VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator

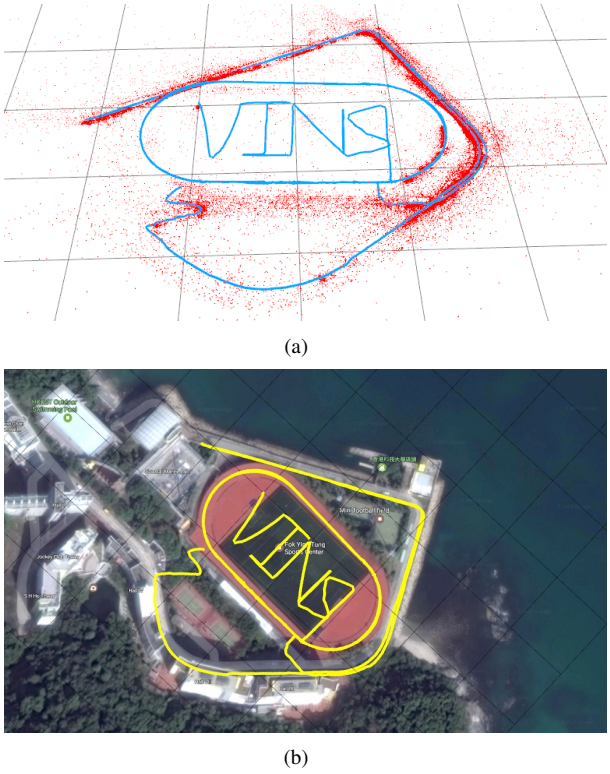
Tong Qin, Peiliang Li, Zhenfei Yang, Shaojie Shen

***Abstract*—This paper presents a monocular visual-inertial system (VINS) for localization in complex environments. The monocular camera and a low-cost inertial measurement unit (IMU) constitute the minimum sensor suite for six degree-of-freedom state estimation in various environments. Our algorithm optimizes the visual and inertial measurements in a bounded sliding window iteratively to perform accurate state estimation. Visual structure is maintained by keyframes in the sliding window, while inertial metric measurements are kept by pre-integration between keyframes. Our system is robust to initialization with unknown states, online camera-IMU extrinsic parameter calibration, unified reprojection error defined on the sphere, loop detection, and four degree-of-freedom pose graph optimization. These properties make our system quite practicable and easy to use. We validate the performance of our system on the public dataset and real-world experiment by comparing with other state-of-the-art algorithms. Finally, we perform onboard closed-loop autonomous fight on the MAV platform and port the algorithm to an iOS application. We highlight that the proposed work is a reliable and complete system, which can be easily operated with any intelligent equipment. We open source our implementation code and provide an augment reality (AR) application on iOS mobile devices.** [1]

***Index Terms*—Monocular odometry, visual-inertial system (VINS), state estimation, localization.**

I. Introduction

VISUAL-inertial fusion is currently a hot topic in robotics communities. The accurate state estimation is required in a wide range of applications, such as aerial graphics, autonomous driving, transportation, surveillance, virtual reality (VR) and augment reality (AR). Vision-only algorithms, such as [1]–[5], can estimate pose and construct the structure of environments which are up-to-scale. However, the visual-only method can easily suffer from tracking lost due to illumination change, texture-less area and motion blur caused by aggressive motion or long exposure time. Additionally, absolute scale information is necessary in practice, especially for autonomous robots and AR application. Recently, we have seen a lot of complementary visual-based sensor sets applied to robots, which can increase the robustness of odometry and render the absolute scale. Such as monocular camera and inertial measurement unit (IMU) [6]–[10], stereo cameras and IMU [11]–[13], and RGB-D camera [14]. Among these combinations, one camera and a low-cost IMU constitute the minimum sensor set which can be easily used on any smart devices, even on the mobile phone. Actually, the monocular visual-inertial system (VINS-Mono) is the simplest combination which can provide sufficient self and environment awareness. The camera provides environmental information, which depicts the up-to-scale 3D structure. Inertial measurements complement the scale and render the roll and pitch angle observable.



(a)

(b)

Fig. 1. The out door experiment result of monocular visual-inertial system. The data is collected by hand-hold monocular camera-IMU set in normal walking. Two complete circles inside playground and two semicircles outside playground. (a) The blue line is trajectory and the red points are feature positions. (b) The trajectory aligned with Google Map. The total length is 2.5 km.

Monocular visual-inertial fusion is more challenging compared with stereo or RGB-D based configurations since visual scale cannot be recovered directly. In addition, the accelerated motion is required to initialize the metric scale, which leads to nontrivial but unknown initial state. The whole system, which is a fragile and highly nonlinear process, requires stable feature tracking, accurate camera-IMU extrinsic calibration, and good estimator initialization. Also, the visual-inertial algorithms usually suffer from accumulated drift after a long trajectory since the past states are marginalized to bound the computational complex.

T. Qin, P. Li and S. Shen are with Robotics Institute, Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology. Z. Yang is with Dajiang Innovations Technology Co., Ltd. e-mail: {tong.qin, pliap, zyangag, eeshaojie}@ust.hk

[1]https://github.com/HKUST-Aerial-Robotics

To solve the issues mentioned above, we proposed a tightly-coupled sliding-window based monocular visual-inertial system with robust initialization, online extrinsic calibration, loop detection and pose graph correction. The proposed algorithm is the extension of previous work [7], [9] by improving the initialization procedure, taking IMU bias into consideration, adding loop detection and four-DoF pose graph optimization, making the whole system extremely practical and easy to use.

In this paper, we detail every part of our sliding window based optimization framework to tightly fuse visual and inertial measurement. We identify the contributions of this work as following:

- A robust initialization procedure that is able to bootstrap the system without prior of initial state and environment.
- Unified reprojection error which is defined on sphere make our algorithm suitable for small and large Field-of-View camera.
- A tightly-coupled monocular visual-inertial fusion methodology for accurate state estimation with IMU bias calibration.
- Four degreed-of-freedom (x, y, z and yaw angle) pose graph optimization after loop detection.
- Real-time onboard autonomous flight experiments on MAV platform and augment reality application on iOS mobile devices.
- Open-source code for the community. All the experiments are reproducible with source code and available devices.

The rest of the paper is structured as follows. In Sect. II, we discuss the relevant literature. We give an overview of the complete system pipeline in Sect. III. Measurement preprocessing are presented in Sect. IV. In Sect. V, we discuss the robust initialization procedure without prior of initial state and environment. A tightly-coupled, self-calibrating, nonlinear optimization-based monocular VINS estimator, is presented in Sect. VI. We present the loop closure part in Sect. VII. The implementation details and experimental results are shown in Sect. VIII. Finally, the paper is concluded with a discussion and possible future research in Sect. IX.

II. Related Work

There is a rich body of scholarly work on visual-inertial state estimation. We can categorize solutions to VINS as filtering-based algorithms [15]–[19], or graph optimization-based algorithms [7], [9], [11], [20]. Filtering-based approaches require less computational resources due to the immediate marginalization of past states, but the early fix of linearized errors may lead to overconfident or sub-optimal estimates unless carefully maintain consistency. On the other hand, graph optimization-based approaches which keep a large number of states may improve performance via iteratively re-linearizing current states at the expenses of higher computational demands. Essentially, they are the same optimal probability estimation in the different implementation.

For visual measurements, the algorithm can be divided into two categories according to the error model, direct and geometric models. The direct approaches [2], [3], [13] minimize photometric error while the indirect approaches [9], [11], [19] focus on geometry displacement. The geometry approaches consume extra computational resources on extracting and matching features, while the direct methods require an good initial guess to linearize the image. In practice, the geometry approaches are more robust due to the stable feature tracking. The direct approaches widely used in mapping work since it is easily operated on every pixel.

For IMU measurement processing, heavy and repeated propagation is needed when the starting states changed. Recently, one efficient technique to deal with it is called pre-integration, which avoids repeating integrating IMU measurement by a re-parametrization of the relative motion constraints. [7] considered on-manifold uncertainty propagation of this technique. Furthermore, [21] improves pre-integration theory, which properly addressed the manifold structure of rotation group and modeled the posterior bias correction.

Accurate initial values are necessary to bootstrap the monocular visual-inertial system. A linear estimator initialization method was proposed in [6], [9], which leverages the relative rotations from short-term IMU integration. It is highly influenced by gyroscope bias and fails to model the sensor noise in the raw projection equations. A closed-form solution has been introduced in [22]. 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 [23]. The authors add gyroscope bias calibration in this method. These approaches fail to model the accuracy of inertial integration since they rely on the double integration of IMU measurement over an extended period of time. In [8], a re-initialization and failure recovery algorithm based on SVO [2] is proposed. It is a practical method in the loosely-coupled visual-inertial system. Inertial measurements are used to stabilize the MAV, 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.

Loop closure plays an important role in long-distance estimation, since position, orientation, and scale will accumulatively drift along with trajectory. ORB-SLAM proposed in [4] is able to close loops and reuse the map, which takes the advantage of Bag-of-World [24]. A seven degree-of-freedom [25] (position, orientation, and scale) pose graph optimization is followed loop detection. In fact, for visual-inertial system, the drift only occurs in four degree-of-freedom (position and yaw angle), thanks to IMU. In this paper, we novelly optimize the pose graph with loop constraints in minimum four DoF.

III. Overview

The structure of proposed visual-inertial system is shown in Fig. 2. The first part is measurements processing front-end which extracts, tracks features for each new image frame and pre-integrates all the IMU data between two frames. The second part is initialization procedure, which provides necessary initial values (pose, velocity, gravity vector, gyroscope bias and 3D feature position) to bootstrap the nonlinear system. The third part implements nonlinear graph optimization to solve the states in our sliding window by optimizing all the

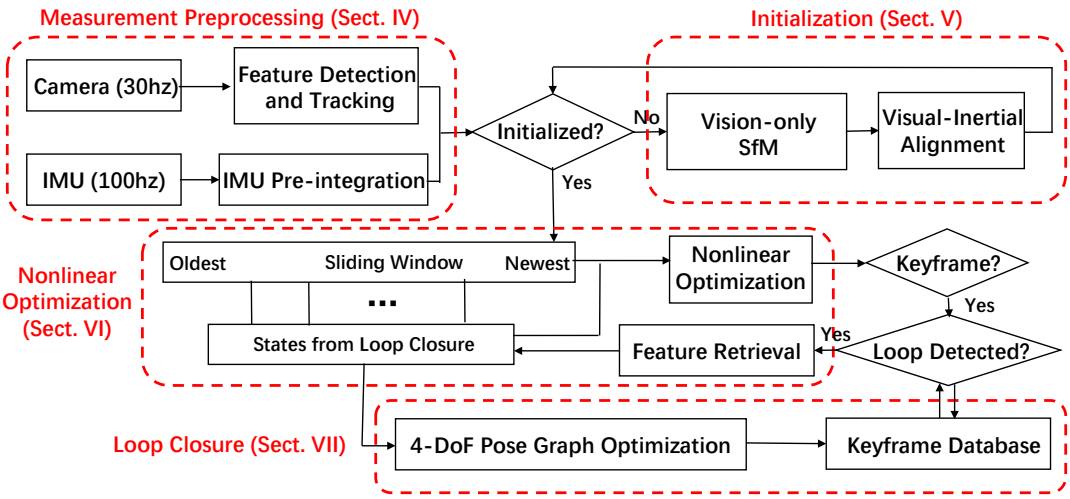


Fig. 2. Block diagram illustrating the full pipeline of the proposed monocular visual-inertial system.

visual, inertial information together. Another part which runs in another thread takes charge of loop detection and pose graph optimization.

Notation: We consider $(\cdot)^w$ as world frame, where gravity vector is along with z axis. $(\cdot)^b$ is body frame which is aligned with IMU frame. $(\cdot)^c$ is camera frame. We use quaternion $\mathbf{q}$ to denote rotation matrix. $\mathbf{q}_b^w, \mathbf{p}_b^w$ are the rotation and translation from body frame to world frame. To simplify the notation, we also use the homogeneous representation, $\mathbf{T} = \begin{pmatrix} \mathbf{R}(\mathbf{q}_b^w) & \mathbf{p}_b^w \\ \mathbf{0} & 1 \end{pmatrix}$, to denote this. The same representation applies to transformations between other frames. b_k is the body frame while taking the k^{th} image. c_k is the camera frame while taking the k^{th} image. $\otimes$ 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.

IV. Measurement Preprocessing

We pre-process visual and inertial measurements in this section. For visual measurements, we track features between consecutive frames and detect new features in latest frame. For IMU measurements, we pre-integrate them between two consecutive frames. Note that IMU measurement is affected by both bias and noise. So we especially take bias into consideration in IMU pre-integration and optimization part, which is essential for low-cost IMU chips.

A. Vision Processing Front-end

For each new image, the existing features are tracked by the KLT sparse optical flow algorithm [26]. Meanwhile, new corner features are detected [27] to maintain a minimum number of features in each image. The detector enforces a uniform feature distribution by setting a minimum separation of pixels between two neighboring features. Features are projected to a unit sphere after passing outlier rejection. Outlier rejection is performed by the RANSAC step in the fundamental matrix test.

Keyframes are also selected in this step. We have two criteria for keyframe selection. One of them is average parallax. If the average parallax of the tracked features is beyond a certain threshold, we treat this image as a keyframe. Note that not only translation but also rotation can cause parallax; however, features cannot be triangulated in the rotation-only motion. To avoid this situation, we use the IMU propagation result to compensate the rotation when calculating the parallax. Another criterion is tracking quality. If the number of tracked features goes below a certain threshold, we treat this frame also as a keyframe.

B. IMU Pre-integration

Given two time instants that correspond to images frame b_k and b_{k+1}, the state variables are constrained by inertial measurements during time interval $[k, k+1]$:

$$\mathbf{p}_{b_{k+1}}^w \approx \mathbf{p}_{b_k}^w + \mathbf{v}_{b_k}^w \Delta t + \iint_{t\in[k,k+1]} (\mathbf{q}_t^w(\hat{\mathbf{a}}_t - \mathbf{b}_a) - \mathbf{g}^w)\, dt^2 \tag{1}$$

$$\mathbf{v}_{b_{k+1}}^w \approx \mathbf{v}_{b_k}^w + \int_{t\in[k,k+1]} (\mathbf{q}_t^w(\hat{\mathbf{a}}_t - \mathbf{b}_a) - \mathbf{g}^w)\, dt \tag{2}$$

$$\mathbf{q}_{b_{k+1}}^w \approx \mathbf{q}_{b_k}^w \otimes \int_{t\in[k,k+1]} \mathbf{q}_t^w \otimes \begin{bmatrix} 0 \\ \frac{1}{2}(\hat{\boldsymbol{\omega}}_t - \mathbf{b}_g) \end{bmatrix} dt \tag{3}$$

where Δt is the duration of time interval $[k, k+1]$. $\hat{\boldsymbol{\omega}}_t$, $\hat{\mathbf{a}}_t$ are IMU measurements, which are are affected by acceleration bias $\mathbf{b}_a$, gyroscope bias $\mathbf{b}_g$, and noise.

It can be seen that the IMU state propagation require the rotation, position and velocity of frame b_k. When these starting states change, we need to re-propagate IMU measurements. Especially in optimization-based algorithm, every time we optimize the poses of some frames, we will need to re-propagate the IMU measurements between them. This propagation strategy is computation consuming. To avoid re-propagation, we adopt pre-integration algorithm.

After change the reference frame for IMU propagation to b_k, we can only pre-integrate the parts which are related to

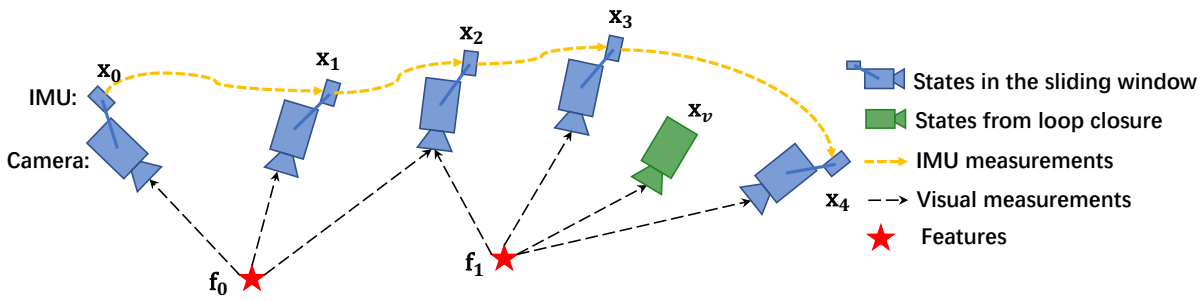


Fig. 3. An illustration of our sliding window tightly coupled with IMU, visual and loop measurements.

linear acceleration $\mathbf{a}$ and angular velocity $\boldsymbol{\omega}$ as follows:

$$\begin{aligned}\boldsymbol{\alpha}_{b_{k+1}}^{b_k} &= \iint_{t\in[k,k+1]} \boldsymbol{\gamma}_{b_t}^{b_k}\hat{\mathbf{a}}_t dt^2 \\ \boldsymbol{\beta}_{b_{k+1}}^{b_k} &= \int_{t\in[k,k+1]} \boldsymbol{\gamma}_{b_t}^{b_k}\hat{\mathbf{a}}_t dt \\ \boldsymbol{\gamma}_{b_{k+1}}^{b_k} &= \int_{t\in[k,k+1]} \boldsymbol{\gamma}_{b_t}^{b_k}\otimes\begin{bmatrix}0\\ \frac{1}{2}\hat{\boldsymbol{\omega}}_t\end{bmatrix} dt,\end{aligned} \tag{4}$$

It can be seen that the pre-integration part (4) can be obtained solely with IMU measurements by taking b_k instead of w as the base frame. $\boldsymbol{\alpha}_{b_{k+1}}^{b_k}, \boldsymbol{\beta}_{b_{k+1}}^{b_k}, \boldsymbol{\gamma}_{b_{k+1}}^{b_k}$ are only related to IMU bias. When the estimation of bias is changed, if the change is small, we adjust $\boldsymbol{\alpha}_{b_{k+1}}^{b_k}, \boldsymbol{\beta}_{b_{k+1}}^{b_k}, \boldsymbol{\gamma}_{b_{k+1}}^{b_k}$ by its first-order approximation with respect bias, otherwise we do re-propagation. This strategy saves a lot of computation resource for optimization-based algorithm since we don't need to propagate IMU measurements again and again.

Now we are able to write down the IMU measurements function:

$$\begin{bmatrix}\boldsymbol{\alpha}_{b_{k+1}}^{b_k}\\ \boldsymbol{\beta}_{b_{k+1}}^{b_k}\\ \boldsymbol{\gamma}_{b_{k+1}}^{b_k}\\ \mathbf{b}_{a b_k}\\ \mathbf{b}_{g b_k}\end{bmatrix} = \begin{bmatrix}\mathbf{q}_w^{b_k}(\mathbf{p}_{b_{k+1}}^w - \mathbf{p}_{b_k}^w + \frac{1}{2}\mathbf{g}^w\Delta t_k^2) - \mathbf{q}_w^{b_k}\mathbf{v}_{b_k}^w\Delta t_k\\ \mathbf{q}_w^{b_k}(\mathbf{v}_{b_{k+1}}^w + \mathbf{g}^w\Delta t_k) - \mathbf{q}_w^{b_k}\mathbf{v}_{b_k}^w\\ \mathbf{q}_{b_k}^{w^{-1}}\otimes\mathbf{q}_{b_{k+1}}^w\\ \mathbf{b}_{a b_{k+1}}\\ \mathbf{b}_{g b_{k+1}}\end{bmatrix}. \tag{5}$$

The details about mean, covariance propagation and first order of $\boldsymbol{\alpha}_{b_{k+1}}^{b_k}, \boldsymbol{\beta}_{b_{k+1}}^{b_k}, \boldsymbol{\gamma}_{b_{k+1}}^{b_k}$ approximation with respect to bias can be found in Appendix A.

V. Initialization

Monocular tightly-coupled visual-inertial optimization is a highly non-linear system. Since the scale is not directly observable from a monocular camera, it is hard to directly fuse these two measurements without a good initial guess. In usual, under the stationary assumption, the average of IMU measurements in first few seconds is treated as gravity vector and IMU propagation is treated as the initial poses. However, this treatment is improper when IMU measurements are influenced by non-trivial bias or accelerated movement occurs at the beginning. To improve the success rate of the monocular visual-inertial system, a robust initialization procedure is required.

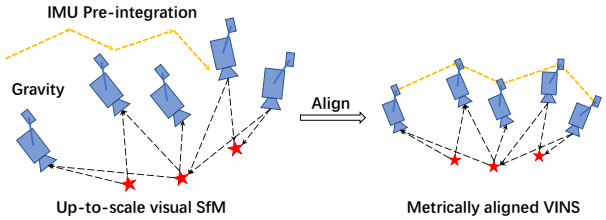


Fig. 4. An illustration of aligning IMU pre-integration with visual structure.

We adopt a loosely-coupled sensor fusion method to get the initial values. We find that vision-only SLAM, or Structure from Motion (SfM), has a good property of initialization. In most cases, a visual-only system can bootstrap itself by a derived initial guess from relative motion method, such as the Eight-point [28], Five-point [29], homogeneous and fundamental method. Through align the metric IMU pre-integration into the visual-only structure, we can roughly recover the scale, gravity, velocity, and even bias, which is of benefit for bootstrapping a nonlinear system, as shown in Fig. 4.

A. Visual SfM in Sliding Window

The initialization procedure starts with a vision-only structure, which estimates a graph of up-to-scale camera poses and feature positions.

Firstly, we check the feature correspondences between the latest frame and previous frames. If we can find a previous frame which has more than 30 tracked features and the average parallax is more than 20 pixel between the latest frame, we recover the relative rotation and up-to-scale translation between these two frames by Five-point method [29]. Otherwise, we keep the latest frame in the window and wait for new frames. If Five-point method success, we 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 whole window. Finally, a global full Bundle Adjustment [30] is applied to minimize the total re-projection error of all feature observations. Since we don't know the absolute world frame, we set the first camera frame $(\cdot)^{c_0}$ as the base frame. All frame poses $(\bar{\mathbf{p}}_{c_k}^{c_0}, \mathbf{q}_{c_k}^{c_0})$ and feature positions are represented with respect to $(\cdot)^{c_0}$. According to the prior of the extrinsic parameter $(\mathbf{p}_b^c, \mathbf{q}_b^c)$ between camera frame and IMU (body)

frame, we can translate the poses from camera center to body center,

$$\begin{aligned} \mathbf{q}_{b_k}^{c_0} &= \mathbf{q}_{c_k}^{c_0} \otimes \mathbf{q}_b^c \\ s\bar{\mathbf{p}}_{b_k}^{c_0} &= s\bar{\mathbf{p}}_{c_k}^{c_0} + \mathbf{q}_{c_k}^{c_0}\mathbf{p}_b^c, \end{aligned} \tag{6}$$

where s aligns the visual structure to the absolute scale, which will be solved in the following section.

B. Visual-Inertial Alignment

1) Gyroscope Bias Calibration: Considering two consecutive frames b_k and b_{k+1} in the window, we know the rotation $\mathbf{q}_{b_k}^{c_0}$ and $\mathbf{q}_{b_{k+1}}^{c_0}$ from the visual structure, as well as relative constraint $\hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k}$ from the IMU pre-integration. We linearize the IMU pre-integration term with respect to gyroscope bias and minimizing the of following equation:

$$\begin{aligned} &\min_{\delta \mathbf{b}_g} \sum_{k\in\mathcal{B}} \left\| {\mathbf{q}_{b_{k+1}}^{c_0}}^{-1} \otimes \mathbf{q}_{b_k}^{c_0} \otimes \boldsymbol{\gamma}_{b_{k+1}}^{b_k} \right\|^2 \\ &\boldsymbol{\gamma}_{b_{k+1}}^{b_k} \approx \hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}\frac{\partial \boldsymbol{\gamma}_{b_{k+1}}^{b_k}}{\partial \mathbf{b}_g}\delta\mathbf{b}_g \end{bmatrix}, \end{aligned} \tag{7}$$

where $\mathcal{B}$ indexes all frames in the window. The second equation is first order approximation of $\hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k}$ with respect to the gyroscope bias in its error-state presentation. In such way, we calibrate the gyroscope bias $\mathbf{b}_g$. Then we update $\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k}, \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k}$ with respect to $\mathbf{b}_g$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} &\leftarrow \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} + \frac{\partial \boldsymbol{\beta}_{b_{k+1}}^{b_k}}{\partial \mathbf{b}_g}\delta\mathbf{b}_g \\ \hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k} &\leftarrow \hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k} + \frac{\partial \boldsymbol{\gamma}_{b_{k+1}}^{b_k}}{\partial \mathbf{b}_g}\delta\mathbf{b}_g. \end{aligned} \tag{8}$$

The derivative of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ with respect to $\mathbf{b}_g$ can be found in Appendix A.

2) Velocity, Gravity Vector and Metric Scale Initialization: We define the variables that we estimate in initialization step as

$$\mathcal{X}_I = \left[\mathbf{v}_{b_0}^{c_0}, \mathbf{v}_{b_1}^{c_0}, \cdots \mathbf{v}_{b_n}^{c_0}, \mathbf{g}^{c_0}, s\right], \tag{9}$$

Considering two consecutive frames b_k and b_{k+1} in the window, the IMU measurement model for position and velocity, expressed in the local frame b_k. can be written as:

$$\begin{aligned} \boldsymbol{\alpha}_{b_{k+1}}^{b_k} &= \mathbf{q}_{c_0}^{b_k} s(\bar{\mathbf{p}}_{b_{k+1}}^{c_0} - \bar{\mathbf{p}}_{b_k}^{c_0}) + \frac{1}{2}\mathbf{q}_{c_0}^{b_k}\mathbf{g}^{c_0}\Delta t_k^2 - \mathbf{q}_{c_0}^{b_k}\mathbf{v}_{b_k}^{c_0}\Delta t_k \\ \boldsymbol{\beta}_{b_{k+1}}^{b_k} &= \mathbf{q}_{c_0}^{b_k}\mathbf{v}_{b_{k+1}}^{c_0} + \mathbf{q}_{c_0}^{b_k}\mathbf{g}^{c_0}\Delta t_k - \mathbf{q}_{c_0}^{b_k}\mathbf{v}_{b_k}^{c_0}. \end{aligned} \tag{10}$$

We can rewrite the above equations in the following linear form:

$$\begin{aligned} \hat{\mathbf{z}}_{b_{k+1}}^{b_k} &= \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k} \\ \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} \end{bmatrix} = \mathbf{H}_{b_{k+1}}^{b_k}\mathcal{X}_I + \mathbf{n}_{b_{k+1}}^{b_k} \\ &\approx \begin{bmatrix} -\mathbf{q}_{c_0}^{b_k}\Delta t_k & \mathbf{0} & \frac{1}{2}\mathbf{q}_{c_0}^{b_k}\Delta t_k^2 & \mathbf{q}_{c_0}^{b_k}(\bar{\mathbf{p}}_{b_{k+1}}^{c_0} - \bar{\mathbf{p}}_{b_k}^{c_0}) \\ -\mathbf{q}_{c_0}^{b_k} & \mathbf{q}_{c_0}^{b_k} & \mathbf{q}_{c_0}^{b_k}\Delta t_k & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_{b_k}^{c_0} \\ \mathbf{v}_{b_{k+1}}^{c_0} \\ \mathbf{g}^{c_0} \\ s \end{bmatrix}. \end{aligned} \tag{11}$$

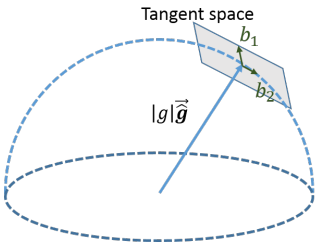


Fig. 5. Illustration of two degree-of-freedom parameterization of gravity. Since the magnitude of gravity is known, $\mathbf{g}$ lies on a sphere with the radius $|g|$. The gravity is parameterized around current estimate as $g\cdot\bar{\mathbf{g}}+w_1\mathbf{b}_1+w_2\mathbf{b}_2$, where $\mathbf{b}_1$ and $\mathbf{b}_2$ are two orthogonal basis spanning the tangent space.

In the above formula, $\mathbf{q}_{b_k}^{c_0}, \bar{\mathbf{p}}_{b_k}^{c_0}, \bar{\mathbf{p}}_{b_{k+1}}^{c_0}$ are obtained from the visual structure. $\mathbf{q}_{c_0}^{b_k}$ is the inverse rotation of $\mathbf{q}_{b_k}^{c_0}$. Δt_k is the time interval between two consecutive frames. By solving the this least square problem:

$$\min_{\mathcal{X}_I} \sum_{k\in\mathcal{B}} \left\| \hat{\mathbf{z}}_{b_{k+1}}^{b_k} - \mathbf{H}_{b_{k+1}}^{b_k}\mathcal{X}_I \right\|^2, \tag{12}$$

we can get the velocities and the gravity vector in the visual base frame $(\cdot)^{c_0}$, as well as the scale parameter. The translational components $\bar{\mathbf{p}}^{c_0}$ from the visual structure will be scaled to the metric units.

3) Gravity Refinement: The gravity vector obtained from the previous step can be refined by constraining the magnitude constraint. In most cases, the magnitude of the gravity vector is known. However, if we directly add this norm constraint into the optimization problem in (12), it will become nonlinear and hard to solve. Here, we use a linear method to enforce this constraint by optimizing the 2D error states 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 $g\cdot\bar{\mathbf{g}} + w_1\mathbf{b}_1 + w_2\mathbf{b}_2$, where g is the magnitude of gravity, $\bar{\mathbf{g}}$ 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, as shown in Fig. 5. We can use Gram-Schmidt process to find one set of $\mathbf{b}_1$, $\mathbf{b}_2$ easily. Then we substitute $\mathbf{g}$ in (11) by $g\cdot\bar{\mathbf{g}} + w_1\mathbf{b}_1 + w_2\mathbf{b}_2$ 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 frame $(\cdot)^{c_0}$ 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.

VI. Tightly-coupled Nonlinear Optimization

After state initialization, we proceed with a sliding window based nonlinear estimator for high-accuracy state estimation. An illustration of sliding window is shown in Fig. 3. Ceres Solver [31] is used for solving this non-linear optimization problem.

A. Formulation

The full state vector in the sliding window is defined as (the transpose is ignored for simplicity of representation):

$$\begin{aligned}\mathcal{X} &= \left[\mathbf{x}_0,\, \mathbf{x}_1,\, \cdots \mathbf{x}_n,\, \mathbf{x}_c^b,\, \lambda_0,\, \lambda_1,\, \cdots \lambda_m\right] \\ \mathbf{x}_k &= \left[\mathbf{p}_{b_k}^w,\, \mathbf{v}_{b_k}^w,\, \mathbf{q}_{b_k}^w,\, \mathbf{b}_a,\, \mathbf{b}_g\right], k \in [0, n] \\ \mathbf{x}_c^b &= \left[\mathbf{p}_c^b,\, \mathbf{q}_c^b\right],\end{aligned} \tag{13}$$

where $\mathbf{x}_k$ is the k^{th} frame state, which contains position, velocity, and orientation in the world frame and acceleration bias and gyroscope bias in the body frame. n is the number of keyframes and m is the number of features in the sliding window. λ_l is the inverse depth of the l^{th} feature from its first observation on the unit sphere.

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

$$\begin{aligned}\min_{\mathcal{X}} \Bigg\{ &\|\mathbf{r}_p - \mathbf{H}_p\mathcal{X}\|^2 + \sum_{k\in\mathcal{B}} \left\| r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k},\, \mathcal{X}) \right\|_{\mathbf{P}_{b_{k+1}}^{b_k}}^2 + \\ &\sum_{(l,j)\in\mathcal{C}} \left\| r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j},\, \mathcal{X}) \right\|_{\mathbf{P}_l^{c_j}}^2 \Bigg\},\end{aligned} \tag{14}$$

where $r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k},\, \mathcal{X})$ and $r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j},\, \mathcal{X})$ are residuals for the IMU and visual models respectively. $\mathcal{B}$ is the set of all IMU measurements, $\mathcal{C}$ is the set of feature which has been observed at least 2 times in the current sliding window. $\{\mathbf{r}_p,\, \mathbf{H}_p\}$ is the prior information from marginalization. Corresponding models are defined in the following sections.

B. IMU Model

Considering two consecutive frames b_k and b_{k+1} in the window, according to the IMU measurement function eq. 5, the residual of IMU measurement model can be defined as:

$$\begin{aligned} r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k},\, \mathcal{X}) &= \begin{bmatrix} \delta\boldsymbol{\alpha}_{b_{k+1}}^{b_k} \\ \delta\boldsymbol{\beta}_{b_{k+1}}^{b_k} \\ \delta\boldsymbol{\theta}_{b_{k+1}}^{b_k} \\ \delta\mathbf{b}_a \\ \delta\mathbf{b}_g \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{q}_w^{b_k}(\mathbf{p}_{b_{k+1}}^w - \mathbf{p}_{b_k}^w + \frac{1}{2}\mathbf{g}^w\Delta t_k^2) - \mathbf{q}_w^{b_k}\mathbf{v}_{b_k}^w\Delta t_k - \hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k} \\ \mathbf{q}_w^{b_k}(\mathbf{v}_{b_{k+1}}^w + \mathbf{g}^w\Delta t_k) - \mathbf{q}_w^{b_k}\mathbf{v}_{b_k}^w - \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k} \\ 2\left[\mathbf{q}_{b_{k+1}}^{w^{-1}} \otimes \mathbf{q}_{b_k}^w \otimes \hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k}\right]_{xyz} \\ \mathbf{b}_{a_{b_{k+1}}} - \mathbf{b}_{a_{b_k}} \\ \mathbf{b}_{g_{b_{k+1}}} - \mathbf{b}_{g_{b_k}} \end{bmatrix}, \end{aligned} \tag{15}$$

where $\left[\cdot\right]_{xyz}$ extracts the vector part of the quaternion $\mathbf{q}$, which is the approximation of error state representation. $[\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k},\, \hat{\boldsymbol{\beta}}_{b_{k+1}}^{b_k},\, \hat{\boldsymbol{\gamma}}_{b_{k+1}}^{b_k}]^{\mathrm{T}}$ is the pre-integrated IMU measurement using only noisy accelerometer and gyroscope measurements, which is related to accelerometer and gyroscope bias and independent of initial velocity and attitude. $\delta\boldsymbol{\theta}_{b_{k+1}}^{b_k}$ is the minimum error-state representation of quaternion.

The corresponding covariance matrix $\mathbf{P}_{b_{k+1}}^{b_k}$ of IMU measurements can be found in Appendix A.

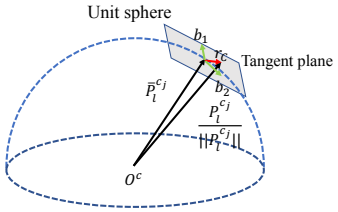


Fig. 6. An illustration of the visual residual on unit sphere. $\bar{\mathcal{P}}_l^{c_j}$ is the ray vector of l^{th} feature observed in the j^{th} frame , $\mathcal{P}_l^{c_j}$ is the transformed ray vector from the i^{th} frame. The residual is defined on the tangent plane of $\bar{\mathcal{P}}_l^{c_j}$.

C. Vision Model

We define the camera measurement residual on unified unit sphere, which also suitable for large FOV camera, such as fisheye and omni-directional camera. As shown in Fig. 6, the camera residual for the observation of the l^{th} feature in the j^{th} image is defined as,

$$\begin{aligned} r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j},\, \mathcal{X}) &= \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2\end{bmatrix}^T \cdot (\bar{\mathcal{P}}_l^{c_j} - \frac{\mathcal{P}_l^{c_j}}{\|\mathcal{P}_l^{c_j}\|}) \\ \bar{\mathcal{P}}_l^{c_j} &= \pi_c^{-1}(\begin{bmatrix}\hat{u}_l^{c_j} \\ \hat{v}_l^{c_j}\end{bmatrix}) \\ \mathcal{P}_l^{c_j} &= \begin{bmatrix}{}^P x_l^{c_j} \\ {}^P y_l^{c_j} \\ {}^P z_l^{c_j}\end{bmatrix} = \mathbf{T}_c^{b^{-1}} \cdot \mathbf{T}_{b_j}^{w^{-1}} \cdot \mathbf{T}_{b_i}^w \cdot \mathbf{T}_c^b \cdot \frac{1}{\lambda_l} \cdot \pi_c^{-1}(\begin{bmatrix}u_l^{c_i} \\ v_l^{c_i}\end{bmatrix}), \end{aligned} \tag{16}$$

where $[u_l^{c_i},\, v_l^{c_i}]$ is the first observation of the l^{th} feature that happens in the i^{th} image. $[\hat{u}_l^{c_j},\, \hat{v}_l^{c_j}]$ is the observation of the same feature in the j^{th} image. To simply the representation, we omit homogeneous term in above equations. $\mathbf{T}_c^b$ is the extrinsic transformation from the camera frame to the IMU (body) frame, and its inverse transforms from the IMU frame to the body frame. $\mathbf{T}_{b_i}^w$ transforms from the ith IMU frame to the world frame. π_c^{-1} is the back projection model which outputs the unit vector in 3D space. Since the degree-of-freedom of the vision residual is two, we project the residual vector onto the tangent plane. $\mathbf{b}_1, \mathbf{b}_2$ are two arbitrarily selected orthogonal bases which span the tangent plane of $\bar{\mathcal{P}}_l^{c_j}$, as shown in Fig. 6.

$\mathbf{P}_l^{c_j}$ is the standard covariance of a fixed length in tangent space.

D. Marginalization

In order to bound the computational complexity of optimization-based methods, marginalization is incorporated. We selectively marginalize out IMU states $\mathbf{x}_k$ and features λ_l from the sliding window, meanwhile convert measurements corresponding to the marginalized states into a prior.

As shown in the Fig. 7, when the second latest frame is a keyframe, it will stay in the window and the oldest frame states are marginalized out with its corresponding measurement. Otherwise, if the second latest frame is a non-keyframe, we throw the visual measurements and keep the IMU measurements in the window, instead of marginalizing out all measurements. This strategy will maintain the sparsity of the system. The

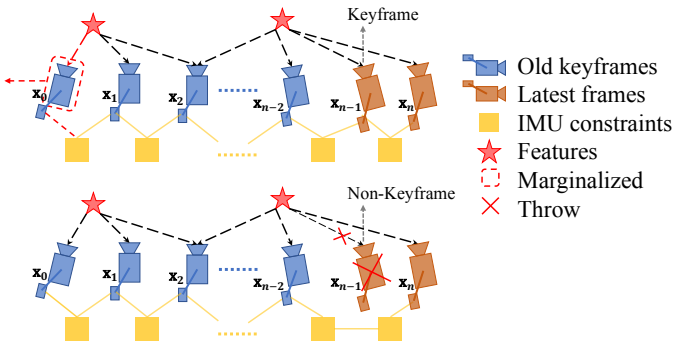


Fig. 7. An illustration of marginalization strategy. If the second latest frame is a keyframe, we will keep it in the window, and marginalize the oldest frame and its corresponding visual and inertial measurements. Marginalized measurements are used to construct a prior. If the second latest frame is non-keyframe, we will throw it and its corresponding visual measurements, and keep the inertial measurements in the window.

marginalization scheme can keep spatial keyframes in the window, meanwhile bound the uncertainty for pre-integrated IMU measurements.

We construct a new prior based on marginalized measurements related to the removed state. The marginalization is carried out using the Schur complement. Intuitively, by marginalization, important information of the removed states is kept and computation complexity is bounded.

E. IMU Propagation for High Frequency Output

Note that the IMU measurements come at a much higher rate than visual measurements. The frequency of our nonlinear optimization estimator is limited by visual measurements. To benefit the performance for real-time control, the outputs of the estimator are directly propagated with the newest IMU measurements, which serves as the high-frequency feedback in the control loop.

F. Failure Detection and Recovery

Sometimes failure is unavoidable due to violent illumination change or severely aggressive motion. Active failure detection and recovery strategy can improve the practicability of proposed system. Failure detection is an independent module, which detect unpractical outputs of estimator. We have several criterion for failure detection:

- Tracked feature number of latest frame less than 5;
- A big discontinuity in position or rotation between last two outputs;
- A big value in bias or extrinsic parameters estimation;

If failure is detected, the proposed system switch into initial status, which will try to reinitialize the system. The reinitialize will start at the last possible pose. Meanwhile, the keyframe database will be kept, which will be used for loop closure.

VII. Loop Closure

We detect loop and maintain a pose-only graph in loop thread. Since the sliding window lacks absolute position and yaw observation, we attach it at the end of pose graph.

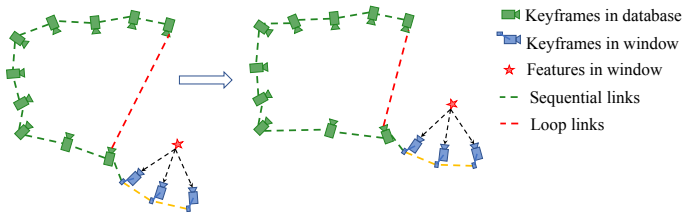


Fig. 8. Local window state estimation is attach to the end of pose graph, and is separate from the pose graph. The green cameras are keyframes in database and the blue cameras are keyframes in current sliding window. Whenever the database optimizes the pose graph, the local window will shift to the new end of pose graph.

Whenever a loop closure occurs, the pose graph will adjust with loop constraints. The sliding window will shift to the end of pose graph, as shown in Fig. 8. Such that the loop closure model is separate from real-time state estimation in local window. The IMU measurements render roll and pitch angle fully observable, so the accumulated drift only occurs in four degree-of-freedom (x, y, z and yaw angle). To avoid importing spurious information, we directly optimize pose graph on these four degree-of-freedom.

A. Loop Detection

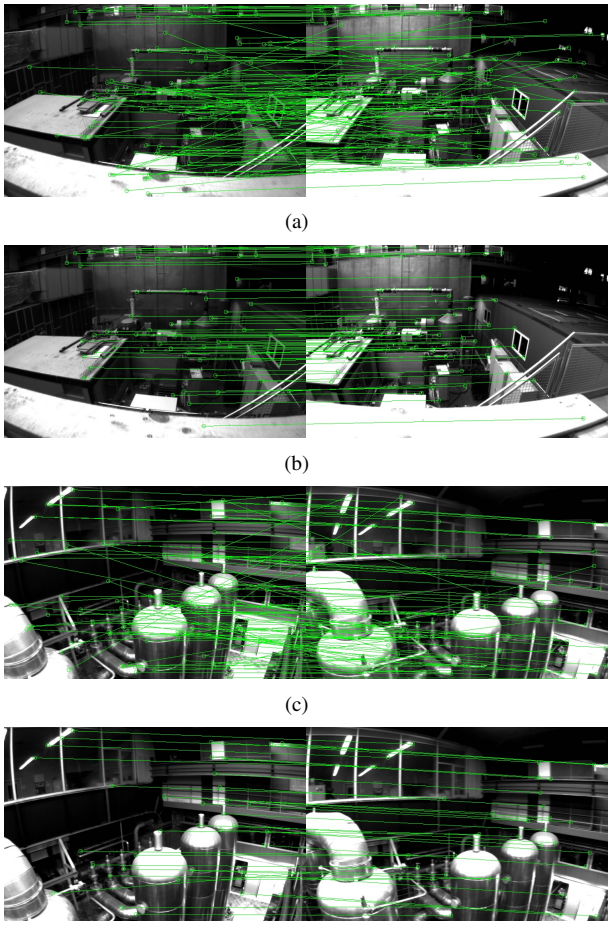
We utilize bag-of-word place recognition which is introduced in DBoW2 [24] with BRIEF descriptors [32] to perform loop detection. FAST coner are extracted and BRIEF descriptors are calculated to describe observed features when a new keyframe is coming. The descriptors are treated as the visual word to query the visual vocabulary. If loop happens, the DBoW2 will return the best loop candidate after temporal and geometrical consistency check. Only the BRIEF descriptors and pose of each keyframe is saved, so we can throw the raw images to save memory.

B. Feature Retrieving

The connection between new keyframe and its loop candidate is established by retrieved features. Feature retrieving is performed by BRIEF descriptor matching. Directly descriptor matching can cause a lot of outliers. In order to reduce outliers, we limit search area within the neighbor of the same position when matching in another frame. Then a geometry consistency check by fundamental matrix test with RANSAC is performed to remove outliers, as shown in Fig. 9. When the number of feature correspondences beyond a certain threshold, we treat this candidate as a right loop detection and fuse its information in the following.

C. Compute Relative Transformation

We use m and v to denote a new keyframe and its looped keyframe. Since new keyframe m will be put into the sliding window, and all frames share the same features in sliding window, we can easily connect looped keyframe v with sliding window by these retrieved features. The relative transformation is calculated by jointly optimizing the looped keyframe v into the siding window bundle. The looped keyframe is treated



(a) (b) (c) (d)

Fig. 9. (a)(c) Directly BRIEF feature matching results. (b)(d) Results of limited-area BRIEF feature matching and geometric RANSAC rejection. Apparently, outliers are efficiently rejected by limited-area and geometric RANSAC.

as an addition measurement frame in sliding window, which only contains visual constraints without IMU constraint. We can easily write a same visual model for retrieved features in looped frame v as eq. 16, and add this term into whole nonlinear cost function eq. 14:

$$\min_{\mathcal{X},\mathbf{q}_v^w,\mathbf{p}_v^w}\left\{\|\mathbf{r}_p-\mathbf{H}_p\mathcal{X}\|^2+\sum_{k\in\mathcal{B}}\left\|r_{\mathcal{B}}(\hat{\mathbf{z}}_{b_{k+1}}^{b_k},\mathcal{X})\right\|^2_{\mathbf{P}_{b_{k+1}}^{b_k}}+\sum_{(l,j)\in\mathcal{C}}\left\|r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{c_j},\mathcal{X})\right\|^2_{\mathbf{P}_l^{c_j}}+\sum_{l\in\mathcal{L}}\|r_{\mathcal{C}}(\hat{\mathbf{z}}_l^{v},\mathcal{X},\mathbf{q}_v^w,\mathbf{p}_v^w)\|^2_{\mathbf{P}_l^{c_v}}\right\}. \tag{17}$$

Set $\mathcal{L}$ is the set of retrieved features.

The loop closure factor only increases two variables and several constraints in the cost function of the sliding window. It is solved along with window optimization without much extra computation. After minimizing the cost function, we get the relative pose of looped frame v with respect to the local window. We add the relative pose (position $\mathbf{p}_{vm}^v$, yaw angle ψ_{vm}) between frame m and v as loop link into the pose graph.

D. 4-DoF Pose Graph Optimization

Since the absolute scale, roll and pitch angle are fully observable in the visual-inertial system, accumulate drifts only occur in position (x, y, z) and yaw angle. To take the advantage of observability characteristics and avoid spurious information, we only optimize 4-DoF pose graph when loop detection occurs instead of 6 or 7-DoF optimization in traditional visual-only SLAM.

We maintain two kinds of links in the pose graph. One is sequential links from pure odometry. One keyframe will establish several sequential links to its neighbor keyframes. The relative pose $(\hat{\mathbf{p}}_{ij}^i,\hat{\psi}_{ij})$ is established by origin odometry. The other one is loop link, which connects one keyframe with its looped frame. The relative pose is calculated by VII-C.

We define the residual between frames i and j minimally as:

$$r_{i,j}(\mathbf{p}_i^w,\psi_i,\mathbf{p}_j^w,\psi_j)=\begin{bmatrix}\mathbf{q}_i^w(\psi_i)^{-1}(\mathbf{p}_j^w-\mathbf{p}_i^w)-\hat{\mathbf{p}}_{ij}^i\\ \psi_i-\psi_j-\hat{\psi}_{ij}\end{bmatrix}. \tag{18}$$

The first row is relative position error, and the second row is relative yaw angle error.

The whole graph of sequential constraints and loop constraints is optimized by:

$$\min_{\mathbf{p},\psi}\left\{\sum_{(i,j)\in\mathcal{S}}\|r_{i,j}\|^2_{\Omega_s}+\sum_{(i,j)\in\mathcal{L}}h(\|r_{i,j}\|_{\Omega_l})\right\}, \tag{19}$$

where $\mathcal{S}$ is the set of all sequential edges and $\mathcal{L}$ is all loop edges. $h(\cdot)$ is huber robust cost function. The reason we use huber cost function for loop constraints is that sometimes false loop detection occurs. Huber cost function can relieve the influence of potential false loop detection. For sequential edges, we don't have such worry since they are extracted from incremental odometry without outliers.

E. Database Management

Along with the increase of trajectory, the database becomes larger and larger. Loop detection and pose optimization graph time becomes longer and longer. Although we take the minimum storage strategy (only save pose and features descriptor without raw image), the calculation time will limit real-time performance when working several hours. To this end, we maintain the database in a limited size. We downsample the database according to distribution density when the number of keyframes beyond certain threshold. We remove keyframes in angle and pose centralized area, and remain the frames which have a minimum distance or angle difference with its neighbor.

VIII. Experimental Results

We perform three experiments and two applications to evaluate proposed algorithm. In the first experiment, we compare the proposed algorithm with other state-of-art algorithm on public datasets. We do the numerical analysis to show the accuracy of our system. Secondly, we test our system in the indoor environment to evaluate the performance in repeated scenes. Then a large-scale experiment is carried out to illustrate the long-time practicability of our system.

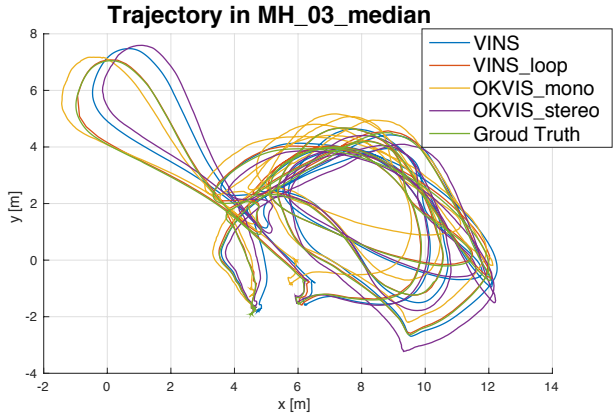


Fig. 10. Trajectory in MH_03_median, compared with OKVIS.

Additionally, we apply proposed system into two devices. The one is the MAV platform. We use VINS as position feedback to control the MAV following designed trajectory. The other one is mobile phone platform. We transplant our implementation on the iOS mobile device and compare with other professional commercial device.

A. Dataset Comparison

We evaluate proposed method with ASL MAV Visual-Inertial Datasets [33]. The datasets are collected onboard a micro aerial vehicle, which contain 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 left camera from stereo images set. Nontrivial IMU bias and illumination change are included in the datasets, which make it representative and challenging.

In these experiments, we compare proposed method with OKVIS [11], which is the state-of-the-art visual-inertial odometry working with monocular and stereo cameras. OKVIS is another optimization-based sliding-window algorithm. Our algorithm is different with OKVIS in every detail, and our system is more complete with robust initialization and loop closure. We choose two sequences, MH_03_median, MH_05_difficult to show the performance of proposed method. For simplifying notation, we use VINS to denote our pure odometry and VINS_loop to denote the refined pose graph after loop detection. We use OKVIS_mono and OKVIS_stereo to denote the OKVIS's result working with monocular image and stereo image respectively. To fairly compare the results from two algorithms, for each result, we throw the first 100 output and use the following 150 outputs to calculate the transformation towards the ground truth and align all outputs to the ground truth through this transformation, because monocular system need a little more time to converge.

For the sequence MH_03_median, the trajectory is shown in Fig. 10. Since little rotation movement occurs in this sequences, we mainly compare translation error. The x, y, z error along with time and the translation error along with distance is shown in Fig. 11. In the error plot, the proposed method

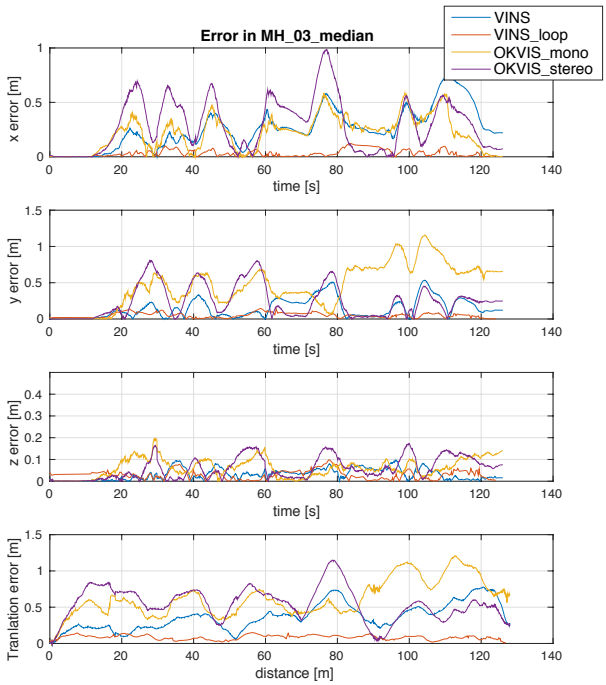


Fig. 11. Translation error plot in MH_03_median.

with loop function has the smallest translation error. The result is same in MH_03_difficult. The proposed method with loop function has the smallest translation error. The translation and rotation error is shown in Fig. 13. Since the movement is smooth without much yaw angle change in these sequence, only position drifts occurs in this dataset. Obviously, the loop closure function efficiently bound the accumulated drifts. For the rotation error, OKVIS seems more stable. Honestly, OKVIS performs better in roll and pitch angle estimation. The possible reason may be that proposed method use the pre-integration technique which is the first-order approximation of IMU pre-integration to save computation resource. Another possible reason is that OKVIS marginalize IMU constraints in a higher frequency.

Our proposed method performs well in all EuRoC datasets, even in the most challenging sequence, V1_03_difficult, the one includes aggressive motion, texture-less area and great illumination change. Our proposed method can initialize quickly in V1_03_difficult, due to the robust initialization procedure. The robust initialization procedure can improve the success rate in practice.

Admittedly, for pure odometry, both proposed method and OKVIS are much accurate, it is hard to distinguish which one is better. It is impossible to further increase the accuracy by orders of magnitude for tightly-coupled optimization-based fusion. Our proposed method is a complete system compare with OKVIS since we have robust initialization and loop closure function to assist pure odometry.

B. Indoor Experiment

In the indoor experiment, we choose our laboratory environment as the experiment area. The sensor set we use is shown in

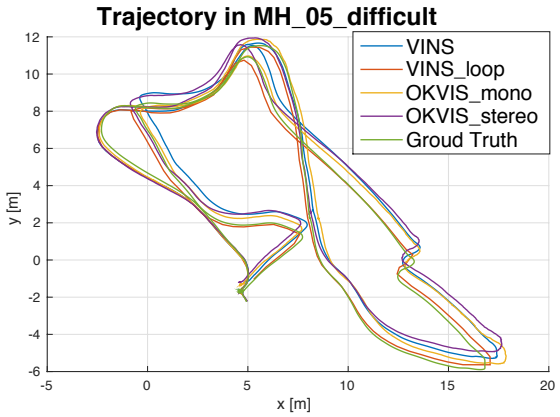


Fig. 12. Trajectory in MH_05_difficult, compared with OKVIS..

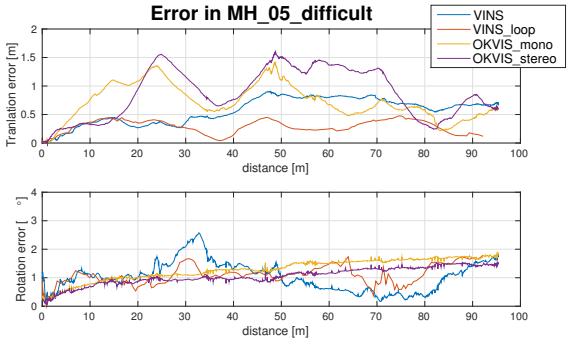


Fig. 13. Translation error and rotation error plot in MH_05_difficult.

Fig. 14, a monocular camera (20Hz) and IMU (100Hz) inside DJI A3 controller[2]. We hold the sensor set by hand and walk in the laboratory around and around. We encounter pedestrians, darkness, low-texture area, glass and reflection, as shown in Fig. 15, which represents the normal daily life. Details can be found in the supplementary video.

We compare our result with OKVIS, as shown in Fig. 16. The Fig. 16(a) is the odometry from OKVIS. Fig. 16(b) is the pure odometry from proposed method without loop closure. Fig. 16(c) is the result of proposed method with loop closure. Apparently, nontrivial drifts occur when we walk indoor around and around. Both OKVIS and proposed pure odometry accumulate great drifts in x, y, z, and yaw angle. Our loop closure function can correct drifts efficiently.

C. Large-scale Environment

1) Go out of lab: We test proposed algorithm from indoor environment to outdoor environment. The sensor that we use is the same with that used in indoor environment, bluefox camera with A3. We start from the seat of our laboratory, and go around our laboratory. Then we go down the stairs and walk around the play ground outside the building. Next, we go back to the building and go up stairs. Finally, we return to the seat in laboratory. The whole trajectory is more than 700 meters and last ten minutes. The details can be found in supplementary video.

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

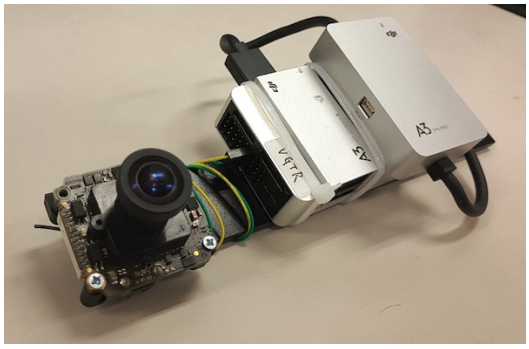
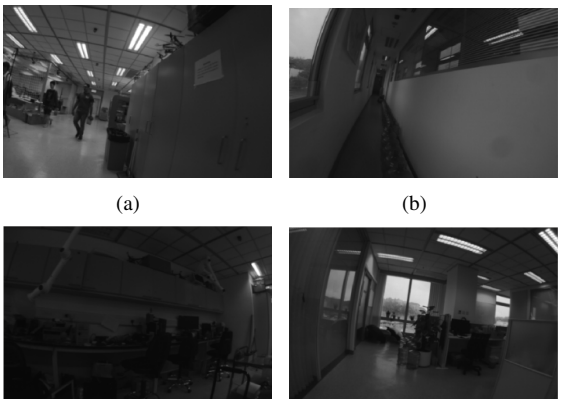
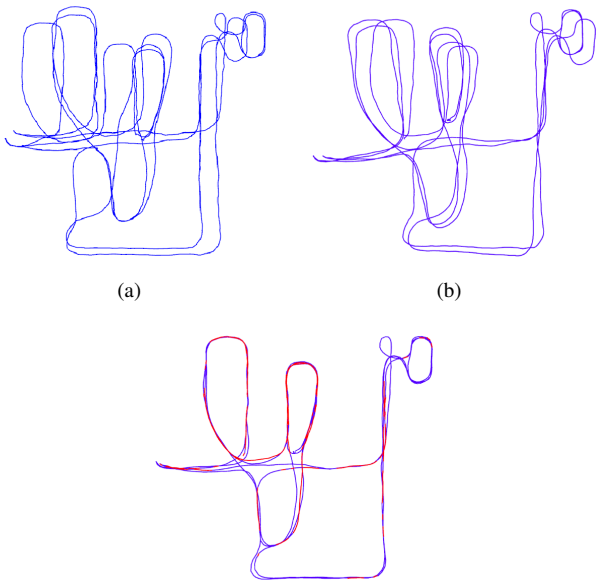


Fig. 14. The device used in the indoor experiments. One forward-looking global shutter camera (MatrixVision mvBlueFOX-MLC200w) with 752×480 resolution. Inertial measurement unit (IMU, ADXL278 and ADXRS290, 100Hz) inside DJI A3 flight controller.



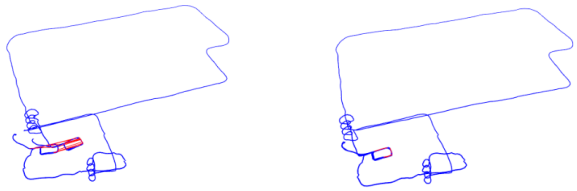
(a) (b) (c) (d)

Fig. 15. (a) Pedestrians. (b) Low-texture area. (c) Darkness. (4) Glass and reflection.



(a) (b) (c)

Fig. 16. (a) Trajectory of OKVIS. (b) Trajectory of propose method without loop closure. (b) Trajectory of propose method with loop closure. The red path is the places where loop are detected.



(a) (b)

Fig. 17. Trajectory of going out of laboratory. (a) The trajectory without loop closure. (b) The trajectory with loop closure. The red lines are loop links. The spiral lines are going up and down stairs.

Fig. 18. The trajectory of going out of lab aligned with Google Map. The yellow line is the trajectory of proposed method. The red line is loop detected area.

The trajectory is shown in Fig. 17. The trajectory without loop closure is shown in Fig. 17(a), while the trajectory with loop closure is shown in Fig. 17(b). The trajectory is aligned with Google Map in Fig. 18.

Without loop closure, the distance between start point and end point is [-5.47, 2.76. -0.29] in x, y and z axis, which occupies 0.88% in whole length. With loop correction. the distance between start point and end point is [-0.032, 0.09, -0.07], which is trivial compared with whole length. In addition, the optimized trajectory is smooth, which can be precisely aligned with satellite map.

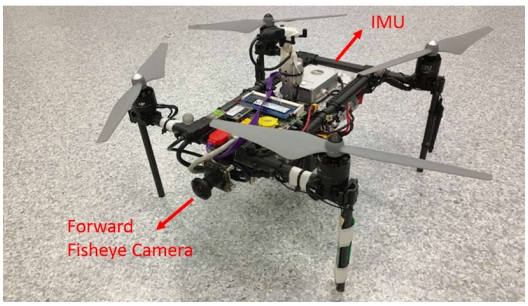
2) Go around campus: This large-scale experiment was recorded with the handheld VI-Sensor[3] walking around HKUST campus, which is around 710m in length, 240m in width and 60m in height. The whole path length is 5.62km. The data contains 25Hz image and 200Hz IMU lasting for 1 hour and 34 minutes. It's a good long-time experiment to test the stability and durability of the system.

In this large-scale test, We set the keyframe database size as 2000, which can provide sufficient loop information and achieve real-time performance. Every time when the loop is detected, we optimize the whole pose graph once. We run this test with Intel i7-4790 CPU, 3.60GHz. The timing statistics is show in Table.I. The trajectory is aligned with Google map in Fig.19. Compared with Google map, we can see our results are almost drift-free in such a long run.

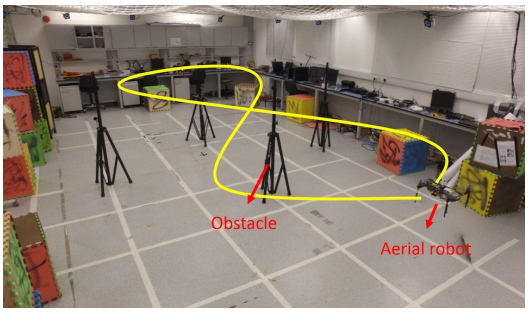
[3]http://www.skybotix.com/

TABLE I
TIMING STATISTICS

Tread	Modules	Time (ms)	Rate (Hz)
1	Harris detector	15	25
	KLT tracker	5	25
2	Window optimization	50	10
3	Loop detection	100	
	Pose graph optimization	130	



(a)



(b)

Fig. 20. (a) The self-developed aerial robot with one forward-looking fisheye camera (MatrixVision mvBlueFOX-MLC200w, 190 FOV) and IMU (ADXL278 and ADXRS290 inside in DJI A3, 100Hz). (b) The designed trajectory in onboard experiment. Four position-known obstacles are put in the test ground. The yellow line is the design eihgt-figure trajectory which the aerial robot should follow. The robot follows the trajectory four times without loop closure. Details are shown in the supplementary video.

D. Application I: Onboard Aerial Robot

We apply our algorithm onboard an aerial robot, as shown in Fig. 20(a). 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 is used both as the inertial measurement unit (IMU, ADXL278 and ADXRS290, 100Hz) and attitude stabilization control. The onboard computation resource is an Intel i7-5500U CPU running at 3.00 GHz.

In this experiment, we test the performance of autonomous trajectory tracking under the VINS estimation. No loop closure is used in this experiment. The quadrotor is commanded to track a figure eight pattern with each circle being 1.0 meters in radius, as shown in Fig. 20(b). The four obstacles are put around the trajectory to verify the accuracy of monocular VINS without loop closure. The quadrotor follows this trajectory four times continuously during the experiment. The 100 Hz state estimation results achieve the real-time feedback

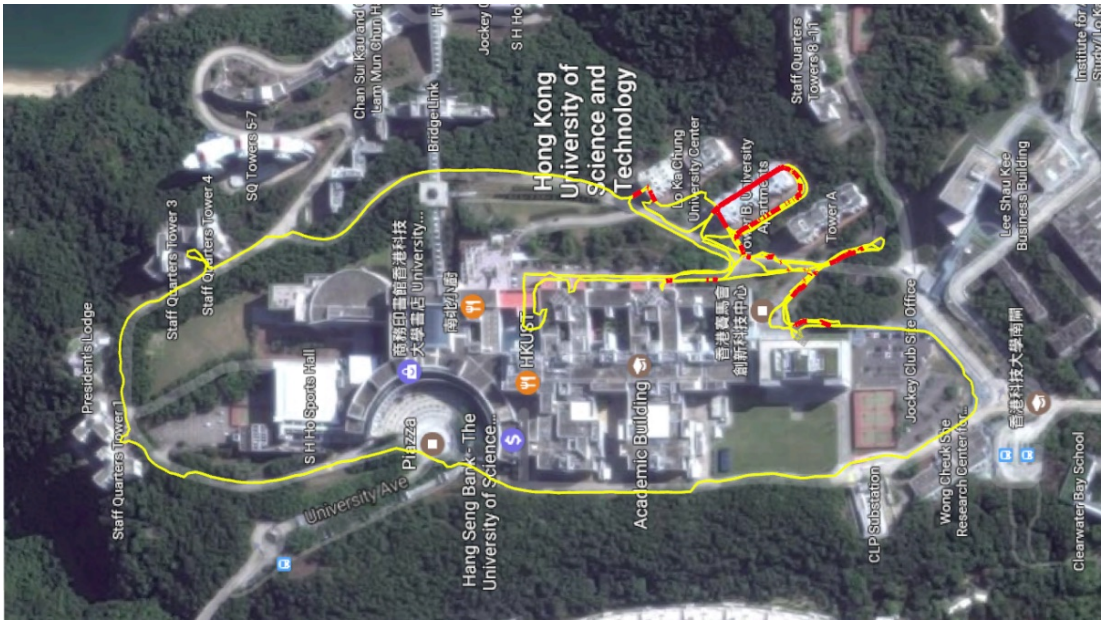


Fig. 19. The trajectory of large-scale environment test aligned with Google map. The yellow line is the trajectory of proposed method. The red line is loop detected area.

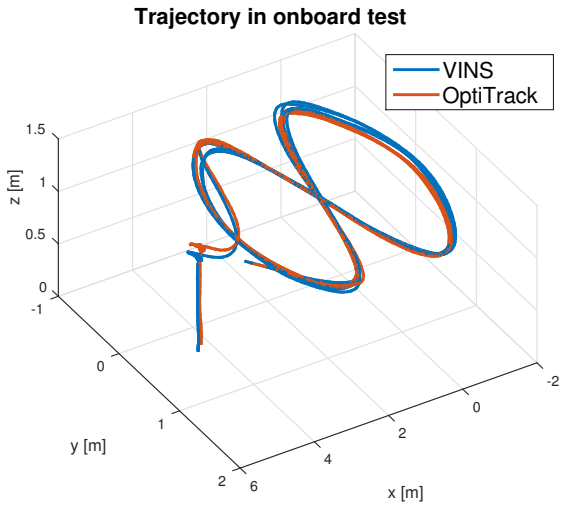


Fig. 21. The trajectory of VINS on the MAV platform without loop closure compared with ground truth. The MAV follows the trajectory four times. VINS outputs are used as real-time feedback in the real-time control loop. The ground truth are provided by OptiTrack. Total length is $61.97m$. Final drift is $0.18m$.

to control the quadrotor.

The robustness and accuracy are of vital importance to this real-time onboard experiment. The final drift is [0.08, 0.09, 0.13] m, relative to the total 61.97 m path length, by comparing with OptiTrack[4] without loop closure. The final percentage of drift is 0.29%. The details of the translation and rotation as well as their corresponding error are shown in Fig. 22.

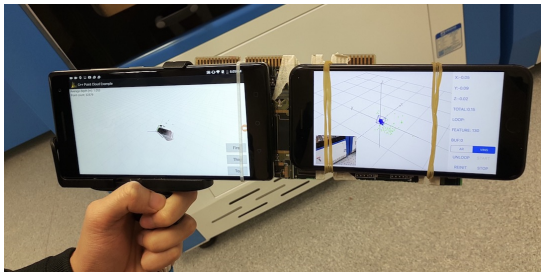


Fig. 23. The simple holder that we used to mount the Tango device (left) and the iPhone7 Plus (right) on which our algorithm runs.

E. Application II: Mobile Device

To validate the practicability of the proposed algorithm, we port the whole system to mobile devices and present a simple AR application to show its accuracy and robustness. We compare the performance of VINS with Google Tango device[5], which is one of the best commercial augmented reality solutions on mobile platforms in the current stage.

In this experiment, we implement the whole VINS system on iPhone7 Plus. we use 30 Hz images with 640 $\times$ 480 resolution captured by the iPhone, and IMU data at 100 Hz obtained by the built-in InvenSense MP67B 6-axis gyroscope and accelerometer. As Fig. 14 shows, we mount the iPhone with the Tango-Enabled smartphone Lenovo Phab 2 Pro which uses the fisheye camera, synchronized IMU, to estimate the state and percept the environment. Firstly, we insert a virtual cube on the plane which is extracted from estimated visual features as shown in Fig. 24(a). Then we hold the two devices and walk inside and outside the room in a normal pace. After loop detected, we use the 4-DoF pose graph optimization (Sect. VII-D) to correct the x, y, z and yaw drift for all the

[4]http://optitrack.com/

[5]http://shopap.lenovo.com/hk/en/tango/

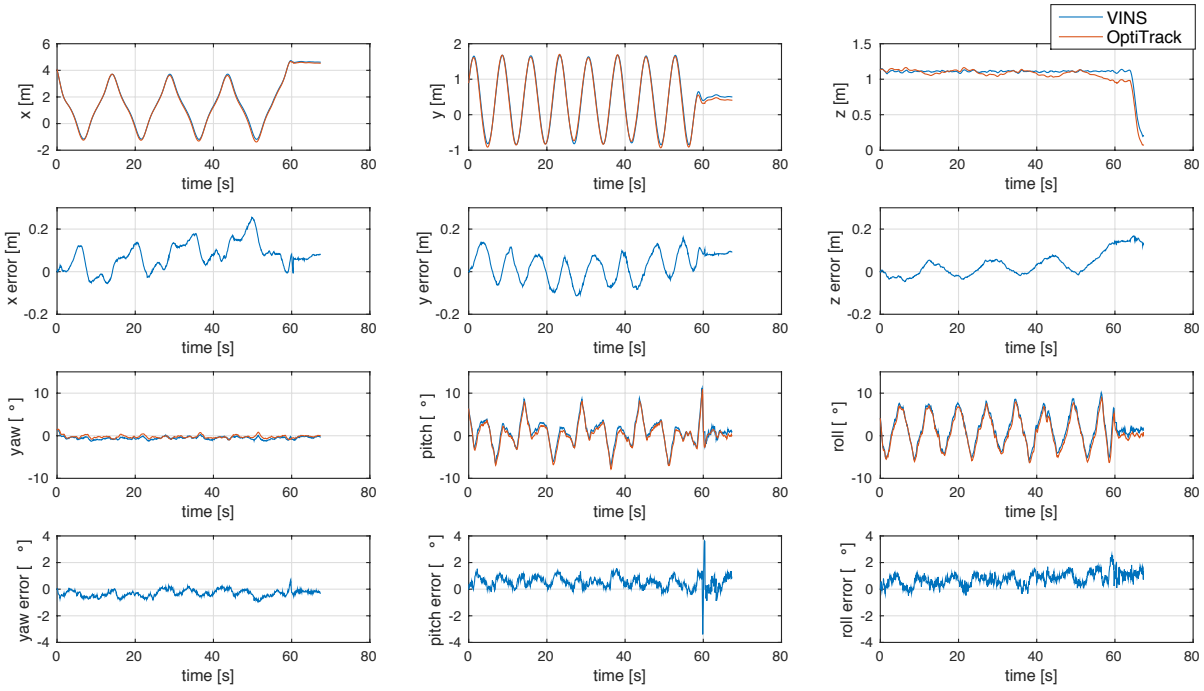


Fig. 22. Position, orientation and their corresponding error of proposed system compared with OptiTrack in the onboard experiment. An impulse in pitch error at the 60s is caused by a violent break at the end of designed trajectory.

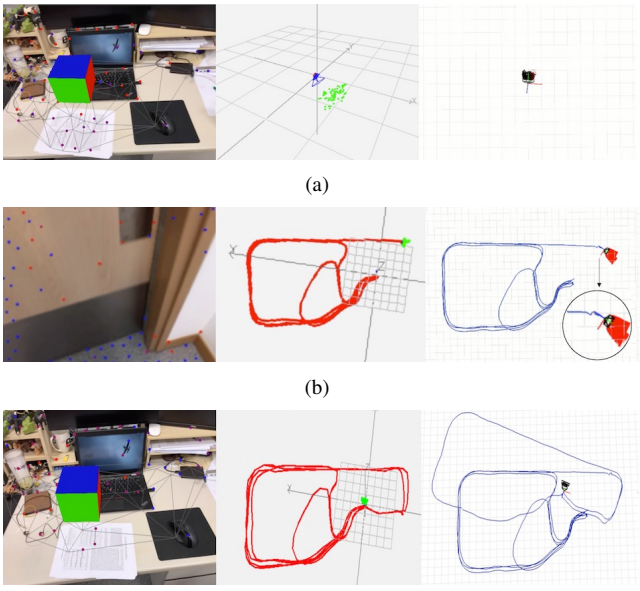
keyframes between the looped old frame and the current frame in keyframe database. As illustrated in Fig. 24(b), an obvious drift occurs in Tango while our system still works well in this challenging case. After traveling about 264m, we return to the start location. The whole result can be seen in Fig. 24(c), the trajectory of tango occurs drift in the last lap while our VINS returns to the start point and the drift in total trajectory is eliminated due to the 4-DoF pose graph optimization. This is also evidenced by the fact that the cube is registered to the same place on the image comparing to the beginning. All the experiment details can be found in the supplementary video.

IX. Conclusion

In this paper, we perform a complete monocular visual-inertial system for 6-DoF state estimation. Our system is robust to initialization with unknown states, online camera-IMU extrinsic parameter calibration, loop detection and 4-DoF pose graph optimization. These characters make our system practicable and easy-to-use. Two applications prove that our system has a great potentiality to be extended for other platforms.

Appendix A
IMU pre-integration

Consider two time instants that correspond to images frame k and $k+1$, several IMU measurements exist in this period, $[b_k..., i-1, i, i+1...b_{k+1}]$. Our goal is to integrate this measurements together independently of the starting pose. α, β, γ are propagated in the local frame k. ba, bg are acceleration and gyroscope bias, which are propagated in body frame.



(a)

(b)

(c)

Fig. 24. From left to right: The AR image of VINS. The trajectory of VINS. The trajectory of Tango device. (a) The VINS and Tango are initialized at the start location and a virtual box is inserted on the plane which extracted from estimated features. (b) A challenging case in which the cameras watch the moving door. The drift of Tango trajectory is highlighted. (c) The whole trajectories of VINS and Tango during we walk inside and outside the room. The total length is about 264m.

In discrete-time implementation, various numerical integration methods such as Euler, Mid-point, RK4 integration can be applied. Here Euler integration is chose to demonstrate the procedure for easy understanding.

Assuming that the noise in acceleration and gyroscope measurements are Gaussian white noise, $n_a \sim \mathcal{N}(0, {\sigma_a}^2)$, $n_w \sim \mathcal{N}(0, {\sigma_w}^2)$. The bias of acceleration and gyroscope measurements are random walk, whose derivatives are Gaussian white noise, $n_{ba} \sim \mathcal{N}(0, {\sigma_{ba}}^2)$, $n_{bg} \sim \mathcal{N}(0, {\sigma_{bg}}^2)$.

At beginning, $\alpha_{b_k}^{b_k}, \beta_{b_k}^{b_k}$ is $\mathbf{0}$, $\gamma_{b_k}^{b_k}$ is identity matrix. The mean of $\alpha, \beta, \gamma, ba, bg$ is propagate step by step as follows,

$$\begin{aligned}
\alpha_{i+1}^{b_k} &= \alpha_i^{b_k} + \beta_i^{b_k}\delta t + \frac{1}{2}\gamma_i^{b_k}(a_i + n_a - ba_i)\delta t^2 \\
\gamma_{i+1}^{b_k} &= \gamma_i^{b_k} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}(a_i + n_a - ba_i)\delta t \end{bmatrix}
\end{aligned} \tag{20}$$

$$\begin{aligned}
\beta_{i+1}^{b_k} &= \beta_i^{b_k} + q_i^{b_k}(w_i + n_w - bg_i)\delta t \\
ba_{i+1} &= ba_i + n_{ba}\delta t \\
bg_{i+1} &= bg_i + n_{bg}\delta t.
\end{aligned}$$

Then we deal with the covariance propagation. For simplicity, we denote $x_i = [\alpha_i^{b_k}, \beta_i^{b_k}, \gamma_i^{b_k}, ba_i, bg_i]^T$, $n = [n_a, n_w, n_{ba}, n_{bg}]^t$, and $Q = diag({\sigma_a}^2, {\sigma_w}^2, {\sigma_{ba}}^2, {\sigma_{bg}}^2)$. The beginning covariance $P_{b_k}^{b_k}$ of state x_{b_k} is $\mathbf{0}$, we need to propagate it to time b_{k+1}. Firstly, we linearize the above-mentioned equations eq. 20 in its error-state representation:

$$\delta x_{i+1} = F_i \delta x_i + G_i n. \tag{21}$$

$$\begin{aligned}
F_i &= \begin{bmatrix}
I & \frac{1}{2}(-q_i^{b_k}\lfloor a_i - ba_i \rfloor_\times)\delta t^2 & \delta t & -\frac{1}{2}q_i^{b_k}\delta t^2 & 0 \\
0 & I - \lfloor w_i - bg_i \rfloor_\times \delta t & 0 & 0 & -\delta t \\
0 & -q_i^{b_k}\lfloor a_i - ba_i \rfloor_\times \delta t & I & -q_i^{b_k}\delta t & 0 \\
0 & 0 & 0 & I & 0 \\
0 & 0 & 0 & 0 & I
\end{bmatrix} \\
G_i &= \begin{bmatrix}
\frac{1}{2}q_i^{b_k}\delta t^2 & 0 & 0 & 0 \\
0 & \delta t & 0 & 0 \\
q_i^{b_k}\delta t & 0 & 0 & 0 \\
0 & 0 & \delta t & 0 \\
0 & 0 & 0 & \delta t
\end{bmatrix}.
\end{aligned} \tag{22}$$

where $\lfloor \cdot \rfloor_\times$ is the skew-matrix operate. Details about error-state representation of quaternion are in [34]. Then the covariance matrix of time i is propagated to the time $i+1$,

$$\begin{aligned}
P_{i+1}^{b_k} &= F_i^T P_i^{b_k} F_i + G_i^T Q G_i \\
J_{i+1} &= \frac{\partial \delta x_{i+1}}{\partial \delta x_i} = F_i.
\end{aligned} \tag{23}$$

Meanwhile, F_i is also the Jacobian matrix J_{i+1} of δx_{i+1} with respect to δx_i.

By propagating the IMU measurements between $[b_k, b_{k+1}]$ in chain, we obtain the mean $[\alpha_{b_{k+1}}^{b_k}, \beta_{b_{k+1}}^{b_k}, \gamma_{b_{k+1}}^{b_k}, ba_{b_{k+1}}, bg_{b_{k+1}}]$, covariance $P_{b_{k+1}}^{b_k}$ and Jacobian $J_{b_{k+1}}$ finally,

$$\begin{aligned}
P_{b_{k+1}}^{b_k} &= \cdots F_{i+1}^T(F_i^T P_i^{b_k} F_i + G_i^T Q_i G_i)F_{i+1} + G_{i+1}^T Q G_{i+1} \cdots \\
J_{b_{k+1}} &= \cdots \frac{\partial \delta x_{i+2}}{\partial \delta x_{i+1}} \frac{\partial \delta x_{i+1}}{\partial \delta x_i} \frac{\partial \delta x_i}{\partial \delta x_{i-1}} \cdots \\
&= \cdots F_{i+1} F_i F_{i-1} \cdots
\end{aligned} \tag{24}$$

The first order approximation of $\alpha_{b_{k+1}}^{b_k}, \beta_{b_{k+1}}^{b_k}, \gamma_{b_{k+1}}^{b_k}$ with respect to the bias can be write as:

$$\begin{aligned}
\alpha_{b_{k+1}}^{b_k} &\approx \hat{\alpha}_{b_{k+1}}^{b_k} + J_{ba}^{\alpha}\delta ba_k + J_{bg}^{\alpha}\delta bg_k \\
\beta_{b_{k+1}}^{b_k} &\approx \hat{\beta}_{b_{k+1}}^{b_k} + J_{ba}^{\beta}\delta ba_k + J_{bg}^{\beta}\delta bg_k \\
\gamma_{b_{k+1}}^{b_k} &\approx \hat{\gamma}_{b_{k+1}}^{b_k} + J_{bg}^{\gamma}\delta bg_k,
\end{aligned} \tag{25}$$

where J_{ba}^{α} and is the sub-block matrix in $J_{b_{k+1}}$ whose location is corresponding to $\frac{\delta \alpha_{b_{k+1}}^{b_k}}{\delta ba_k}$. The same meaning for $J_{bg}^{\alpha}, J_{ba}^{\beta}, J_{bg}^{\beta}, J_{bg}^{\gamma}$. Note that $\gamma_{b_{k+1}}^{b_k}$ is only related to bg_k.

ACKNOWLEDGMENT

Thanks Lin Yonggen for large-scale environment dataset collection.

REFERENCES

[1] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. IEEE, 2007, pp. 225–234.

[2] 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.

[3] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer International Publishing, 2014, pp. 834–849.

[4] R. Mur-Artal, J. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[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, Jul. 2015.

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

[23] 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.

[24] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012.

[25] H. Strasdat, J. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular slam," *Robotics: Science and Systems VI*, 2010.

[26] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the Intl. Joint Conf. on Artificial Intelligence*, Vancouver, Canada, Aug. 1981, pp. 24–28.

[27] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 593–600.

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

[29] 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.

[30] 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.

[31] S. Agarwal, K. Mierle, and Others, "Ceres solver," http://ceres-solver.org.

[32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *Computer Vision–ECCV 2010*, pp. 778–792, 2010.

[33] 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.

[34] N. Trawny and S. I. Roumeliotis, "Indirect kalman filter for 3d attitude estimation," *University of Minnesota, Dept. of Comp. Sci. & Eng., Tech. Rep*, vol. 2, p. 2005, 2005.