Accurate Monocular Visual-inertial SLAM using a Map-assisted EKF Approach

Meixiang Quan, Songhao Piao, Minglang Tan, Shi-Sheng Huang

***Abstract*—This paper presents a novel tightly-coupled monocular visual-inertial Simultaneous Localization and Mapping algorithm, which provides accurate and robust localization within the globally consistent map in real time on a standard CPU. This is achieved by firstly performing the visual-inertial extended kalman filter(EKF) to provide motion estimate at a high rate. However the filter becomes inconsistent due to the well known linearization issues. So we perform a keyframe-based visual-inertial bundle adjustment to improve the consistency and accuracy of the system. In addition, a loop closure detection and correction module is also added to eliminate the accumulated drift when revisiting an area. Finally, the optimized motion estimates and map are fed back to the EKF-based visual-inertial odometry module, thus the inconsistency and estimation error of the EKF estimator are reduced. In this way, the system can continuously provide reliable motion estimates for the long-term operation. The performance of the algorithm is validated on public datasets and real-world experiments, which proves the superiority of the proposed algorithm.**

***Index Terms*—Simultaneous localization and mapping, Visual-inertial odometry, Visual-inertial sensor Fusion, State estimation, Optimization.**

I. Introduction

Concurrent motion estimation and map reconstruction by combining visual and inertial measurements has received significant interest in the field of Robotics and Computer Vision communities. This visual-inertial sensor suite can serve as an ideal alternative to GPS in environments where GPS is denied, since both sensors are small, lightweight, cheap enough and complementary. On the one hand, Visual SLAM can provide good tracking and rich map information in visually distinguishable environments. However due to the sensor limitation, visual simultaneous localization and mapping(SLAM) is sensitive to motion blur, occlusions and illumination changes. On the other hand, inertial sensors are able to provide self-motion information at high frequency, so inertial navigation is robust to aggressive motion and can provide absolute scale for the motion. Whereas the result of inertial navigation is noisy and diverges even in a few seconds. Therefore, fusing measurements from the inertial sensor to the visual SLAM in a tightly-coupled way, both the robustness and the accuracy of motion tracking can be dramatically improved. The advantage of visual-inertial sensor fusion is the most obvious in a monocular visual-inertial setup, because the scale of the motion estimation and map structure computed from monocular SLAM is ambiguous and liable to drift over time.

In this paper, we aim to build a system which enables the accurate and robust motion tracking within the accurately and consistently reconstructed map. Firstly, we employ the EKF-based visual-inertial odometry(VIO) to track the 3D motion of the IMU body frame in real time. Whereas due to the linearization errors, the estimator tends to be inconsistent, which results in large estimation errors and divergence. So we intend to perform bundle adjustment(BA) to reduce the inconsistency of the estimator. It is demonstrated in [1] that the BA techniques can achieve better accuracy than filtering techniques, because optimization methods can iteratively relinearize measurement equations to better deal with their nonlinearity. However, the optimization method leads to a high computational cost. In addition, increasing the number of feature correspondences and keyframes in the window of local BA will lead to significant increase in accuracy, whereas feature extraction and matching, optimization for the local BA are also time consuming. So if we increase the number of features, and perform the local BA after the EKF for each frame to increase the accuracy of the system, the system will incapable of real time operation. Thus, in order to ensure both real time operation and accurate global map reconstruction, we extract a vast number of features only for those selected keyframes, we use these features to construct global map and perform the local BA in a parallel thread. In this way, we can properly solve the loss of accuracy due to the linearization in real-time. Besides, if the system is unable to close loops, the drift of the estimated trajectory will accumulate without bound, even if the sensor is continuously revisiting the same place. Therefore, we add a loop closure module in a new parallel thread for reducing the accumulated drift when returning to an already mapped area.

In summary, we propose a tightly-coupled monocular visual-inertial SLAM(VISLAM) system, which is able to perform real-time, accurate, robust and long term localization and map reconstruction. Our approach operates in three parallel threads, one thread is used to perform the EKF VIO, and the other two threads, one for BA and the other one for loop closing, are used to compensate for the error growth of the estimated trajectory and construct an accurate and consistent global map. In VIO thread, since the computational cost of the EKF is quadratic in the number of features, for the real-time state estimation we extract the appropriate number of features for each frame, and extract a lot of features only in selected keyframes for performing local BA and loop closure. Finally, we design a feedback mechanism to increase both the consistency and accuracy of the EKF estimator, which is achieved by feeding back the optimized state and globally

S. Piao(piaosh@hit.edu.cn) and S. Huang(shishenghuang.net@gmail.com) are the corresponding authors

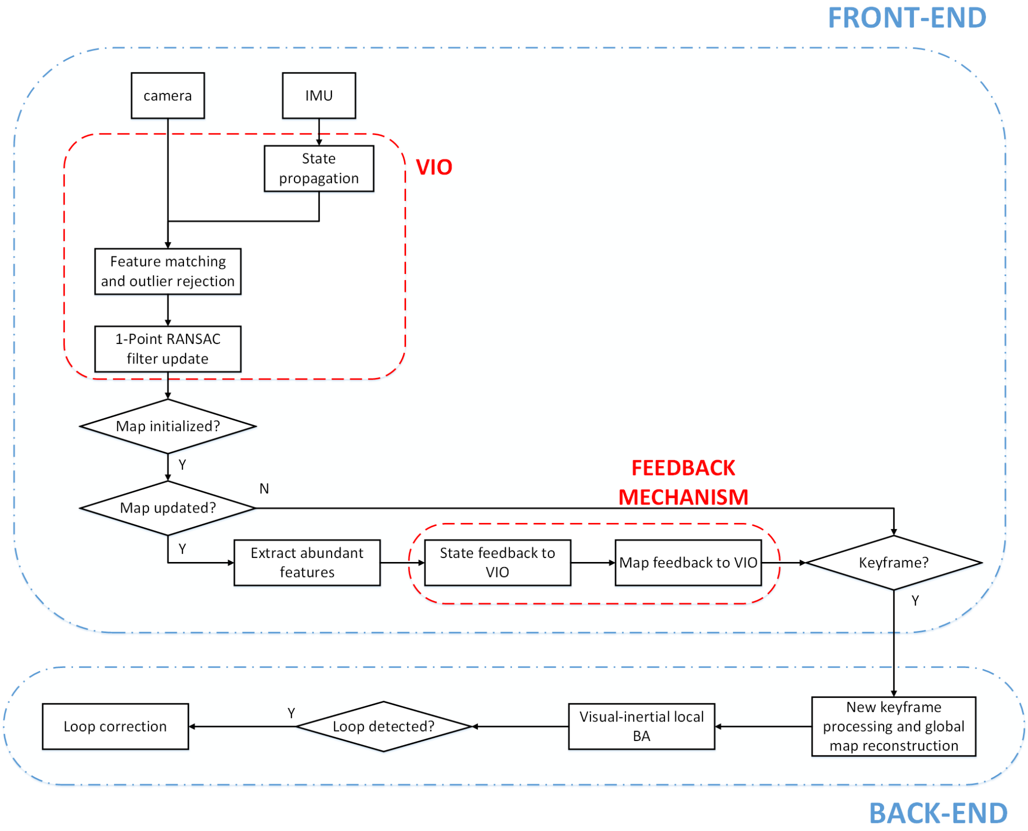

Fig. 1. An overview of the proposed monocular VISLAM algorithm, which contains two main components: EKF based VIO front-end and nonlinear optimization and loop closure back-end.

consistent map to update the state vector of the EKF VIO module.

In experiments, the results demonstrate the benefits of our system towards the EKF based VIO. We also compare to the state-of-the-art VIO and VISLAM approaches, and demonstrate the superior performance of our method.

Our monocular visual-inertial SLAM algorithm is shown in Fig.1. The system is complete and drift-free in large scale environments. The remainder of the paper is organized as follows. In Section II, we describe the relevant literature. Notations are given in Section III. In Section IV, EKF based VIO is presented and the Jacobian matrices of EKF are given in appendix. A tightly-coupled joint visual-inertial nonlinear optimization is introduced In Section V. Section VI introduce our tightly-coupled VISLAM approach. Experiments results are shown in Section VII. Finally, the paper is concluded in Section VIII.

II. Related Work

There are a vast amount of research towards visual SLAM problem, we refer to the review paper [2] for the progress made in the past few decades. In this section, we will discuss the most relevant works on monocular VIO and VISLAM system.

The fusion for visual and inertial measurements is usually divided into two classes. Loosely-coupled approaches, e.g. [3] [4], process the visual and inertial measurements separately. Therefore the accumulated drift in vision module cannot be eliminated from the usability of inertial measurements, which leads the resulted estimate to be sub-optimal. Tightly-coupled ones interested in this work, e.g. [5]–[19], perform VIO or VISLAM system by considering the tight interaction between visual and inertial measurements to optimally exploit the both sensing cues, thus achieve higher precision at the expensive of additional computational complexity. Besides, for tightly-coupled VIO/VISLAM solutions, two methodologies have been prevalent: filtering-based methods [5]–[13] and BA-based methods [14]–[20].

Historically, the monocular SLAM problem has been addressed with filtering method, which operates on the mean and covariance of the probabilistic distribution in a kalman filtering framework. Filtering based approaches require fewer computational resources due to the continuous marginalization of past state, however the system get slightly lower accuracy due to the linearization error. According to the way processing the measurement information, the recursive filtering approaches can be classified into two main categories: extended kalman filter (EKF) based methods [5]–[9] and sliding window filtering approaches [10]–[13]. The state vector of EKF-based SLAM algorithms include both the pose of the platform and a set of feature positions, so as long as these features are continuously observed and contained in the state vector, the estimated pose relative to these features will not drift. However it have high computational complexity (quadratic in the number of features in the state vector), therefore only currently observable landmarks are tracked to ensure real-time operation. In contrast, sliding window filtering approaches maintain a sliding window of past camera poses in the state vector, and use the feature measurements to impose probabilistic constraints on these poses, therefore keep computational complexity only linear in the number of features by excluding point features from filter state vector. Generally, the VIO problem has four unobservable directions, but since the linearization errors, the system only have three unobservable directions, which renders the filter inconsistent. So papers [8], [9], [11]–[13] were proposed a series of methods, e.g. first-estimates Jacobian and constraint of system observability, to improve the consistency of the system. If the measurement models were linear, both methods yield the same result equal to the MAP estimate.

In [1], it was shown that nonlinear optimization-based approaches provide better accuracy than filtering approaches by its capability to relinearize the state at each iteration, therefore avoiding integrated error from linearization, however it leads to higher computational demands. In following, we introduce several classic tightly-coupled BA-based VIO and VISLAM system. OKVIS [14] presented an approach to tightly integrate inertial measurements into keyframe-based visual SLAM, which makes the nonlinear cost function comprised of IMU error term with the landmark reprojection error term to be jointly optimized. Additionally, marginalization of old state is used to maintain a bounded-sized optimization window, therefore OKVIS has achieved increased accuracy and robustness in real-time operation. However, the system needs to repeatedly compute the IMU integration when the linearization point changes. To eliminate this repeated computation, Foster et al. provided a preintegration theory for inertial measurements in [15] that properly address the manifold structure of the rotation group based on the work of [21]. Then, the preintegrated IMU model and structureless visual model are seamlessly integrated in a fully probabilistic manner to build a much more computationally efficient optimization method for state estimation. Therefore, the system achieves better accuracy than Project Tango [22] by using SVO as front-end and the visual-inertial joint optimization in back-end. Tightly-coupled visual-inertial odometry methods mentioned above all lack the capability to close loops and reuse an already reconstructed map, due to the marginalization of past states to maintain a constant computational cost or the use of full smoothing. Thus, ORB-VISLAM [17] presented a real-time tightly-coupled monocular visual-inertial SLAM system, which enables the loop closure and the reuse of previously estimated 3D map. The system achieved higher accuracy than the fully direct, stereo visual-inertial odometry [18], because there is no drift accumulation for localization in already mapped areas. Recently, a novel real-time, tightly-coupled, sliding window optimization based versatile monocular visual-inertial odometry [19] [20] was proposed, in which, the state of the system and a representation of the environment are estimated by local BA in one thread, and loops are closed in lightweight manner in parallel thread.

Both filtering-based methods and BA-based methods have their merits, so in this work, we tightly fuse both methods to achieve the best accuracy, robustness and efficiency.

III. Notations

Throughout the paper, we denote the world reference frame as $(\cdot)^W$, and denote the IMU body frame and camera frame for the k^{th} image as $(\cdot)^{B_k}$ and $(\cdot)^{C_k}$ respectively. In addition, we employ $\mathbf{R}_{\mathcal{F}_2}^{\mathcal{F}_1} \in \mathbf{SO}(3)$ to represent rotation from frame $\{\mathcal{F}_2\}$ to $\{\mathcal{F}_1\}$, $\mathbf{p}_{\mathcal{F}_2}^{\mathcal{F}_1} \in \mathbb{R}^3$ and $\mathbf{v}_{\mathcal{F}_2}^{\mathcal{F}_1} \in \mathbb{R}^3$ to describe the 3D position and velocity of frame $\{\mathcal{F}_2\}$ with respect to the frame $\{\mathcal{F}_1\}$. Besides, the rotation and translation between the rigidly mounted camera-IMU sensor are denoted as $\mathbf{R}_C^B$ and $\mathbf{p}_C^B$, which was computed from the calibration.

For the over-parameterized rotation matrix, a vector $\boldsymbol{\xi} \in \mathbb{R}^3$ can be computed from the tangent space $\mathfrak{so}(3)$ of manifold $\mathbf{SO}(3)$ to provide a minimal representation. The Lie algebra $\mathfrak{so}(3)$ and Lie group $\mathbf{SO}(3)$ are related through the logarithm map and exponential map:

$$\boldsymbol{\xi} = \mathrm{Log}(\mathbf{R}) = \log(\mathbf{R})^{\vee} \tag{1}$$

$$\mathbf{R}(\boldsymbol{\xi}) = \mathrm{Exp}(\boldsymbol{\xi}) = \exp(\boldsymbol{\xi}^{\wedge}) \tag{2}$$

where $(\cdot)^{\wedge}$ operator maps a vector in $\mathbb{R}^3$ to a 3×3 skew symmetric matrix, and $(\cdot)^{\vee}$ is the inverse operator. The formula for $\log(\cdot)$ and $\exp(\cdot)$ can be found in [23].

Furthermore, the uncertainty of rotation $\mathbf{R} \in \mathbf{SO}(3)$ and $\boldsymbol{\xi} \in \mathbb{R}^3$ are described as:

$$\mathbf{R} = \hat{\mathbf{R}} \oplus \delta\boldsymbol{\xi} = \hat{\mathbf{R}}\mathrm{Exp}(\delta\boldsymbol{\xi}) \tag{3}$$

$$\boldsymbol{\xi} = \hat{\boldsymbol{\xi}} \oplus \delta\boldsymbol{\xi} = \mathrm{Log}(\mathrm{Exp}(\hat{\boldsymbol{\xi}})\mathrm{Exp}(\delta\boldsymbol{\xi})) \tag{4}$$

where $\hat{\mathbf{R}}$ and $\hat{\boldsymbol{\xi}}$ are the mean estimate of $\mathbf{R}$ and $\boldsymbol{\xi}$ respectively, and $\delta\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a normally distributed perturbation.

For visual measurements, we consider a projection function $\pi : \mathbb{R}^3 \rightarrow \Omega \subset \mathbb{R}^2$, which projects the l^{th} map point expressed in the current camera frame $\mathbf{f}_l^{C_k} = [x_l^{C_k} \; y_l^{C_k} \; z_l^{C_k}]^{\mathrm{T}} \in \mathbb{R}^3$ onto 2D points on the image plane $\mathbf{z}_{kl} = [u_{kl} \; v_{kl}]^{\mathrm{T}} \in \Omega$:

$$\begin{aligned} \widetilde{\mathbf{z}}_{kl} &= \mathbf{z}_{kl} + \boldsymbol{\sigma}_{kl} \\ &= \pi(\mathbf{f}_l^{C_k}) + \boldsymbol{\sigma}_{kl} \\ &= \begin{bmatrix} f_x \frac{x_l^{C_k}}{z_l^{C_k}} + c_x \\ f_y \frac{y_l^{C_k}}{z_l^{C_k}} + c_y \end{bmatrix} + \boldsymbol{\sigma}_{kl} \end{aligned} \tag{5}$$

where $\widetilde{\mathbf{z}}_{kl}$ is the corresponding feature measurement, and $\boldsymbol{\sigma}_{kl}$ is the 2×1 measurement noise with covariance $\boldsymbol{\Sigma}_{\sigma_{kl}}$ associated to the feature scale. In addition, f_x, f_y are focal length and c_x, c_y are principle point, which are known from calibration.

IV. VIO Description

In this section, we describe the EKF VIO system, which is based on the work of [24]. An overview of the algorithm is given in VIO part of Fig. 1. Inertial measurements are used to predict the motion movement in the prediction stage, then the state is updated using the matched visual features. In this way, we can estimate the state of the body frame efficiently.

A. Full State Vector

The state vector to be estimated comprises the IMU state $\mathbf{X}_{B_k}$ and a set of landmark parameters $\mathbf{X}_{L_k}$:

$$\mathbf{X}_k = [\mathbf{X}_{B_k}^{\mathrm{T}} \ \mathbf{X}_{L_k}^{\mathrm{T}}]^{\mathrm{T}} \tag{6}$$

The IMU state is formulated by the vector:

$$\mathbf{X}_{B_k} = [\boldsymbol{\xi}_{B_k}^{W\ \mathrm{T}} \ \mathbf{p}_{B_k}^{W\ \mathrm{T}} \ \mathbf{v}_{B_k}^{W\ \mathrm{T}} \ \mathbf{b}_{a_k}^{\mathrm{T}} \ \mathbf{b}_{g_k}^{\mathrm{T}}]^{\mathrm{T}} \tag{7}$$

where $\boldsymbol{\xi}_{B_k}^{W} \in \mathbb{R}^3$ is the Lie algebra of orientation $\mathbf{R}_{B_k}^{W} \in \mathbf{SO}(3)$ from frame $\{B_k\}$ to $\{W\}$, $\mathbf{p}_{B_k}^{W} \in \mathbb{R}^3$ and $\mathbf{v}_{B_k}^{W} \in \mathbb{R}^3$ are the 3D position and velocity of frame $\{B_k\}$ with respect to $\{W\}$, as well as $\mathbf{b}_a \in \mathbb{R}^3$ and $\mathbf{b}_g \in \mathbb{R}^3$ are additive accelerometer and gyroscope biases respectively. Following (7), the IMU error state vector is defined as:

$$\delta\mathbf{X}_{B_k} = [\delta\boldsymbol{\xi}_{B_k}^{\mathrm{T}} \ \delta\mathbf{p}_{B_k}^{W\ \mathrm{T}} \ \delta\mathbf{v}_{B_k}^{W\ \mathrm{T}} \ \delta\mathbf{b}_{a_k}^{\mathrm{T}} \ \delta\mathbf{b}_{g_k}^{\mathrm{T}}]^{\mathrm{T}} \tag{8}$$

where we use the standard additive error for the 3D position, velocity and biases, while for rotation, the error is defined as (4).

Assuming that m features are included in the map at time-step k, then the coordinates of features are:

$$\mathbf{X}_{L_k} = [\mathbf{f}_1^{W\mathrm{T}} \ \cdots \ \mathbf{f}_m^{W\mathrm{T}}]^{\mathrm{T}} \tag{9}$$

The position of the l^{th} landmark $\mathbf{f}_l^W$ is paramterized in inverse depth coordinates as:

$$\mathbf{f}_l^W = [x_l \ y_l \ z_l \ \theta_l \ \phi_l \ \rho_l]^{\mathrm{T}} \tag{10}$$

where $(x_l, y_l, z_l)^{\mathrm{T}}$ is the camera position, in which the l^{th} landmark was firstly observed. θ_l and ϕ_l are the azimuth and elevation angle defining unit ray(expressed in the world frame) that goes from the camera center $(x_l, y_l, z_l)^{\mathrm{T}}$ to the l^{th} landmark, and ρ_l is its inverse depth along the unit ray.

Therefore the EKF error state vector is expressed as:

$$\delta\mathbf{X}_k = [\delta\mathbf{X}_{B_k}^{\mathrm{T}} \ \delta\mathbf{f}_1^{W\mathrm{T}} \ \cdots \ \delta\mathbf{f}_m^{W\mathrm{T}}]^{\mathrm{T}} \tag{11}$$

where we use the standard additive error for the landmark.

B. IMU Propagation Model

Different to many other visual-inertial methods, we define the state propagation model directly in discrete-time to make the derivatives required for the EKF prediction are available in close-form. Using the measured acceleration $\widetilde{\mathbf{a}}_{k-1}$ and angular velocity $\widetilde{\boldsymbol{\omega}}_{k-1}$ obtained from IMU, the discrete-time propagation model $\mathbf{X}_{B_{k|k-1}} = \mathbf{f}_k(\mathbf{X}_{B_{k-1}})$ is:

$$\begin{aligned}
\boldsymbol{\xi}_{B_{k|k-1}}^{W} &= \mathrm{Log}\left(\mathrm{Exp}(\boldsymbol{\xi}_{B_{k-1}}^{W})\mathrm{Exp}\left((\widetilde{\boldsymbol{\omega}}_{k-1} - \mathbf{b}_{g_{k-1}} - \mathbf{n}_{gd})\Delta t\right)\right) \\
\mathbf{p}_{B_{k|k-1}}^{W} &= \mathbf{p}_{B_{k-1}}^{W} + \mathbf{v}_{B_{k-1}}^{W}\Delta t \\
\mathbf{v}_{B_{k|k-1}}^{W} &= \mathbf{v}_{B_{k-1}}^{W} + (\mathbf{R}_{B_{k-1}}^{W}(\widetilde{\mathbf{a}}_{k-1} - \mathbf{b}_{a_{k-1}} - \mathbf{n}_{ad}) + \mathbf{g}^W)\Delta t \\
\mathbf{b}_{a_{k|k-1}} &= \mathbf{b}_{a_{k-1}} \\
\mathbf{b}_{g_{k|k-1}} &= \mathbf{b}_{g_{k-1}}
\end{aligned} \tag{12}$$

where $\mathbf{R}_{B_{k-1}}^{W} = \mathrm{Exp}(\boldsymbol{\xi}_{B_{k-1}}^{W})$, $\mathbf{n}_{gd} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_g/\Delta t)$ and $\mathbf{n}_{ad} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_a/\Delta t)$ are discrete-time white Gaussian noise for inertial measurements, and $\mathbf{g}^W$ is the constant gravity. In this work, we ignore the slow random walk of the inertial biases, so the biases are considered fixed and estimated as part of the state. The linearized discrete-time IMU error state propagation model is represented as:

$$\delta\mathbf{X}_{B_{k|k-1}} = \boldsymbol{\Phi}_k \delta\mathbf{X}_{B_{k-1}} + \mathbf{G}_k \mathbf{n}_B \tag{13}$$

where $\mathbf{n}_B = [\mathbf{n}_{ad}{}^{\mathrm{T}} \ \ \mathbf{n}_{gd}{}^{\mathrm{T}}]^{\mathrm{T}}$ is the system noise with covariance $\mathcal{Q} = \begin{bmatrix} \boldsymbol{\Sigma}_a/\Delta t & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \boldsymbol{\Sigma}_g/\Delta t \end{bmatrix}$. The matrices $\boldsymbol{\Phi}_k$ and $\mathbf{G}_k$ in (13) are Jacobians of $\mathbf{f}_k(\cdot)$ with respect to IMU state and the system noise, these representation can be found in appendix A.

Therefore, the covariance matrix is propagated as follows:

$$\begin{aligned}
\mathbf{P}_{k|k-1} &= \begin{bmatrix} \mathbf{P}_{B_{k|k-1}} & \mathbf{P}_{BL_{k|k-1}} \\ \mathbf{P}_{LB_{k|k-1}} & \mathbf{P}_{L_{k|k-1}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Phi}_k \mathbf{P}_{B_{k-1}} \boldsymbol{\Phi}_k{}^{\mathrm{T}} + \mathbf{G}_k \mathcal{Q} \mathbf{G}_k{}^{\mathrm{T}} & \boldsymbol{\Phi}_k \mathbf{P}_{BL_{k-1}} \\ \mathbf{P}_{LB_{k-1}} \boldsymbol{\Phi}_k{}^{T} & \mathbf{P}_{L_{k-1}} \end{bmatrix}
\end{aligned} \tag{14}$$

C. Measurement Model

The inverse depth parameterization is used for features to (1) enhance the degree of linearity for measurement equations, and (2) better deal with the low parallax features. The Euclidean XYZ coordinates of the l^{th} feature in world frame $\mathbf{y}_l^W$ can be transformed from its inverse depth representation $\mathbf{f}_l^W$ as:

$$\mathbf{y}_l^W = \begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix} + \frac{1}{\rho_l}\mathbf{m}(\theta_l, \phi_l) \tag{15}$$

$$\mathbf{m}(\theta_l, \phi_l) = \begin{bmatrix} \cos\phi_l \sin\theta_l \\ -\sin\phi_l \\ \cos\phi_l \cos\theta_l \end{bmatrix} \tag{16}$$

Thus the measurement model representing the projection of the l^{th} landmark to the k^{th} image is:

$$\begin{aligned}
\widetilde{\mathbf{z}}_{kl} &= \mathbf{h}_{kl}(\mathbf{X}_{B_{k|k-1}}, \mathbf{f}_l^W) + \boldsymbol{\sigma}_{kl} \\
&= \pi\left(\mathbf{f}_l^{C_k}\right) + \boldsymbol{\sigma}_{kl}
\end{aligned} \tag{17}$$

where $\mathbf{f}_l^{C_k} = \mathbf{R}_W^{C_{k|k-1}}\left(\rho_l\left(\begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix} - \mathbf{p}_{C_{k|k-1}}^W\right) + \mathbf{m}(\theta_l, \phi_l)\right)$, $\mathbf{R}_W^{C_{k|k-1}} = (\mathbf{R}_{B_{k|k-1}}^W \mathbf{R}_C^B)^{\mathrm{T}}$ and $\mathbf{p}_{C_{k|k-1}}^W = \mathbf{p}_{B_{k|k-1}}^W + \mathbf{R}_{B_{k|k-1}}^W \mathbf{p}_C^B$. From the measurement model, we compute the reprojection error and its linearized approximation as:

$$\begin{aligned}
\mathbf{r}_{kl} &= \widetilde{\mathbf{z}}_{kl} - \mathbf{h}_{kl}(\mathbf{X}_{B_{k|k-1}}, (\mathbf{X}_{L_k})_l) \\
&\simeq \mathbf{H}_{B_{kl}} \delta\mathbf{X}_{B_{k|k-1}} + \mathbf{H}_{f_{kl}} \delta\mathbf{f}_l^W + \boldsymbol{\sigma}_{kl}
\end{aligned} \tag{18}$$

where the matrices $\mathbf{H}_{B_{kl}}$ and $\mathbf{H}_{f_{kl}}$ are derivatives of the measurement model with respect to the IMU state estimate and the l^{th} feature position respectively, their expressions are given in appendix B.

Therefore, we can obtain the measurement Jacobian matrix as:

$$\mathbf{H}_{kl} = [\mathbf{H}_{B_{kl}} \ \mathbf{0} \ \cdots \ \mathbf{H}_{f_{kl}} \ \mathbf{0} \ \cdots] \tag{19}$$

D. Filter Update

To perform an update of the estimated state, we stack the m individual measurement residual $\mathbf{r}_{kl}$ at time-step k together to form a single $2m \times 1$ residual vector $\mathbf{r}_k = \left[\mathbf{r}_{k1}^{\mathrm{T}} \cdots \mathbf{r}_{kl}^{\mathrm{T}} \cdots \mathbf{r}_{km}^{\mathrm{T}}\right]^{\mathrm{T}}$. In the same way, the measurement Jacobians are also combined to a single $2m \times n$ measurement matrix $\mathbf{H}_k = \left[\mathbf{H}_{k1}^{T} \cdots \mathbf{H}_{kl}^{\mathrm{T}} \cdots \mathbf{H}_{km}^{T}\right]^{\mathrm{T}}$. Then the kalman gain is computed as:

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^{\mathrm{T}}(\mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^{\mathrm{T}} + \boldsymbol{\Sigma})^{-1} \tag{20}$$

where $\boldsymbol{\Sigma}$ is the stacked $2m \times 2m$ measurement uncertainty. Finally, the full state and covariance are updated by:

$$\begin{aligned} \mathbf{X}_{k|k} &= \mathbf{X}_{k|k-1} \circ \mathbf{K}_k\mathbf{r}_k \\ \mathbf{P}_{k|k} &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k|k-1} \end{aligned} \tag{21}$$

where $\circ$ operator is equal to the $\oplus$ operator in (4) for the orientation and the addition of vector for other state.

Based on the predicted IMU state, features used to update the state are searched with the optical flow method. Prior to using each feature's measurement to update, for each matched feature, the Mahalanobis distance $d = \mathbf{r}_{kl}^{\mathrm{T}}(\mathbf{H}_{kl}\mathbf{P}_{k|k-1}\mathbf{H}_{kl}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\sigma_{kl}})^{-1}\mathbf{r}_{kl}$ is firstly computed to reject outliers. If d is smaller than a threshold given by the 95-th percentile of the χ^2 distribution, the feature is accepted as an inlier, and used for the filter update. We perform the filter update by combining 1-point RANSAC method as in [25] to find reliable inliers.

E. State Augmentation

Once a new feature l is needed at time-step k, the initial position $\mathbf{f}_l^W = [\breve{x}_l \; \breve{y}_l \; \breve{z}_l \; \breve{\theta}_l \; \breve{\phi}_l \; \breve{\rho}_l]^{\mathrm{T}}$ of the new feature is computed as follows. The camera position is computed by:

$$[\breve{x}_l \; \breve{y}_l \; \breve{z}_l]^{\mathrm{T}} = \mathbf{R}_{B_{k|k}}^W \mathbf{p}_C^B + \mathbf{p}_{B_{k|k}}^W \tag{22}$$

Besides, from the observation $[u_{kl} \; v_{kl}]^{\mathrm{T}}$ of the new feature in the image, the angles $\breve{\theta}_l$ and $\breve{\phi}_l$ defining its direction are calculated as:

$$\begin{bmatrix} \breve{\theta}_l \\ \breve{\phi}_l \end{bmatrix} = \begin{bmatrix} \arctan(x_{kl}^W, z_{kl}^W) \\ \arctan(-y_{kl}^W, \sqrt{{x_{kl}^W}^2 + {z_{kl}^W}^2}) \end{bmatrix} \tag{23}$$

$$\boldsymbol{\tau}_{kl}^W = \begin{bmatrix} x_{kl}^W \\ y_{kl}^W \\ z_{kl}^W \end{bmatrix} = \mathbf{R}_{B_{k|k}}^W \mathbf{R}_C^B \begin{bmatrix} \frac{u_{kl} - c_x}{f_x} \\ \frac{v_{kl} - c_y}{f_y} \\ 1 \end{bmatrix} \tag{24}$$

The initial value for $\breve{\rho}_i$ and its standard deviation σ_ρ are set as in [26]. Then the initial position for new feature is appended to the state vector, and the state covariance matrix is also augmented accordingly:

$$\mathbf{P}_{k|k} \leftarrow \mathbf{J} \begin{bmatrix} \mathbf{P}_{k|k} & \mathbf{0}_{(15+6m)\times 6} \\ \mathbf{0}_{6\times(15+6m)} & \boldsymbol{\Sigma}_{h\rho} \end{bmatrix} \mathbf{J}^{\mathrm{T}} \tag{25}$$

where $\boldsymbol{\Sigma}_{h\rho} = \begin{bmatrix} \boldsymbol{\Sigma}_{\sigma_{kl}} & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{1\times 2} & {\sigma_\rho}^2 \end{bmatrix}$ denotes the uncertainty of the visual measurements and the initial inverse depth, for the Jacobian $\mathbf{J}$ we refer the reader to appendix C.

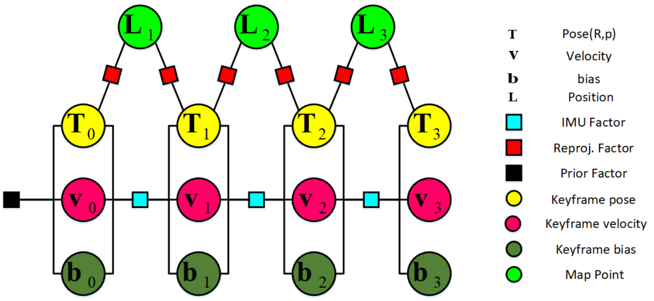

Fig. 2. Factor graph representing the tightly-coupled visual-inertial optimization problem in a sliding window. The states are shown as circles and factors are shown as squares. IMU factors are represented as blue squares, which is connected to the state of the previous keyframe. Red squares denote visual factors corresponding to camera observations, and black squares denote prior factors.

V. VISUAL-INERTIAL BUNDLE ADJUSTMENT

Once a frame processed by EKF based VIO is selected as a keyframe, we apply a nonlinear optimization in a sliding window to improve the accuracy of the estimated state. In this section, we combine the visual and inertial measurements in an unified formulation.

A. Bundle Adjustment Representation

We formulate a joint optimization problem to optimally estimate the full state in a sliding window using all the available inertial and visual measurements. Full state in sliding window contain a set of successive keyframes from i to j and n landmarks visible by the keyframes in sliding window, which is denoted as:

$$\boldsymbol{\mathcal{X}} = \{\mathbf{X}_{B_i}, \; \cdots, \; \mathbf{X}_{B_j}, \; \mathbf{L}_1^W, \cdots, \; \mathbf{L}_n^W\} \tag{26}$$

where $\mathbf{L}_l^W$ is the position of the 3D map point expressed in Euclidean XYZ coordinates. We denote the keyframes and visual measurements in sliding window as $\mathcal{K}$ and $\mathcal{C}$ respectively. Then the energy function that we want to minimize is given by:

$$\begin{aligned} f(\boldsymbol{\mathcal{X}}) = \|\mathbf{r}_p\|_{\boldsymbol{\Sigma}_p}^2 &+ \sum_{k\in\mathcal{K}} \rho\left(\|\mathbf{r}_{\mathcal{I}_{k-1k}}\|_{\boldsymbol{\Sigma}_{\mathcal{I}_{k-1k}}}^2\right) \\ &+ \sum_{k\in\mathcal{K},l\in\mathcal{C}} \rho\left(\|\mathbf{r}_{\mathcal{C}_{kl}}\|_{\boldsymbol{\Sigma}_{\mathcal{C}_{kl}}}^2\right) \end{aligned} \tag{27}$$

where $\mathbf{r}_p$, $\mathbf{r}_{\mathcal{I}_{k-1k}}$, $\mathbf{r}_{\mathcal{C}_{kl}}$ are prior error, temporal IMU error and reprojection error respectively, as well as $\boldsymbol{\Sigma}_p$, $\boldsymbol{\Sigma}_{\mathcal{I}_{k-1k}}$, $\boldsymbol{\Sigma}_{\mathcal{C}_{kl}}$ are the corresponding covariance matrices, and ρ is the Huber robust cost function. The optimization problem can be interpreted as a factor graph shown in Fig. 2.

Therefore the best estimate for variable $\boldsymbol{\mathcal{X}}$ can be obtained by minimizing the objective function on manifold:

$$\boldsymbol{\mathcal{X}}^* = \underset{\boldsymbol{\mathcal{X}}\in\mathcal{M}}{\operatorname{argmin}} \, f(\boldsymbol{\mathcal{X}}) \tag{28}$$

Detailed IMU and visual residuals are provided in the following subsections. The least squares problem are solved by Gauss-Newton method implemented in g2o [27] or ceres solver.

B. Inertial Measurement Model

IMU measurements arrive at a much higher frequency than the visual measurements. So in order to avoid the frequent integration whenever the linearization point changes, we adopt the IMU preintegration approach proposed in [15]. The IMU preintegraton is independent of the initial conditions, and can incorporate the change of IMU biases. The concept was firstly proposed in [21].

We integrate all the IMU measurements $\{\widetilde{\mathbf{a}}_k,\ \widetilde{\boldsymbol{\omega}}_k\}$ between keyframes i and j to compute the IMU preintegration $\Delta\widetilde{\mathbf{I}}_{ij} = [\Delta\widetilde{\mathbf{R}}_{ij}, \Delta\widetilde{\mathbf{p}}_{ij}, \Delta\widetilde{\mathbf{v}}_{ij}]$ on manifold as:

$$
\begin{aligned}
\Delta\widetilde{\mathbf{R}}_{ij} &= \prod_{k=i}^{j-1} \mathrm{Exp}\left((\widetilde{\boldsymbol{\omega}}_k - \mathbf{b}_{g_i})\Delta t\right) \\
\Delta\widetilde{\mathbf{p}}_{ij} &= \sum_{k=i}^{j-1} \left(\Delta\widetilde{\mathbf{v}}_{ik}\Delta t + \frac{1}{2}\Delta\widetilde{\mathbf{R}}_{ik}(\widetilde{\mathbf{a}}_k - \mathbf{b}_{a_i})\Delta t^2\right) \\
\Delta\widetilde{\mathbf{v}}_{ij} &= \sum_{k=i}^{j-1} \Delta\widetilde{\mathbf{R}}_{ik}(\widetilde{\mathbf{a}}_k - \mathbf{b}_{a_i})\Delta t
\end{aligned} \tag{29}
$$

Furthermore, given a bias update $\delta\mathbf{b}$ and using the first-order expansion, the preintegrated IMU measurement can be updated as:

$$
\begin{aligned}
\Delta\widetilde{\mathbf{R}}_{ij}(\mathbf{b}_{g_i}) &= \Delta\widetilde{\mathbf{R}}_{ij}(\bar{\mathbf{b}}_{g_i})\mathrm{Exp}\left(\frac{\partial\Delta\bar{\mathbf{R}}_{ij}}{\partial\mathbf{b}_g}\delta\mathbf{b}_g\right) \\
\Delta\widetilde{\mathbf{p}}_{ij}(\mathbf{b}_{g_i}, \mathbf{b}_{a_i}) &= \Delta\widetilde{\mathbf{p}}_{ij}(\bar{\mathbf{b}}_{g_i}, \bar{\mathbf{b}}_{a_i}) + \frac{\partial\Delta\bar{\mathbf{p}}_{ij}}{\partial\mathbf{b}_g}\delta\mathbf{b}_g + \frac{\partial\Delta\bar{\mathbf{p}}_{ij}}{\partial\mathbf{b}_a}\delta\mathbf{b}_a \\
\Delta\widetilde{\mathbf{v}}_{ij}(\mathbf{b}_{g_i}, \mathbf{b}_{a_i}) &= \Delta\widetilde{\mathbf{v}}_{ij}(\bar{\mathbf{b}}_{g_i}, \bar{\mathbf{b}}_{a_i}) + \frac{\partial\Delta\bar{\mathbf{v}}_{ij}}{\partial\mathbf{b}_g}\delta\mathbf{b}_g + \frac{\partial\Delta\bar{\mathbf{v}}_{ij}}{\partial\mathbf{b}_a}\delta\mathbf{b}_a
\end{aligned} \tag{30}
$$

where Jacobians $\frac{\partial\Delta(\cdot)}{\partial\mathbf{b}_{\cdot}}$ describe how a change in the bias estimate effects the preintegrated IMU measurements. The derivation of the Jacobians can be found in [15]. From geometric constraints, we get the IMU measurement as:

$$
\begin{aligned}
\Delta\widetilde{\mathbf{R}}_{ij}(\mathbf{b}_{g_i}) &= \mathbf{R}_{B_i}^{W\,\mathrm{T}}\mathbf{R}_{B_j}^{W}\mathrm{Exp}(\delta\boldsymbol{\xi}_{ij}) \\
\Delta\widetilde{\mathbf{p}}_{ij}(\mathbf{b}_{g_i}, \mathbf{b}_{a_i}) &= \mathbf{R}_{B_i}^{W\,\mathrm{T}}(\mathbf{p}_{B_j}^{W} - \mathbf{p}_{B_i}^{W} - \mathbf{v}_{B_i}^{W}\Delta t_{ij} - \frac{1}{2}\mathbf{g}^{W}\Delta t_{ij}^2) \\
&\quad + \delta\mathbf{p}_{ij} \\
\Delta\widetilde{\mathbf{v}}_{ij}(\mathbf{b}_{g_i}, \mathbf{b}_{a_i}) &= \mathbf{R}_{B_i}^{W\,\mathrm{T}}(\mathbf{v}_{B_j}^{W} - \mathbf{v}_{B_i}^{W} - \mathbf{g}^{W}\Delta t_{ij}) + \delta\mathbf{v}_{ij}
\end{aligned} \tag{31}
$$

where $[\delta\boldsymbol{\xi}_{ij}^{\mathrm{T}}\ \delta\mathbf{t}_{ij}^{\mathrm{T}}\ \delta\mathbf{v}_{ij}^{\mathrm{T}}]^{\mathrm{T}} \in \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{I}_{ij}})$ is zero-mean white Gaussian noise. Given the above measurement model, the IMU preintegration residual $\mathbf{r}_{\Delta_{ij}} = [\mathbf{r}_{\Delta\mathbf{R}_{ij}}^{\mathrm{T}}\ \mathbf{r}_{\Delta\mathbf{p}_{ij}}^{\mathrm{T}}\ \mathbf{r}_{\Delta\mathbf{v}_{ij}}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^9$ is:

$$
\begin{aligned}
\mathbf{r}_{\Delta\mathbf{R}_{ij}} &= \mathrm{Log}\left(\left(\Delta\widetilde{\mathbf{R}}_{ij}(\bar{\mathbf{b}}_{g_i})\mathrm{Exp}\left(\frac{\partial\Delta\bar{\mathbf{R}}_{ij}}{\partial\mathbf{b}_g}\delta\mathbf{b}_g\right)\right)^{\mathrm{T}}\mathbf{R}_{B_i}^{W\,\mathrm{T}}\mathbf{R}_{B_j}^{W}\right) \\
\mathbf{r}_{\Delta\mathbf{p}_{ij}} &= \mathbf{R}_{B_i}^{W\,\mathrm{T}}(\mathbf{p}_{B_j}^{W} - \mathbf{p}_{B_i}^{W} - \mathbf{v}_{B_i}^{W}\Delta t_{ij} - \frac{1}{2}\mathbf{g}^{W}\Delta t_{ij}^2) \\
&\quad - \left[\Delta\widetilde{\mathbf{p}}_{ij}(\bar{\mathbf{b}}_{g_i}, \bar{\mathbf{b}}_{a_i}) + \frac{\partial\Delta\bar{\mathbf{p}}_{ij}}{\partial\mathbf{b}_g}\delta\mathbf{b}_g + \frac{\partial\Delta\bar{\mathbf{p}}_{ij}}{\partial\mathbf{b}_a}\delta\mathbf{b}_a\right] \\
\mathbf{r}_{\Delta\mathbf{v}_{ij}} &= \mathbf{R}_{B_i}^{W\,\mathrm{T}}(\mathbf{v}_{B_j}^{W} - \mathbf{v}_{B_i}^{W} - \mathbf{g}^{W}\Delta t_{ij}) \\
&\quad - \left[\Delta\widetilde{\mathbf{v}}_{ij}(\bar{\mathbf{b}}_{g_i}, \bar{\mathbf{b}}_{a_i}) + \frac{\partial\Delta\bar{\mathbf{v}}_{ij}}{\partial\mathbf{b}_g}\delta\mathbf{b}_g + \frac{\partial\Delta\bar{\mathbf{v}}_{ij}}{\partial\mathbf{b}_a}\delta\mathbf{b}_a\right]
\end{aligned} \tag{32}
$$

The corresponding covariance matrix $\boldsymbol{\Sigma}_{\Delta_{ij}}$ can be calculated by incrementally propagating the preintegration noise from keyframe i to j. The detailed derivatives about Jaocbians and the uncertainty propagation on manifold can refer to paper [15].

C. Bias Model

Biases are slowly time-varying, so for the biases between consecutive keyframes i and j, we have:

$$
\mathbf{b}_{g_j} = \mathbf{b}_{g_i} + \boldsymbol{\eta}_{b_{gd}}, \quad \mathbf{b}_{a_j} = \mathbf{b}_{a_i} + \boldsymbol{\eta}_{b_{ad}} \tag{33}
$$

where $\boldsymbol{\eta}_{b_{gd}}$ and $\boldsymbol{\eta}_{b_{ad}}$ are the discretized bias random walk with covariance $\boldsymbol{\Sigma}_{b_{gd}}$ and $\boldsymbol{\Sigma}_{b_{ad}}$. Therefore, we express the bias error $\mathbf{r}_b = [\mathbf{r}_g^{\mathrm{T}}\ \mathbf{r}_a^{\mathrm{T}}] \in \mathbb{R}^6$ as:

$$
\begin{aligned}
\mathbf{r}_g &= \mathbf{b}_{g_j} - \mathbf{b}_{g_i} \\
\mathbf{r}_a &= \mathbf{b}_{a_j} - \mathbf{b}_{a_i}
\end{aligned} \tag{34}
$$

D. Visual Measurement Model

Through the measurement model in (5), the reprojection residual $\mathbf{r}_{\mathcal{C}_{kl}} \in \mathbb{R}^2$ for the l^{th} map point seen by the k^{th} keyframe is:

$$
\mathbf{r}_{\mathcal{C}_{kl}} = \pi\left(\mathbf{R}_C^{B\,\mathrm{T}}\left(\mathbf{R}_{B_i}^{W\,\mathrm{T}}(\mathbf{L}_l^W - \mathbf{p}_{B_i}^W) - \mathbf{p}_C^B\right)\right) - \widetilde{\mathbf{z}}_{kl} \tag{35}
$$

The corresponding covariance matrix $\boldsymbol{\Sigma}_{\mathcal{C}_{kl}}$ is equal to $\boldsymbol{\Sigma}_{\sigma_{kl}}$.

E. Error Term Representation

In this section, we give detailed representation for the IMU error term $\|\mathbf{r}_{\mathcal{I}_{ij}}\|^2_{\boldsymbol{\Sigma}_{\mathcal{I}_{ij}}}$ and the reprojection error term $\|\mathbf{r}_{\mathcal{C}_{kl}}\|^2_{\boldsymbol{\Sigma}_{\mathcal{C}_{kl}}}$ in (27). Given the inertial measurement model in Section V-B and bias model in Section V-C, the IMU error term is:

$$
\|\mathbf{r}_{\mathcal{I}_{ij}}\|^2_{\boldsymbol{\Sigma}_{\mathcal{I}_{ij}}} = \mathbf{r}_{\Delta_{ij}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\Delta_{ij}}^{-1}\mathbf{r}_{\Delta_{ij}} + \mathbf{r}_b^{\mathrm{T}}\boldsymbol{\Sigma}_{b_d}^{-1}\mathbf{r}_b \tag{36}
$$

where $\boldsymbol{\Sigma}_{b_d} = \begin{bmatrix} \boldsymbol{\Sigma}_{b_{gd}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{b_{ad}} \end{bmatrix}$.

In addition, given the visual measurement model in Section V-D, the reprojection error term is:

$$
\|\mathbf{r}_{\mathcal{C}_{kl}}\|^2_{\boldsymbol{\Sigma}_{\mathcal{C}_{kl}}} = \mathbf{r}_{\mathcal{C}_{kl}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathcal{C}_{kl}}^{-1}\mathbf{r}_{\mathcal{C}_{kl}} \tag{37}
$$

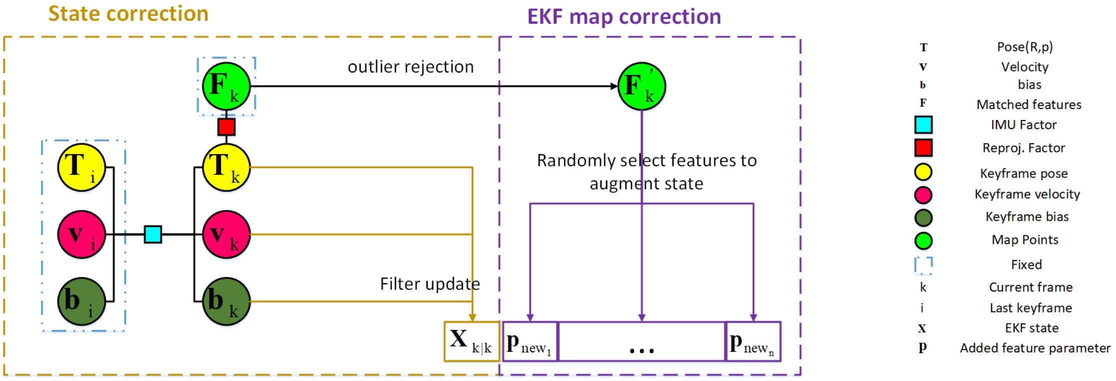

Fig. 3. A diagram shows how the feedback mechanism works for the current frame k when the last keyframe i just updated. The state of the current frame k is firstly corrected by the state correction, and then new features are added to the state vector of the EKF by EKF map correction.

VI. TIGHTLY-COUPLED MONOCULAR VISLAM

In this section, we introduce our tightly-coupled visual-inertial monocular SLAM approach, which combines the EKF-based VIO front-end with the BA and loop closure back-end to provide the accurate and robust state estimation. An overview of our system is shown in Fig.1. When new inertial sensor is used, we firstly perform the EKF VIO alone with initial bias value of zero to obtain a good initial bias estimates before starting the all system, then set it as the initial bias value of our system.

A. Map Initialization

The map initialization is in charge of constructing an initial set of map points for the subsequent nonlinear optimization and loop closure. The initial map is created according to the estimated state from the EKF VIO. Since the accuracy of the initially created map will pose big influence on the accuracy of the whole system, we create initial map after running the EKF system about 10 seconds to converge.

Firstly, we extract ORB features in the current frame k, and search for feature matches with the reference frame r. If sufficient feature correspondences are found, we perform the next step, else set the current frame as reference frame. The second step is to check the parallax of each correspondence, and pick out a set of feature matches $\mathcal{F}$ which have sufficient parallax. When the size of $\mathcal{F}$ is greater than a threshold, using the estimated pose from the EKF based VIO, we triangulate the matched features $\mathcal{F}$. Then, if enough map points are successfully created, a full BA combining reprojection error term and temporal IMU error term is applied to refine the initial map. Finally, using the optimized state and map to correct the EKF state according to the feedback mechanism described in Section VI-D.

B. FRONT-END

*1) **State Estimation**:* We firstly perform the VIO described in Section IV to estimate the state of current frame $\mathbf{X}_{B_k} = \{\boldsymbol{\xi}^W_{B_k}\ \mathbf{p}^W_{B_k}\ \mathbf{v}^W_{B_k}\ \mathbf{b}_{a_k}\ \mathbf{b}_{g_k}\}$. In order to ensure the real-time performance, we only remain those features visible in current frame to limit the number of keypoints in the state vector. In our work, we maintain 50 features in the EKF state. Finally, if the map is updated in the back-end, the state of the current frame will be corrected according to the feedback mechanism described in Section VI-D. In this way, the front-end can always provide reliable state estimates even if we run for long periods of time.

*2) **Keyframe Selection**:* After the state of a frame is estimated by the VIO, we adopt three criteria to determine whether this frame is a keyframe. (1) The time interval from last keyframe is beyond a certain threshold. This criteria ensures the accuracy of the system. Because the IMU just provide valuable information in short-time, if the time interval from the last keyframe is too long, the IMU constraint between two keyframes will become inaccurate. (2) The nonlinear optimization in back-end is finished. This criteria makes as many keyframes as possible to enhance the motion tracking accuracy. (3) The rotation angle from last keyframe is beyond a certain threshold. This criteria ensures the reconstruction of the globally consistent map.

Once a frame is selected as a keyframe, we extract the ORB features for the new keyframe and trigger the back-end to make the pose estimation more accurate.

C. BACK-END

Once a new keyframe is inserted, the nonlinear optimization described in Section V is performed to optimization the local map in a parallel thread. After the local BA is finished, some redundant keyframes will be culled to make the factor graph more concise. In loop closure thread, place recognition is performed. Once a loop is detected, a $\mathbf{Sim}(3)$ optimization and a full BA is performed to eliminate the accumulated drift. We refer the interested readers to papers [17] [28] for more details.

D. FEEDBACK MECHANISM

The EKF-based VIO can estimate the frame state efficiently. However since the accumulation of the linearization errors and the absence of the loop closure, the error of the state estimate will accumulate as time goes on. If the state provided in the front-end drift too much, the local BA will be hard to find best estimates. Therefore in order to constrain the error of the state estimated from VIO, we provide following feedback mechanism. The feedback mechanism is invoked whenever the map in the back-end is updated and contains two steps:

*1) **State Correction**:* After the BA and loop closure are performed in the back-end, the state estimation of the last keyframe i is accurate enough. The key observation of the state correction is to improve the state estimation of the current frame k by leveraging the optimized state of the last keyframe i. Therefore, given the estimated state $\mathbf{X}_{B_k}$ from VIO, we optimize the state of the current frame by performing the nonlinear optimization as shown in Fig. 3, that is minimizing the following objective function:

$$\mathbf{X}^*_{B_k} = \underset{\mathbf{X}^*_{B_k}}{\arg\min}\, \rho\left(\|\mathbf{r}_{\mathcal{I}_{ik}}\|^2_{\mathbf{\Sigma}_{\mathcal{I}_{ik}}}\right) + \sum_{l \in F_k} \rho\left(\|\mathbf{r}_{\mathcal{C}_{kl}}\|^2_{\mathbf{\Sigma}_{\mathcal{C}_{kl}}}\right) \tag{38}$$

where F_k denotes the features matched with the map points in current frame k. The form of $\|\mathbf{r}_{\mathcal{I}_{ik}}\|^2_{\mathbf{\Sigma}_{\mathcal{I}_{ik}}}$ and $\|\mathbf{r}_{\mathcal{C}_{kl}}\|^2_{\mathbf{\Sigma}_{\mathcal{C}_{kl}}}$ is the same as (36) and (37) respectively. Then the optimized state $\mathbf{X}^*_{B_k}$ and its covariance matrix $\mathbf{\Upsilon}_{B_k}$ obtained from the optimization are used to update the EKF state of the current frame as:

$$\begin{aligned}
\mathbf{H}^*_k &= \begin{bmatrix} \mathbf{I}_{15\times15} & \mathbf{0}_{15\times6m} \end{bmatrix} \\
\mathbf{r}^*_k &= \mathbf{X}^*_{B_k} - \mathbf{X}_{B_{k|k}} \\
\mathbf{K}^*_k &= \mathbf{P}_{k|k}\mathbf{H}^{*T}_k(\mathbf{H}^*_k\mathbf{P}_{k|k}\mathbf{H}^{*T}_k + \mathbf{\Upsilon}_{B_k})^{-1} \\
\mathbf{X}^*_{k|k} &= \mathbf{X}_{k|k} + \mathbf{K}^*_k\mathbf{r}^*_k \\
\mathbf{P}^*_{k|k} &= (\mathbf{I}_{15+6m} - \mathbf{K}^*_k\mathbf{H}^*_k)\mathbf{P}_{k|k}
\end{aligned} \tag{39}$$

*2) **EKF Map Correction**:* It is well known that the estimated pose relative to the map in the state vector is not drift. Therefore, as long as the position of map points in the EKF state vectorF is consistent with the optimized global map, the accuracy of the estimated state will accordingly increase. Therefore after the IMU state in EKF state vector is updated by state correction, we will add features in the optimized consistent map to the state vector. We denote $F^{'}_k$ as a set of features matched with the optimized map in current frame k, the outliers are removed based on the optimized state. If we need to add n new features to the EKF state, we randomly select n features in $F^{'}_k$, then compute their initial position and add it to the filter state vector. For selected new feature l, the initial position $\mathbf{p}^{W*}_l = [x^*_l\ y^*_l\ z^*_l\ \theta^*_l\ \phi^*_l\ \rho^*_l]^T$ is set as follows. $[x^*_l\ y^*_l\ z^*_l]^T$ is computed as (22) and $[\theta^*_l\ \phi^*_l]^T$ is computed as (23)(24) using the updated state $\mathbf{X}^*_{k|k}$ from (39) and the observation of the selected new features in the current frame. In addition, for computing ρ^*_l, we firstly transform the map point in world frame $\mathbf{L}^W_l$ to the current camera frame $\mathbf{L}^{C_k}_l$:

$$\mathbf{L}^{C_k}_l = \begin{bmatrix} x_L \\ y_L \\ z_L \end{bmatrix} = (\mathbf{R}^{W\ *}_{B_k}\mathbf{R}^B_C)^T \left(\mathbf{L}^W_l - (\mathbf{R}^{W\ *}_{B_k}\mathbf{R}^B_C + \mathbf{p}^{W\ *}_{B_k})\right) \tag{40}$$

Then the initial inverse depth is obtained by $\rho^*_l = \frac{1}{\|\mathbf{L}^{C_k}_l\|}$, and its variance is set as follows:

$$\sigma^*_{\rho_l} = \mathbf{J}_L * \mathbf{\Sigma}_L * \mathbf{J}^T_L + \mathbf{J}_{Rt} * \mathbf{\Sigma}_{Rt} * \mathbf{J}^T_{Rt} \tag{41}$$

where $\mathbf{J}_L = -\frac{1}{\|\mathbf{L}^{C_k}_l\|^3}\mathbf{L}^{C_k\,T}_l(\mathbf{R}^{W\ *}_{B_k}\mathbf{R}^B_C)^T$ and $\mathbf{J}_{Rt} = -\frac{1}{\|\mathbf{L}^{C_k}_l\|^3}\mathbf{L}^{C_k\,T}_l\left[\mathbf{R}^{B\,T}_C\left(\mathbf{R}^{W\ *T}_{B_k}(\mathbf{L}^W_l - \mathbf{p}^{W\ *}_{B_k})\right)^\wedge\ -\mathbf{R}^{B\,T}_C\right]$ are the Jacobians of ρ^*_l with respect to $\mathbf{L}^W_l$ and $[\mathbf{R}^{W\ *}_{B_k}\ \mathbf{p}^{W\ *}_{B_k}]$ on manifold. Besides, $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_{Rt}$ are the covariance matrix for $\mathbf{L}^W_l$ and $[\mathbf{R}^{W\ *}_{B_k}\ \mathbf{p}^{W\ *}_{B_k}]$, which is computed from the BA in the back-end.

VII. EXPERIMENTS

We make a complete evaluation of the proposed algorithm qualitatively and quantitatively on the EuRoC dataset [29]. The dataset contains 11 data sequences, which was recorded from a flying MAV in two different $30m^2$ indoor rooms and a $300m^2$ industrial environment. Depending on the illumination, texture and motion dynamics, the data sequences are classified as easy, medium and difficult levels. The dataset provides synchronized global shutter WVGA stereo images at 20Hz, IMU measurements at 200Hz and ground truth state at 200Hz. We only use images from the left camera. Firstly, we evaluate the proposed algorithm qualitatively and quantitatively on the EuRoC dataset to show the accuracy of our system. Then we compare our method with other state-of-the-art approaches on the EuRoC dataset. Finally, the performance of our algorithm is validated again by indoor real-world experiments using the sensor of Intel RealSense ZR300. The experiments are performed on a laptop with Intel Core i5 2.2GHz CPU and an 8GB RAM.

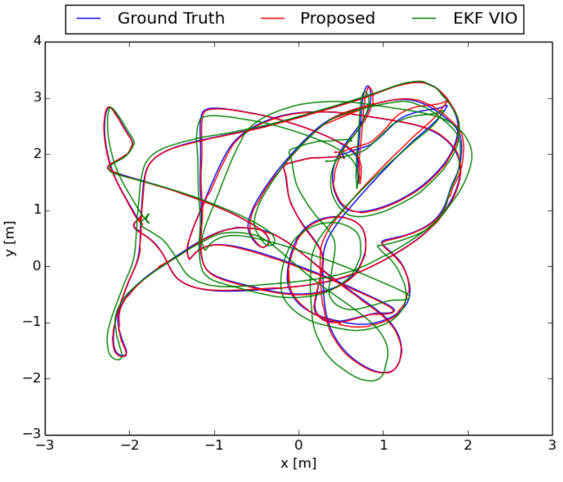

Fig. 4. Comparisons of the ground truth, the trajectories estimated by our algorithm and EKF VIO on V1_02_medium sequence, which is viewed from the gravity direction.

A. Algorithm Evaluation

We successfully perform our algorithm on all 11 sequences of EuRoC dataset in real-time. Fig. 4 and Fig. 6 show the comparisons of the ground truth, the trajectory estimated by our algorithm and EKF VIO on V1_02_medium and MH_02_easy sequence respectively. The corresponding x,y,z translation error versus time is shown in Fig. 5 and Fig. 7.

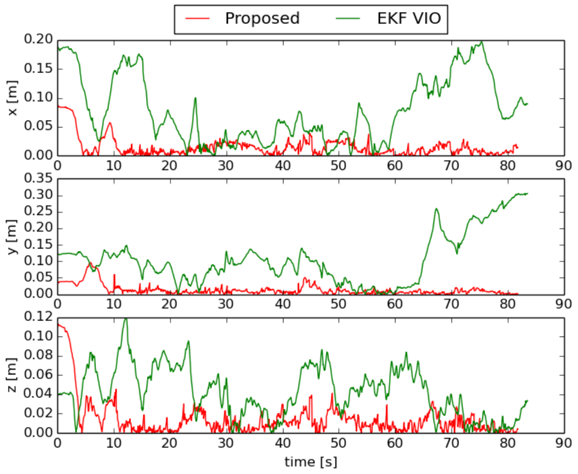

Fig. 5. Translation error of our algorithm and EKF VIO on V1_02_medium sequence.

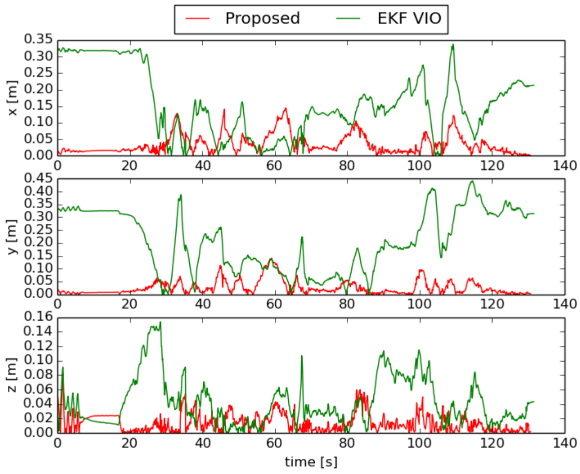

Fig. 7. Translation error of our algorithm and EKF VIO on MH_03_medium sequence.

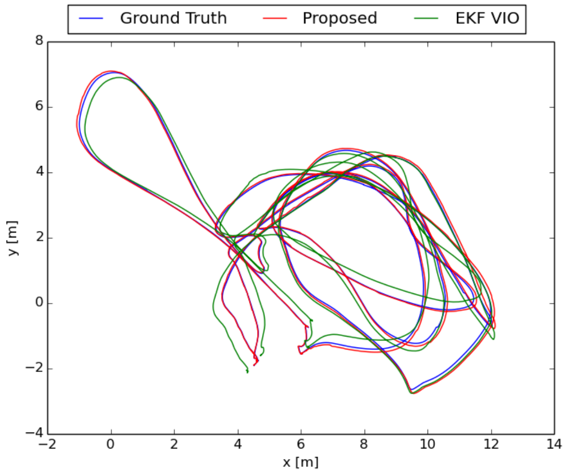

Fig. 6. Comparisons of the ground truth, the trajectories estimated by our algorithm and EKF VIO on MH_03_medium sequence, which is viewed from the gravity direction.

TABLE I
TRANSLATION RMSE OF THE TRAJECTORIES ESTIMATED FROM THE PROPOSED METHOD AND EKF VIO ON THE EUROC MAV DATASET

Sequence	Our Method With Loop	Our Method Without Loop	EKF
V1_01_easy	0.080	0.080	0.087
V1_02_medium	0.043	0.099	0.170
V1_03_difficult	0.124	0.245	0.301
V2_01_easy	0.052	0.052	0.082
V2_02_medium	0.042	0.042	0.191
V2_03_difficult	0.074	0.275	0.368
MH_01_easy	0.021	0.021	0.175
MH_02_easy	0.071	0.071	0.277
MH_03_medium	0.061	0.061	0.307
MH_04_difficult	0.064	0.064	0.309
MH_05_difficult	0.048	0.056	0.529

The estimated trajectories are aligned with the ground truth using the method of Horn [30]. As evident, the translation error of our approach is smaller than the error of EKF VIO, thus proving the superiority of our algorithm towards the EKF VIO method.

For quantitative analysis, table I shows the translation Root Mean Square Error(RMSE) of the estimated trajectory for each sequence, as proposed in [31]. The proposed method has achieved the average translation RMSE of 0.082m, 0.056m and 0.053m for V1, V2 and MH sequences with respect to 0.186m, 0.213m, 0.319m of EKF VIO system, which illustrates that our method reduced the error of 55%, 73%, 83% for V1, V2 and MH sequences. From the third and fourth columns of the table I, we can know that adding BA and feedback mechanism to EKF VIO system, the RMSE of sequences are much reduced, this is since (1) BA is able to relinearize measurement models to properly deal with the nonlinearity of the system, (2) increasing the number of feature matches in the window of local BA can greatly improve the accuracy, and (3) feedback of the optimized state and map to VIO can improve the consistency and accuracy of the EKF system. However, the advantage of BA and feedback mechnism is not well demonstrated in V1_02_medium, V1_03_difficult and V2_03_difficult sequences, because in which fast rotation and low texture are frequently happened. Both fast rotation and low texture result in fewer feature correspondences, and thereby make less constraints in local BA, thus the accuracy of the system is not improved greatly. Therefore, in these sequences, adding a loop closure achieved much better accuracy by eliminating the accumulated error when revisiting an already mapped area.

While adding BA, loop closure and feedback mechanism improves accuracy of the system, it increases the computational cost of the system due to the need of (1) extracting a lot of features for BA and loop closure, and (2) pose optimization and EKF update of the feedback mechanism. The computational increase only occurs in selected keyframes, so the

TABLE II
TRANSLATION RMSE OF THE TRAJECTORIES ESTIMATED FROM DIFFERENT APPROACHES ON THE EUROC MAV DATASET. THE BEST RESULTS ARE GIVEN IN BOLD.

sequence	Our Method With Loop	Our Method Without Loop	VINS-MONO With Loop	VINS-MONO Without Loop	ORB-VIN	OKVIS	ROVIO
V1_01_easy	0.080	0.080	0.081	0.088	**0.027**	0.089	1.412
V1_02_medium	0.043	0.099	0.042	0.068	**0.028**	0.141	0.160
V1_03_difficult	**0.124**	0.245	0.156	0.160	X	0.262	0.170
V2_01_easy	0.052	0.052	0.063	0.068	**0.032**	0.135	0.236
V2_02_medium	0.042	0.042	0.066	0.084	**0.041**	0.155	0.408
V2_03_difficult	**0.074**	0.275	0.157	0.159	**0.074**	0.279	0.213
MH_01_easy	**0.021**	**0.021**	0.098377	0.301814	0.075	0.309	0.354
MH_02_easy	**0.071**	**0.071**	0.152	0.249	0.084	0.293	0.594
MH_03_medium	**0.061**	**0.061**	0.080	0.173	0.087	0.310	0.310
MH_04_difficult	**0.064**	**0.064**	0.129	0.323	0.217	0.360	1.058
MH_05_difficult	**0.048**	0.063	0.077	0.257	0.082	0.404	1.241

proposed method will not incur too much computational cost. Our algorithm requires approximately 27 msec for processing each image, so it can still run in real time, at about 40Hz.

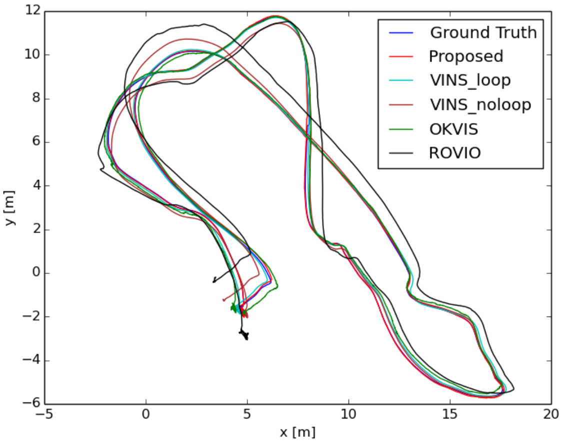

Fig. 8. Comparisons of the ground truth, the trajectories estimated by our algorithm and state-of-the-art methods on MH_04_difficult sequence, which is viewed from the gravity direction.

B. Comparison to State-of-the-art Algorithms

We compare the proposed method with the state-of-the-art VINS-MONO [19] [20], OKVIS [14], ROVIO [32] and ORB-VISLAM [17] method. VINS-MONO, OKVIS and ROVIO are open-source and contain the default parameters for the EuRoC dataset, for fair comparison, we only use left image. ORB-VISLAM shows its results in EuRoC dataset, which allowing for a direct comparison.

A comparison of the translation RMSE of the estimated trajectories on EuRoC dataset are shown in table II, X means the concerned method fails to run in the sequence. From these results, we can draw following conclusions. ORB-VISLAM and our algorithm have obtained the best accuracy, this is because local BA in both methods is performed in a parallel thread, so more feature correspondences are used in local BA. Besides, both methods can close loop to eliminate the accumulated error. Howerver, ORB-VISLAM fails to track the V1_03_difficult sequence. In comparison, VINS-MONO and OKVIS perform BA in the thread of tracking, so the number of features contained in local BA must be limited to ensure the real-time performance, therefore leading to slightly worse accuracy. VINS-MONO with loop can achieve better accuracy than OKVIS, due to its capability to close loop. In addition, ROVIO is an EKF method and not able to close loop, its the linearization error and the error accumulation make the method obtain diminished localization error. However, since it is a direct method, it can achieve minor drift in fast motion sequences of V1_03_difficult and V2_03_difficult.

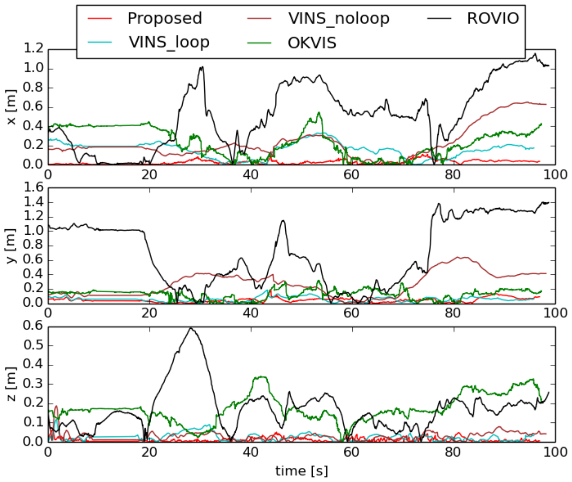

Fig. 9. Translation error of the estimated trajectories for MH_04_difficult sequence.

For sequence MH_04_difficult, the estimated trajectories are shown in Fig. 8, and translation errors versus time are shown in Fig. 9. In this sequence, our loop closure is not triggered, however in the error plot, our method still achieved smallest translation error, which can prove the superiority of our algorithm again.

C. Indoor Real-world Experiment

We perform the indoor experiment in an $60m^2$ office environment using the monocular-inertial Realsense ZR 300 sensor suite that provides images at the frequency of 20 Hz and

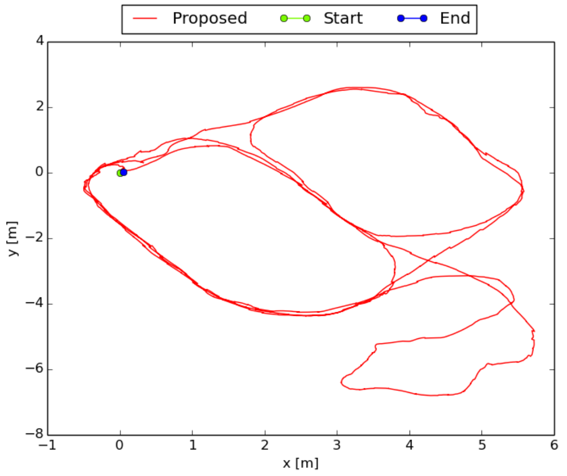

Fig. 10. The estimated trajectory of the indoor real-world experiment.

IMU measurements at 200 Hz. As shown in the accompanying video[1], we hold the sensor suite by hand and walk in normal pace in the office, and we starts and ends at the same location. Fig. 10 shows the estimated trajectory, from which we can know there is no noticeable drifts occurred when we circle indoor. The end-to-end error is 0.055m with respect to the total length of 82m, it is just the 0.067% of the total trajectory length.

VIII. Conclusion and Future Work

In this paper, we have presented a tightly-coupled monocular VISLAM system, which robustly tracks camera motion by EKF VIO, and perform non-linear optimization and loop closure to solve the linearization issues of the EKF system and eliminate the accumulated error. We also proposed a feedback mechanism to directly improve the consistency and accuracy of the EKF system. Therefore, our algorithm has achieved high accuracy, the performance of the proposed method is validated through experiments.

Point feature-based monocular VISLAM is prone to fail in poorly textured scenes or motion blurred images. Therefore in the future, we inted to deal with these specific situations for better accuracy and robustness. We also aim to build dense map to assist the understanding of the environment.

Acknowledgment

This paper is supported by National Science Foundation of China[grant number 61375081]; a special fund project of Harbin science and technology innovation talents research [grant number RC2013XK010002].

Appendix

A. *Matrices $\mathbf{\Phi}_k$ and $\mathbf{G}_k$*

The Jacobian matrix $\mathbf{\Phi}_k$ in (13) is:

$$\mathbf{\Phi}_k = \begin{bmatrix} \mathbf{\Phi}_{\xi\xi} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{\Phi}_{\xi b_g} \\ \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} & \mathbf{I}_{3\times3}\Delta t & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{\Phi}_{v\xi} & \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} & \mathbf{\Phi}_{vb_a} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} \end{bmatrix} \tag{42}$$

where $\mathbf{I}_{3\times3}$ is the 3×3 identity matrix, and

$$\begin{aligned} \mathbf{\Phi}_{\xi\xi} &= \mathbf{J}_r^{-1}(\boldsymbol{\xi}^W_{B_{k|k-1}})\mathrm{Exp}\left((\widetilde{\boldsymbol{\omega}}_{k-1} - \mathbf{b}_{g_{k-1}} - \mathbf{n}_{gd})\Delta t\right)^T \\ \mathbf{\Phi}_{\xi b_g} &= -\mathbf{J}_r^{-1}(\boldsymbol{\xi}^W_{B_{k|k-1}})\mathbf{J}_r\left((\widetilde{\boldsymbol{\omega}}_{k-1} - \mathbf{b}_{g_{k-1}} - \mathbf{n}_{gd})\Delta t\right)\Delta t \\ \mathbf{\Phi}_{v\xi} &= -\mathbf{R}^W_{B_{k-1}}\left((\widetilde{\mathbf{a}}_{k-1} - \mathbf{b}_{a_{k-1}} - \mathbf{n}_{ad})\Delta t\right)^{\wedge} \\ \mathbf{\Phi}_{vb_a} &= -\mathbf{R}^W_{B_{k-1}}\Delta t \end{aligned}$$

In addition, the Jacobian matrix $\mathbf{G}_k$ in (13) is:

$$\mathbf{G}_k = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{\Phi}_{\xi n_g} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{\Phi}_{vn_a} & \mathbf{0}_{3\times3} \\ \mathbf{I}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} \end{bmatrix} \tag{43}$$

with $\mathbf{\Phi}_{\xi n_g} = \mathbf{\Phi}_{\xi b_g}$ and $\mathbf{\Phi}_{vn_a} = \mathbf{\Phi}_{vb_a}$.

B. *Measurement matrices*

The Jacobian of the measurement model with respect to the IMU state in (18) is represented as:

$$\mathbf{H}_{B_{kl}} = [-\mathbf{H}_{h\xi} \;\; -\mathbf{H}_{hp} \;\; \mathbf{0}_{2\times9}] \tag{44}$$

with:

$$\mathbf{H}_{h\xi} = \frac{\partial \mathbf{h}_{kl}}{\partial \mathbf{f}_l^{C_k}} \mathbf{R}_C^{B\,T} \cdot$$

$$\left({\mathbf{R}^W_{B_{k|k-1}}}^T \left(\rho_l \left(\begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix} - \mathbf{p}^W_{B_{k|k-1}} \right) + \mathbf{m}(\theta_l, \phi_l) \right) \right)^{\wedge}$$

$$\mathbf{H}_{hp} = -\rho_l \frac{\partial \mathbf{h}_{kl}}{\partial \mathbf{f}_l^{C_k}} \mathbf{R}_W^{C_{k|k-1}}$$

$$\frac{\partial \mathbf{h}_{kl}}{\partial \mathbf{f}_l^{C_k}} = \begin{bmatrix} \frac{f_x}{z_l^{C_k}} & 0 & -\frac{f_x x_l^{C_k}}{{z_l^{C_k}}^2} \\ 0 & \frac{f_y}{z_l^{C_k}} & -\frac{f_y y_l^{C_k}}{{z_l^{C_k}}^2} \end{bmatrix}$$

where $\mathbf{f}_l^{C_k} = [x_l^{C_k} \;\; y_l^{C_k} \;\; z_l^{C_k}]^T$. In addition, the Jacobian of the measurement model with respect to the l^{th} feature position in (18) is:

$$\mathbf{H}_{f_{kl}} = [-\mathbf{H}_{hxyz} \;\; -\mathbf{H}_{h\theta\phi} \;\; -\mathbf{H}_{h\rho}] \tag{45}$$

with:

$$\mathbf{H}_{hxyz} = \rho_l \frac{\partial \mathbf{h}_{kl}}{\partial \mathbf{f}_l^{C_k}} \mathbf{R}_W^{C_{k|k-1}}$$

$$\mathbf{H}_{h\theta\phi} = \frac{\partial \mathbf{h}_{kl}}{\partial \mathbf{f}_l^{C_k}} \mathbf{R}_W^{C_{k|k-1}} \begin{bmatrix} cos\phi_l cos\theta_l & -sin\phi_l sin\theta_l \\ 0 & -cos\phi_l \\ -cos\phi_l sin\theta_l & -sin\phi_l cos\theta_l \end{bmatrix}$$

$$\mathbf{H}_{h\rho} = \frac{\partial \mathbf{h}_{kl}}{\partial \mathbf{f}_l^{C_k}} \mathbf{R}_W^{C_{k|k-1}} \left(\begin{bmatrix} x_l \\ y_l \\ z_l \end{bmatrix} - \mathbf{p}^W_{C_{k|k-1}} \right)$$

[1] https://youtu.be/5_G8jUOjtN0

C. State Augmentation Jacobian

The Jacobian matrix used to augment the covariance matrix of the state vector is:

$$\mathbf{J} = \begin{bmatrix} \mathbf{I}_{15+6m} & \mathbf{0}_{(15+6m)\times 6} \\ \mathbf{J}_X & \mathbf{J}_{h\rho} \end{bmatrix} \tag{46}$$

with:

$$\mathbf{J}_X = \begin{bmatrix} -\mathbf{R}^W_{B_{k|k}} \mathbf{p}^{B\wedge}_C & \mathbf{I}_3 & \mathbf{0}_{3\times 9} & \mathbf{0}_{3\times 6m} \\ \frac{\partial \breve{\theta}_l \breve{\phi}_l}{\partial \boldsymbol{\xi}_{B_{k|k}}} & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 9} & \mathbf{0}_{3\times 6m} \\ \mathbf{0}_{1\times 3} & & & \end{bmatrix} \tag{47}$$

where:

$$\frac{\partial \breve{\theta}_l \breve{\phi}_l}{\partial \boldsymbol{\xi}_{B_{k|k}}} = -\frac{\partial \breve{\theta}_l \breve{\phi}_l}{\partial \boldsymbol{\tau}^W_{lk}} \mathbf{R}^W_{B_{k|k}} \left(\mathbf{R}^B_C \begin{bmatrix} \frac{u_{lk}-c_x}{f_x} \\ \frac{v_{lk}-c_y}{f_y} \\ 1 \end{bmatrix} \right)^\wedge$$

$$\frac{\partial \breve{\theta}_l \breve{\phi}_l}{\partial \boldsymbol{\tau}^W_{lk}} = \begin{bmatrix} \frac{z^W_{lk}}{\zeta} & 0 & -\frac{x^W_{lk}}{\zeta} \\ \frac{x^W_{lk} y^W_{lk}}{\varsigma} & -\frac{x^{W\,2}_{lk} + z^{W\,2}_{lk}}{\varsigma} & \frac{y^W_{lk} z^W_{lk}}{\varsigma} \end{bmatrix}$$

$$\zeta = x^{W\,2}_{lk} + z^{W\,2}_{lk}$$

$$\varsigma = (x^{W\,2}_{lk} + y^{W\,2}_{lk} + z^{W\,2}_{lk}) \sqrt{x^{W\,2}_{lk} + z^{W\,2}_{lk}}$$

and

$$\mathbf{J}_{h\rho} = \begin{bmatrix} \mathbf{0}_{3\times 2} & \mathbf{0}_{3\times 1} \\ \frac{\partial \breve{\theta}_l \breve{\phi}_l}{\partial \boldsymbol{\tau}^W_{lk}} \mathbf{R}^W_{B_{k|k}} \mathbf{R}^B_C \begin{bmatrix} \frac{1}{f_x} & 0 \\ 0 & \frac{1}{f_y} \\ 0 & 0 \end{bmatrix} & \mathbf{0}_{2\times 1} \\ 0 & 1 \end{bmatrix} \tag{48}$$

References

[1] Hauke Strasdat, J. M. M. Montiel, and Andrew J. Davison. Visual slam: Why filter? *Image and Vision Computing*, 30(2):65–77, 2012.

[2] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jos Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

[3] Kurt Konolige, Motilal Agrawal, and Joan Sol. Large-scale visual odometry for rough terrain. In *Robotics Research - the International Symposium*, pages 201–212, Nov 2010.

[4] S Weiss, M. W Achtelik, S Lynen, and M Chli. Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*, pages 957–964, May 2012.

[5] Pedro Pinies, Todd Lupton, Salah Sukkarieh, and Juan D. Tardos. Inertial aiding of inverse depth slam using a monocular camera. pages 2797–2802, 2007.

[6] Markus Kleinert and Sebastian Schleith. Inertial aided monocular slam for gps-denied navigation. In *2010 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 20–25, Sep 2010.

[7] Eagle S Jones and Stefano Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *International Journal of Robotics Research*, 30(4):407–430, 2011.

[8] Guoquan P. Huang, Anastasios I. Mourikis, and Stergios I. Roumeliotis. A first-estimates jacobian ekf for improving slam consistency. *Experimental Robotics. Springer, Berlin, Heidelberg*, pages 373–382, 2009.

[9] J. A Hesch and S. I Roumeliotis. Consistency analysis and improvement for single-camera localization. In *Computer Vision and Pattern Recognition Workshops*, pages 15–22, 2013.

[10] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings IEEE International Conference on Robotics and Automation*, pages 3565–3572, April 2007.

[11] Dimitrios G. Kottas, Joel A. Hesch, Sean L. Bowman, and Stergios I. Roumeliotis. On the consistency of vision-aided inertial navigation. *Experimental Robotics. Springer International Publishing*, pages 303–317, 2013.

[12] Mingyang Li and A. I. Mourikis. High-precision, consistent ekf-based visual-inertial odometry. 32(6):690–711, 2013.

[13] Joel A. Hesch, Dimitrios G. Kottas, Sean L. Bowman, and Stergios I. Roumeliotis. Consistency analysis and improvement of vision-aided inertial navigation. *IEEE Transactions on Robotics*, 30(1):158–176, 2017.

[14] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research*, 34(3):314–334, 2015.

[15] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. *Georgia Institute of Technology*, 2015.

[16] V. Kumar A. Concha, G. Loianno and J. Civera. Visual-inertial direct slam. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1331–1338, May 2016.

[17] R. Mur-Artal and J. D. Tards. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017.

[18] V. Usenko, J. Engel, J. Stckler, and D. Cremers. Direct visual-inertial odometry with stereo cameras. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1885–1892, May 2016.

[19] Peiliang Li, Tong Qin, Botao Hu, Fengyuan Zhu, and Shaojie Shen. Monocular visual-inertial state estimation for mobile augmented reality. In *2017 IEEE International Symposium on Mixed and Augmented Reality(ISMAR)*, pages 11–21, Oct 2017.

[20] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *arXiv*, abs/1708.03852, 2017.

[21] T. Lupton and S. Sukkarieh. Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. *IEEE Transactions on Robotics*, 28(1):61–76, Feb 2012.

[22] Google, projecttango, url https://www.google.com/atap/projecttango/.

[23] G. S. Chirikjian. Stochastic models, information theory, and lie groups, volume 2: Analytic methods and modern applications(applied and numerical harmonic analysis)]. *Birkhauser*, 2012.

[24] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, June 2007.

[25] J. Civera, O. G. Grasa, A. J. Davison, and J. M. M. Montiel. 1-point ransac for ekf filtering. application to real-time structure from motion and visual odometry. In *J. Field Rob*, volume 27, pages 609–631, 2010.

[26] J. Civera, A. J. Davison, and J. M. M. Montiel. Inverse depth parametrization for monocular slam. *IEEE Transactions on Robotics*, 24(5):932–945, Oct 2008.

[27] R. Kuemmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613, May 2011.

[28] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, Oct 2015.

[29] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *International Journal of Robotics Research*, 35(10):1157–1163, 2016.

[30] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987.

[31] J Sturm, N Engelhard, F Endres, and W Burgard. A benchmark for the evaluation of rgb-d slam systems. In *Ieee/rsj International Conference on Intelligent Robots and Systems*, pages 573–580, 2012.

[32] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct ekf-based approach. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 298–304, Sept 2015.