2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Tightly-coupled Fusion of Global Positional Measurements in Optimization-based Visual-Inertial Odometry

Giovanni Cioffi, Davide Scaramuzza

***Abstract*— Motivated by the goal of achieving robust, drift-free pose estimation in long-term autonomous navigation, in this work we propose a methodology to fuse global positional information with visual and inertial measurements in a tightly-coupled nonlinear-optimization–based estimator. Differently from previous works, which are loosely-coupled, the use of a tightly-coupled approach allows exploiting the correlations amongst all the measurements. A sliding window of the most recent system states is estimated by minimizing a cost function that includes visual re-projection errors, relative inertial errors, and global positional residuals. We use IMU preintegration to formulate the inertial residuals and leverage the outcome of such algorithm to efficiently compute the global position residuals. The experimental results show that the proposed method achieves accurate and globally consistent estimates, with negligible increase of the optimization computational cost. Our method consistently outperforms the loosely-coupled fusion approach. The mean position error is reduced up to 50% with respect to the loosely-coupled approach in outdoor Unmanned Aerial Vehicle (UAV) flights, where the global position information is given by noisy GPS measurements. To the best of our knowledge, this is the first work where global positional measurements are tightly fused in an optimization-based visual-inertial odometry algorithm, leveraging the IMU preintegration method to define the global positional factors.**

I. Introduction

In order to achieve accurate and globally consistent pose estimates in autonomous robot navigation, different sensors are required. In recent years, many algorithms have been proposed, which use visual and inertial information to achieve accurate and high-rate pose estimates [1], [2]. However, such algorithms accumulate drift over time due to sensor noise and modeling errors, and are not suitable for long-term navigation. As a consequence, global measurements are needed to achieve accurate estimates for long trajectories since their errors do not depend on the distance travelled. They can be used together with visual and inertial measurements to achieve high-rate, both locally and globally consistent estimates. The Global Positioning System (GPS) is an example of global position measurements widely used for localization in outdoor applications. However, GPS measurements are noisy and not reliable to be used as the only sensor modality for accurate localization. More accurate GPS systems, such as differential GPS, are possible but they require the availability of ground stations which limits the number of use cases.

The authors are with the Robotics and Perception Group, Dep. of Informatics, University of Zürich, and Dep. of Neuroinformatics, University of Zürich and ETH Zürich, Switzerland `http://rpg.ifi.uzh.ch`.
This research was supported by the European Union's Horizon2020 research and innovation program through the AERIAL-CORE project (H2020-2019-871479).

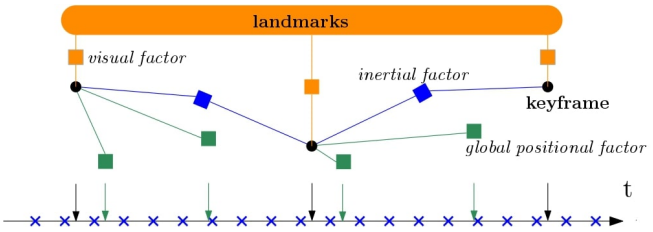

Fig. 1. Representation of the proposed optimization-based multi-sensor fusion. We distinguish three types of factors: visual (orange), inertial (blue), and global positional factors (green). The optimization variables are the states of the keyframes in the current sliding window and the visible landmarks. In the bottom part of the figure, IMU measurements are depicted with crosses on the temporal line, while keyframes and global positional measurements are depicted with black and green arrows, respectively.

Global position measurements were first fused with Visual-Inertial Odometry (VIO) estimates in a pose-graph optimization in [3] and [4]. However, such systems were *loosely-coupled*, meaning that the relative pose updates were estimated by the VIO algorithm independently of the global position information and only then aligned to the global frame via pose-graph optimization. In this way, the correlations amongst all the sensor measurements are automatically discarded resulting in sub-optimal results.

In this work we propose an optimization-based *tightly-coupled* approach to fuse visual, inertial, and global position measurements. The global position measurements are used to define new factors in the optimization graph as depicted in Fig. 1. We define a keyframe-based sliding-window optimization as proposed in [5], where the main difference with respect to [5] is the addition of the global position factors, since the number of states in the optimization does not change. These new error terms can be efficiently computed using the IMU preintegration algorithm [6], [7]. We take advantage of the IMU preintegrated terms, already computed to define the inertial error between consecutive keyframes, to create the constraints between the position of the keyframes in the sliding window and the global position measurements. We show in Section IV that, thanks to the proposed formulation of the global error terms, the increase of the computational time needed to compute and minimize the new cost function is negligible compared to the visual-inertial case. Another important question for the multi-sensors fusion problem addressed in this work is how the number of global positional factors affects the estimates. In Section IV, we run experiments for different numbers of global position measurements included in the estimation

process.

In all experiments, we compare our tightly-coupled approach to a loosely-coupled based on the method proposed in [3]. The results validate our method and show that our approach can be a step towards the target of achieving high-rate locally and globally consistent pose estimates in long-range navigation. To the best of our knowledge, this is the first work that proposes a tightly-coupled approach to fuse global with visual and inertial measurements in an optimization-based algorithm, using the IMU preintegration method to efficiently derive the global positional error terms.

The paper is structured as following: Section I-A contains recent work on algorithms for visual, inertial and global measurements fusion. Section II shows our formulation of the sliding-window optimization problem. Section III introduces the IMU preintegration algorithm and shows how it can be used to derive the global positional residuals. Section IV contains experiments and discussions. Section V concludes the paper.

A. Related Work

Two major approaches can be found in the literature to address the visual, inertial, and global position fusion problem: *filtering methods* and *smoothing methods*.

Filtering methods: Filtering methods carry out efficient estimation by only updating the latest state. Many filter-based approaches involving visual and inertial measurements are inspired by the work in [8], where an Extended Kalman Filter (EKF) was proposed to perform visual-inertial odometry. In [9], an EKF was proposed to fuse inertial data, GPS measurements and vision-based pose estimates. In this case, the poses estimated by an independent (i.e., loosely-coupled) visual odometry algorithm were fused with inertial and GPS measurements in a subsequent estimation step. In [10], the EKF includes online calibration of IMU-GPS extrinsics and time offset.

Smoothing methods: Smoothing algorithms are classified as *full-* or *fixed-lag* smoothers. *Full-lag* smoothers estimate the complete history of the states. They guarantee the highest accuracy but incur high computational cost. In [11], it was proposed to use incremental smoothing technique [12] and IMU preintegration to reduce the computational cost of the full-batch optimization. In [13], achieving high accuracy was prioritized over an online implementation. This work was subsequently extended in [14] to include an extended version of the IMU preintegration algorithm, incorporating gravity and Earth rotation in the IMU model. *Fixed-lag* smoothers (or sliding window estimators) estimate a window of the latest states while marginalizing out the previous states [5]. This approach is more computational efficient than *full-lag* smoothers but less accurate due to accumulation of linearization errors in the marginalization [15].

In [3], the global position measurements were fused with poses estimated by a VIO algorithm in a sliding window pose-graph optimization of the most recent robot states. Similarly in [4], an independent VIO algorithm provided pose estimates that were successively fused with GPS measurements in a pose-graph optimization. Differently from [3], in [4] the pose-graph contains an additional node representing the origin of the local coordinate frame in order to constrain the absolute orientation. However, both these approaches were loosely-coupled, i.e. the relative pose estimates were provided by an independent VIO algorithm. Differently from [3], [4], we propose a tightly-coupled approach where all the measurements are included in a common optimization problem thus considering the correlations amongst them. In [5], it was shown that for the visual-inertial case considering all measurement correlations is crucial for high precision estimates. In [16], it was proposed a tightly-coupled sliding window optimization for visual and inertial measurements with loosely-coupled GPS refinement. The GPS measurements were assumed to be available at low-rate and they were given the same time stamp of the temporally closest image in order to be included in the sliding window. Differently from [16], we tightly couple the global position measurements using the IMU preintegration algorithm to efficiently derive the global positional factors. This allows to add multiple global factors per keyframe in the sliding window with negligible extra computational cost.

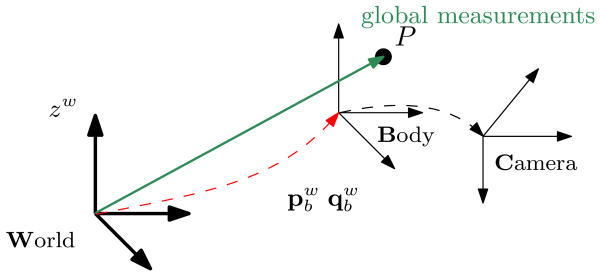

Fig. 2. Reference frames used in this work.

II. Problem Formulation

A. Notation

The coordinate frames used are depicted in Fig. 2. W represents the world frame. We assume the direction of the gravity aligned to z^w axis. B is the body frame and corresponds to the IMU frame. The camera frame is denoted by C. We use the notation $(\cdot)^w$ to represent a quantity in the world frame W. Similar notation applies for every reference frame. We use $\mathbf{p}_{b_k}^w$ and $\mathbf{q}_{b_k}^w$ to represent the position and orientation of B with respect to W at time t_k. The rotation matrix representation is $\mathbf{R}_{b_k}^w$. The velocity of B expressed in W at time t_k is $\mathbf{v}_{b_k}^w$. Global position measurements are given by $\mathbf{p}_{p_k}^w$, where P is a point rigidly attached to B by $\mathbf{p}_p^b$. For example, the point P could represent the position of the receiver antenna in the case of GPS measurements. The value of $\mathbf{p}_p^b$ can be obtained from the calibration of the system. The notation $\hat{(\cdot)}$ is used to represent noisy measurements.

The keyframe-based sliding window optimization variables are $\mathcal{X} = \{\mathcal{L}, \mathcal{X}_B\}$, where $\mathcal{L}$ comprises the position of the 3D landmarks visible in the sliding window and $\mathcal{X}_B = [\mathbf{x}_1, \cdots, \mathbf{x}_K]$ comprises the system states, with K

the total number of keyframes in the sliding window. The system state $\mathbf{x}_k$ at time t_k is given by the body position $\mathbf{p}^w_{b_k}$, the body orientation quaternion $\mathbf{q}^w_{b_k}$, the body velocity $\mathbf{v}^w_{b_k}$, accelerometer $\mathbf{b}_{a_k}$, and gyroscope biases $\mathbf{b}_{g_k}$: $\mathbf{x}_k = [\mathbf{p}^w_{b_k}, \mathbf{q}^w_{b_k}, \mathbf{v}^w_{b_k}, \mathbf{b}_{a_k}, \mathbf{b}_{g_k}]$.

B. Optimization-based Visual, Inertial, and Global Information Fusion

The keyframe-based visual-inertial localization and mapping problem is formulated as a joint nonlinear optimization which solves for the maximum a posteriori estimate of $\mathcal{X}$. Using the problem formulation as proposed in [5] with some minor changes, the cost function to minimize is written as

$$J_{VI}(\mathcal{X}) = \sum_{k=0}^{K-1} \sum_{j \in \mathcal{J}_k} \left\| \mathbf{e}^{j,k}_{\mathbf{v}} \right\|^2_{\mathbf{W}^{j,k}_{\mathbf{v}}} + \sum_{k=0}^{K-1} \left\| \mathbf{e}^{k}_{\mathbf{i}} \right\|^2_{\mathbf{W}^{k}_{\mathbf{i}}} + \left\| \mathbf{e}_{\mathbf{p}} \right\|^2 . \tag{1}$$

$J_{VI}(\mathcal{X})$ contains the weighted visual $\mathbf{e_v}$, inertial $\mathbf{e_i}$, and marginalization residuals $\mathbf{e_p}$.

The visual residuals are $\mathbf{e}^{j,k}_{\mathbf{v}} = \mathbf{z}^{j,k} - h(\mathbf{l}^w_j)$, which describe the re-projection error of the landmark $\mathbf{l}^w_j \in \mathcal{J}_k$, where $\mathcal{J}_k$ is the set containing all the visible landmarks from the keyframe k in the sliding window. The function $h(\cdot)$ denotes the camera projection model and $\mathbf{z}^{j,k}$ the 2D image measurement. We refer to [5] for additional details. The inertial residuals $\mathbf{e_i}$ are formulated using the IMU preintegration algorithm as proposed in [7], [17]. The derivation of the global positional residuals is inspired by the IMU preintegration algorithm as we describe in Section III. The error term $\mathbf{e_p}$ denotes the prior information obtained from marginalization. We adopt the marginalization strategy proposed in [5]. Namely, when a new frame is inserted in the sliding window we distinguish two cases. In the case the oldest state in the sliding window is not a keyframe, it is marginalized out and all its landmarks are dropped to keep sparsity. In the case the oldest state is a keyframe, the landmarks visible from such frame but not in the most recent keyframe are also marginalized out.

Global positional residuals are added to (1) to derive the cost function proposed in this work, as

$$J(\mathcal{X}) = J_{VI}(\mathcal{X}) + \sum_{k=0}^{K-1} \sum_{j \in \mathcal{G}_k} \left\| \mathbf{e}^{j,k}_{\mathbf{g}} \right\|^2_{\mathbf{W}^{k}_{\mathbf{g}}} , \tag{2}$$

where $\mathcal{G}_k$ contains the global positional measurements connected to the state $\mathbf{x}_k$ by an error term.

Next, we derive the global residual terms $\mathbf{e}^{j,k}_{\mathbf{g}}$ leveraging the outcome of the IMU preintegration algorithm and infer the residual weights $\mathbf{W}^k_{\mathbf{g}}$.

III. Derivation of Global Position Residuals

A. IMU Preintegration

In this section, we review the IMU preintegration algorithm focusing on the derivation of the quantities then utilized in Section III-B for the formulation of the global positional residuals. We use the IMU preintegration derivation proposed in [17], which is based on the continuous-time quaternion-based formulation in [18] and includes the manipulation of IMU biases as in [7].

IMU residuals are formulated as relative constraints between consecutive states using accelerometer $\hat{\mathbf{a}}_t = \mathbf{a}_t + \mathbf{b}_{a_t} + \mathbf{R}^t_w \mathbf{g}^w + \mathbf{n}_a$, and gyroscope $\hat{\mathbf{w}}_t = \mathbf{w}_t + \mathbf{b}_{w_t} + \mathbf{n}_w$ measurements. The accelerometer and gyroscope additive noises are modeled as additive Gaussian noise $\mathbf{n}_a \sim \mathcal{N}(\mathbf{0}, \sigma^2_a \cdot \mathbf{I})$ and $\mathbf{n}_w \sim \mathcal{N}(\mathbf{0}, \sigma^2_w \cdot \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Biases are modeled as random walks $\dot{\mathbf{b}}_{a_t} = \boldsymbol{\eta}_{b_a}$ and $\dot{\mathbf{b}}_{w_t} = \boldsymbol{\eta}_{b_w}$, with $\boldsymbol{\eta}_{b_a} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{b_a} \cdot \mathbf{I})$ and $\boldsymbol{\eta}_{b_w} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{b_w} \cdot \mathbf{I})$.

Given the time interval $[t_k, t_{k+1}]$, $\mathbf{p}^w_{b_k}$, $\mathbf{v}^w_{b_k}$, and $\mathbf{q}^w_{b_k}$ can be propagated in such time interval by using the accelerometer and gyroscope measurements. The propagation in the world frame requires the knowledge of the initial state. This implies that every time the estimate of the initial state changes, e.g. when it is updated in an optimization step, repropagation is needed. The main benefit of the IMU preintegration algorithm is to avoid the need of repropagation at every optimization step, which results in saving of valuable computational resources.

The propagation is executed in the local frame B_k instead of the world frame as

$$\begin{aligned}
\mathbf{R}^{b_k}_w \mathbf{p}^w_{b_{k+1}} &= \mathbf{R}^{b_w}_w (\mathbf{p}^w_{b_k} + \mathbf{v}^w_{b_k} \Delta t_k - \frac{1}{2} \mathbf{g}^w \Delta t^2_k) + \boldsymbol{\alpha}^{b_k}_{b_{k+1}} \\
\mathbf{R}^{b_k}_w \mathbf{v}^w_{b_{k+1}} &= \mathbf{R}^{b_k}_w (\mathbf{v}^w_{b_k} - \mathbf{g}^w \Delta t_k) + \boldsymbol{\beta}^{b_k}_{b_{k+1}} \\
\mathbf{q}^{b_k}_w \otimes \mathbf{q}^w_{b_{k+1}} &= \boldsymbol{\gamma}^{b_k}_{b_{k+1}},
\end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}^{b_k}_{b_{k+1}}$, $\boldsymbol{\beta}^{b_k}_{b_{k+1}}$, and $\boldsymbol{\gamma}^{b_k}_{b_{k+1}}$ are the preintegration terms, which only depend on the inertial measurements as well as biases.

In the discrete-time case using Euler numerical integration method, the mean of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be computed recursively as

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^{b_k}_{i+1} &= \hat{\boldsymbol{\alpha}}^{b_k}_i + \hat{\boldsymbol{\beta}}^{b_k}_i \delta t + \frac{1}{2} \mathbf{R}(\hat{\boldsymbol{\gamma}}^{b_k}_i)(\hat{\mathbf{a}}_i - \mathbf{b}_{a_i}) \delta t^2 \\
\hat{\boldsymbol{\beta}}^{b_k}_{i+1} &= \hat{\boldsymbol{\beta}}^{b_k}_i + \mathbf{R}(\hat{\boldsymbol{\gamma}}^{b_k}_i)(\hat{\mathbf{a}}_i - \mathbf{b}_{a_i}) \delta t \\
\hat{\boldsymbol{\gamma}}^{b_k}_{i+1} &= \hat{\boldsymbol{\gamma}}^{b_k}_i \otimes \begin{bmatrix} 1 \\ \frac{1}{2}(\hat{\boldsymbol{w}}_i - \mathbf{b}_{w_i}) \delta t \end{bmatrix},
\end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}^{b_k}_{b_k} = \boldsymbol{\beta}^{b_k}_{b_k} = 0$ and $\boldsymbol{\gamma}^{b_k}_{b_k}$ is equal to the identity quaternion. $\mathbf{R}(\hat{\boldsymbol{\gamma}}^k_j)$ is the rotation matrix representation of $\hat{\boldsymbol{\gamma}}^k_j$. The covariance $\mathbf{P}^{b_k}_{b_{k+1}}$ can be also calculated recursively, we refer to [17] for the derivation.

As proposed in [7], we update $\boldsymbol{\alpha}^{b_k}_{b_{k+1}}, \boldsymbol{\beta}^{b_k}_{b_{k+1}}, \boldsymbol{\gamma}^{b_k}_{b_{k+1}}$ using their first-order approximation with respect to the biases if the change in the estimate of the biases is small. Otherwise, propagation is redone. The inertial residuals $\mathbf{e}^k_{\mathbf{i}}$ are derived from (3) as

$$\mathbf{e}^k_{\mathbf{i}} = \begin{bmatrix} \mathbf{R}^{b_w}_w (\mathbf{p}^w_{b_{k+1}} - \mathbf{p}^w_{b_k} - \mathbf{v}^w_{b_k} \Delta t_k + \frac{1}{2} \mathbf{g}^w \Delta t^2_k) - \hat{\boldsymbol{\alpha}}^{b_k}_{b_{k+1}} \\ \mathbf{R}^{b_w}_w (\mathbf{v}^w_{b_{k+1}} - \mathbf{v}^w_{b_k} + \mathbf{g}^w \Delta t_k) - \hat{\boldsymbol{\beta}}^{b_k}_{b_{k+1}} \\ 2[(\mathbf{q}^w_{b_k})^{-1} \otimes \mathbf{q}^w_{b_{k+1}} \otimes (\hat{\boldsymbol{\gamma}}^{b_k}_{b_{k+1}})^{-1}]_{xyz} \\ \mathbf{b}_{a_{k+1}} - \mathbf{b}_{a_k} \\ \mathbf{b}_{w_{k+1}} - \mathbf{b}_{w_k} \end{bmatrix} . \tag{5}$$

We leverage the formulation of the position error term in (5) to derive the global positional residuals as formulated in the next section.

B. Global Position Residuals

The global positional measurements are given by $\{\mathbf{p}_{p_j}^w\}$ at time $\{t_j\}$. We model the measurement uncertainty with additive Gaussian noise so that

$$\hat{\mathbf{p}}_{p_j}^w = \mathbf{p}_{p_j}^w + \mathbf{n}_p, \tag{6}$$

where $\mathbf{n}_p \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \cdot \mathbf{I})$. Given a state in the current sliding window $\mathbf{x}_k$ at time t_k and a measurement $\hat{\mathbf{p}}_{p_j}^w$ at time $t_j \in [t_k, t_{k+1})$ the global position residual is defined as

$$\mathbf{e}_{\mathbf{g}}^{j,k} = \mathbf{R}_w^{b_k}(\hat{\mathbf{p}}_{b_j}^w - \mathbf{p}_{b_k}^w - \mathbf{v}_{b_k}^w \Delta t_k + \frac{1}{2}\mathbf{g}^w \Delta t_k^2) - \hat{\boldsymbol{\alpha}}_{b_j}^{b_k}, \tag{7}$$

where the measurement $\hat{\mathbf{p}}_{p_j}^w$ is transformed in $\hat{\mathbf{p}}_{b_j}^w$ as

$$\hat{\mathbf{p}}_{b_j}^w = \hat{\mathbf{p}}_{p_j}^w - \mathbf{R}_{b_j}^w \mathbf{p}_p^b, \tag{8}$$

with $\mathbf{R}_{b_j}^w = \mathbf{R}_{b_k}^w \hat{\boldsymbol{\gamma}}_j^k$.

To define the global residuals, the state position is propagated using inertial measurements in the time interval $[t_k, t_j]$. We express the error term in (7) in the reference frame B_k and take advantage of the computation of the preintegration terms in (4). In fact, $\hat{\boldsymbol{\alpha}}_{b_j}^{b_k}$ can be efficiently obtained during the recursive calculation of $\hat{\boldsymbol{\alpha}}_{b_{k+1}}^{b_k}$, in (4), since $t_j < t_{k+1}$ and imu measurements are buffered in $[t_k, t_{k+1}]$. The same applies for $\hat{\boldsymbol{\gamma}}_j^k$. This allows to minimize the computational time required to include the error term (7) in the cost function (2). To derive the residual weights $\mathbf{W}_{\mathbf{g}}^k$, we rewrite (7) as

$$\begin{aligned}\mathbf{e}_{\mathbf{g}}^{j,k} =& \mathbf{R}_w^{b_k}(-\mathbf{p}_{b_k}^w - \mathbf{v}_{b_k}^w \Delta t_k + \frac{1}{2}\mathbf{g}^w \Delta t_k^2) \\ & - \hat{\boldsymbol{\alpha}}_{b_j}^{b_k} + \mathbf{R}_w^{b_k} \hat{\mathbf{p}}_{p_j}^w - \mathbf{R}(\hat{\boldsymbol{\gamma}}_j^k)\mathbf{p}_p^b. \end{aligned} \tag{9}$$

In (9), $\hat{\boldsymbol{\alpha}}_{b_j}^{b_k}$, $\hat{\mathbf{p}}_{p_j}^w$ and $\hat{\boldsymbol{\gamma}}_j^k$ are the noisy measurements. The covariance of $\hat{\boldsymbol{\gamma}}_j^k$ depends on gyroscope noise and bias. Since gyroscope noise is already considered in the computation of $\hat{\boldsymbol{\alpha}}_{b_j}^{b_k}$ (the reader can refer to [17] and [7] for additional details) and it is usually smaller than accelerometer noise, we omit $\hat{\boldsymbol{\gamma}}_j^k$ in the derivation of $\mathbf{W}_{\mathbf{g}}^k$.

As consequence, the residual weights depend on the covariance of $\hat{\boldsymbol{\alpha}}_{b_j}^{b_k}$ and $\hat{\mathbf{p}}_{p_j}^w$ as

$$\mathbf{W}_{\mathbf{g}}^k = {}_{\hat{\boldsymbol{\alpha}}}\mathbf{P}_{b_j}^{b_k} + \mathbf{R}_w^{b_k}(\sigma_p^2 \cdot \mathbf{I})(\mathbf{R}_w^{b_k})^t, \tag{10}$$

where ${}_{\hat{\boldsymbol{\alpha}}}\mathbf{P}_{b_j}^{b_k}$ is the top-left 3x3 part of $\mathbf{P}_{b_j}^{b_k}$. The covariance $\mathbf{P}_{b_j}^{b_k}$ differs from the covariance of the inertial residuals (i.e., $\mathbf{P}_{b_{k+1}}^{b_k}$) since it is derived from a sub-set of the inertial measurements in $[t_k, t_{k+1}]$.

When a state connected to global position residuals needs to be marginalized, the global residuals are transformed in the prior linear error term together with inertial and visual residuals.

Sampling Strategy: We define $\mathcal{G}_k$ as the set containing the global position measurements in the time interval $[t_k, t_{k+1})$, which are connected to the state $\mathbf{x}_k$ by error term in (7). N is the cardinality of $\mathcal{G}_k$ such that N = $|\mathcal{G}_k|$. Since we use the recursive formulation of the IMU preintegrated terms in (4) to compute $\hat{\alpha}_{b_j}^{b_k}$ in (7), increasing N only has a minor affect on the optimization computational cost. As soon as a new measurement is available, it is included in $\mathcal{G}_k$ and a new residual (7) is added to the optimization. In Section IV, we evaluate how different maximum values of N affect the pose estimates.

IV. Experiments

We evaluated our approach on two visual-inertial datasets with global position measurements: an indoor one (the EuRoC dataset [19], Section IV-A) and an outdoor one (from [4], Section IV-B). Since the EuRoC dataset provides global position measurements from a motion capture system, we corrupted the motion capture system measurements with Gaussian noise to simulate noisy global position measurements. The second dataset, instead, provides global position measurements from a GPS and ground-truth from a total station.

As a vision front-end, we used the one of SVO [20]. The vision front-end deals with feature detection and tracking from images. Features correctly tracked in subsequent frames are then triangulated and added to the sliding-window optimization. The vision front-end is also responsible for the selection of the keyframes. We limited the number of keyframes in the sliding-window to 10. We used the Ceres Solver [21] to solve the optimization problem. The vision front-end and the sliding-window solver run in two separate threads. All the experiments ran on a laptop equipped with a 2.60GHz Intel Core i7 CPU.

A. EuRoC Dataset

1) Setup: The EuRoC dataset contains eleven sequences recorded from a hex-rotor helicopter. Five sequences are recorded in an industrial machine hall and six in an office room. The sequences recorded in the industrial hall are labeled as MH_ and those in the office room as V_. Every sequence is classified as easy, medium, or hard depending on illumination conditions, scene texture and vehicle motion. Hard sequences contain challenging illumination conditions and fast motion. Hardware synchronized stereo images and IMU measurements are available at a rate of 20 Hz and 200 Hz, respectively. We ran the experiments in a monocular setup using only images from the left camera. In every sequence, a motion capture system was used to record ground-truth. The ground-truth measurements were corrupted with zero-mean Gaussian noise to simulate noisy global position measurements. The Gaussian noise was defined as $\mathbf{n}_{mc} \sim \mathcal{N}(\mathbf{0}, \sigma_{mc}^2 \cdot \mathbf{I})$, σ_{mc} = 20 cm. The additive Gaussian noise $\mathbf{n}_p$ in (6) was set equal to $\mathbf{n}_{mc}$. For initialization, we set the initial position equal to the corresponding noisy motion capture system measurement.

2) Results: The proposed method was evaluated in terms of estimated trajectory accuracy and solver time. We used the trajectory evaluation toolbox in [22] to compute the evaluation metrics. Each state $\mathbf{x}_k$ in the optimization window is connected to N = $|\mathcal{G}_k|$ global positional residuals, where $\mathcal{G}_k$ is the set containing the global positional measurements in the time interval $[t_k, t_{k+1})$. We were interested in how the cardinality of $\mathcal{G}_k$ influences the trajectory estimates and for

TABLE I

ATE [M] ON THE EUROC SEQUENCES.

Sequence	VIO-only (no GP)	Loosely-coupled ([3])	Tightly-coupled (proposed)			
			N=1	N=2	N=3	N=4
MH01	0.188	0.081	0.031	0.029	0.022	0.022
MH02	0.140	0.085	0.036	0.032	0.027	0.025
MH03	0.133	0.110	0.048	0.039	0.034	0.033
MH04	0.186	0.119	0.068	0.058	0.051	0.048
MH05	0.306	0.115	0.056	0.044	0.039	0.039
V101	0.061	0.081	0.041	0.036	0.034	0.034
V102	0.103	0.097	0.048	0.042	0.036	0.035
V103	0.179	0.099	0.068	0.050	0.047	0.042
V201	0.065	0.087	0.038	0.027	0.026	0.026
V202	0.103	0.127	0.046	0.038	0.036	0.033
V203	0.232	0.177	0.098	0.074	0.057	0.057

the experiments in this section we evaluated $|\mathcal{G}_k|$ = [1, 2, 3, 4]. We also included as reference the VIO estimates, in this case no global residual terms are included in the optimization and the sliding window cost function is (1).

Accuracy: Table I contains the absolute trajectory error (ATE) [22], [23] obtained on all the EuRoC sequences. We included the results for the VIO-only case, i.e. without fusion of global positional (GP) measurements, the loosely-coupled approach, and our proposed tightly-coupled approach with N $\in$ [1,2,3,4]. For the VIO-only case, the estimated trajectory is aligned to the ground-truth using the *posyaw* alignment method in [22]. When N $\geq$ 1 (i.e., when GP measurements are considered), no alignment is applied. Each configuration was run three times and the ATE median value is reported. By comparing the VIO-only to the loosely- and tightly-coupled results, we can see that the ATE decreases when global positional factors are included in the sliding window. In our proposed tightly-coupled approach, the residual in (7) constrains the position estimate at t_k to be consistent with the global positional measurement allowing to reduce the error that accumulates in the visual-inertial estimates. The largest improvement between the VIO-only case and the tightly-coupled approach with N=1 is in sequence MH05, where the ATE decreases from 0.306 m to 0.056 m. The estimate accuracy is improved by adding more global residuals per keyframe (i.e., N>1). Increasing N from 1 to 2 allows to reduce the ATE in every sequence, with the largest and smallest improvements in sequence V203 and MH01, respectively. The average decrease of the ATE for all the EuRoC sequences is equal to 0.010 m. Increasing N from 2 to 3 improves the ATE in every sequence but the benefit is smaller than the previous case (i.e., increasing N from 1 to 2). The average decrease of the ATE on all the EuRoC sequences is equal to 0.005 m. A further increase of N from 3 to 4 has a minor effect on the estimate accuracy. The ATE remains the same for the sequence MH01, MH05, V101, V201, and V203, and it slightly decreases for other sequences.

In Fig. 3, we show the relative pose error, computed as proposed in [24], in the sequence V203, which is labeled as difficult. The top-view of the trajectory is in Fig 4. We see in the top plot of Fig. 3 that the relative translation error

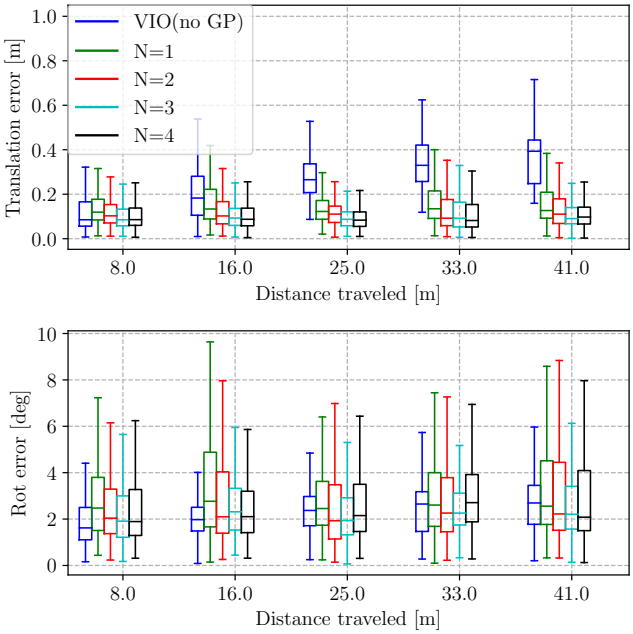

Fig. 3. Relative translation and rotation error in EuRoC V203 difficult. Each plot contains evaluation for different values of N = $|\mathcal{G}_k|$ as well as VIO-only estimates, i.e. global positional (GP) measurements are not included in the estimation process.

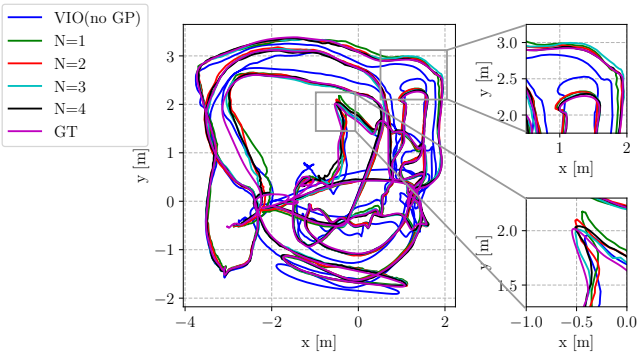

Fig. 4. Trajectory top-view of the sequence V203 difficult. The two zoomed-in sections in the right column highlight how the drift accumulated in the VIO-only case is corrected depending on the number of global residuals N per keyframe included in the sliding window.

visibly decreases when global residuals are included in the estimator. Adding more than one error term per keyframe (i.e., N $>$ 1) helps improving the estimates. Increasing N from 1 to 2 reduces the ATE from 0.098 m to 0.074 m as shown in the bottom line of Table I. The ATE decreases by 0.017 m with further increase of N to 3. Increasing N from 3 to 4 does not provide any improvement.

The rotation error is also decreased by the addition of the global positional measurements in the sliding window optimization as shown in the bottom plot of Fig. 3. However, the effect is smaller compared to the improvement achieved on the translation error. This result was expected since the estimator has only access to global positional information

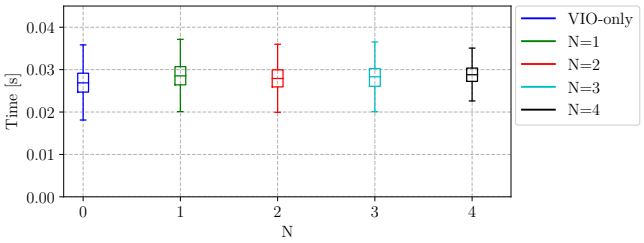

Fig. 5. Optimizer time with respect to N. The value on the y axis represents the processing time, which is the duration between the time at which the front-end receives an image and the time at which its optimized pose is available from the sliding window optimization.

(i.e., no global orientation).

Timing: Increasing the value of N only slightly affects the processing time. The processing time was defined as the duration between the time at which the front-end receives an image and the time at which its optimized pose is available from the sliding window optimization. In Fig. 5, we evaluate how the processing time varies with respect to N and compare to the VIO-only case. The median is 26.2 ms for the VIO-only case and 27.7 ms for N = 1. The increase of just 1.3 ms shows that the formulation of the global residuals in (7) allows to efficiently include new measurements in the sliding window optimization. As observed in Fig. 5, adding more residual terms has a very negligible impact on the processing time.

Comparison to loosely-coupled: The proposed tightly-coupled fusion was compared to a loosely-coupled approach based on the method proposed in [3]. The loosely-coupled pose-graph optimization runs on a sliding window that contains the most recent keyframes selected by the VIO algorithm and the global position measurements. The newest frame in the sliding window corresponds to the most recent frame processed by the VIO pipeline. Each keyframe is connected to one global measurement. At every optimization step, the transformation between the VIO local frame and the global frame is estimated. Global position measurements are expressed with respect to such global frame. This transformation is applied to the most recent VIO output to obtain drift-free global pose estimates at the same rate of the VIO estimates. We refer to [3] for more details on the loosely-coupled approach. Fig. 6 shows the results of the two methods on sequence V203. Our tightly-coupled method outperforms the loosely-coupled approach in terms of both translation and rotation error with N $\in$ [1,2,3,4] in every EuRoC sequence, as shown in Table I. Due to the noise in the global position measurements, the loosely-coupled fusion provides estimates less accurate than the VIO pipeline for sequence V202.

B. Outdoor Dataset with GPS Measurements

1) Setup: In this second set of experiments, we evaluated our approach on the dataset kindly provided by the authors of [4]. This dataset contains three flight sequences from an UAV equipped with a commercial stereo visual-inertial

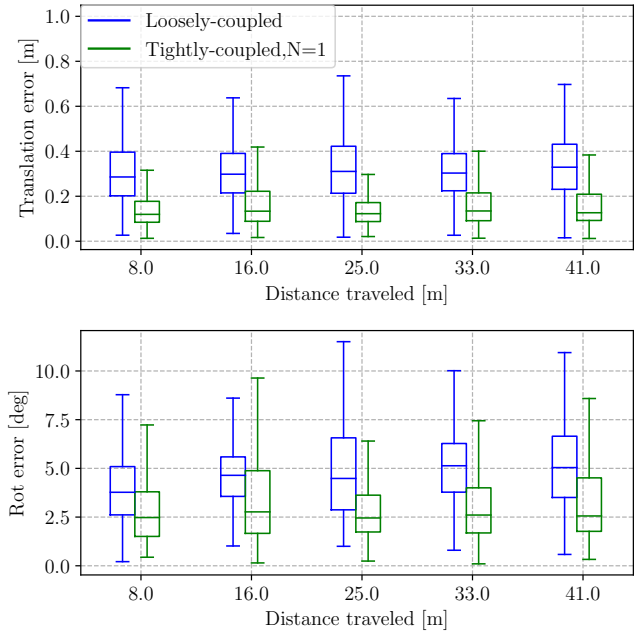

Fig. 6. Relative translation and rotation error in EuRoC V203 difficult. Comparison between the tightly-coupled fusion approach proposed in this work and the loosely-coupled method based on [3]. We used N=1 in the tightly-coupled method.

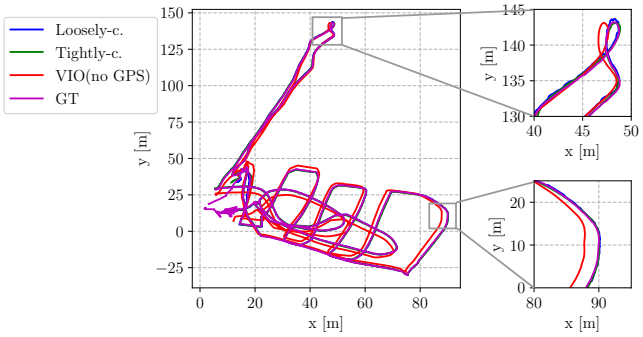

Fig. 7. Top-view trajectory in flight sequence 3. Ground truth (GT), VIO (no GPS), loosely-coupled, and tightly-coupled trajectories are depicted.

sensor, GPS, and ground-truth from a Leica total station. The three flight sequences have a travelled distance of 404.1 m, 483.3 m, and 1033.3 m, respectively. The GPS data, acquired at 5 Hz, provides the global positional measurements. For our monocular visual front-end, we only used images from the left camera. Due to the unavailability of ground-truth orientation, we only reported the position error (mean and standard deviation) after the alignment. Fig. 7 shows the top-view of the third flight sequence, which contains a 1033.3 m long trajectory. For initialization, the initial position corresponds to the first GPS measurement.

2) Results: As observed in Table II, our tightly-coupled approach gives more accurate position estimates than the loosely-coupled. The largest improvement is 50% in the second flight sequence and the smallest is 31% in the third

TABLE II

POSITION ERROR FOR THE OUTDOOR DATASET

Flight	Position Error [m]	VIO (no GPS)	Loosely-coupled ([3])	GOMSF ([4])	Tightly-coupled N=1 (proposed)
1	mean	0.83	0.64	0.33	**0.28**
	std	0.40	0.24	0.16	**0.13**
2	mean	1.28	0.35	0.29	**0.24**
	std	0.63	0.17	0.13	**0.09**
3	mean	3.63	0.45	0.43	**0.38**
	std	1.59	0.20	0.20	**0.18**

flight sequence.

In the same table, we also included the best results of the loosely-coupled method proposed in [4], named GOMSF. GOMSF differs from [3] by the addition of a virtual node representing the local coordinate frame. VIO estimates are expressed in such coordinate frame. We can observe that our method also improves the mean position error with respect to GOMSF by 18%, 14%, and 12%, respectively, in all three flight sequences. The difference in improvement with respect to [4] is very likely due to the different front-end utilized: while we use the monocular SVO front-end, GOMFS uses the stereo OKVIS front-end [5].

V. CONCLUSION

Visual and inertial measurements are suitable to obtain locally accurate pose estimates but accumulate large drift in long-term navigation. To achieve high-rate, accurate, locally and globally consistent estimates, global positional information can be fused with visual and inertial measurements. We proposed in this paper a tightly-coupled optimization-based methodology to solve the multi-sensors fusion problem. We formulated the fusion problem as a keyframe-based sliding window optimization where the global measurements are employed to derive the new global factors. We leveraged the computation of the IMU preintegrated terms to include the global positional factors in the optimization with negligible increase of the computational cost compared to the visual-inertial case. Experimental results showed that the proposed approach efficiently achieves accurate and globally consistent position estimates and consistently outperforms the state-of-the-art loosely-coupled approach.

REFERENCES

[1] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 2502–2509.

[2] G. Huang, "Visual-inertial navigation: A concise review," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019.

[3] T. Qin, S. Cao, J. Pan, and S. Shen, "A general optimization-based framework for global pose estimation with multiple sensors," *arXiv e-prints*, 2019, retrieved on March 1st, 2020. [Online]. Available: https://arxiv.org/abs/1901.03642v1

[4] R. Mascaro, L. Teixeira, T. Hinzmann, R. Siegwart, and M. Chli, "Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 1421–1428.

[5] 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, 2015.

[6] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 61–76, 2011.

[7] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, 2016.

[8] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2007, pp. 3565–3572.

[9] S. Weiss, M. W. Achtelik, M. Chli, and R. Siegwart, "Versatile distributed pose estimation and sensor self-calibration for an autonomous mav," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2012, pp. 31–38.

[10] W. Lee, K. Eckenhoff, P. Geneva, and G. Huang, "Intermittent gps-aided vio: Online initialization and calibration," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020.

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

[12] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "isam2: Incremental smoothing and mapping using the bayes tree," *Int. J. Robot. Research*, vol. 31, no. 2, pp. 216–235, 2012.

[13] D. A. Cucci, M. Rehak, and J. Skaloud, "Bundle adjustment with raw inertial observations in uav applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 1–12, 2017.

[14] D. A. Cucci and J. Skaloud, "On raw inertial measurements in dynamic networks," *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, 2019.

[15] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Camera-imu-based localization: Observability analysis and consistency improvement," *Int. J. Robot. Research*, vol. 33, no. 1, pp. 182–201, 2014.

[16] Y. Yu, W. Gao, C. Liu, S. Shen, and M. Liu, "A gps-aided omnidirectional visual-inertial state estimator in ubiquitous environments," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2019, pp. 7750–7755.

[17] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018.

[18] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft mavs," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2015, pp. 5303–5310.

[19] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *Int. J. Robot. Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[20] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, 2017.

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

[22] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2018, pp. 7244–7251.

[23] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2012, pp. 573–580.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, 2012, pp. 3354–3361.