Robust Initialization of Monocular Visual-Inertial Estimation on Aerial Robots

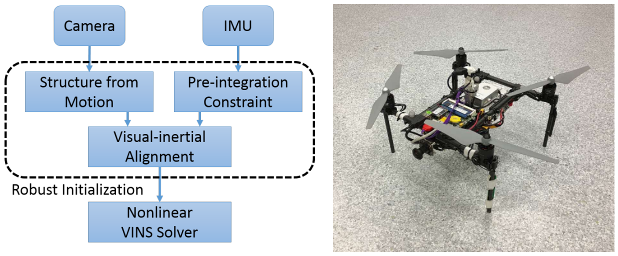
Tong Qin and Shaojie Shen

***Abstract*— In this paper, we propose a robust on-the-fly estimator initialization algorithm to provide high-quality initial states for monocular visual-inertial systems (VINS). Due to the non-linearity of VINS, a poor initialization can severely impact the performance of filtering-based or graph-based methods. Our approach starts with a vision-only structure from motion (SfM) to build the up-to-scale structure of camera poses and feature positions. By loosely aligning this structure with pre-integrated IMU measurements, our approach recovers the metric scale, velocity, gravity vector, and gyroscope bias which are treated as the initial values to bootstrap the nonlinear tightly-coupled optimization framework. We highlight that our approach can perform on-the-fly initialization in various scenarios without using any prior information about system states and movement, and show that the initial values obtained through this process can be efficiently used for launching nonlinear visual-inertial state estimator. The performance of the proposed approach is verified through the public UAV dataset and real-time onboard experiment. We make our implementation open source[1].**

I. Introduction

There are increasing demands in using small and agile aerial robots. A lot of applications are shown recently, such as aerial video, inspection, search and rescue missions. The accurate state estimation is the core foundation for autonomous flight. Many localization algorithms, which are based on monocular cameras [1]–[5], stereo cameras [6]–[8], RGB-D camera [9] and laser scanner [10] have been successfully applied on aerial robots. In some applications, the number of onboard sensors is limited due to constraints on payload and power. The monocular visual-inertial system, which consists of only one camera and one low-cost inertial measurement unit (IMU), has become an attractive sensor choice because of its small size, light weight, and low power consumption. In addition, some sensor suites, such as stereo cameras and RGB-D camera, will degenerate to a monocular camera in large scale environments. The monocular visual-inertial system, which has the ability of autonomous flight in GPS-denied environments, is a valuable research topic.

Robust state estimation is the core capability for autonomous robots operating in complex environments. Due to the nonlinearity of visual-inertial systems, the performance of monocular estimators [2, 4, 6, 11]–[13] heavily rely on the accuracy of initial values (gravity, velocity, bias, and depth of features). A poor initialization will decrease convergence speed or even lead to totally incorrect estimates. Especially for aerial robots which have full six degrees of freedom, accurate initial values are crucial. However, it is hard to obtain accurate initial states for the monocular visual-inertial system. On one hand, the metric scale of a monocular camera is not directly observable. On the other hand, non-zero acceleration motion is required to initialize the metric scale. This leads to nontrivial but unknown initial attitude (gravity vector) and velocity. In particular, during time-limited search and rescue missions, careful initialization with the MAV sitting stationary or moving along certain pattern is often infeasible. It is desirable to launch the MAV quickly and initialize the estimator without any prior knowledge about dynamical motion. In addition, vision algorithm is fragile during fast motion or under strong illumination change. The estimator will easily fail when the visual tracking is lost. This suggests that the development of on-flight automatically re-initialization is necessary. All these issues drive us to find a robust system which is capable of on-the-fly initialization to recover all critical states.



(a) (b)

Fig. 1. (a) The main structure of our initialization procedure. (b) The self-developed quadrotor with one forward-looking camera which is used in the indoor closed-loop experiment.

In this paper, we propose an approach which allows a monocular visual-inertial system to be initialized on-the-fly. Initial velocity, gravity vector, scale as well as gyroscope bias are calibrated in the initialization procedure. We first perform vision-only structure from motion (SfM), then loosely align IMU measurements with SfM results to get the metric initial states. The performance of our approach is proved by public dataset and real-time onboard experiments in indoor and outdoor environments.

We highlight that our contribution in twofold:

- The proposed approach is capable of on-the-fly initialization in various scenarios which can provide accurate initial states to bootstrap the nonlinear optimization system.
- At the system level, we applied the proposed approach

All authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China. tong.qin@connect.ust.hk, eeshaojie@ust.hk

[1] https://github.com/qintony/M-VINS

into sliding-window based monocular visual-inertial system. The practicability is verified by onboard closed-loop autonomous flight experiment.

The rest of the paper is structured as follows. In Sect. II, we discuss the relevant literature. The motivation and system overview are discussed in Sect. III. In Sect. IV, we present the methodology. Implementation details and experimental evaluations are presented in Sect. V. Finally, the paper is finished with the conclusion in Sect. VI.

II. Related Work

There are a large number of studies on visual-inertial state estimation problem. Traditional solutions with either monocular or stereo cameras are classified into two categories, filtering-based frameworks [11, 13]–[16] and graph-based optimization frameworks [2, 4, 6, 12]. Filtering-based approaches have advantage in faster processing since it continuously marginalizes past states. However, linearizing the states early may lead to sub-optimal results. Graph-based approaches benefit from the capacity of iterative re-linearization but they usually suffer from computational requirement. In general, some of the initial states (velocity, gravity orientation, and IMU bias) are assumed to be known or neglected, or the system should stay stationary and horizontal before launch. Without prior information, most methods are not suitable for dynamically taking off or on-the-fly initialization.

Our earlier works [1, 4] proposed an optimization-based linear estimator initialization method by leveraging the known relative rotations from short-term gyroscope integration. This method performs well in indoor environments. However, it fails in environments where feature depths are distributed throughout a wide range (e.g. outdoor environments) due to the geometric treatment of visual observations and the incapability of modeling the sensor noise in the raw projective formulation. Also, our earlier work does not take bias into consideration. Recently, a closed-form solution has been introduced in [17]. This closed-form solution is sensitive to noisy sensor data, and cannot be used in actual applications. Later, a revision of this closed-form solution is proposed in [18]. The authors add gyroscope bias calibration in this method. However, this changes the original formulation into a nonlinear and non-convex form. In addition, in both works, the authors fail to model the accuracy of inertial integration at different time durations. It is known that the accuracy of inertial integrated accuracy drops significantly as the time duration increases. In [3], a re-initialization and failure recovery algorithm based on SVO [19] is proposed. It is a practical method in the loosely-coupled visual-inertial system. Inertial measurements are used first to stabilize the MAV's attitude, then the SVO is launched for position feedback. This work assumes that the drone should be held nearly horizontally at beginning. Also, another distance sensor, TeraRanger, is used for height measurement. [20] proposed another initialization algorithm for loosely-coupled filtering system, which uses optical flow between two consecutive frames to extract velocity and dominant terrain plane. This method also requires no or little motion in initialization step since the initial attitude should be aligned with gravity.

For IMU measurement processing, one efficient technique is called pre-integration, which avoids repeating integrating IMU measurement by a reparametrization of the relative motion constraints. This algorithm was first proposed in [21]. Our previous work [2] considered on-manifold uncertainty propagation of this technique. Furthermore, [22] improves preintegration theory, which properly addressed the manifold structure of rotation group and modeled the noise propagation and posterior bias correction. Refer to [22], we add bias correction into our framework.

Vision-only structure from motion (SfM) techniques are able to recover the relative rotation and translation up to an unknown scale factor within multiple cameras poses [23]. Such methods are currently used in state-of-the-art visual navigation for MAVs [3, 19]. However, the unavailability of metric scale and absolute attitude can result in instability in autonomous flight. [24] proposed a method to compute the gravity vector and scale factor, and provided initial metric values for the state estimation filter. This method was based on analyzing the delta-velocities between SfM and inertial integration. Visual delta-velocities are obtained from the differentiation of up-to-scale camera poses, which is sensitive to noises especially in distanced scenes. Also, no bias calibration in their algorithm. We become aware that [25] presents a visual-inertial initialization algorithm which is similar to our framework. The difference is that we ignore acceleration bias in the initial step, since we find out that acceleration bias coupled with gravity usually lacks observability. Details about acceleration bias calibration is discussed in Sect. IV-E. Furthermore, we test our algorithm on the most challengeable public UAV dataset, which the former cannot deal with. In addition, real-time closed-loop onboard experiments verify the practicability of our algorithm.

III. Overview

The motivation of our algorithm is that no matter loosely coupled method or tightly coupled method, an accurate initial guess is required to bootstrap the monocular visual-inertial nonlinear system. On one hand, absolute scale and velocity are not available in monocular camera. On the other hand, non-zero acceleration motion is required to initialize the metric scale, which leads to nontrivial but unknown initial attitude (gravity vector) and velocity. Also, the scale information hidden in the IMU integration is easily influenced by noise and bias. Visual and initial measurements are two complementary resources, the one represents up-to-scale global structure, the other one contains metric incremental information. It is hard to directly fuse these two factors together without a good initial guess. Usually, short period movement cannot drive the whole system fully observable. The filtering-based method usually does the fusion work for a while until converge, while optimization-based method keeps long time measurements in a bundle and then optimizes these states together. A bad initial value will lead the filtering-based method diverging, and lead the optimization-based method to a local minimum. In usual, the average of IMU measurements in first few seconds is treated as gravity vector and IMU integration result is treated as the initial guess. However, this treatment is improper when IMU measurements are influenced by non-trivial bias or in accelerated movements. To improve the success rate of the monocular visual-inertial system, a robust initialization procedure is required.

Monocular only visual SLAM or structure from motion (SfM) is much more stable than visual-inertial fusion. In fact, constructing the visual-only up-to-scale structure doesn't rely on any initial states mentioned above. The essential initial values can be extracted by aligning IMU measurements with this structure. So we adopt a loosely coupled visual-inertial initialization procedure to get the initial states. The pipeline of our proposed method is shown in Fig. 1(a). We construct the visual-only structure firstly, then align this structure with IMU pre-integrations to recover the initial values.

IV. Methodology

We start with defining notations. We consider $(\cdot)^w$ as world frame, where gravity vector is along with z axis. $(\cdot)^v$ is the base frame in SfM, which is arbitrary fixed frame in visual structure, irrelevant to inertial measurement. $(\cdot)^w_b$ is body frame with respect to world frame. We treat the IMU frame as the body frame, which means IMU frame is aligned with body frame. b_k is the body frame while taking the k^{th} image. $(\cdot)^v_c$ is camera frame with respect to the visual base frame. c_k is the camera frame while taking the k^{th} image. We use $\hat{(\cdot)}$ to denote sensor measurements, which may be affected by noise and bias. We use $\bar{(\cdot)}$ to denote up-to-scale parameters in SfM structure. We use quaternion $\mathbf{q}$ to denote rotation. $\otimes$ is the two quaternion multiplication operation. Quaternion directly multiplies a vector means rotating this vector by the corresponding rotation matrix. $\mathbf{g}^w = [0, 0, g]^T$ is the gravity vector in the world frame. $\mathbf{g}^v$ is the gravity vector in visual base frame.

We assume that the intrinsic calibration of the camera and extrinsic calibration between the camera and IMU is known in the initialization step. In fact, we do not need a very precise extrinsic calibration since we will continuously refine it in the nonlinear optimization (Sect. IV-D).

A. Vision-Only Structure

The initialization procedure starts with a vision-only structure, which estimates a graph of up-to-scale camera poses and feature positions. Our method is based on the sliding-window based method [2, 4], which maintains several spacial-separate image frames. The spacial frames are selected by enough parallax near the neighbor. Sparse features are extracted and tracked among these frames. The feature correspondences are used to construct the visual structure inside the window.

As a common technique used in computer vision, we first choose two frames which contain sufficient feature parallax. Then Five-point method [26] is used to recover the relative rotation and up-to-scale translation between these two frames. Then fix the scale of this translation and triangulate all the features observed in these two frames. Based on these triangulated features, Perspective-n-Point (PnP) method is performed to estimate poses of other frames in the window. Finally, a global full Bundle Adjustment [27] is applied to minimize the total re-projection error of all feature observations. After that, we get all frame poses ($\bar{\mathbf{p}}^v_{c_k}$,$\mathbf{q}^v_{c_k}$) and feature positions. Here, position is up-to-scale. Assume that we have the prior of the extrinsic parameter ($\mathbf{p}^c_b, \mathbf{q}^c_b$) between camera frame and IMU (body) frame, all variables can be translated from camera frame to the IMU frame,

$$\begin{aligned} \mathbf{q}^v_{b_k} &= \mathbf{q}^v_{c_k} \otimes \mathbf{q}^c_b \\ s\bar{\mathbf{p}}^v_{b_k} &= s\bar{\mathbf{p}}^v_{c_k} + \mathbf{q}^v_{c_k}\mathbf{p}^c_b \end{aligned} \tag{1}$$

s is unknown scale, which will be solved in the next.

B. IMU Pre-Integration

Note that the IMU measurements run at a higher frequency than visual measurements. Usually, dozens of IMU measurements exist between two consecutive visual frames. We pre-integrate these IMU measurements, and treat this integration result as the incremental metric constraint. The IMU pre-integration is an efficient technique which can avoid repeating integration. In the paper, we follow the work proposed in [22] which properly addressed the manifold structure.

We denote IMU measurements (angular velocity and acceleration) as $\hat{\boldsymbol{\omega}}^b$, $\hat{\mathbf{a}}^b$. These measurements are affected by bias $\mathbf{b}$ and noise $\boldsymbol{\eta}$,

$$\begin{aligned} \hat{\boldsymbol{\omega}}^b(t) &= \boldsymbol{\omega}^b(t) + \mathbf{b}_g + \boldsymbol{\eta}_g \\ \hat{\mathbf{a}}^b(t) &= \mathbf{q}^w(t)^{\mathrm{T}}(\mathbf{a}^w(t) + \mathbf{g}^w) + \mathbf{b}_a + \boldsymbol{\eta}_a. \end{aligned} \tag{2}$$

Given two time instants that correspond to images frame b_k and b_{k+1}, we can pre-integrate linear acceleration and angular velocity in the local frame b_k:

$$\begin{aligned} \boldsymbol{\alpha}^{b_k}_{b_{k+1}} &= \iint_{t\in[k,k+1]} \boldsymbol{\gamma}^{b_k}_{b_t}\hat{\mathbf{a}}(t)dt^2 \\ \boldsymbol{\beta}^{b_k}_{b_{k+1}} &= \int_{t\in[k,k+1]} \boldsymbol{\gamma}^{b_k}_{b_t}\hat{\mathbf{a}}(t)dt \\ \boldsymbol{\gamma}^{b_k}_{b_{k+1}} &= \int_{t\in[k,k+1]} \boldsymbol{\gamma}^{b_k}_{b_t} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}\hat{\boldsymbol{\omega}}(t) \end{bmatrix} dt, \end{aligned} \tag{3}$$

$\boldsymbol{\alpha}^{b_k}_{b_{k+1}}, \boldsymbol{\beta}^{b_k}_{b_{k+1}}, \boldsymbol{\gamma}^{b_k}_{b_{k+1}}$ represent relative position, velocity, and rotation constraints respectively. It can be seen that the inertial measurements are typically integrated to form relative motion constraints which is independent of initial position and velocity.

C. Visual-Inertial Alignment

Now, we have the up-to-scale camera poses from SfM (Sect. IV-A), and the metric measurements from IMU pre-integration (Sect. IV-B). In this section, we detail our approach to align these two factors.

1) Gyroscope Bias Calibration: Considering two consecutive frames b_k and b_{k+1} in the window, we have the relative rotation $\mathbf{q}^v_{b_k}$ and $\mathbf{q}^v_{b_{k+1}}$ from the visual structure, as well as relative constraint $\hat{\gamma}^{b_k}_{b_{k+1}}$ from the IMU pre-integration. We estimate the gyroscope bias by minimizing the error between these two terms:

$$\begin{aligned} &\min_{b_g} \sum_{k\in\mathcal{B}} \left\| \mathbf{q}^{b_k}_{b_0} \otimes \mathbf{q}^{b_0}_{b_{k+1}} - \gamma^{b_k}_{b_{k+1}} \right\|^2 \\ &\gamma^{b_k}_{b_{k+1}} \approx \hat{\gamma}^{b_k}_{b_{k+1}} + \frac{\partial \gamma^{b_k}_{b_{k+1}}}{\partial \mathbf{b}_g} \delta \mathbf{b}_g, \end{aligned} \tag{4}$$

where $\mathcal{B}$ indexes the all frames in the window. In the second equation, we linearize the rotation constraint with respect to gyroscope bias. Aligning rotation in visual structure with relative constraint $\boldsymbol{\gamma}$, we can get the estimation of $\mathbf{b}_g$. Then we update $\hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}}, \hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}}$ with respect to $\mathbf{b}_g$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}} &\leftarrow \hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}} + \frac{\partial \boldsymbol{\beta}^{b_k}_{b_{k+1}}}{\partial \mathbf{b}_g} \delta \mathbf{b}_g \\ \hat{\boldsymbol{\gamma}}^{b_k}_{b_{k+1}} &\leftarrow \hat{\boldsymbol{\gamma}}^{b_k}_{b_{k+1}} + \frac{\partial \boldsymbol{\gamma}^{b_k}_{b_{k+1}}}{\partial \mathbf{b}_g} \delta \mathbf{b}_g. \end{aligned} \tag{5}$$

2) Initializing Velocity, Gravity Vector and Metric Scale: We define the variables that we would like to estimate as

$$\mathcal{X}_I = \left[\mathbf{v}^v_{b_0}, \, \mathbf{v}^v_{b_1}, \, \cdots \, \mathbf{v}^v_{b_n}, \, \mathbf{g}^v, \, s \right], \tag{6}$$

where s is the scale parameter that aligns the visual structure to the actual metric scale implicitly provided by IMU measurements. The following equation describes the relationship between metric position and velocity constraint with visual structure,

$$\begin{aligned} \hat{\mathbf{z}}^{b_k}_{b_{k+1}} &= \begin{bmatrix} \hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}} \\ \hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}} \end{bmatrix} = \mathbf{H}^{b_k}_{b_{k+1}} \mathcal{X}_I + \mathbf{n}^{b_k}_{b_{k+1}} \\ &\approx \begin{bmatrix} \mathbf{q}^{b_k}_v \Delta t_k & \mathbf{0} & \frac{1}{2}\mathbf{q}^{b_k}_v \Delta t_k^2 & \mathbf{q}^{b_k}_v (\bar{\mathbf{p}}^v_{b_{k+1}} - \bar{\mathbf{p}}^v_{b_k}) \\ -\mathbf{q}^{b_k}_v & \mathbf{q}^{b_k}_v & \mathbf{q}^{b_k}_v \Delta t_k & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}^v_{b_k} \\ \mathbf{v}^v_{b_{k+1}} \\ \mathbf{g}^v \\ s \end{bmatrix}. \end{aligned} \tag{7}$$

In the above formula, $\mathbf{q}^v_{b_k}, \bar{\mathbf{p}}^v_{b_k}, \bar{\mathbf{p}}^v_{b_{k+1}}$ are obtained from the visual structure. $\mathbf{q}^{b_k}_v$ is the inverse rotation of $\mathbf{q}^v_{b_k}$. Δt_k is the time interval between two consecutive frames. By solving the this least square problem:

$$\min_{\mathcal{X}_I} \sum_{k\in\mathcal{B}} \left\| \hat{\mathbf{z}}^{b_k}_{b_{k+1}} - \mathbf{H}^{b_k}_{b_{k+1}} \mathcal{X}_I \right\|^2, \tag{8}$$

we can get the velocities and the gravity vector in the visual base frame $(\cdot)^v$, as well as the scale parameter. The translational components $\bar{\mathbf{p}}^v$ from the visual structure will be scaled to the metric units. The estimated gravity will undergo another round of refinement by enforcing the norm constraint.

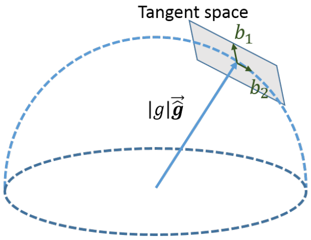


Fig. 2. Since the magnitude of gravity is known, the degree of freedom of the gravity is two. As shown in the figure, $\mathbf{g}$ lies on a sphere where the radius is the known magnitude $|g|$. We parameterize the gravity around current estimate as $g \cdot \bar{\hat{\mathbf{g}}} + w_1 \mathbf{b}_1 + w_2 \mathbf{b}_2$, where $\mathbf{b}_1$ and $\mathbf{b}_2$ are two orthogonal basis spanning the tangent space.

3) Gravity Refinement: The gravity vector obtained from the previous step can be refined by constraining the magnitude of the gravity vector. In most cases, the magnitude of the gravity vector is known. However, if we directly add this norm constraint into the optimization problem in (8), it will become nonlinear and hard to solve. Here, we use a method to enforce the gravity norm by optimizing the 2D error state on its tangent space. Since the magnitude of gravity is known, the degree of freedom of the gravity is two and we can parameterize the gravity with two variables on its tangent space. We parameterize the gravity as $g \cdot \bar{\hat{\mathbf{g}}} + w_1 \mathbf{b}_1 + w_2 \mathbf{b}_2$, where g is the magnitude of gravity, $\bar{\hat{\mathbf{g}}}$ is the direction vector of current estimation, $\mathbf{b}_1$ and $\mathbf{b}_2$ are two orthogonal basis spanning the tangent plane. w_1 and w_2 are the corresponding displacements towards $\mathbf{b}_1$ and $\mathbf{b}_2$, respectively. We can use Gram-Schmidt process to find one set of $\mathbf{b}_1$, $\mathbf{b}_2$ easily. In this way, we reparameterize gravity by two states on its tangent space. Then we substitute $\mathbf{g}$ in (7) by $g \cdot \bar{\hat{\mathbf{g}}} + w_1 \mathbf{b}_1 + w_2 \mathbf{b}_2$ and it is also in linear form. This process iterates several times until $\hat{\mathbf{g}}$ converges.

After refining gravity vector, we rotate all variables from visual base frame $(\cdot)^v$ to the world frame $(\cdot)^w$ according to the gravity vector. At this point, the initialization procedure is completed and these metric values will be fed for a tightly-coupled nonlinear visual-inertial estimator.

D. Nonlinear VINS Estimator

After obtaining all essential initial values, we can launch our tightly-coupled VINS estimator [2, 28]. Here, we briefly describe our graph optimization-based solution to the nonlinear visual-inertial system.

The definition of the full states in a sliding window with N IMU frames and M features are (the transpose is ignored):

$$\begin{aligned} \mathcal{X} &= \left[\mathbf{x}_0, \, \mathbf{x}_1, \, \cdots \, \mathbf{x}_n, \, \mathbf{x}^b_c, \, \lambda_0, \, \lambda_1, \, \cdots \, \lambda_m \right] \\ \mathbf{x}_k &= \left[\mathbf{p}^w_{b_k}, \, \mathbf{v}^w_{b_k}, \, \mathbf{q}^w_{b_k}, \, \mathbf{b}_a, \, \mathbf{b}_g \right], k \in [0, n] \\ \mathbf{x}^b_c &= \left[\mathbf{p}^b_c, \, \mathbf{q}^b_c \right], \end{aligned} \tag{9}$$

where the k-th IMU state consists of the position $\mathbf{p}^w_{b_k}$, velocity $\mathbf{v}^w_{b_k}$, orientation $\mathbf{q}^w_{b_k}$ of body frame b_k with respect to world frame w, and IMU bias $\mathbf{b}_a$, $\mathbf{b}_g$. 3D features are parameterized by their inverse depth λ when first observed in camera frame, and $\mathbf{x}^b_c$ is the extrinsic transformation from camera frame c to body frame b. The estimation is formulated as a nonlinear least-square problem,

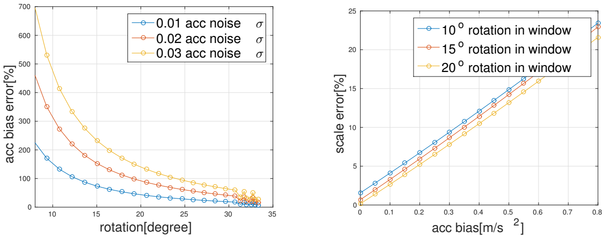
$$\min_{\mathcal{X}} \left\{ \sum_{k \in \mathcal{B}} \left\| r_{\mathcal{B}}(\hat{\mathbf{z}}^{b_k}_{b_{k+1}}, \mathcal{X}) \right\|^2_{\mathbf{P}^{b_k}_{b_{k+1}}} + \sum_{(l,j) \in \mathcal{C}} \left\| r_{\mathcal{C}}(\hat{\mathbf{z}}^{c_j}_{l}, \mathcal{X}) \right\|^2_{\mathbf{P}^{c_j}_{l}} \right\}, \tag{10}$$

where $r_{\mathcal{B}}(\hat{\mathbf{z}}^{b_k}_{bk+1}, \mathcal{X})$ and $r_{\mathcal{C}}(\hat{\mathbf{z}}^{c_j}_{l}, \mathcal{X})$ are nonlinear residual functions for inertial and visual measurements. $|| \cdot ||$ is the Mahalanobis distance weighted by covariance $\mathbf{P}$. To be specific, $r_{\mathcal{B}}$ is the residual of IMU factor which connects pair of consecutive frames b_k and b_{k+1} by the integration of inertial measurements $\hat{\mathbf{z}}^{b_k}_{b_{k+1}}$. $r_{\mathcal{C}}$ is the residual of vision factor which builds the connection between landmark measurements $\hat{\mathbf{z}}^{c_j}_{l}$ and states through re-projection function. The detailed optimization can be found at [28].

E. Discussions

To achieve full observability of the monocular VINS except for the global position shift and the yaw angle, sufficient excitation in both vision and IMU factor is required. The observability of the vision module can be ensured by selecting a number of spacial-separated frames which contains sufficient parallax. However, the IMU measurements within the SfM window may not render the whole system observable. For a rotorcraft MAV, degenerate motions such as rectilinear trajectories or zero-acceleration motions are unavoidable. This is a common issue when trying to use monocular VINS on aerial robots. Intuitively, we can reject small acceleration motion by checking the variation of $\hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}}$, $\hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}}$. The initialization procedure only starts when sufficient excitation is included in the windowed IMU measurements.

Acceleration bias is difficult to calibrate in initialization procedure, since acceleration is usually coupled with gravity under small rotation. To figure out the observability of acceleration bias along with movement, we design the following simulated experiment. In the simulation environment, the aerial robot does sufficient accelerated movement with different levels of rotation. The acceleration bias is constant [0.1,0.1,0.1]$m \cdot s^{-2}$ with noise whose standard deviation is from 0.01 to 0.03$m \cdot s^{-2}$. 15 spatially separated frames are kept in the window to extract the visual structure. Image noise is not included, which means an accurate visual structure. Also, gyroscope bias and noise are not included, which can eliminate the influence of gyroscope. To calibrate the acceleration bias in proposed framework, we can linearize position and velocity constraints $\hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}}, \hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}}$ with respect to acceleration bias, as the same step in eq. 4. Then take the acceleration bias into eq. 7. The calibration results of acceleration bias are shown in 3(a). The x-axis is the average rotation change along three axes in the window, and the y-axis is the magnitude of error of calibrated bias. From the figure, we can see that at least 30 degree rotation change is needed if we want to fully calibrated acceleration bias in such a short initialization procedure. In the real scenario, the more sufficient rotation is needed to fully calibrate acceleration bias in the beginning, which is infeasible in practice due to the dynamical constraints of the robotic platform.



(a) (b)

Fig. 3. (a) The x-axis represents the average rotation change in the visual window. The y-axis represents the estimation error of acceleration bias under different measurement noise. It can be seen that it is difficult to distinguish acceleration bias unless sufficient rotation in movement. (b) The x-axis represents different acceleration bias. The y-axis represents the estimated scale error of proposed algorithm without consideration of acceleration bias. It can be seen that limited acceleration bias will not severely destroy the algorithm. The margin of bearable acceleration bias is up to 0.3 m/s^{-2} in this simulation if we can accept 10% visual scale error in initial guess.

In another simulation, we test the performance of visual-inertial alignment under the influence of neglecting non-trivial acceleration bias. In this simulation, we take image noise (0.5 pixel in σ), gyroscope bias ([0.003,0.02,0.08]$rad \cdot s^{-1}$) and gyroscope noise (0.0024$rad \cdot s^{-1}$in σ) into consideration. The error of scale along with acceleration bias is shown in 3(b). The x-axis is the magnitude of acceleration bias, and the y-axis is the percentage error of scale. The influence caused by acceleration is linear in this scope. The tolerance of acceleration bias is up to 0.3$m \cdot s^{-2}$ if we can accept 10% percent scale error. 0.3$m \cdot s^{-2}$ bias is an extreme value for normal IMU in usual.

From the simulation, we can see that it is hard to distinguish acceleration bias from gravity unless sufficiently excited rotation is executed. This is hard to achieve in practice due to the dynamical constraints of the robotic platform. Neglecting acceleration bias will not pose significant negative impact on the initialization result. To this end, we leave the estimation of the acceleration bias estimation to the nonlinear optimization (Sect. IV-D).

V. Experimental Results

We first validate our algorithm with the publicly available MAV Visual-Inertial Datasets in the ASL Dataset [29]. Then we apply our method to real-world indoor and outdoor environments.

A. Performance on Public Datasets

The MAV Visual-Inertial Datasets in ASL Dataset are collected onboard a micro aerial vehicle. The dataset contains stereo images (Aptina MT9V034 global shutter, WVGA monochrome, 20 FPS), synchronized IMU measurements (ADIS16448, angular rate, and acceleration, 200 Hz), and ground truth states (VICON and Leica MS50). We only use one camera from stereo images set.

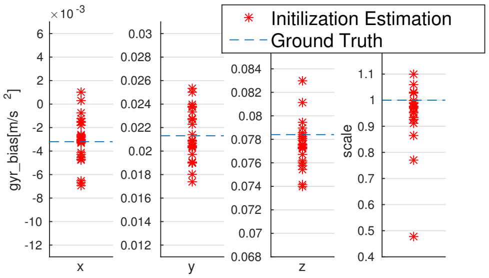
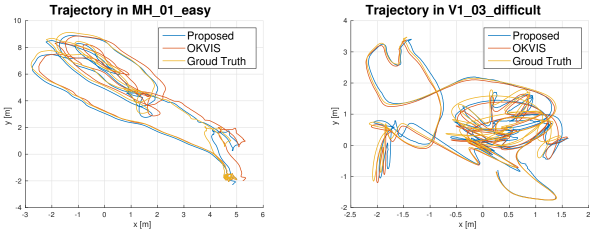


Fig. 4. Gyroscope bias and scale recovered in the initialization procedure in MH_01_easy dataset. The figure contains the results in 25 tests with different start time.



(a)

(b)

Fig. 5. Trajectory in MH_01_easy and V1_01_difficult respectively. Our proposed method is compared with state-of-art stereo visual-initial algorithm, OKVIS.

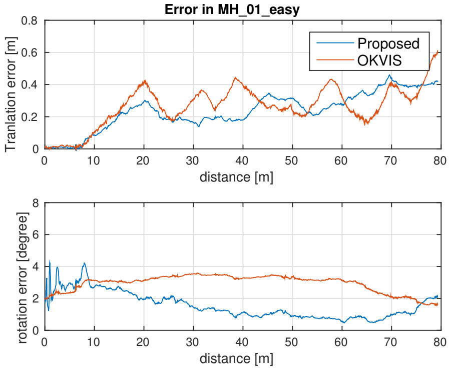


Fig. 6. Translation and rotation error in MH_01_easy.

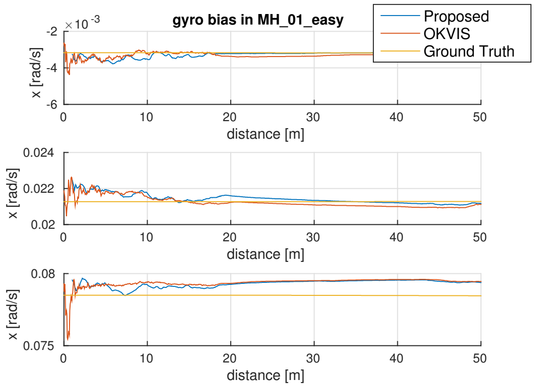


Fig. 7. Gyroscope bias in MH_01_easy.

1) Initial Values Recovery: We use the MH_01_easy dataset for evaluation. In this dataset, the gyroscope bias and accelerometer bias are approximately [-0.0032, 0.021, 0.078]$rad \cdot s^{-1}$ and [-0.0032, 0.026, 0.076]$m \cdot s^{-2}$ in x, y and z axes respectively, which are nontrivial. In our algorithm, we maintain at least 15 frames for initial visual structure. In general, the first few seconds are used for our initialization. We estimate gyroscope bias in the initialization phase, while leaving the accelerometer bias estimation to the following nonlinear optimization. To verify the capability of on-the-fly initialization, we randomly select start times in the dataset, which means our algorithm starts without any prior information when the drone is flying. Fig. 4 shows the gyroscope bias and scale calibration performance in 25 tests with the different start times in MH_01_easy dataset. The four sub-figures are gyroscope bias in xyz axis and scale respectively. The average error of gyroscope bias in two dominant direction x and y are [8.59,1.82]% respectively. And the average scale error is 8.09%. If we define the scale error less than 10% is successful initialization, our procedure performs 84% success rate in this dataset. In fact, the nonlinear estimator can be successfully bootstrapped even the initial scale error is over 30%.

2) Overall Performance: We choose two datasets, MH_01_easy, V1_03_difficult to show the whole performance of our visual-inertial odometry. In these experiments, we compare proposed method with OKVIS [6], which is the state-of-art visual-inertial algorithm working with stereo cameras. The whole trajectories are shown in Fig. 5. Our monocular system can achieve the same accuracy as the stereo system.

The error plot of MH_01_easy is shown in Fig. 6. The x-axis is distance, while the y-axis represents translation error and rotation error respectively. Proposed algorithm achieve nearly the same accurate result as the stereo system in translation. The gyroscope bias estimation is shown in Fig. 7. In the beginning, proposed method presents a stable gyroscope bias estimation because of a good initial guess from initialization procedure. Fig. 8 shows the acceleration bias estimation. Although we neglect the acceleration bias in the initialization procedure, the bias estimation converges gradually along with the movement.

Our algorithm also performs well in V1_03_difficult, which

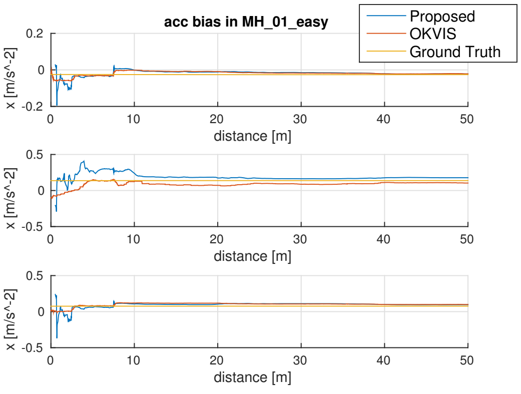


Fig. 8. Acceleration bias in MH_01_easy.

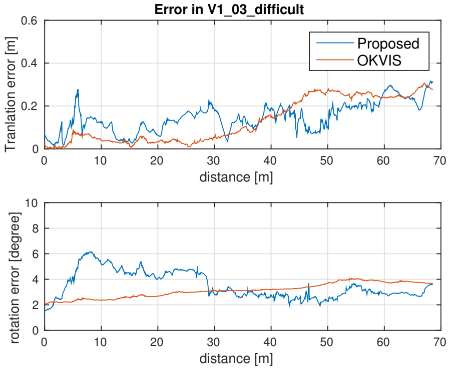


Fig. 9. Translation error and rotation error in V1_03_difficult.

is the most challenging dataset with aggressive motion and great illumination change. Few monocular system can survive in this dataset. The good initial guess from proposed method can compensate the negative influence from unstable feature tracking. The error plot compared with stereo OKVIS is shown in 9. Our system approximately achieves the same accuracy with this stereo algorithm.

B. *Real World Experiment*

We also test our algorithm onboard an aerial robot, as shown in Fig. 1(b). One forward-looking global shutter camera (MatrixVision mvBlueFOX-MLC200w) with 752×480 resolution. It is equipped with a 190-degree fisheye lens. A DJI A3 flight controller[2] is used both as the inertial measurement unit (IMU, ADXL278 and ADXRS290, 100Hz) and attitude stabilization control. The onboard computation resource includes an Intel i7-5500U CPU running at 3.00 GHz. The details about our platform and experimental can be found in the supplementary video.

1) Indoor Closed-Loop Control: We perform real-time closed-loop control in this indoor experiment. The visual-inertial odometry serves as position, attitude, and velocity feedback in the control loop. To test the dynamic initialization capability, we start the visual-inertial system when the aerial robot is flying in the air, instead of launching the system on the ground stably. After launching the visual-inertial system, we add the VINS odometry into the control loop. Finally, we switch the aerial robot into autonomous flight mode, which will follow a designed trajectory.

The trajectory is shown in Fig. 10. The position along with time is shown in Fig. 11. At first, we manually control the quadrotor flying. At 7.5s, we launch the visual-inertial system when the quadrotor is flying in the air. The estimator outputs the odometry after initialization within one second. At 19.0s, we add the visual-inertial estimator into control loop, which helps stabilize the quadrotor in the air. We manually control the quadrotor to the start point. Finally, we switch the aerial robot into autonomous flight mode at 31.0s. The drone autonomously flies, following the trajectory under the visual-inertial feedback.

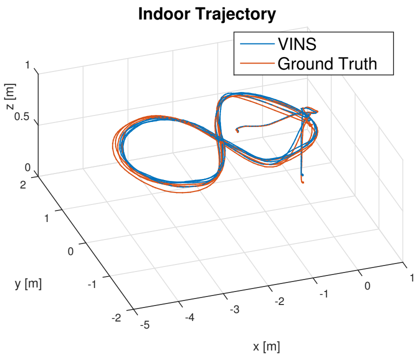


Fig. 10. The trajectory of the indoor onboard closed-loop experiment. At first, quadrotor is flying under manual control in the middle. After the visual-inertial system initialization on the fly, we manually control it to the boundary and switch to the autonomous flight mode. The quadrotor follows the designed trajectory. The designed trajectory is the figure of eight.

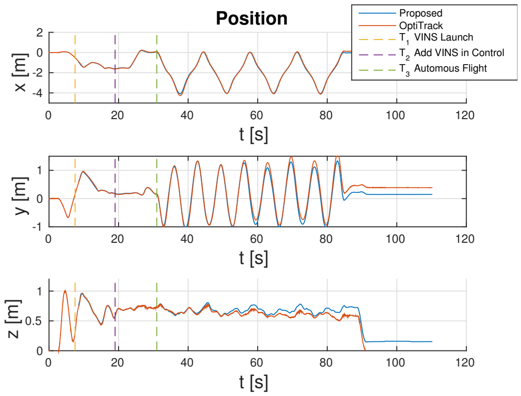


Fig. 11. Position of proposed system compared with OptiTrack in the indoor experiment. The three dash lines mean the time stamp of launching the visual-inertial system, adding odometry into control loop, and switching to autonomous flight.

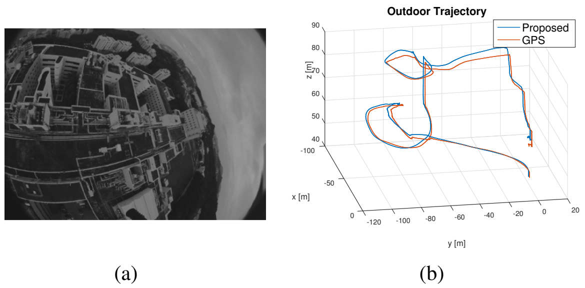
We compare our results with the ground truth which is provided by the OptiTrack system[3]. The blue line is the result from proposed estimator. The red line is the ground truth. Three dash lines represent three timestamps, when the proposed estimator launch, join into control loop, and the quadrotor starts autonomous flight respectively. The whole length in this indoor experiment is 58.12m. The final drift is [-0.13, -0.24, -0.16]m along x, y, and z respectively, which is 0.55% in percentage.

2) Outdoor Environments: Large scale environment is challenging for the monocular visual-inertial system. For visual measurement, a long movement which guarantees sufficient parallax between frames is required, which means a long time interval exists between spatial frames. On the contrary, the long time interval integration will seriously destroy the accuracy of IMU measurements. Also, sufficient excitation is required to fully recover the scale. However, when the drone flies smoothly in the high attitude, the IMU measurements will degenerate and output nearly zero readings besides the gravity.

To avoid long period integration, we maintain all the frames instead of only spatial frames in the initialization

[2] http://www.dji.com/a3

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



(a) (b)

Fig. 12. (a)The view of the drone flying at high altitude with 45^o tiled-front looking fisheye camera. (b)The trajectory of our proposed method compared with GPS.

procedure. Also, we reject degenerated movement and launch the proposed procedure only with sufficient excitation in inertial measurements. These adaptions make our algorithm perform in large scale environments.

We verify our algorithm in the $100m$ by $80m$ outdoor area with altitude ranging from $40m$ to $90m$. We initialize the visual-inertial system when the drone is flying at 40 meters high, as shown in Fig. 12. The trajectory is compared with GPS. The total length is 575m, and the final drift is [-3.03, 0.18, 2.07]m along x, y, and z respectively, which is 0.64% in percentage.

VI. Conclusion

In this paper, we propose a novel algorithm for the initialization of monocular visual-inertial estimators. Our initialization procedure provides initial guess (velocity, gravity vector, gyroscope bias, and depth of features) for nonlinear VINS estimator. These initial guesses are helpful to improve the performance of VINS by making it capable for on-the-fly initialization. We use real-world data in indoor closed-loop control and challenging outdoor environments to validate the practicability our proposed approach.

One potential drawback of proposed initialization procedure is that we neglect acceleration bias. We use simulation results to show that it is difficult to calibrate acceleration bias with smooth motion. Also, ignoring acceleration bias will not pose significant negative impact to other initial quantities. As shown in experiment, Fig. 8, the acceleration bias is gradually calibrated along with movement. In [25], the author tries to calibrate this bias in the initialization step. Dozens of seconds are needed until the bias converges, which usually is unbearable in real implementations. Also, no reliable criteria, which can indicate when acceleration is observable, is shown in [25]. As a future work, we would like to investigate an elegant initialization procedure, which takes acceleration bias into consideration.

References

[1] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Initialization-free monocular visual-inertial estimation with application to autonomous MAVs," in *Proc. of the Int. Sym. on Exp. Robot.*, Marrakech, Morocco, 2014.

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

[3] M. Faessler, F. Fontana, C. Forster, and D. Scaramuzza, "Automatic re-initialization and failure recovery for aggressive flight with a monocular vision-based quadrotor," in *Proc. of the IEEE Int. Conf. on Robot. and Autom.* IEEE, 2015, pp. 1722–1729.

[4] Z. Yang and S. Shen, "Monocular visual–inertial state estimation with online initialization and camera–imu extrinsic calibration," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 1, pp. 39–51, 2017.

[5] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst.* IEEE, 2015, pp. 298–304.

[6] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *Int. J. Robot. Research*, vol. 34, no. 3, pp. 314–334, Mar. 2014.

[7] Y. Ling, T. Liu, and S. Shen, "Aggressive quadrotor flight using dense visual-inertial fusion," in *Proc. of the IEEE Int. Conf. on Robot. and Autom.* IEEE, 2016, pp. 1499–1506.

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

[9] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an RGB-D camera," in *Proc. of the Int. Sym. of Robot. Research*, Flagstaff, AZ, Aug. 2011.

[10] S. Shen, N. Michael, and V. Kumar, "Autonomous indoor 3D exploration with a micro-aerial vehicle," in *Proc. of the IEEE Int. Conf. on Robot. and Autom.*, Saint Paul, MN, May 2012, pp. 9–15.

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

[12] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *J. Field Robot.*, vol. 27, no. 5, pp. 587–608, Sep. 2010.

[13] 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.

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

[15] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *Proc. of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst.* IEEE, 2013, pp. 3923–3929.

[16] M. Li and A. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *Int. J. Robot. Research*, vol. 32, no. 6, pp. 690–711, May 2013.

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

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

[19] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. of the IEEE Int. Conf. on Robot. and Autom.*, Hong Kong, China, May 2014.

[20] S. Weiss, R. Brockers, S. Albrektsen, and L. Matthies, "Inertial optical flow for throw-and-go micro air vehicles," in *Proc. of the IEEE Int. Conf. on Applications of Comput. Vis.* IEEE, 2015, pp. 262–269.

[21] 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.

[22] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," in *Proc. of Robot.: Sci. and Syst.*, Rome, Italy, Jul. 2015.

[23] A. Heyden and M. Pollefeys, "Multiple view geometry," *Emerging Topics in Computer Vision*, 2005.

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

[25] R. Mur-Artal and J. D. Tardos, “Visual-inertial monocular SLAM with map reuse,” *arXiv preprint arXiv:1610.05949*, 2016.

[26] D. Nistér, “An efficient solution to the five-point relative pose problem,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[27] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustmenta modern synthesis,” in *International workshop on vision algorithms*. Springer, 1999, pp. 298–372.

[28] Z. Yang and S. Shen, “Tightly-coupled visual-inertial sensor fusion based on IMU pre-integration,” Hong Kong University of Science and Technology, Tech. Rep., 2016, URL: http://www.ece.ust.hk/~eeshaojie/vins2016zhenfei.pdf.

[29] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, “The euroc micro aerial vehicle datasets,” *The International Journal of Robotics Research*, 2016.