2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Daejeon Convention Center
October 9-14, 2016, Daejeon, Korea

Self-Calibrating Multi-Camera Visual-Inertial Fusion for Autonomous MAVs

Zhenfei Yang, Tianbo Liu, and Shaojie Shen

***Abstract*— We address the important problem of achieving robust and easy-to-deploy visual state estimation for micro aerial vehicles (MAVs) operating in complex environments. We use a sensor suite consisting of multiple cameras and an IMU to maximize perceptual awareness of the surroundings and provide sufficient redundancy against sensor failures. Our approach starts with an online initialization procedure that simultaneously estimates the transformation between each camera and the IMU, as well as the initial velocity and attitude of the platform, without any prior knowledge about the mechanical configuration of the sensor suite. Based on the initial calibrations, a tightly-coupled, optimization-based, generalized multi-camera-inertial fusion method runs onboard the MAV with online camera-IMU calibration refinement and identification of sensor failures. Our approach dynamically configures the system into monocular, stereo, or other multi-camera visual-inertial settings, with their respective perceptual advantages, based on the availability of visual measurements. We show that even under random camera failures, our method can be used for feedback control of the MAVs. We highlight our approach in challenging indoor-outdoor navigation tasks with large variations in vehicle height and speed, scene depth, and illumination.**

I. Introduction

There has been increasing interests in the robotics community and industry to provide miniaturized consumer aerial robots with autonomy in GPS-denied environments. Visual-inertial systems (VINSs), which consist of a low-cost MEMS IMU and cameras, have been very attractive sensor choices due to their superior size, weight, and power (SWaP) characteristics. Lots of efforts have been made toward using monocular VINSs with successful applications in autonomous flight of MAVs [1, 2] and smartphone localization [3, 4].

However, monocular VINSs have the fundamental disadvantage of scale degeneracy in rotation-only or constant velocity motions due to the lack of direct distance measurements. Naturally, stereo [5] VINSs may solve the problem by imposing rigid stereo constraints. However, the performance of stereo-based systems still largely depends on the accuracy of extrinsic stereo calibration and existence of nearby (large baseline-to-depth ratio) features. No simple camera setup is able to achieve the required level of robustness during complex missions. Extending to multi-camera VINSs can potentially provide sufficient redundancy in complex tasks, but one still has to overcome challenges in extrinsic calibration.

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

This work was supported by HKUST project R9341. The authors would also like to thank the equipment support from DJI.

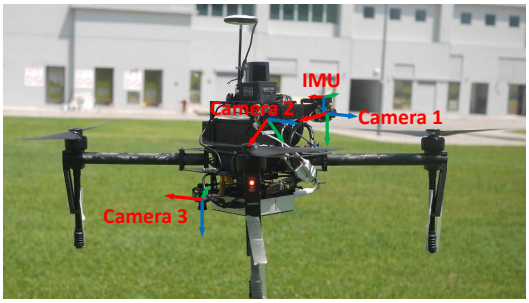


Fig. 1. Our platform, equipped with three cameras and an IMU with unknown camera-IMU extrinsic calibration. Colored coordinate axes are plotted for all sensors (Red: x axis, Green: y axis, Blue: z axis). Note that camera 2 and 3 are set up in a non-orthogonal configuration intentionally. Both rotation and translation offsets are unknown to the estimator and have to be calibrated online. A video of the experiment can be found at http://www.ece.ust.hk/~eeshaojie/iros2016zhenfei.mp4

Careful handling of distanced features is also crucial to make the best use of visual information and overcome cases where rigid stereo configurations fail.

In this work, we address the problem of robust and easy-to-use multi-camera VINS state estimation from the ground up. We realize the necessity to allow nonprofessional users to reconfigure the onboard sensor on-site to adapt to mission requirements. Therefore, we propose an approach to perform extrinsic calibration of the cameras and IMU online, without any artificial calibration objects and without any prior knowledge of the mechanical configuration of the system. To make the best use of the information from multiple visual sensors, we propose the concept of generalized multi-camera-inertial fusion using a tightly-coupled, optimization-based framework. With only one camera, the system is converted into a monocular VINS with online calibration, as in our earlier work [6], which works well given sufficient motion excitation. With multiple non-overlapping cameras, it can be treated as multiple monocular VINSs, or considered a metric scale recovery problem using the continuously refined camera-IMU calibrations [7]. In the case of two or more cameras with a nonzero baseline sharing an overlapping field-of-view, our system naturally, and implicitly, turns into a stereo-based approach that works without any motion, provided with nearby features for stereo triangulation [5]. Our system dynamically switches between the above cases depending on the availability of visual measurements, thus making the best use of all visual information. It becomes clear that our approach naturally provides redundancy in state estimation and is robust against sensor failures.

This work is built on our earlier works [1, 6, 8][1], with

[1]Early access to [6] can be found at http://www.ece.ust.hk/~eeshaojie/ssrr2015zhenfei.pdf

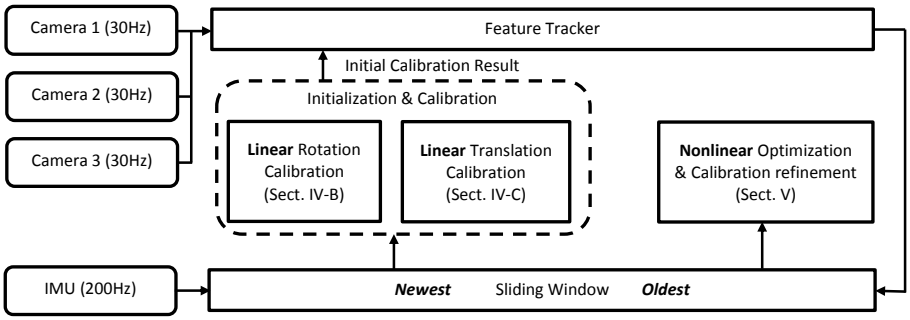


Fig. 2. A block diagram illustrating the full pipeline of the proposed approach.

extension to the generalized multi-camera configuration. We present online experiments to show that our method achieves accurate state estimation and extrinsic camera-IMU calibration. In particular, we demonstrate the necessity of our multi-camera fusion mechanism in challenging indoor-outdoor navigation tasks with large variations in vehicle height and speed, scene depth, and illumination. It represents a substantial step towards robust plug-and-play sensory systems for autonomous flight. Hereby, we identify the contributions of this work as follows:

- An initialization procedure that is able to recover camera-IMU calibrations as well as initial velocity, attitude, and visual scale without any prior knowledge of the mechanical configuration and without any artificial calibration objects.
- A tightly-coupled, optimization-based, and self-calibrating generalized multi-camera-inertial fusion framework that makes the best use of all visual information by implicitly imposing both spatial and temporal feature matching and triangulation from multiple cameras.
- Extensive evaluation in challenging real-world experiments.

The rest of this paper is organized as follows: In Sect. II, we discuss the relevant literatures. We give an overview of the complete system pipeline in Sect. III. , and detail our linear initialization and camera-IMU calibration procedure in Sect. IV. A tightly-coupled, nonlinear optimization-based, generalized multi-camera VINS estimator is presented in Sect. V. We discuss implementation details and present experimental results in Sect. VI. The paper is concluded with a discussion of possible future directions in Sect. VII.

II. Related Work

The scholarly works on visual-inertial state estimation with different sensor configurations, such as monocular [2]–[4], stereo [5], or RGB-D cameras [9], are extensive. Popular mathematical tools for solving VINSs are filtering-based methods [2]–[4, 10], or graph optimization-/bundle adjustment-based methods [1, 5, 11]. Filtering approaches usually save computational resources, while optimization-based approaches may enjoy higher accuracy using iterative re-linearization. VINSs are also realized in loosely-coupled [2] or tightly-coupled [3]–[5, 10, 11] methods, depending on trade-offs between complexity and accuracy. Due to space limitations, we refer readers to a more comprehensive review of related works in [12]. We formulate our method in a tightly-coupled, graph optimization-based framework, as it achieves higher accuracy and offers more flexibility when incorporating different types of sensor measurements.

It is well known that the performance of VINSs relies on accurate initialization of navigational state and camera-IMU extrinsic calibration. The initialization problem has been addressed using both geometric [13]–[15] and probabilistic methods [8, 16]. A simultaneous initialization and calibration approach for monocular VINSs is proposed in our earlier work [6], which is further extended in this work for handling multi-camera VINS calibration.

For multi-camera systems without an IMU, [17, 18] explore inter-camera calibration of a camera rig by running SLAM on each camera independently, and then with the help of wheel odometry, obtain the inter-camera calibration throughout map merging. [19] aims to operate at high altitude with a small baseline-to-depth ratio for visual features. A constrained bundle adjustment is proposed to utilize a prior, approximate stereo extrinsic calibration. However, the performance of scale recovery still heavily relies on the accuracy of the prior extrinsic calibration.

Most similar to our work is the self-calibrating multi-camera system proposed in [20]. However, [20] requires that cameras are formed as calibrated stereo pairs, while in our approach we do not impose any restrictions on the configuration of the cameras. Our system functions properly even without any stereo feature observations.

III. Overview

Our proposed multi-camera visual-inertial system consists of three phases, as illustrated in Fig. 2. The first two phases (Sect. IV-B and Sect. IV-C) aim to initialize the estimator in a linear fashion, as well as to get initial values of camera-IMU calibrations without any prior knowledge of the mechanical configuration of the sensor suite. In phase 3 (Sect. V), a tightly-coupled state estimator using nonlinear optimization is performed, utilizing the initial values from previous phases. Our initialization procedure also provides enough information about the spatial configuration of cameras, which is the basis on which to build our generalized multi-camera model.

We begin by defining notations. We consider $(\cdot)^w$ as the earth's inertial frame, $(\cdot)^b$ as the current IMU body

frame, and $(\cdot)^{c_k^i}$ as the body frame of camera i while taking the k^{th} image. We further note $(\cdot)^{b_k}$ as the IMU body frame while the camera is taking the k^{th} image. The IMU runs at a much higher rate than the cameras, and multiple IMU measurements may exist in the interval $[k, k+1]$. IMU measurements are combined together using IMU pre-integration (Sect. IV-A). $\mathbf{p}_Y^X$, $\mathbf{v}_Y^X$, and $\mathbf{R}_Y^X$ are the 3D position, velocity, and rotation of frame Y with respect to frame X. We also use quaternions ($\mathbf{q} = [q_x, q_y, q_z, q_w]$), following Hamilton notation, to represent rotation. With a slight abuse of notation, $\mathbf{p}_t^X$ represents the position of the IMU body frame at time t with respect to frame X. A similar convention follows for other parameters. The transformation between the i^{th} camera and the IMU is an unknown constant that we denote as $(\mathbf{p}_{c^i}^b, \mathbf{R}_{c^i}^b)$. $\mathbf{g}^w = [0, 0, g]^{\mathrm{T}}$ is the gravity vector in the world frame, and $\mathbf{g}^{b_k}$ is the gravity vector expressed in the IMU body frame during the k^{th} image capture. We assume that all sensors are synchronized, where all cameras capture images at the same time, with one IMU measurement matching the time instance.

IV. Estimator Initialization and Camera-IMU Extrinsic Calibration

Here we present an initialization approach, which builds on top of our prior work with monocular VINS [6], to simultaneously recover velocity, attitude (gravity vector), depth of features, and camera-IMU calibrations. Here we execute the initialization procedure as multiple independent monocular VINS settings. Note that we cannot incorporate inter-camera feature matching at this stage as we do not know how the cameras are mechanically configured.

A. IMU Pre-Integration

Given two time instants corresponding to two image captures, the IMU propagation model for position and velocity in the earth's inertial frame can be written 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 from the IMU, and Δt is the time difference $[k, k+1]$ between two image captures. In (1), the rotation between the world frame and the body frame is required for state propagation. Knowledge about the initial attitude is required to recover this global rotation. However, as introduced in [16], if the reference frame for IMU propagation is changed to the first state of interest (i.e., the first block of pose, velocity, and attitude that we are trying to estimate), (1) can be rewritten as follows:

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

where

$$\begin{aligned}\boldsymbol{\alpha}_{b_{k+1}}^{b_k} &= \iint_{t\in[k,k+1]} \mathbf{R}_t^{b_k} \mathbf{a}_t^b dt^2 \\ \boldsymbol{\beta}_{b_{k+1}}^{b_k} &= \int_{t\in[k,k+1]} \mathbf{R}_t^{b_k} \mathbf{a}_t^b dt\end{aligned} \tag{3}$$

can be obtained solely with IMU measurements within $[k, k+1]$. $\mathbf{R}_{b_k}^{b_0}$ is the change in rotation since the first state (the 0^{th} image), and $\mathbf{R}_{b_{k+1}}^{b_k}$ is the incremental rotation between two image captures. The dependency on global orientation is removed, and multiple IMU measurements can be integrated without any prior state initialization, enabling efficient fusion of high-rate IMU measurements with lower-rate visual measurements (Sect. V). It also enables linear estimator initialization as discussed in Sect. IV-C.

This technique is known as IMU pre-integration, which was first proposed in [16], and is advanced to consider on-manifold uncertainty propagation in [1, 12]. Further improvement to incorporate IMU biases and integrate with full SLAM framework is proposed in [21].

B. Initialization of Camera-IMU Rotation

The initialization of camera-IMU rotation can be treated as a linear hand-eye calibration problem which aims to align rotation sequences obtained from two sensors. The classic 5-point algorithm [22] with RANSAC-based outlier rejection is used to estimate the incremental rotation $\mathbf{R}_{c_{k+1}^i}^{c_k^i}$ between successive images k and $k+1$. Corresponding incremental rotation $\mathbf{R}_{b_{k+1}}^{b_k}$ of the IMU is obtained by short-term integration of gyroscope measurements.

The following equation for the i^{th} camera holds:

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

We can write (4) as linear equations using a quaternion representation for rotation:

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

where $\mathcal{Q}_1(\mathbf{q}_{b_{k+1}}^{b_k})$ and $\mathcal{Q}_2(\mathbf{q}_{c_{k+1}^i}^{c_k^i})$ represent the matrix multiplication form of quaternion multiplication. Stacking constraints from multiple time intervals forms an over-constrained linear system. Due to space limitation, we refer readers to [6] for details of outlier-resistant solutions to (5) as well as termination criteria for the procedure.

C. Initialization of Velocity, Attitude, and Camera-IMU Translation

Based on the camera-IMU rotation obtained in Sect. IV-B, we extend the linear sliding window estimator proposed in [6] to our multi-camera system, for simultaneous initialization of velocity, attitude, and camera-IMU translations ($\mathbf{p}_{c^1}^b \cdots \mathbf{p}_{c^L}^b$). The state vector is written as follows (the transpose is ignored for simplicity of presentation):

$$\begin{aligned}\mathcal{X} &= \left[\mathbf{x}_n \cdots \mathbf{x}_{n+N},\, \mathbf{p}_{c^1}^b \cdots \mathbf{p}_{c^L}^b,\, d_m \cdots d_{m+M}\right] \\ \mathbf{x}_k &= \left[\mathbf{p}_{b_k}^{b_0},\, \mathbf{v}_{b_k}^{b_k},\, \mathbf{g}^{b_k}\right],\end{aligned} \tag{6}$$

where N is the number of IMU states in the sliding window, L is the number of cameras, and M is the number of features that have sufficient parallax within the sliding window, n and m are starting indexes of the sliding window, $\mathbf{x}_k$ is the k^{th} IMU state, and d_l is the depth of the l^{th} feature in the camera frame of its first observation.

We solve for the maximum likelihood estimation of the system by formulating a linear system using all inertial and visual measurements within the sliding window:

$$\min_{\mathcal{X}} \Bigg\{ \|\mathbf{r}_p - \mathbf{\Lambda}_p \mathcal{X}\|^2 + \sum_{k\in\mathcal{B}} \left\| \hat{\mathbf{z}}_{b_{k+1}}^{b_k} - \mathbf{H}_{b_{k+1}}^{b_k} \mathcal{X} \right\|^2_{\mathbf{P}_{b_{k+1}}^{b_k}} + \sum_{i\in\mathcal{L}} \sum_{(l,j)\in\mathcal{C}^i} \left\| \hat{\mathbf{z}}_l^{c_j^i} - \mathbf{H}_l^{c_j^i} \mathcal{X} \right\|^2_{\mathbf{P}_l^{c_j^i}} \Bigg\}, \tag{7}$$

where $\mathcal{B}$ is the set of all IMU measurements, $\mathcal{L}$ is the set of cameras. $\mathcal{C}^i$ is the set of all observations by camera i, and j, l are time and feature indexes, respectively. Detailed derivation of the measurement matrices $\mathbf{H}_{b_{k+1}}^{b_k}$, $\mathbf{H}_l^{c_j^i}$, the (optional) prior matrix $\mathbf{\Lambda}_p$, and the covariance $\mathbf{P}_{b_{k+1}}^{b_k}$, $\mathbf{P}_l^{c_j^i}$ can be found in [6].

Note that (7) does not solve for any rotational components besides implicit attitude initialization from the gravity vector. Given the camera-IMU rotations $\mathbf{R}_{c^i}^b$ from Sect. IV-B, and incremental rotations $\mathbf{R}_{b_{k+1}}^{b_k}$ from the short-term gyroscope integration, we can solve all quantities in the state vector (6) in a linear fashion, as suggested by the IMU pre-integration formulation (2). To this end, the initialization procedure is exact, linear, and able to be solved in a non-iterative way. This is in contrast to the nonlinear formulation in Sect. V, which can only be solved iteratively. We refer readers to [6] for the information-based termination criteria for the initialization procedure.

V. Tightly-Coupled Nonlinear Optimization with Online Calibration Refinement

We proceed with a nonlinear estimator for high accuracy state estimation and multi-camera-IMU calibration refinement. This is an extension of our earlier work [1] by processing multi-camera measurements in a unified way.

A. Generalized Multi-Camera Visual-Inertial System

Our initialization procedure (Sect. IV) recovers the mechanical configuration of each camera with respect to the IMU. Based on this, we can establish both intra- (temporal) and inter- (spatial) camera feature tracking, depending on the relative pose between cameras. In our formulation, both temporal and spatial feature tracking are formulated in a unified way (Sect. V-C). Intuitively, cameras with overlapping fields-of-view enable spatial triangulation of features. Our formulation takes this into account and turns it into implicit stereo constraints, thus allowing the estimator to observe the visual scale without any motion. The online camera-IMU calibration refinement improves the "stereo rigidity" as more measurements are incorporated into the system.

On the other hand, if there are no overlapping fields-of-view between cameras or features are too far away, the system will degenerate into multiple monocular VINS configurations. Fortunately, due to the tightly-coupled fusion of visual and inertial measurements, our approach still allows metric scale recovery, subject to certain motion requirements. We further utilize the two-way marginalization scheme proposed in [1] for handling degenerate hovering motion. Our formulation is in contrast to that in [19] and [20], where stereo triangulation is necessary to have the system be operational. The multi-camera model is illustrated in Fig. 3.

B. Formulation of Nonlinear Sliding Window Estimator

The definition of the full state vector is similar to the linear case, with a few exceptions: 1) full 6-degree-of-freedom (DOF) camera-IMU transformations $\mathbf{x}_{c^i}^b$ are included in the state vector; 2) the gravity vector is replaced with the quaternion $\mathbf{q}_{b_j}^{b_i}$ for joint optimization of translation and rotation; and 3) inverse depth formulation is used, where $\lambda_l = 1/d_l$.

$$\begin{aligned} \mathcal{X} &= \left[\mathbf{x}_n \cdots \mathbf{x}_{n+N},\, \mathbf{x}_{c^1}^b \cdots \mathbf{x}_{c^L}^b,\, \lambda_m \cdots \lambda_{m+M}\right] \\ \mathbf{x}_k &= \left[\mathbf{p}_{b_k}^{b_0},\, \mathbf{v}^{b_k},\, \mathbf{q}_{b_k}^{b_0}\right], \quad \mathbf{x}_{c^i}^b = \left[\mathbf{p}_{c^i}^b,\, \mathbf{q}_{c^i}^b\right] \end{aligned} \tag{8}$$

We solve for the maximum a posteriori estimation using all measurements:

$$\min_{\mathcal{X}} \Bigg\{ \|\mathbf{r}_p - \mathbf{\Lambda}_p \mathcal{X}\|^2 + \sum_{k\in\mathcal{B}} \left\| r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k},\, \mathcal{X}) \right\|^2_{\mathbf{P}_{b_{k+1}}^{b_k}} + \sum_{i\in\mathcal{L}} \sum_{(l,j)\in\mathcal{C}^i} \left\| r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j^i},\, \mathcal{X}) \right\|^2_{\mathbf{P}_l^{c_j^i}} \Bigg\}, \tag{9}$$

where $r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k},\, \mathcal{X})$ and $r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j^i},\, \mathcal{X})$ are measurement residuals for inertial and visual measurements, respectively. We define the generalized multi-camera measurement model in Sect. V-C. The IMU measurement model is defined following the spirit of IMU pre-integration, as in Sect. IV-A. We refer readers to our prior work [1] for detailed derivation of IMU pre-integration with on-manifold uncertainty propagation.

We use error-state representation for linearizing the nonlinear system (9) and solving it using the Gauss-Newton method. The residual for position, velocity, and camera-IMU translations can be trivially defined as

$$\mathbf{p} = \hat{\mathbf{p}} + \delta\mathbf{p}, \quad \mathbf{v} = \hat{\mathbf{v}} + \delta\mathbf{v}, \quad \lambda = \hat{\lambda} + \delta\lambda. \tag{10}$$

The definition of the rotation residual is more involved, and it is formulated using an on-manifold formulation similar to that in [5]:

$$\mathbf{q} = \hat{\mathbf{q}} \otimes \delta\mathbf{q}, \quad \delta\mathbf{q} \approx \begin{bmatrix} \frac{1}{2}\delta\boldsymbol{\theta} \\ 1 \end{bmatrix}, \tag{11}$$

where $\otimes$ is quaternion multiplication and $\delta\boldsymbol{\theta}$ is the minimum-dimensional representation of the rotation residual. The full error-state vector then becomes

$$\begin{aligned} \delta\mathcal{X} &= \left[\delta\mathbf{x}_n \cdots \delta\mathbf{x}_{n+N},\, \delta\mathbf{x}_{c^1}^b \cdots \delta\mathbf{x}_{c^L}^b,\, \delta\lambda_m \cdots \delta\lambda_{m+M}\right] \\ \delta\mathbf{x}_k &= \left[\delta\mathbf{p}_{b_k}^{b_0},\, \delta\mathbf{v}^{b_k},\, \delta\boldsymbol{\theta}_{b_k}^{b_0}\right], \quad \delta\mathbf{x}_{c^i}^b = \left[\delta\mathbf{p}_{c^i}^b,\, \delta\boldsymbol{\theta}_{c^i}^b\right], \end{aligned} \tag{12}$$

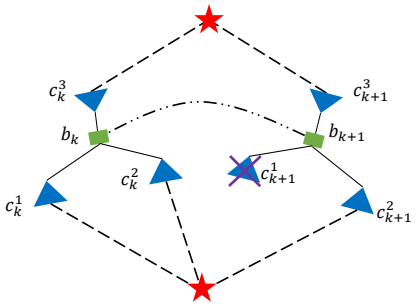


Fig. 3. A system with an IMU (green), together with a pair of forward-facing cameras (blue) and a backward-facing camera (blue). Both temporal and spatial visual constraints are extracted by a front-end feature tracker and incorporated into the MAP estimator (Sect. V-B) using the generalized multi-camera measurement model (Sect. V-C).

and the linearized cost function with respect to the current state estimate $\hat{\mathcal{X}}$ can be written as

$$\min_{\delta\mathcal{X}} \left\{ \left\| \mathbf{r}_p - \mathbf{\Lambda}_p \hat{\mathcal{X}} \right\|^2 + \sum_{k\in\mathcal{D}} \left\| r_{\mathcal{D}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k}, \hat{\mathcal{X}}) + \mathbf{G}_{b_{k+1}}^{b_k} \delta\mathcal{X} \right\|_{\mathbf{P}_{b_{k+1}}^{b_k}}^2 + \sum_{i\in\mathcal{L}} \sum_{(l,j)\in\mathcal{C}^i} \left\| r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j^i}, \hat{\mathcal{X}}) + \mathbf{G}_l^{c_j^i} \delta\mathcal{X} \right\|_{\mathbf{P}_l^{c_j^i}}^2 \right\}, \tag{13}$$

where $\mathbf{G}_l^{c_j^i}$ is the Jacobian of visual measurements, which will be defined in Sect. V-C. For each Gauss-Newton iteration, (13) is solved as a linear system in the following form:

$$\left(\mathbf{\Lambda}_p + \mathbf{\Lambda}_{\mathcal{B}} + \mathbf{\Lambda}_{\mathcal{C}}\right) \delta\mathcal{X} = \left(\mathbf{b}_p + \mathbf{b}_{\mathcal{B}} + \mathbf{b}_{\mathcal{C}}\right), \tag{14}$$

where $\mathbf{\Lambda}_p$, $\mathbf{\Lambda}_{\mathcal{B}}$, and $\mathbf{\Lambda}_{\mathcal{C}}$ are information matrices from prior, inertial, and visual measurements, respectively. The error-state is updated as

$$\hat{\mathcal{X}} \leftarrow \hat{\mathcal{X}} \oplus \delta\mathcal{X}, \tag{15}$$

where $\oplus$ can be either vector addition or quaternion multiplication, depending on the specific residual definition. In fact, (11) and (15) define the logarithm and exponential maps for rotational components.

C. Generalized Multi-Camera Measurement Model

Based on the discussion in Sect. V-A, we present a generalized multi-camera model that formulates both temporal feature matching with the same camera across different frames, and spatial feature matching between different cameras at the same time instant. The two kinds of matching scheme can mixed together to form redundancy against failures of arbitrary cameras, as shown in Fig. 3.

$$\begin{aligned} r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j^i}, \mathcal{X}) &= \pi\left(\mathbf{f}_l^{c_j^i}, \begin{bmatrix} \hat{u}_l^{c_j^i} \\ \hat{v}_l^{c_j^i} \end{bmatrix} \right) \\ \mathbf{f}_l^{c_j^i} &= g_{c^i}^{b^{-1}} \left(g_{b_q}^{b_j} \left(g_{c^p}^{b} \left(\frac{1}{\lambda_l} \begin{bmatrix} u_l^{c_q^p} \\ v_l^{c_q^p} \\ 1 \end{bmatrix} \right) \right) \right), \end{aligned} \tag{16}$$

where $[\hat{u}_l^{c_j^i}, \hat{v}_l^{c_j^i}]^{\mathrm{T}}$ represents the feature observation in the normalized image coordinates of the l^{th} feature in the j^{th} image captured by the i^{th} camera, $[u_l^{c_q^p}, v_l^{c_q^p}]^{\mathrm{T}}$ represents the noiseless first observation of the l^{th} feature, which occurs at the q^{th} image captured by the p^{th} camera, $\pi(\cdot)$ is the projection function, and $g_b^a(\cdot)$ is the invertible frame transform function:

$$g_b^a(\mathbf{f}) = \mathbf{R}_b^a \cdot \mathbf{f}^b + \mathbf{p}_b^a. \tag{17}$$

The Jacobian matrix $\mathbf{G}_l^{c_j^i}$ varies depending on whether the observation involves a feature first observed by another camera. To this end, temporal feature matching ($i = p$) has the Jacobian matrix in the following form:

$$\mathbf{G}_l^{c_j^i} = \frac{\partial \pi}{\partial \mathbf{f}_l^{c_j^i}} \left[\frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_q}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_j}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_{c^i}^b}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\lambda_l} \right]. \tag{18}$$

On the other hand, if a feature is first observed by another camera ($i \neq p$), which represents temporal-spatial feature matching, the Jacobian matrix involves two camera-IMU transformations:

$$\mathbf{G}_l^{c_j^i} = \frac{\partial \pi}{\partial \mathbf{f}_l^{c_j^i}} \left[\frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_q}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_j}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_{c^i}^b}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\mathbf{x}_{c^p}^b}, \frac{\partial \mathbf{f}_l^{c_j^i}}{\partial \delta\lambda_l} \right]. \tag{19}$$

The temporal-spatial feature matching, if exists, implicitly forms stereo constraints, thus allowing the estimator to operate without any motion excitation.

VI. Experimental Results

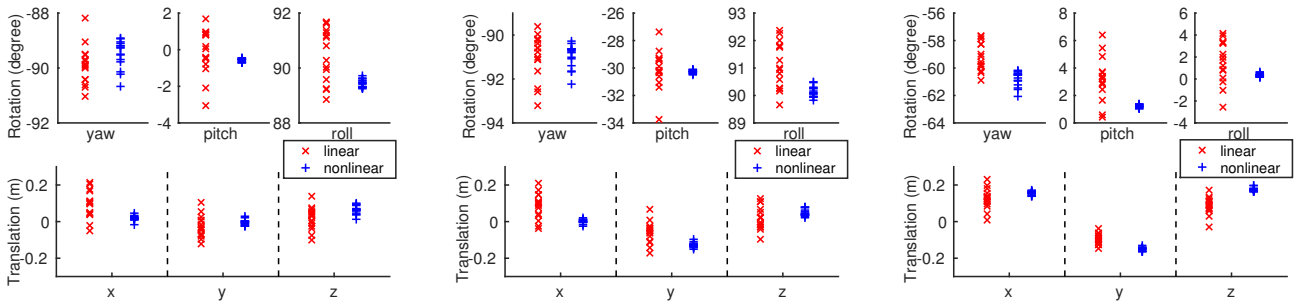
A. Implementation Details

We choose DJI Matrice 100[2] as our flight platform. We equip the platform with a Microstrain 3DM-GX4 IMU and three mvBlueFox-MLC200w grayscale HDR cameras with wide-angle lenses, as shown in Fig. 1. Cameras 1 and 2 are forward-facing, while camera 3 faces downward. In order to test the calibration performance, cameras 2 and 3 are intentionally set up in non-orthogonal configurations. The three cameras have significant translational offset with respect to the IMU.

Our algorithm runs real-time onboard an Intel NUC computer (Intel i5-4250U, 16GB RAM) with a multi-thread implementation.

The first thread performs feature tracking all the time at 30 Hz, and spatial feature matching at 10 Hz after initial camera-IMU calibration is obtained (Sect. IV). For each new image, we track existing features temporally using the KLT optical flow. Every three images, we detect new FAST features, maintaining a maximum of 100 points with at least 30-pixel separation for each camera, match features between cameras spatially using the KLT optical flow, and then supply the result to the estimator. In the estimator, only features with at least 10-pixel rotation-compensated parallax are used. We set a new keyframe in the estimator if the average parallax is larger than 10 pixels.

[2] https://developer.dji.com/matrice-100/



(a) Camera 1 (b) Camera 2 (c) Camera 3

Fig. 4. Calibration results given by linear initialization (Sect. IV) and nonlinear refinement (Sect. V).

The second thread runs the proposed 3-phase approach: rotation initialization (Sect. IV-B), online initialization (Sect. IV-C), and nonlinear optimization (Sect. V).

An additional thread is used for determination of the initialization termination criteria at 2 Hz (Sect. IV-C).

B. Camera-IMU Calibration Performance

In this experiment, we demonstrate the performance of our online camera-IMU calibration method. As shown in Fig. 1, we intentionally rotate cameras 2 and 3 for an unknown fixed angle. The performance of the proposed method is verified by convergence and repeatability across multiple trials.

We conduct 15 trials in a typical office environment with only natural features. For each trial, we start our approach without any prior knowledge of the mechanical configuration. We walk along in a random path holding the sensor suite by hand and start the estimator from a non-stationary state. The linear calibration of rotation (Sect. IV-B) and translation (Sect. IV-C) are recorded as the terminating values of phases 1 and 2, respectively.

Fig. 4 summarizes the results from all 15 trials. It can be clearly seen that the linear calibration (Sect. IV) provides reasonable initial values, and the nonlinear optimization (Sect. V) continuously refines the results. For all of the calibration trials, the results converge to a small range of $[-0.02, 0.02]$ (units in meters) for translation and $[-1, 1]$ (units in degrees) for rotation.

This suggests that the proposed system can function properly with customized sensor configurations without prior initialization.

C. Autonomous Flight with Random Camera Failures

In this experiment, we show the robustness of our approach under random camera failures. Camera failure here is used to simulate common situations during autonomous flight where visual observations are not available, for example, facing texture-less scene, under extreme illumination condition, and covered by moving objects. The flight is performed autonomously using onboard state estimates. We use a standard PD controller to track a pre-generated trajectory at an average speed of 1.5 m/s. We change the sensor combination during flight by intentionally turning off some of the cameras using external commands. Note that the controller is not aware of the changes in sensing modality. The performance of the proposed method is evaluated with a motion capture system. As illustrated in Fig. 5, our onboard estimation compares consistently well with the ground truth during the whole experiment, even with sensor failures. In particular, the standard deviation of the velocity error is $[0.0283, 0.0327, 0.0229]$ (units in m/s) in the x, y, and z axes, respectively.

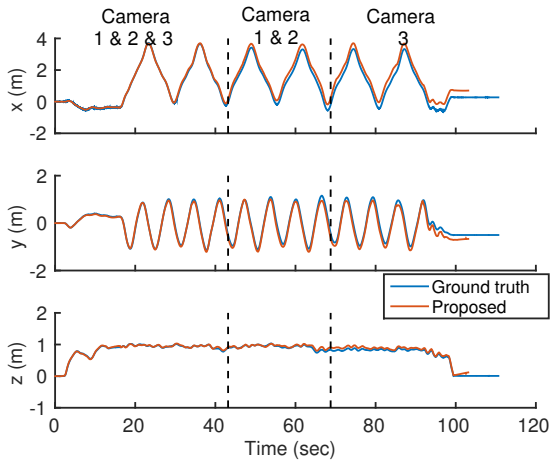
D. Performance in Large-Scale Complex Environments

In this section, we evaluate the accuracy and robustness of the proposed approach in large-scale complex environments. The testing spans a variety of challenging cases for VINSs, including a narrow passageway (Fig. 6(a)), open space (Fig. 6(b)), high-speed flight (Fig. 6(c)), and large rotation (Fig. 6(d)).

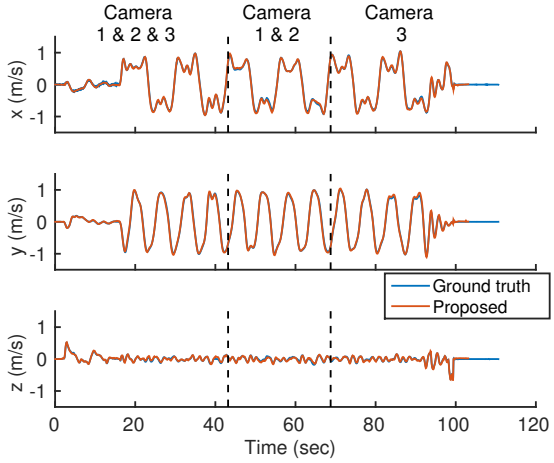
The total flight time is approximately 8 minutes, and the vehicle travels 642 meters, with a highest speed of 8.9 m/s and a height range of $[-6.6, 4.5]$ (units in meters). The trajectory is aligned with an aerial map using GPS measurements as the position reference (Fig. 7). The proposed method ends up with a final position drift of 1.7 m, which is 0.28% of the total trajectory length.

We emphasize situations where simple configurations such as monocular or stereo VINSs fail: distanced scene in Fig. 6(b) limits the ability of the forward-facing stereo cameras to recover the visual scale (thus it degenerates to a forward-facing monocular VINS), and the high-speed flight near the ground with high-frequency texture shown in Fig. 6(c) reduces the performance of the downward-facing VINS. Since we use a generalized model to ensure that multiple cameras work in a complementary way, these situations, which are common during flight but can not be resolved by simple monocular or stereo configurations, are successfully handled by the proposed method, as expected.

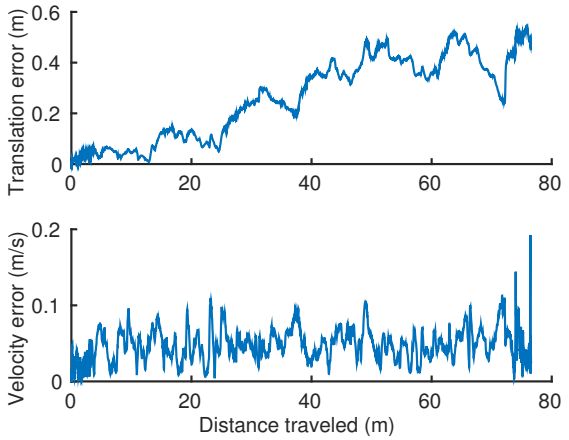
A comparison of our method using different sensor configurations is given in Fig. 8. We repeat our approach with one or two cameras. The forward-facing monocular version with only camera 1 demonstrates the largest final position drift due to the lack of direct scale measurements. With the stereo version, with cameras 1 and 2, we benefit from the direct



(a) Position

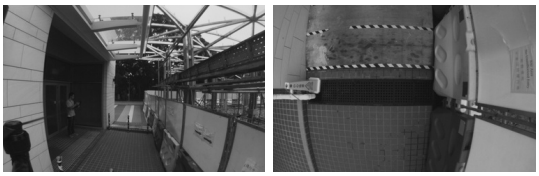


(b) Velocity



(c) Position and velocity error

Fig. 5. The MAV is flown autonomously using its onboard state estimates. We show a performance comparison against a motion capture system. The two dashed lines in each plot indicate the time of changes in camera combinations. It can be seen that the speed of position drift keeps nearly the same during random sensor failures.



(a) Narrow passageway

(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 breaking (largest pitch 35 degrees)

Fig. 6. Onboard images during an experiment in large-scale complex environments. For each case, two images captured by cameras 1 and 3 are shown.

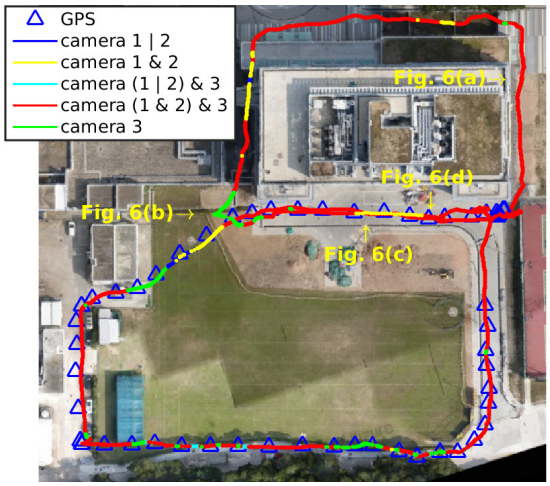


Fig. 7. Vehicle trajectory aligned with aerial photo. The total travel distance is 642 meters. GPS references are not available when flying close to a building. Different colors indicate different combinations of visual measurement availability. Sensor availability is measured based on whether the number of well-estimated features is higher than 20. Well-estimated features are those with average reprojection error lower than 3 pixels.

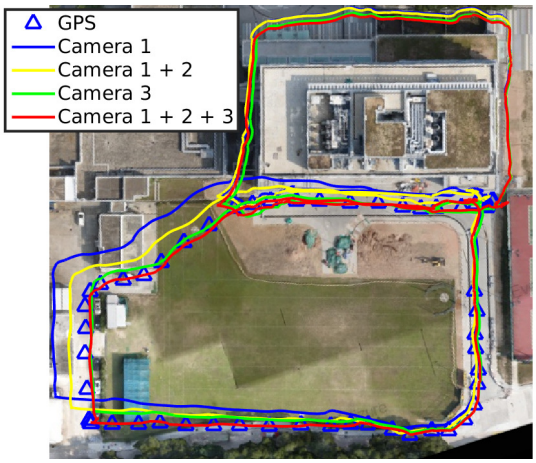


Fig. 8. Comparison between different sensor configurations using the same dataset as in Fig 7. The total travel distance is 642 meters. For the four methods, final position drifts are 1.39%, 0.63%, 0.47%, and 0.28% of the traveled distance, respectively. We observe that the forward-facing cameras accumulate more drift in the x-y direction, while the downward-facing camera accumulates more drift in the z direction (1.9 meters). This shows the necessity of our multi-camera configuration.

observation of the visual scale. Although the monocular version with only camera 3 seems from the figure to perform well, we stress that the downward version results in the largest drift in height (1.9 meters).

VII. Conclusion and Future Work

In this paper, we propose a multi-camera-inertial navigation system with online self-calibration. Our approach initializes camera-IMU calibration as well as all essential navigational states (position, velocity, attitude) during free motion in natural environments without any prior knowledge. Based on the initial values, a tightly-coupled generalized multi-camera fusion framework is proposed to make the best use of visual information from both intra- (temporal) and inter- (spatial) feature matchings. We present extensive online experimental results obtained in complex indoor and outdoor environments.

In the future, we will look into more general online camera-IMU calibrations for VINSs, including intrinsic camera calibration and temporal alignment between sensors. We are also interested in pursuing long-duration and large-scale field experiments in different environments.

References

[1] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs," in *Proc. of the IEEE Intl. Conf. on Robot. and Autom.*, Seattle, WA, May 2014.

[2] D. Scaramuzza, M. Achtelik, L. Doitsidis, F. Fraundorfer, E. Kosmatopoulos, A. Martinelli, M. Achtelik, M. Chli, S. Chatzichristofis, L. Kneip, D. Gurdan, L. Heng, G. Lee, S. Lynen, L. Meier, M. Pollefeys, A. Renzaglia, R. Siegwart, J. Stumpf, P. Tanskanen, C. Troiani, and S. Weiss, "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, 2014.

[3] 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.

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

[5] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visual-inertial SLAM using nonlinear optimization," in *Proc. of Robot.: Sci. and Syst.*, Berlin, Germany, June 2013.

[6] Z. Yang and S. Shen, "Monocular visual-inertial fusion with online initialization and camera-IMU calibration," in *Proc. of the IEEE Intl. Sym. on Safty, Security, and Rescure Robotics*, West Lafayette, IN, USA, Oct. 2015, URL http://www.ece.ust.hk/~eeshaojie/ssrr2015zhenfei.pdf.

[7] T. Kazik, L. Kneip, J. Nikolic, M. Pollefeys, and R. Siegwart, "Real-time 6D stereo visual odometry with non-overlapping fields of view," in *Proc. of the IEEE Intl. Conf. on Pattern Recognition*, Providence, RI, 2012, pp. 529–1536.

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

[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 Intl. Sym. of Robot. Research*, Flagstaff, AZ, Aug. 2011.

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

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

[12] S. Shen, "Autonomous navigation in complex indoor and outdoor environments with micro aerial vehicles," Ph.D. dissertation, University of Pennsylvania, Philadelphia, PA, Aug. 2014.

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

[14] 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. of Mediterranean Conf. on Control and Automation*, Platanias-Chania, Crete, Greece, June 2013, pp. 1261–1266.

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

[16] 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.

[17] G. Carrera, A. Angeli, and A. J. Davison, "SLAM-based automatic extrinsic calibration of a multi-camera rig," in *Proc. of the IEEE Intl. Conf. on Robot. and Autom.*, Shanghai, China, May 2011, pp. 20–25.

[18] L. Heng, M. Brki, G. H. Lee, P. Furgale, R. Siegwart, and M. Pollefeys, "Infrastructure-based calibration of a multi-camera rig," in *Proc. of the IEEE Intl. Conf. on Robot. and Autom.*, Shanghai, China, May 2011, pp. 1793–1800.

[19] M. Warren and B. Upcroft, "High altitude stereo visual odometry," in *Proc. of Robot.: Sci. and Syst.*, Berlin, Germany, June 2013.

[20] 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. of Robot.: Sci. and Syst.*, Berkeley, CA, 2014.

[21] 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, 2015.

[22] D. Nister, "An efficient solution to the five-point relative pose problem," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 2, Madison, WI, June 2003, pp. 195–202.