{C}}\}$ are information matrices and vectors for IMU and visual measurements, respectively, and $\{\boldsymbol{\Lambda}_p = {\mathbf{H}_p}^T\mathbf{H}_p, \mathbf{b}_p = {\mathbf{H}_p}^T\mathbf{r}_p\}$ is the (optional) prior. Due to the known incremental and relative rotations, the cost is linear with respect to the states, and the system in (21) has a unique solution, even the prior $\{\boldsymbol{\Lambda}_p, \mathbf{b}_p\}$ is only used to lock the first position ($\mathbf{p}^{b_0}_{b_0} = [0,0,0]$). This suggests that our method is able to recover all quantities in the full state vector, including the camera–IMU translation, without any initial guess of those values.

5) Termination Criteria: The covariance matrix (inverse of the information matrix) $\left(\boldsymbol{\Lambda}_p + \boldsymbol{\Lambda}_{\mathcal{B}} + \boldsymbol{\Lambda}_{\mathcal{C}}\right)^{-1}$ naturally tells the uncertainty of the linear initialization estimator. The block that corresponds to the camera–IMU translation in the covariance matrix represents the uncertainty of the calibration parameters. We use the maximum singular value $\sigma^{\max}_{\mathbf{p}}$ of the block as the convergence indicator and terminate the linear initialization process if $\sigma^{\max}_{\mathbf{p}} < \sigma_{\mathbf{p}}$, where $\sigma_{\mathbf{p}}$ is a threshold. A convergence plot can be found in Fig. 6.

As computing the matrix inverse is much slower than solving the linear system (29) using Cholesky decomposition, we run the termination check at a slower rate in another thread. This will slightly delay the termination time but will not harm overall performance. After this point, the whole initialization and calibration process is completed.

VI. Tightly Coupled Nonlinear Optimization With Calibration Refinement

After state initialization and obtaining camera–IMU calibration (Section V), we proceed with a sliding window nonlinear estimator, as illustrated in Fig. 3, for high-accuracy state estimation and calibration refinement. This is an extension of [7] and [8] by including camera–IMU calibration in the nonlinear optimization.

Since a large number of parameters in the nonlinear optimization share the same physical meaning as those in the linear initialization (Section V), here we introduce a slight abuse of notations by reusing symbols to represent state vectors ($\mathcal{X}$), Jacobian matrices ($\mathbf{H}, \mathbf{F}, \mathbf{G}$), covariance matrices ($\mathbf{P}, \mathbf{Q}$), and information matrices ($\boldsymbol{\Lambda}$).

A. Formulation

The definition of the full state is similar to the linear case, with the exception that the full 6-DOF camera–IMU transformation $\mathbf{x}^b_c$ is included in the state vector. The gravity vector is also replaced with the quaternion $\mathbf{q}^{b_i}_{b_j}$ for joint optimization of IMU translation and rotation (the transpose is again ignored)

$$\begin{aligned}\mathcal{X} &= [\mathbf{x}_n, \mathbf{x}_{n+1}, \cdots \mathbf{x}_{n+N}, \mathbf{x}^b_c, \lambda_0 \cdots \lambda_{m+M}]\\ \mathbf{x}_k &= \left[\mathbf{p}^{b_0}_{b_k}, \mathbf{v}^{b_k}_{b_k}, \mathbf{q}^{b_0}_{b_k}\right]\\ \mathbf{x}^b_c &= \left[\mathbf{p}^b_c, \mathbf{q}^b_c\right].\end{aligned} \tag{22}$$

We minimize the sum of the Mahalanobis norm of all measurement residuals to obtain a maximum *a posteriori* estimation

$$\min_{\mathcal{X}} \left\{ \|\mathbf{r}_p - \mathbf{H}_p \mathcal{X}\|^2 + \sum_{k\in\mathcal{B}} \left\| r_{\mathcal{B}}\left(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \mathcal{X}\right) \right\|^2_{\mathbf{P}_{b_{k+1}}^{b_k}} + \sum_{(l,j)\in\mathcal{C}} \left\| r_{\mathcal{C}}\left(\hat{\mathbf{z}}_l^{c_j}, \mathcal{X}\right) \right\|^2_{\mathbf{P}_l^{c_j}} \right\} \tag{23}$$

where $r_{\mathcal{B}}(\hat{\mathbf{z}}_{bk+1}^{b_k}, \mathcal{X})$ and $r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j}, \mathcal{X})$ are measurement residuals for the IMU and the camera, respectively. Corresponding measurement models are defined in Sections VI-B and VI-C.

We use error-state representation [6] to linearize the nonlinear system (23) and solve it with the Gauss–Newton algorithm with the Huber norm [30] for robust outlier rejection. The residuals for linear components such as position, velocity, and feature depth can be easily defined as an addition to the latest state estimates

$$\mathbf{p} = \hat{\mathbf{p}} + \delta\mathbf{p}, \quad \mathbf{v} = \hat{\mathbf{v}} + \delta\mathbf{v}, \quad \lambda = \hat{\lambda} + \delta\lambda. \tag{24}$$

The residual for rotation is instead modeled as the perturbation in the tangent space of the rotation manifold. The error quaternion term $\delta\mathbf{q}$ is defined as the small difference between the estimated and the true quaternions

$$\mathbf{q} = \hat{\mathbf{q}} \otimes \delta\mathbf{q}, \quad \delta\mathbf{q} \approx \begin{bmatrix} \frac{1}{2}\delta\boldsymbol{\theta} \\ 1 \end{bmatrix} \tag{25}$$

where $\otimes$ is the quaternion multiplication operator. The 3-D error vector $\delta\boldsymbol{\theta}$ is the minimal presentation of rotation residual. We also have the equivalent formulation in the form of rotation matrix to provide a simple linearization form for the rotation residual

$$\mathbf{R} \approx \hat{\mathbf{R}} \cdot (\mathbb{I} + \lfloor \delta\boldsymbol{\theta} \times \rfloor). \tag{26}$$

The full error-state vector then becomes

$$\begin{aligned} \delta\mathcal{X} &= \left[\delta\mathbf{x}_n, \ldots \delta\mathbf{x}_{n+N}, \delta\mathbf{x}_c^b, \delta\lambda_m \ldots \delta\lambda_{m+M}\right] \\ \delta\mathbf{x}_k &= \left[\delta\mathbf{p}_{b_k}^{b_0}, \delta\mathbf{v}_{b_k}^{b_k}, \delta\boldsymbol{\theta}_{b_k}^{b_0}\right] \\ \delta\mathbf{x}_c^b &= \left[\delta\mathbf{p}_c^b, \delta\boldsymbol{\theta}_c^b\right]. \end{aligned} \tag{27}$$

In each Gauss–Newton iteration, (23) is linearized at the current state estimation $\hat{\mathcal{X}}$ with respect to the error state $\delta\mathcal{X}$

$$\min_{\delta\mathcal{X}} \left\{ \|\mathbf{r}_p - \mathbf{H}_p \mathcal{X}\|^2 + \sum_{k\in\mathcal{B}} \left\| r_{\mathcal{B}}\left(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \hat{\mathcal{X}}\right) + \mathbf{H}_{b_{k+1}}^{b_k} \delta\mathcal{X} \right\|^2_{\mathbf{P}_{b_{k+1}}^{b_k}} + \sum_{(l,j)\in\mathcal{C}} \left\| r_{\mathcal{C}}\left(\hat{\mathbf{z}}_l^{c_j}, \hat{\mathcal{X}}\right) + \mathbf{H}_l^{c_j} \delta\mathcal{X} \right\|^2_{\mathbf{P}_l^{c_j}} \right\} \tag{28}$$

where $\mathbf{H}_{b_{k+1}}^{b_k}$ and $\mathbf{H}_l^{c_j}$ are Jacobians of IMU and visual measurements, respectively. Minimizing the linearized cost function (28) is equivalent to solving the following linear system:

$$(\boldsymbol{\Lambda}_p + \boldsymbol{\Lambda}_{\mathcal{B}} + \boldsymbol{\Lambda}_{\mathcal{C}})\delta\mathcal{X} = (\mathbf{b}_p + \mathbf{b}_{\mathcal{B}} + \mathbf{b}_{\mathcal{C}}) \tag{29}$$

where $\boldsymbol{\Lambda}_p$, $\boldsymbol{\Lambda}_{\mathcal{B}}$, and $\boldsymbol{\Lambda}_{\mathcal{C}}$ are information matrices from the prior, IMU, and visual measurements, respectively. Incremental construction of the prior $\{\boldsymbol{\Lambda}_p = \mathbf{H}_p{}^T\mathbf{H}_p, \mathbf{b}_p = \mathbf{H}_p{}^T\mathbf{r}_p\}$ is discussed in Section VII. Note that (29) is different from (21) in the sense that (29) solves for the increments for the nonlinear optimization, while (21) directly recovers the initial values.

The error state is updated as

$$\hat{\mathcal{X}} = \hat{\mathcal{X}} \oplus \delta\mathcal{X} \tag{30}$$

where $\oplus$ is the compound operator that has the form of simple addition for position, velocity, and feature depth as in (24), but is formulated as quaternion multiplication for rotations as in (25).

B. IMU Measurement Model

Following (3), the residual of a preintegrated IMU measurement is

$$r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \mathcal{X}) = \begin{bmatrix} \delta\boldsymbol{\alpha}_{b_{k+1}}^{b_k} \\ \delta\boldsymbol{\beta}_{b_{k+1}}^{b_k} \\ \delta\boldsymbol{\theta}_{b_{k+1}}^{b_k} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{b_0}^{b_k}\left(\mathbf{p}_{b_{k+1}}^{b_0} - \mathbf{p}_{b_k}^{b_0} + \mathbf{g}^{b_0}\Delta t^2/2\right) - \mathbf{v}_{b_k}^{b_k}\Delta t - \hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k} \\ \mathbf{R}_{b_0}^{b_k}\left(\mathbf{R}_{b_{k+1}}^{b_0}\mathbf{v}_{b_{k+1}}^{b_{k+1}} + \mathbf{g}^{b_0}\Delta t\right) - \mathbf{v}_{b_k}^{b_k} - \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} \\ 2\left[\hat{\mathbf{q}}_{b_{k+1}}^{b_k{}^{-1}} \otimes \mathbf{q}_{b_k}^{b_0{}^{-1}} \otimes \mathbf{q}_{b_{k+1}}^{b_0}\right]_{xyz} \end{bmatrix} \tag{31}$$

where $[\mathbf{q}]_{xyz}$ extracts the vector part of the quaternion $\mathbf{q}$, which forms the rotation error vector as in (25), and $[\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k}, \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k}, \hat{\mathbf{q}}_{b_{k+1}}^{b_k{}^{-1}}]^T$ is the preintegrated IMU measurement obtained following (2) without knowing the initial velocity and attitude, using only noisy accelerometer and gyroscope measurements

$$\begin{bmatrix} \hat{\mathbf{a}}_t^b \\ \hat{\boldsymbol{\omega}}_t^b \end{bmatrix} = \begin{bmatrix} \mathbf{a}_t^b \\ \boldsymbol{\omega}_t^b \end{bmatrix} + \begin{bmatrix} {}^a\mathbf{n}_t \\ {}^\omega\mathbf{n}_t \end{bmatrix}, \begin{bmatrix} {}^a\mathbf{n}_t \\ {}^\omega\mathbf{n}_t \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t). \tag{32}$$

Taking the derivative of the residual $r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \mathcal{X})$ with respect to error-state vector $\delta\mathcal{X}$ gives the Jacobian of IMU measurements $\mathbf{H}_{b_{k+1}}^{b_k}$ in (28).

To obtain the covariance matrix $\mathbf{P}_{b_{k+1}}^{b_k}$ of the preintegrated IMU measurement, we form the linearized continuous dynamics of the error state of the IMU measurement using (2) and (26)

$$\begin{bmatrix} \delta\dot{\boldsymbol{\alpha}}_t^{b_k} \\ \delta\dot{\boldsymbol{\beta}}_t^{b_k} \\ \delta\dot{\boldsymbol{\theta}}_t^{b_k} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\hat{\mathbf{R}}_t^{b_k}\lfloor \hat{\mathbf{a}}_t^b \times \rfloor \\ \mathbf{0} & \mathbf{0} & -\lfloor \hat{\boldsymbol{\omega}}_t^b \times \rfloor \end{bmatrix} \begin{bmatrix} \delta\boldsymbol{\alpha}_t^{b_k} \\ \delta\boldsymbol{\beta}_t^{b_k} \\ \delta\boldsymbol{\theta}_t^{b_k} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -\hat{\mathbf{R}}_t^{b_k} & \mathbf{0} \\ \mathbf{0} & -\mathbb{I} \end{bmatrix} \begin{bmatrix} {}^a\mathbf{n}_t \\ {}^\omega\mathbf{n}_t \end{bmatrix} = \mathbf{F}_t \cdot \delta\mathbf{z}_t^{b_k} + \mathbf{G}_t \cdot \mathbf{n}_t. \tag{33}$$

The covariance matrix $\mathbf{P}_{b_{k+1}}^{b_k}$ can then be calculated by first-order discrete-time propagation within the time interval

$[k, k+1]$, with initial covariance $\mathbf{P}_{b_k}^{b_k} = \mathbf{0}$

$$\mathbf{P}_{t+\delta t}^{b_k} = (\mathbb{I} + \mathbf{F}_t \delta t) \cdot \mathbf{P}_t^{b_k} \cdot (\mathbb{I} + \mathbf{F}_t \delta t)^T + (\mathbf{G}_t \delta t) \cdot \mathbf{Q}_t \cdot (\mathbf{G}_t \delta t)^T. \tag{34}$$

Note how this propagation (33) is different from the dynamics of the linear IMU model (14) as we also consider the error propagation in the rotation component as well as the cross correlation between the rotation and the translation.

C. Camera Measurement Model

The camera measurement model can be formulated similar to the linear initialization (Section V-B), but with the residual being the reprojection error with covariance matrix $\mathbf{P}_l^{c_j}$ due to the feature depth initialization presented in Section V-B

$$r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j}, \mathcal{X}) = \begin{bmatrix} \frac{f x_l^{c_j}}{f z_l^{c_j}} - \hat{u}_l^j \\ \frac{f y_l^{c_j}}{f z_l^{c_j}} - \hat{v}_l^j \end{bmatrix}$$

$$\mathbf{f}_l^{c_j} = \begin{bmatrix} f x_l^{c_j} \\ f y_l^{c_j} \\ f z_l^{c_j} \end{bmatrix} = f_c^{b^{-1}} \left(\mathbf{f}_{b_j}^{b_0^{-1}} \left(\mathbf{f}_{b_i}^{b_0} \cdot \mathbf{f}_c^b \left(\lambda_l \begin{bmatrix} u_l^{c_i} \\ v_l^{c_i} \\ 1 \end{bmatrix} \right) \right) \right) \tag{35}$$

where $\mathbf{f}_c^b(.)$ and its inverse are defined in (19).

The Jacobian of the visual measurement $\mathbf{H}_l^{c_j}$ in (28) is obtained by taking the derivative of the residual $r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j}, \mathcal{X})$ with respect to the error-state vector $\delta\mathcal{X}$.

VII. Two-Way Marginalization

In order to bound the computational complexity of graph optimization-based methods, marginalization is usually used. We selectively marginalize out IMU states $\mathbf{x}_k$ and features λ_l from the sliding window for both the linear (Section V) and nonlinear (Section VI) optimization. Due to the well-known acceleration excitation requirement [13], [14] for scale observability for monocular VINS, a naive strategy that always marginalizes the oldest state may result in unobservable scale in degenerate motions, such as hovering or constant velocity motions.

To avoid this, we employ the two-way marginalization scheme originally proposed in [8] and [9] to selectively remove recent or old IMU states based on a scene parallax test. We add a new IMU state to the sliding window if the time between two IMU states Δt is larger than a threshold. We do not have a notion of spatial keyframes as in vision-only approaches [17], due to the requirement of bounding the uncertainty for every preintegrated IMU measurements. We then select whether to remove the oldest or the most recent IMU states based on a parallax test. As shown in Fig. 4, a recent state is identified as *fixed* only if it has sufficient parallax to the previous fixed state; otherwise, it will be removed in the next marginalization. We refer readers to [9] for details of this selection process.

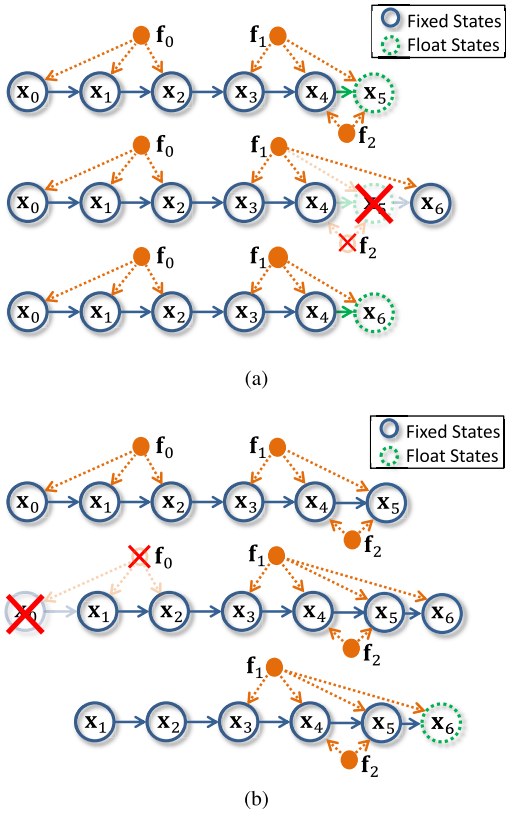

Fig. 4. Example with seven IMU states $\mathbf{x}_k$ and three features $\mathbf{f}_l$. (a) Structure of the full state before, during, and after marginalizing a recent IMU state $\mathbf{x}_5$ after a newer IMU state $\mathbf{x}_6$ is added. Visual and IMU measurements related to IMU state $\mathbf{x}_5$ (denoted by transparent edges) are summarized into a new prior $\mathbf{\Lambda}_p$ in (36). Also, feature f_2 is removed because it has no valid observation (under the assumption that the first observation is noiseless, as described in Section V-B3). (b) Similar marginalization process of the oldest IMU state, where IMU state $\mathbf{x}_0$ and feature f_0 are removed and all involved measurements are used to construct a new prior. Small parallax: hovering or slow motion in (a). Large parallax: fast motion in (b).

We construct a new prior based on all measurements related to the removed states

$$\mathbf{\Lambda}_p = \mathbf{\Lambda}_p + \sum_{k \in \mathcal{B}^-} \mathbf{H}_{b_{k+1}}^{b_k \intercal} \mathbf{P}_{b_{k+1}}^{b_k -1} \mathbf{H}_{b_{k+1}}^{b_k} + \sum_{(l,j) \in \mathcal{C}^-} \mathbf{H}_l^{c_j \intercal} \mathbf{P}_l^{c_j -1} \mathbf{H}_l^{c_j} \tag{36}$$

where $\mathcal{B}^-$ and $\mathcal{C}^-$ are sets of removed IMU and camera measurements, respectively. The marginalization is carried out using the Schur complement [16].

Intuitively, the two-way marginalization keeps removing recent IMU states if the platform has small or no motion. Keeping older IMU states in this case will preserve acceleration information that is necessary to recover the visual scale. The camera–IMU calibration also benefits from this marginalization scheme because it naturally accumulates all measurements to refine the calibration parameters.

VIII. Experimental Results

A. Implementation Details

As shown in Fig. 1, our monocular VINS sensor suite consists of an mvBlueFOX-MLC200w grayscale camera with

TABLE I
TIMING STATISTICS

	Modules	Time (ms)	Rate (Hz)	Thread
1	Harris detector	15	10Hz	1
	KLT tracker	5	30Hz	1
2	Linear initialization	15	10Hz	2
	Termination detection	40	5Hz	3
	Two-way marginalization	5	10Hz	2
3	Nonlinear optimization	60	10Hz	2
	Two-way marginalization	5	10Hz	2

a wide-angle lens that captures 752×480 images at 30 Hz and a Microstrain 3DM-GX4 IMU that runs at 500 Hz. The mount for the sensor suite has significant translation between sensors. The sensors are also purposely mounted in different frames with approximately 90° rotation offsets in roll and yaw to test the performance of camera–IMU calibration.

Our algorithm runs real time on an Intel NUC mini PC with i5-4250U processor. Three threads run in parallel in our implementation. The first thread performs detection of corner features at 10 Hz and KLT tracking [31] at 30 Hz. The second thread performs initialization and calibration, as well as nonlinear optimization at 10 Hz. During the linear translation initialization (Section V-B), a third thread is launched for recovery of state covariance and detection of convergence. We maintain $N = 30$ IMU states (30 images) and $M = 200$ features in the sliding window. For each image, we detect a maximum of 100 new features with a minimum separation of 30 pixels. A tracked feature has to pass a rotation-compensated parallax threshold of 30 pixels before it can be triangulated and added into the optimization. Timing statistics for feature tracking, linear initialization, and nonlinear optimization are given in Table I.

For the linear initialization (Section V), we assume biases of the IMU can be removed by initial subtraction. Therefore, biases are not included in the state vector. Since the initialization phase normally takes only a few seconds, ignoring IMU biases will not lead to noticeable negative effects. For the nonlinear optimization, IMU biases are continuously estimated.

B. Camera–IMU Calibration Performance

In this experiment, we evaluate the performance of our online camera–IMU calibration method. The camera and the IMU, as shown in Fig. 1, are rigidly mounted on a metal bar. The orientation offsets between the camera and the IMU are approximately $[-90, 0, +90]^\circ$ in yaw, pitch, and roll, respectively. The translation between the two sensors is approximately $[0, -0.3, 0]$ m in x, y, and z, respectively. Note that although the sensors are rigidly mounted, we do not know the precise camera–IMU calibration. We use results estimated by Kalibr [1], a state-of-the-art offline camera–IMU calibration tool using fiducial markers, as a reference. The performance of calibration is evaluated by consistency of different methods and repeatability across multiple trials.

During the experiment, the user moves the sensor suite freely in a typical lab environment, with only natural visual features for our method but with a chessboard for Kalibr. We conducted ten experimental trials for both algorithms. The calibration results for linear initialization (Section V), nonlinear optimization (Section VI), and Kalibr are shown in Fig. 5. For both rotation and translation calibration, we can see that the linear method provides a reasonable initialization without any prior knowledge about the mechanical configuration of the system, while the nonlinear optimization further refines the calibration results. We achieve a final calibration accuracy of approximately 1° in rotation and 0.02 m in translation.

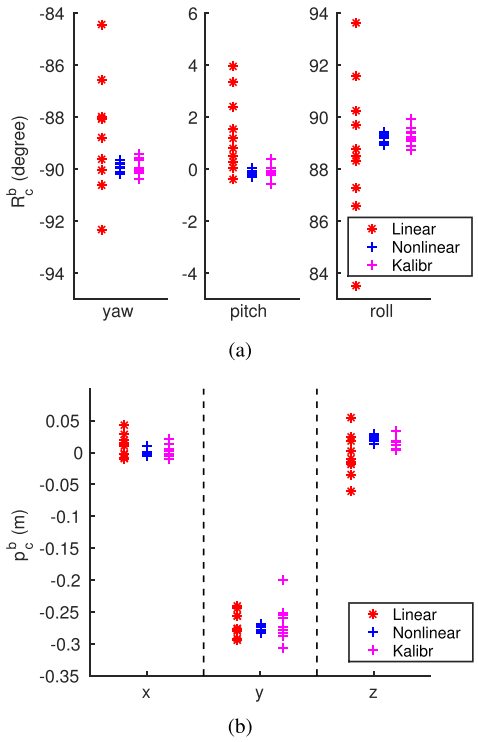

Fig. 5. Performance of camera–IMU calibration for linear (Section V) and nonlinear (Section VI) estimators and Kalibr. The orientation offset between the camera and the IMU is approximately $[-90, 0, +90]$ degrees in yaw, pitch, roll, and $[0, -0.3, 0]$ m in x, y, and z, respectively. Since we do not know the precise camera–IMU calibration, the performance is evaluated by consistency between different methods and repeatability across multiple trials. In all cases, the linear method provides a reasonable initialization without knowing the mechanical configuration of the sensor suite. The nonlinear method further refines the calibration to achieve a final accuracy of 1° in rotation and 0.02 m in translation. (a) Camera–IMU rotation. (b) Camera–IMU translation.

We observe that the standard deviation of the translation in the y-axis (the direction with the largest translational shift) is 0.45 and 2.85 cm for our system and Kalibr, respectively. For other dimensions, we also observe much better error distribution than that of Kalibr. This experiment demonstrates the competitive accuracy of our system against state-of-the-art algorithms. We stress that Kalibr requires fiducial markers and is offline calibration with high computational demands, while our method runs online and does not require any fiducial markers.

Fig. 6 details all phases of the calibration process during one of the trials, with the trajectory of the sensor suite shown in Fig. 7. The calibration starts with linear camera–IMU rotation calibration Phase 1 (Section V-A), during which the camera–IMU rotation is recovered from scratch. Phase 1 terminates when all three nonzero singular values of $\mathbf{Q}_N$ reach

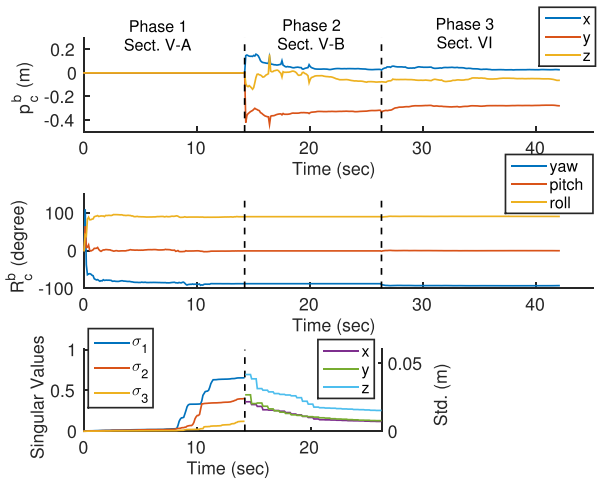

Fig. 6. Detailed illustration of the whole calibration process with system flow shown in Fig. 2 and trajectory shown in Fig. 7. Different phases are separated by dashed lines. In Phase 1, only the camera–IMU rotation is estimated, with the singular value-based termination criteria (Section V-A2) shown in the first segment in the bottom figure. In Phase 2, the camera–IMU translation and other VINS navigation quantities are recovered. The uncertainty-based termination criteria (Section V-B5) are shown in the second segment in the bottom figure. During Phase 2, the camera–IMU rotation remains constant. Phase 3 jointly and continuously optimizes the full 6-DOF camera–IMU calibration.

a high level (Section V-A2). During Phase 1, the translation component is not estimated. In Phase 2 (Section V-B), the velocity, attitude, and feature depth of the VINS, as well as the camera–IMU translation, are estimated simultaneously using a linear sliding window estimator. Here, we show only the camera–IMU translation calibration and defer the discussion of the initialization of other quantities to Section VIII-C. During this phase, the translation component is recovered, again with no prior knowledge about the mechanical configuration. The termination criteria are determined by the uncertainty of the calibration parameters (Section V-B5). Note that Phase 2 may last for an extensive period of time if there is insufficient motion excitation. However, our two-way sliding window marginalization scheme (Section VII) ensures a bounded complexity algorithm that is able to operate reliably until convergence of the calibration parameters. This is the key advantage of our approach compared with [24]. During Phase 2, the camera–IMU rotation is treated as a constant. Phase 3 (Section VI) uses nonlinear optimization to jointly and continuously refine the camera–IMU rotation and translation. Since there are no termination criteria in Phase 3, we do not compute the calibration uncertainty to save computational resources.

C. Motion Estimation Performance

We now compare the motion estimation performance of the overall monocular VINS estimator against that of a ground truth referencing system that consists of eight OptiTrack Flex13 cameras,[1] as shown in Fig. 8. The two dashed lines in each plot indicate the switching between rotation calibration (Section V-A), linear initialization (Section V-B),

[1]http://www.optitrack.com/

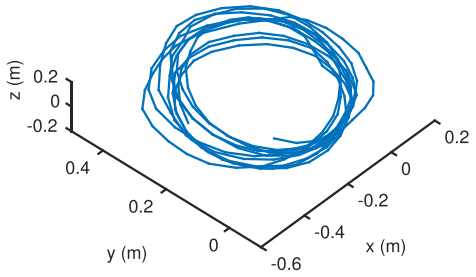

Fig. 7. Trajectory of the sensor suite for the trial shown in Fig. 6.

and nonlinear optimization (Section VI). The whole process starts by moving the sensor suite freely in the space. During the rotation calibration, position and velocity quantities are unavailable. After switching to the linear initialization at approximately 16 s, the estimator recovers the nontrivial velocity on the fly without any initial guesses. After that, it can be seen that the onboard velocity estimates compare well with the ground truth, with a standard deviation of $\{0.0278, 0.0117, 0.0026\}$ (m/s) in x, y, and z axes, respectively. As we do not know the exact starting position of the estimator, the position is aligned to the ground truth data with the initial zero pose from the estimator. Since the global position is unobservable, position drift will occur. However, we can still visually verify that the scale estimation is correct, which indicates the effectiveness of monocular visual–inertial fusion.

D. Performance in Large-Scale Environments

In this experiment, we evaluate the performance of the overall system with challenging data sets in complex indoor and outdoor environments.

In the first scenario, we handhold the sensor suite and move it in a complex library environment. We encounter large rotation [Fig. 9(a)], motion blur [Fig. 9(b)], people walking and view obstruction [Fig. 9(c)], as well as mirrors [Fig. 9(d)]. Fig. 10 shows the position estimation from the overall monocular VINS estimator with challenging cases shown in Fig. 9. The total trajectory length is 247.36 m and the final position drift is 3.28 m. The error is 1.3% of the total trajectory length. However, considering that during the experiment the sensor suite reaches an angular velocity up to 120°/s, which causes significant motion blur, we can still claim that overall estimation accuracy is high.

In the second scenario, we consider aerial navigation tasks in complex outdoor environments. We mount our sensor suite on a quadrotor to show its capability to assist autonomous navigation. Our system setup consists of a forward-facing camera, a downward-facing camera, and an IMU. All cameras capture data at the same time, which allows us to evaluate the performance of the proposed system with two configurations: 1) forward-facing camera + IMU and 2) downward-facing camera + IMU. The testing site spans a variety of cases, including narrow sidewalk [Fig. 11(a)], open space [Fig. 11(b)], high-speed flight [Fig. 11(c)], and large rotation [Fig. 11(d)]. The total flight time is approximately 8 min, and the vehicle travels 642 m with a highest speed of 8.9 m/s and a

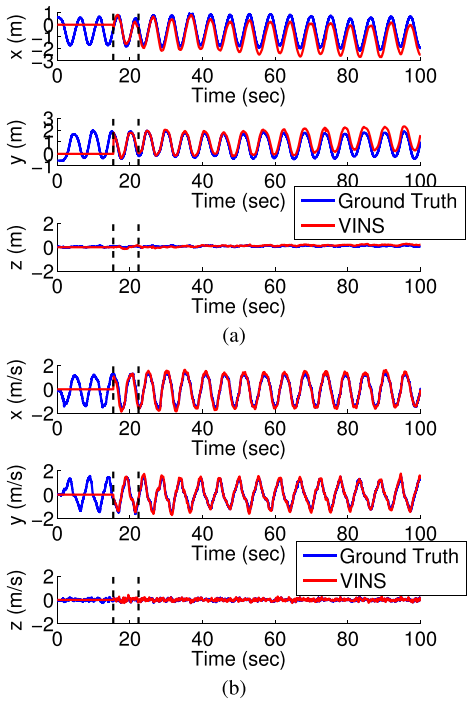

Fig. 8. Comparison of estimator performance against ground truth. The two dashed lines in each plot indicate the switching between camera–IMU rotation calibration (Section V-A), linear initialization (Section V-B), and nonlinear optimization (Section VI). The rotation calibration runs between 0 and 16 s, after which the linear initialization starts and recovers the nontrivial velocity of the platform on the fly. It can be seen that the estimated velocity matches well with the ground truth data. As we do not know the exact starting position, the position is aligned to the ground truth data with its first estimate. Although there is unavoidable position drift, we can still visually verify that the scale estimation is correct, indicating the effectiveness of monocular visual–inertial fusion. (a) Position. (b) Velocity.

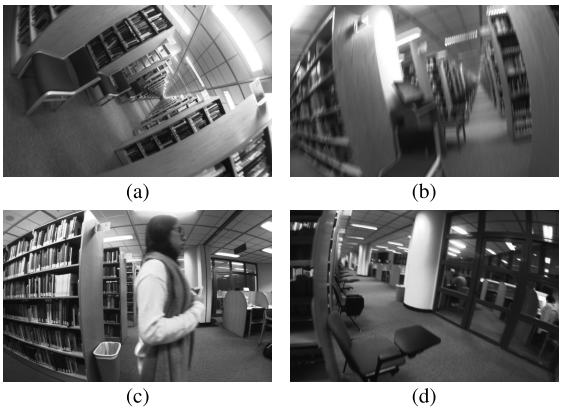

Fig. 9. Onboard images during experiment in indoor environments. (a) Large rotation. (b) Motion blur. (c) People walking. (d) Mirror.

height range of $[-6.6, 4.5]$ (m). The trajectory is aligned with an aerial map using GPS measurements as position reference (Fig. 12). Note that GPS reference is only available when the quadrotor is far away from buildings.

We run the whole system pipeline separately with the two configurations. As shown in Fig. 12, the final position drifts obtained with the forward- and downward-facing configurations are 5.86 and 2.80 m, respectively, which correspond to 0.91% and 0.44% of the total trajectory length. While Fig. 12 reports that the downward-facing version

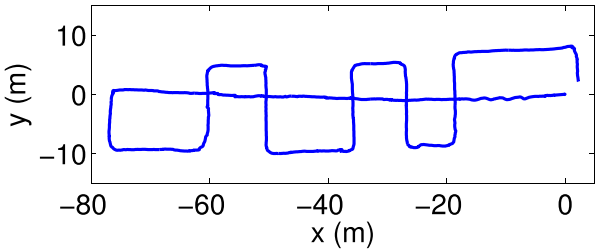

Fig. 10. Position estimation with our monocular VINS estimator in a complex indoor environment. The total trajectory length is 247.36 m and the final position drift is 3.28 m.

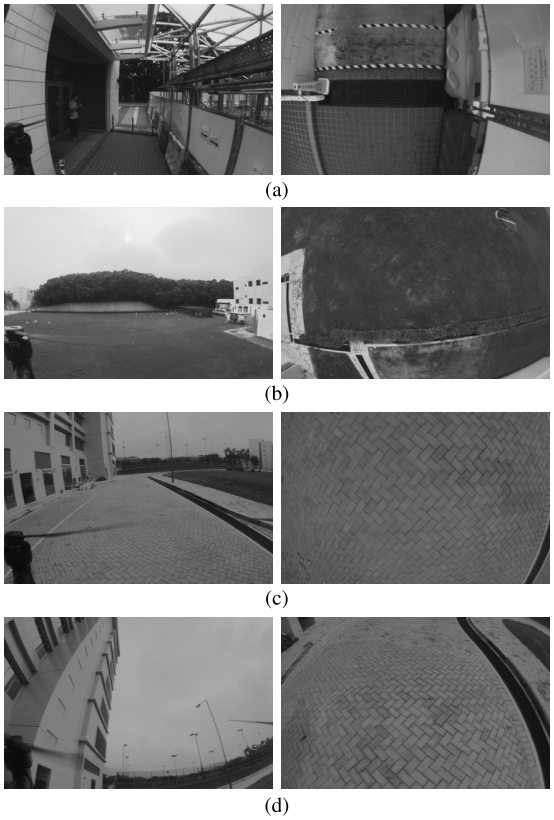

Fig. 11. Onboard images during experiment in outdoor environment. For each case, two images, one captured by the forward-facing camera (left) and the other one by the downward-facing camera (right), are shown. (a) Narrow sidewalk. (b) Hovering in open space (4.5 m above the ground). (c) Low-altitude high-speed flight (highest velocity 8.9 m/s). (d) Aggressive braking (largest pitch 35°).

aligns better with GPS in the xy-direction, we observe that the downward-facing camera accumulates more drift in the z-direction (1.75 m). This result makes sense since it is well known that the estimation performance is worse when the direction of movement is parallel to the camera's optical axis. Another reason that the forward-facing configuration performs worse is that the camera often observes only distant features [Fig. 11(b)], resulting in no features being triangulated for an extended period of time (Section VIII-A). The performance difference indicates the necessity of choosing different sensor configurations to adapt to different applications. We stress again that no explicit calibration is required to launch our system, thanks to self-calibration and online initialization.

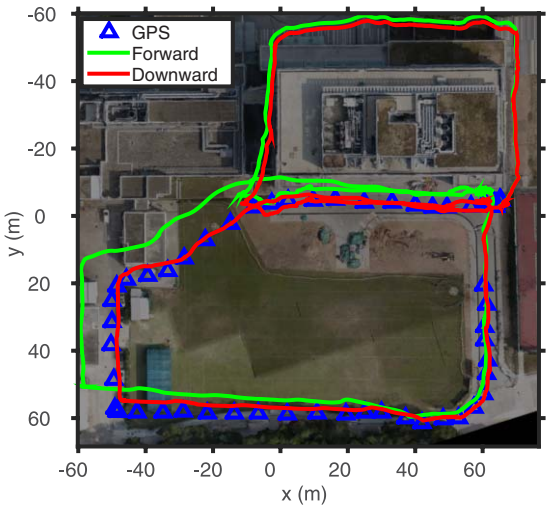

Fig. 12. Comparison between different sensor configurations (forward- or downward-facing monocular camera) in the same flight. The total travel distance is 642 m. Final position drifts are 0.91% and 0.44% of the traveled distance for the two configurations, respectively. We observe that the forward-facing camera accumulates more drift in the x–y direction, while the downward-facing camera accumulates more drift in the z direction (1.75 m). This shows the necessity of rapid mission-dependent sensor reconfiguration.

IX. Conclusion

In this paper, we propose a novel monocular VINS state estimator for real-time state estimation with unknown initialization and camera–IMU calibration. Specifically, our system initializes the velocity, attitude, visual scale, and camera–IMU calibration automatically while the system is performing free motion in natural environments. Our system is able to automatically identify convergence of the calibration parameters and switch between different system modules. After the initialization, a nonlinear optimization runs real time recursively for high-accuracy state estimation. Online experimental results are presented to demonstrate the performance of our approach. We also show competitive accuracy compared with a state-of-the-art offline marker-based camera–IMU calibration method.

A limitation of monocular VINS is the need for motion excitation for scale observability. A multicamera system may solve the problem, but the extrinsic calibration process can be troublesome. To overcome this, we plan to extend our framework into a generalized multicamera-inertial system for elimination of degenerate motions and online calibration of transformations between multiple cameras and the IMU. We are also interested in using omnidirectional or fisheye cameras to eliminate the rotation-only degenerate motion.

X. Acknowledgment

The authors would like to thank DJI for equipment support.

References

[1] 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.*, Nov. 2013, pp. 1280–1286.

[2] A. Bachrach, S. Prentice, R. He, and N. Roy, "RANGE—Robust autonomous navigation in GPS-denied environments," *J. Field Robot.*, vol. 28, no. 5, pp. 644–666, Sep./Oct. 2011.

[3] A. S. Huang *et al.*, "Visual odometry and mapping for autonomous flight using an RGB-D camera," in *Proc. Int. Symp. Robot. Res.*, Flagstaff, AZ, USA, Aug. 2011, pp. 1–16.

[4] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Consistency analysis and improvement of vision-aided inertial navigation," *IEEE Trans. Robot.*, vol. 30, no. 1, pp. 158–176, Feb. 2014.

[5] 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.

[6] 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, Mar. 2014.

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

[8] S. Shen, "Autonomous navigation in complex indoor and outdoor environments with micro aerial vehicles," Ph.D. dissertation, Dept. Eng., Univ. Pennsylvania, Philadelphia, PA, USA, Aug. 2014.

[9] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Initialization-free monocular visual-inertial state estimation with application to autonomous MAVs," in *Proc. Int. Symp. Experim. Robot.*, Marrakesh, Morocco, 2014, pp. 211–227.

[10] Z. Yang and S. Shen, "Monocular visual-inertial fusion with online initialization and camera-IMU calibration," in *Proc. IEEE Int. Symp. Safety, Secur., Rescue Robot.*, West Lafayette, IN, USA, Oct. 2015, pp. 1–8.

[11] D. Scaramuzza *et al.*, "Vision-controlled micro flying robots: From system design to autonomous navigation and mapping in GPS-denied environments," *IEEE Robot. Autom. Mag.*, vol. 21, no. 3, pp. 26–40, Sep. 2014.

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

[13] 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, Jan. 2011.

[14] E. S. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *Int. J. Robot. Res.*, vol. 30, no. 4, pp. 407–430, Apr. 2011.

[15] V. Indelman, S. Williams, M. Kaess, and F. Dellaert, "Information fusion in navigation systems via factor graph based incremental smoothing," *Robot. Auto. Syst.*, vol. 61, no. 8, pp. 721–738, Aug. 2013.

[16] 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./Oct. 2010.

[17] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. IEEE ACM Int. Symp. Mixed Augmented Reality*, Nara, Japan, Nov. 2007, pp. 225–234.

[18] S. Lange, N. Sünderhauf, and P. Protzel, "Incremental smoothing vs. filtering for sensor fusion on an indoor UAV," in *Proc. IEEE Int. Conf. Robot. Autom.*, Karlsruhe, Germany, May 2013, pp. 1773–1778.

[19] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, Saint Paul, MN, USA, May 2012, pp. 957–964.

[20] 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.*, Rome, Italy, Jul. 2015, pp. 1–10.

[21] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, Hong Kong, May/Jun. 2014, pp. 15–22.

[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, Feb. 2012.

[23] L. Kneip, S. Weiss, and R. Siegwart, "Deterministic initialization of metric state estimation filters for loosely-coupled monocular vision-inertial systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, San Francisco, CA, USA, Sep. 2011, pp. 2235–2241.

[24] T. C. Dong-Si and A. I. Mourikis, "Estimator initialization in vision-aided inertial navigation with unknown camera-IMU calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vilamoura, Portugal, Oct. 2012, pp. 1064–1071.

[25] A. Martinelli, "Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 44–60, Feb. 2012.

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

[27] V. Lippiello and R. Mebarki, "Closed-form solution for absolute scale velocity estimation using visual and inertial data with a sliding least-squares estimation," in *Proc. Medit. Conf. Control Autom.*, Chania, Greece, Jun. 2013, pp. 1261–1266.
[28] L. Heng, G. H. Lee, and M. Pollefeys, "Self-calibration and visual SLAM with a multi-camera system on a micro aerial vehicle," in *Proc. Robot., Sci. Syst. (RSS)*, Berkeley, CA, USA, 2014.
[29] D. Nister, "An efficient solution to the five-point relative pose problem," in *Proc. IEEE Int. Conf. Pattern Recognit.*, vol. 2. Madison, WI, USA, Jun. 2003, pp. II-195–II-202.
[30] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964.
[31] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, Vancouver, BC, Canada, Aug. 1981, pp. 674–679.

Zhenfei Yang (S'15) received the B.Eng. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2014. He is currently pursuing the M.Phil. degree with the Hong Kong University of Science and Technology, Hong Kong, under the supervision of Prof. S. Shen.

He focuses on visual-inertial navigation and aims to build aerial robots that are able to perform autonomous flight reliably in both indoor and outdoor environments.

Shaojie Shen (S'10–M'14) received the B.Eng. (Hons.) degree in electronics engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2009, and the M.S. degree in robotics and the Ph.D. degree in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2011 and 2014, respectively.

He joined the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, in 2014, as an Assistant Professor. His current research interests include robotics and unmanned aerial vehicles, with focus on state estimation, sensor fusion, localization, and mapping, and autonomous navigation in complex environments.