Keyframe-Based Visual-Inertial SLAM Using Nonlinear Optimization

Stefan Leutenegger*, Paul Furgale*, Vincent Rabaud†, Margarita Chli*, Kurt Konolige‡ and Roland Siegwart*
* Autonomous Systems Lab (ASL), ETH Zurich, Switzerland
† Willow Garage, Menlo Park, CA 94025, USA ‡ Industrial Perception, Palo Alto, CA 94303, USA

***Abstract*—The fusion of visual and inertial cues has become popular in robotics due to the complementary nature of the two sensing modalities. While most fusion strategies to date rely on filtering schemes, the visual robotics community has recently turned to non-linear optimization approaches for tasks such as visual Simultaneous Localization And Mapping (SLAM), following the discovery that this comes with significant advantages in quality of performance and computational complexity. Following this trend, we present a novel approach to *tightly* integrate visual measurements with readings from an Inertial Measurement Unit (IMU) in SLAM. An IMU error term is integrated with the landmark reprojection error in a fully probabilistic manner, resulting to a joint non-linear cost function to be optimized. Employing the powerful concept of 'keyframes' we partially marginalize old states to maintain a bounded-sized optimization window, ensuring real-time operation. Comparing against both vision-only and loosely-coupled visual-inertial algorithms, our experiments confirm the benefits of tight fusion in terms of accuracy and robustness.**

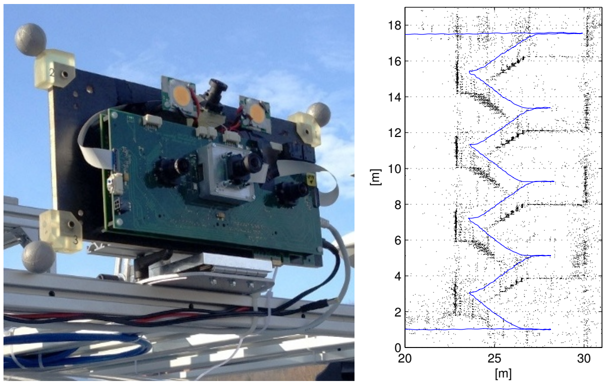

Fig. 1. Synchronized stereo vision and IMU hardware prototype and indoor results obtained walking up a staircase.

I. Introduction

Combining visual and inertial measurements has long been a popular means for addressing common Robotics tasks such as egomotion estimation, visual odometry and SLAM. The rich representation of a scene captured in an image, together with the accurate short-term estimates by gyroscopes and accelerometers present in a typical IMU have been acknowledged to complement each other, with great uses in airborne [6, 20] and automotive [14] navigation. Moreover, with the availability of these sensors in most smart phones, there is great interest and research activity in effective solutions to visual-inertial SLAM.

Historically, the visual-inertial pose estimation problem has been addressed with filtering, where the IMU measurements are propagated and keypoint measurements are used to form updates. Mourikis and Roumeliotis [14] proposed an EKF-based real-time fusion using monocular vision, while Jones and Soatto [8] presented mono-visual-inertial filtering results on a long outdoor trajectory including IMU to camera calibration and loop closure. Both works perform impressively with errors below 0.5% of the distance travelled. Kelly and Sukhatme [9] provided calibration results and a study of observability in the context of filtering-based vision-IMU fusion. Global unobservability of yaw and position, as well as growing uncertainty with respect to an initial pose of reference are intrinsic to the visual-inertial estimation problem; this poses a challenge to the filtering approaches which typically rely on some form of linearization.

In [18] it was shown that in purely visual SLAM optimization-based approaches provide better accuracy for the same computational work when compared to filtering approaches. Maintaining a relatively sparse graph of *keyframes* and their associated landmarks subject to nonlinear optimization, has since been very popular.

The visual-inertial fusion approaches found in the literature can be categorized to follow two approaches. In *loosely-coupled* systems, e.g. [10], the IMU measurements are incorporated as independent inclinometer and relative yaw measurements into the stereo vision optimization. Weiss et al. [20] use vision-only pose estimates as updates to an EKF with indirect IMU propagation. Also in [15, 7], relative stereo pose estimates are integrated into a factor-graph containing inertial terms and absolute GPS measurements. Such methods limit the complexity, but disregard correlations amongst internal states of different sensors. In contrast, *tightly-coupled* approaches jointly estimate all sensor states. In order to be tractable and as an alternative to filtering, Dong-Si and Mourikis [2] propose a fixed-lag smoother, where a window of successive robot poses and related states is maintained, marginalizing out states (following [19]) that go out of scope. A similar approach, but without inertial terms and in the context of planetary landing is used in [16].

With the aim of robust and accurate visual-inertial SLAM, we advocate *tightly-coupled* fusion for maximal exploitation of sensing cues and *nonlinear estimation* wherever possible rather than filtering in order to reduce suboptimality due to linearization. Our method is inspired by [17], where it was proposed to use IMU error terms in batch-optimized SLAM (albeit only during initialization). Our approach is closely related to the fixed-lag smoother proposed in [2], as it combines inertial terms and reprojection error in a single cost function, and old states get marginalized in order to bound the complexity.

In relation to these works, we see a threefold contribution:

1) We employ the keyframe paradigm for drift-free estimation also when slow or no motion at all is present: rather than using an optimization window of time-successive poses, we keep *keyframes* that may be spaced arbitrarily far in time, keeping visual constraints—while still respecting an IMU term. Our formulation of relative uncertainty of keyframes allows for building a pose graph without expressing global pose uncertainty, taking inspiration from RSLAM [13].
2) We provide a fully probabilistic derivation of IMU error terms, including the respective information matrix, relating successive image frames without explicitly introducing states at IMU-rate.
3) At the system level, we developed both the hardware and the algorithms for accurate real-time SLAM, including robust keypoint matching and outlier rejection using inertial cues.

In the remainder of this article, we introduce the inertial error term in batch visual SLAM in II-B, followed by an overview our real-time stereo image processing and keyframe selection in II-C, and the marginalization formalism in II-D. Finally, we show results obtained with our stereo-vision and IMU sensor indoor and outdoor in III.

II. Tightly Coupled Visual-Inertial Fusion

In visual SLAM, a nonlinear optimization is formulated to find the camera poses and landmark positions by minimizing the reprojection error of landmarks observed in camera frames. Figure 2 shows the respective graph representation : it displays measurements as edges with square boxes and estimated quantities as round nodes. As soon as inertial measurements are introduced, they not only create temporal constraints between successive poses, but also between successive speed and IMU bias estimates of both accelerometers and gyroscopes by which the robot state vector is augmented. In this section, we present our approach of incorporating inertial measurements into batch visual SLAM.

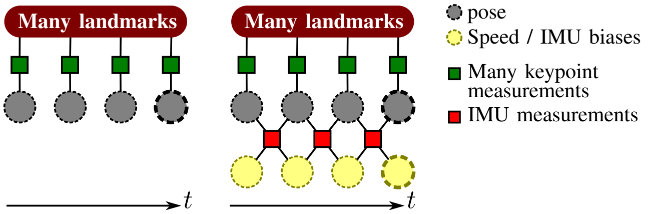

Fig. 2. Graphs of the state variables and measurements involved in the visual SLAM problem (left) versus visual-inertial SLAM (right)

A. Notation and Definitions

1) Notation: We employ the following notation throughout this work: $\underrightarrow{\mathcal{F}}_A$ denotes a reference frame A; vectors expressed in it are written as $\mathbf{p}_A$ or optionally as $\mathbf{p}_A^{BC}$, with B and C as start and end points, respectively. A transformation between frames is represented by a homogeneous transformation matrix $\boldsymbol{T}_{AB}$ that transforms the coordinate representation of homogeneous points from $\underrightarrow{\mathcal{F}}_B$ to $\underrightarrow{\mathcal{F}}_A$. Its rotation matrix part is written as $\mathbf{C}_{AB}$; the corresponding quaternion is written as $\mathbf{q}_{AB} = [\boldsymbol{\epsilon}^T, \eta]^T \in S^3$, $\boldsymbol{\epsilon}$ and η representing the imaginary and real parts. We adopt the notation introduced in Barfoot et al. [1]: concerning the quaternion multiplication $\mathbf{q}_{AC} = \mathbf{q}_{AB} \otimes \mathbf{q}_{BC}$, we introduce a left-hand side compound operator $(.)^+$ and a right-hand side operator $(.)^\oplus$ such that $\mathbf{q}_{AC} = \mathbf{q}_{AB}{}^+ \mathbf{q}_{BC} = \mathbf{q}_{BC}{}^\oplus \mathbf{q}_{AB}$.

2) Frames: The performance of the proposed method is evaluated using a stereo-camera/IMU setup schematically depicted in Figure 3. Inside the tracked body that is represented relative to an inertial frame, $\underrightarrow{\mathcal{F}}_W$, we distinguish camera frames, $\underrightarrow{\mathcal{F}}_{C_i}$, and the IMU-sensor frame, $\underrightarrow{\mathcal{F}}_S$.

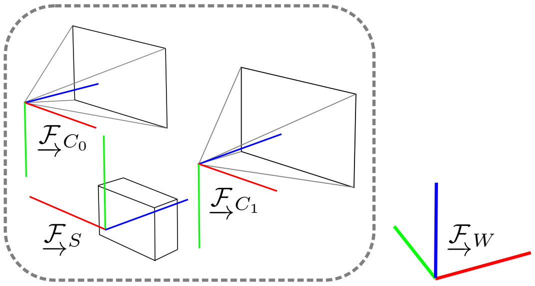
$\underrightarrow{\mathcal{F}}_{C_0}$
$\underrightarrow{\mathcal{F}}_{C_1}$
$\underrightarrow{\mathcal{F}}_S$
$\underrightarrow{\mathcal{F}}_W$

Fig. 3. Coordinate frames involved in the hardware setup used: two cameras are placed as a stereo setup with respective frames, $\underrightarrow{\mathcal{F}}_{C_i}, i \in \{0, 1\}$. IMU data is acquired in $\underrightarrow{\mathcal{F}}_S$. $\underrightarrow{\mathcal{F}}_S$ is estimated with respect to $\underrightarrow{\mathcal{F}}_W$.

3) States: The variables to be estimated comprise the robot states at the image times (index k) $\mathbf{x}_{\mathrm{R}}^{k}$ and landmarks $\mathbf{x}_{\mathrm{L}}^{c}$. $\mathbf{x}_{\mathrm{R}}$ holds the robot position in the inertial frame $\mathbf{p}_{W}^{WS}$, the body orientation quaternion $\mathbf{q}_{WS}$, the velocity in inertial frame $\mathbf{v}_{W}^{WS}$, as well as the biases of the gyroscopes $\mathbf{b}_{\mathrm{g}}$ and the biases of the accelerometers $\mathbf{b}_{\mathrm{a}}$. Thus, $\mathbf{x}_{\mathrm{R}}$ is written as:

$$\mathbf{x}_{\mathrm{R}} := \left[\mathbf{p}_{W}^{WS^T}, \mathbf{q}_{WS}^{T}, \mathbf{v}_{W}^{WS^T}, \mathbf{b}_{\mathrm{g}}^{T}, \mathbf{b}_{\mathrm{a}}^{T}\right]^{T} \in \mathbb{R}^{3} \times S^{3} \times \mathbb{R}^{9}. \quad (1)$$

Furthermore, we use a partition into the pose states $\mathbf{x}_{\mathrm{T}} := [\mathbf{p}_{W}^{WS^T}, \mathbf{q}_{WS}^{T}]^{T}$ and the speed/bias states $\mathbf{x}_{\mathrm{sb}} := [\mathbf{v}_{W}^{WS^T}, \mathbf{b}_{\mathrm{g}}^{T}, \mathbf{b}_{\mathrm{a}}^{T}]^{T}$. Landmarks are represented in homogeneous coordinates as in [3], in order to allow seamless integration of close and very far landmarks: $\mathbf{x}_{\mathrm{L}} := \boldsymbol{l}_{W}^{WL} = [l_x, l_y, l_z, l_w]^{T} \in \mathbb{R}^{4}$.

We use a perturbation in tangent space $\mathfrak{g}$ of the state manifold and employ the group operator $\boxplus$, the exponential exp and logarithm log. Now, we can define the perturbation $\delta\mathbf{x} := \mathbf{x} \boxplus \bar{\mathbf{x}}^{-1}$ around the estimate $\bar{\mathbf{x}}$. We use a minimal coordinate representation $\delta\boldsymbol{\chi} \in \mathbb{R}^{\dim \mathfrak{g}}$. A bijective mapping Φ transforms from minimal coordinates to tangent space:

$$\delta\mathbf{x} = \exp(\Phi(\delta\boldsymbol{\chi})). \quad (2)$$

Concretely, we use the minimal (3D) axis-angle perturbation of orientation $\delta\boldsymbol{\alpha} \in \mathbb{R}^{3}$ which can be converted into its quaternion equivalent $\delta\mathbf{q}$ via the exponential map:

$$\delta\mathbf{q} := \exp\left(\begin{bmatrix} \frac{1}{2}\delta\boldsymbol{\alpha} \\ 0 \end{bmatrix}\right) = \begin{bmatrix} \operatorname{sinc}\left\|\frac{\delta\boldsymbol{\alpha}}{2}\right\| \frac{\delta\boldsymbol{\alpha}}{2} \\ \cos\left\|\frac{\delta\boldsymbol{\alpha}}{2}\right\| \end{bmatrix}. \quad (3)$$

Therefore, using the group operator $\otimes$, we write $\mathbf{q}_{WS} = \delta\mathbf{q} \otimes \bar{\mathbf{q}}_{WS}$. We obtain the minimal robot error state vector

$$\delta\boldsymbol{\chi}_{\mathrm{R}} = \left[\delta\mathbf{p}^{T}, \delta\boldsymbol{\alpha}^{T}, \delta\mathbf{v}^{T}, \delta\mathbf{b}_{\mathrm{g}}^{T}, \delta\mathbf{b}_{\mathrm{a}}^{T}\right]^{T} \in \mathbb{R}^{15}. \quad (4)$$

Analogously to the robot state decomposition $\mathbf{x}_{\mathrm{T}}$ and $\mathbf{x}_{\mathrm{sb}}$, we use the pose error state $\delta\boldsymbol{\chi}_{\mathrm{T}} := [\delta\mathbf{p}^{T}, \delta\boldsymbol{\alpha}^{T}]^{T}$ and the speed/bias error state $\delta\boldsymbol{\chi}_{\mathrm{sb}} := [\delta\mathbf{v}^{T}, \delta\mathbf{b}_{\mathrm{g}}^{T}, \delta\mathbf{b}_{\mathrm{a}}^{T}]^{T}$.

We treat homogeneous landmarks as (non-unit) quaternions with the minimal perturbation $\delta\boldsymbol{\beta}$, thus $\delta\boldsymbol{\chi}_{\mathrm{L}} := \delta\boldsymbol{\beta}$.

B. Batch Visual SLAM with Inertial Terms

We seek to formulate the visual-inertial localization and mapping problem as one joint optimization of a cost function $J(\mathbf{x})$ containing both the (weighted) reprojection errors $\mathbf{e}_{\mathrm{r}}$ and the temporal error term from the IMU $\mathbf{e}_{\mathrm{s}}$:

$$J(\mathbf{x}) := \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j \in \mathcal{J}(i,k)} \mathbf{e}_{\mathrm{r}}^{i,j,k^T} \mathbf{W}_{\mathrm{r}}^{i,j,k} \mathbf{e}_{\mathrm{r}}^{i,j,k} + \sum_{k=1}^{K-1} \mathbf{e}_{\mathrm{s}}^{k^T} \mathbf{W}_{\mathrm{s}}^{k} \mathbf{e}_{\mathrm{s}}^{k}, \quad (5)$$

where i is the camera index of the assembly, k denotes the camera frame index, and j denotes the landmark index. The indices of landmarks visible in the k^{th} frame and the i^{th} camera are written as the set $\mathcal{J}(i,k)$. Furthermore, $\mathbf{W}_{\mathrm{r}}^{i,j,k}$ represents the information matrix of the respective landmark measurement, and $\mathbf{W}_{\mathrm{s}}^{k}$ the information of the k^{th} IMU error.

Inherently, the purely visual SLAM has 6 Degrees of Freedom (DoF) that need to be held fixed during optimization, i.e. the absolute pose. The combined visual-inertial problem has only 4 DoF, since gravity renders two rotational DoF observable. This complicates fixation. We want to freeze yawing around the gravity direction (world z-axis), as well as the position, typically of the first pose (index k_1). Thus, apart from setting position changes to zero, $\delta\mathbf{p}_{W}^{WS^{k_1}} = \mathbf{0}_{3\times 1}$, we also postulate $\delta\boldsymbol{\alpha}^{k_1} = [\delta\alpha_1^{k_1}, \delta\alpha_2^{k_1}, 0]^{T}$.

In the following, we will present the (standard) reprojection error formulation. Afterwards, an overview on IMU kinematics combined with bias term modeling is given, upon which we base the IMU error term.

1) Reprojection Error Formulation: We use a rather standard formulation of the reprojection error adapted with minor modifications from Furgale [3]:

$$\mathbf{e}_{\mathrm{r}}^{i,j,k} = \mathbf{z}^{i,j,k} - \mathbf{h}_{i}\left(\boldsymbol{T}_{C_iS}\boldsymbol{T}_{SW}^{k}\boldsymbol{l}_{W}^{WL,j}\right). \quad (6)$$

Hereby $\mathbf{h}_{i}(\cdot)$ denotes the camera projection model and $\mathbf{z}^{i,j,k}$ stands for the measurement image coordinates. The error Jacobians with respect to minimal disturbances follow directly from Furgale [3].

2) IMU Kinematics: Under the assumption that the measured effects of the Earth's rotation is small compared to the gyroscope accuracy, we can write the IMU kinematics combined with simple dynamic bias models as:

$$\begin{aligned} \dot{\mathbf{p}}_{W}^{WS} &= \mathbf{v}_{W}^{WS}, \\ \dot{\mathbf{q}}_{WS} &= \frac{1}{2}\boldsymbol{\Omega}\left(\tilde{\boldsymbol{\omega}}_{S}^{WS}, \mathbf{w}_{\mathrm{g}}, \mathbf{b}_{\mathrm{g}}\right)\mathbf{q}_{WS}, \\ \dot{\mathbf{v}}_{W}^{WS} &= \mathbf{C}_{WS}\left(\tilde{\mathbf{a}}_{S}^{WS} + \mathbf{w}_{\mathrm{a}} - \mathbf{b}_{\mathrm{a}}\right) + \mathbf{g}_{W}, \\ \dot{\mathbf{b}}_{\mathrm{g}} &= \mathbf{w}_{b_{\mathrm{g}}}, \\ \dot{\mathbf{b}}_{\mathrm{a}} &= -\frac{1}{\tau}\mathbf{b}_{\mathrm{a}} + \mathbf{w}_{b_{\mathrm{a}}}, \end{aligned} \quad (7)$$

where the elements of $\mathbf{w} := [\mathbf{w}_{\mathrm{g}}^{T}, \mathbf{w}_{\mathrm{a}}^{T}, \mathbf{w}_{b_{\mathrm{g}}}^{T}, \mathbf{w}_{b_{\mathrm{a}}}^{T}]^{T}$ are each uncorrelated zero-mean Gaussian white noise processes. $\tilde{\mathbf{a}}_{S}^{WS}$ are accelerometer measurements and $\mathbf{g}_{W}$ the Earth's gravitational acceleration vector. In contrast to the gyro bias modeled as random walk, we use the time constant $\tau > 0$ to model the accelerometer bias as bounded random walk. The matrix $\boldsymbol{\Omega}$ is formed from the estimated angular rate $\boldsymbol{\omega}_{S}^{WS} = \tilde{\boldsymbol{\omega}}_{S}^{WS} + \mathbf{w}_{\mathrm{g}} - \mathbf{b}_{\mathrm{g}}$, with gyro measurement $\tilde{\boldsymbol{\omega}}_{S}^{WS}$:

$$\boldsymbol{\Omega}\left(\tilde{\boldsymbol{\omega}}_{S}^{WS}, \mathbf{w}_{\mathrm{g}}, \mathbf{b}_{\mathrm{g}}\right) := \begin{bmatrix} -\frac{1}{2}\boldsymbol{\omega}_{S}^{WS} \\ 0 \end{bmatrix}^{\oplus}. \quad (8)$$

The linearized error dynamics take the form

$$\delta\dot{\boldsymbol{\chi}}_R \approx \mathbf{F}_c(\mathbf{x}_\mathrm{R})\delta\boldsymbol{\chi}_R + \mathbf{G}(\mathbf{x}_\mathrm{R})\mathbf{w}, \tag{9}$$

where $\mathbf{G}$ is straight-forward to derive and:

$$\mathbf{F}_c = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{1}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{C}_{WS} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \left[\mathbf{C}_{WS}\left(\tilde{\mathbf{a}}_S^{WS} - \mathbf{b}_\mathrm{a}\right)\right]^\times & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & -\mathbf{C}_{WS} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & -\frac{1}{\tau}\mathbf{1}_3 \end{bmatrix} \tag{10}$$

$(.)^\times$ denoting the skew-symmetric cross-product matrix associated with a vector.

Notice that the equations (7) and (10) can be used the same way as in classical EKF filtering for propagation of the mean ($\hat{\mathbf{x}}_\mathrm{R}$) and covariance ($\mathbf{P}_\mathrm{R}$, in minimal coordinates). For the actual implementation, discrete-time versions of these equations are needed, where the index p denotes the p^th IMU measurement. For considerations of computational complexity, we choose to use the simple Euler-Forward method for integration over a time difference Δt. Analogously, we obtain the discrete-time error state transition matrix as

$$\mathbf{F}_\mathrm{d}(\mathbf{x}_\mathrm{R}, \Delta t) = \mathbf{1}_{15} + \mathbf{F}_\mathrm{c}(\mathbf{x}_\mathrm{R})\Delta t. \tag{11}$$

This results in the covariance propagation equation:

$$\mathbf{P}_\mathrm{R}^{p+1} = \mathbf{F}_\mathrm{d}(\hat{\mathbf{x}}_\mathrm{R}^p, \Delta t)\mathbf{P}_\mathrm{R}^p\mathbf{F}_\mathrm{d}(\hat{\mathbf{x}}_\mathrm{R}^p, \Delta t)^T + \mathbf{G}(\hat{\mathbf{x}}_\mathrm{R}^p)\mathbf{Q}\mathbf{G}(\hat{\mathbf{x}}_\mathrm{R}^p)^T\Delta t, \tag{12}$$

where $\mathbf{Q} := \mathrm{diag}(\sigma_\mathrm{g}^2\mathbf{1}_3, \sigma_\mathrm{a}^2\mathbf{1}_3, \sigma_{b_\mathrm{g}}^2\mathbf{1}_3, \sigma_{b_\mathrm{a}}^2\mathbf{1}_3)$ contains all the noise densities σ_m^2 of the respective processes.

3) Formulation of the IMU Measurement Error Term: Figure 4 illustrates the difference in measurement rates with camera measurements taken at time steps k and $k+1$, as well as faster IMU-measurements that are not synchronized with the camera measurements in general. We need the IMU error term $\mathbf{e}_\mathrm{s}^k(\mathbf{x}_\mathrm{R}^k, \mathbf{x}_\mathrm{R}^{k+1}, \mathbf{z}_\mathrm{s}^k)$ to be a function of robot states at steps k and $k+1$ as well as of all the IMU measurements in-between these time instances (comprising accelerometer and gyro readings) summarized as $\mathbf{z}_\mathrm{s}^k$. Hereby we have to assume an approximate normal conditional probability density f for given robot states at camera measurements k and $k+1$:

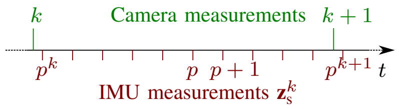

Fig. 4. Different rates of IMU and camera: one IMU term uses all accelerometer and gyro readings between successive camera measurements.

$$f\left(\mathbf{e}_\mathrm{s}^k|\mathbf{x}_\mathrm{R}^k, \mathbf{x}_\mathrm{R}^{k+1}\right) \approx \mathcal{N}\left(\mathbf{0}, \mathbf{R}_\mathrm{s}^k\right). \tag{13}$$

For the state prediction $\hat{\mathbf{x}}_\mathrm{R}^{k+1}\left(\mathbf{x}_\mathrm{R}^k, \mathbf{z}_\mathrm{s}^k\right)$ with associated conditional covariance $\mathbf{P}\left(\delta\hat{\mathbf{x}}_\mathrm{R}^{k+1}|\mathbf{x}_\mathrm{R}^k, \mathbf{z}_\mathrm{s}^k\right)$, the IMU prediction error term can now be written as:

$$\mathbf{e}_\mathrm{s}^k\left(\mathbf{x}_\mathrm{R}^k, \mathbf{x}_\mathrm{R}^{k+1}, \mathbf{z}_\mathrm{s}^k\right) = \begin{bmatrix} \hat{\mathbf{p}}_W^{WS^{k+1}} - \mathbf{p}_W^{WS^{k+1}} \\ 2\left[\hat{\mathbf{q}}_{WS}^{k+1} \otimes \mathbf{q}_{WS}^{k+1^{-1}}\right]_{1:3} \\ \hat{\mathbf{x}}_\mathrm{sb}^{k+1} - \mathbf{x}_\mathrm{sb}^{k+1} \end{bmatrix} \in \mathbb{R}^{15}. \tag{14}$$

This is simply the difference between the prediction based on the previous state and the actual state—except for orientation, where we use a simple multiplicative minimal error.

Next, upon application of the error propagation law, the associated information matrix $\mathbf{W}_\mathrm{s}^k$ is found as:

$$\mathbf{W}_\mathrm{s}^k = \mathbf{R}_\mathrm{s}^{k^{-1}} = \left(\frac{\partial\mathbf{e}_\mathrm{s}^k}{\partial\delta\hat{\boldsymbol{\chi}}_\mathrm{R}^{k+1}}\mathbf{P}\left(\delta\hat{\boldsymbol{\chi}}_\mathrm{R}^{k+1}|\mathbf{x}_\mathrm{R}^k, \mathbf{z}_\mathrm{s}^k\right)\frac{\partial\mathbf{e}_\mathrm{s}^k}{\partial\delta\hat{\boldsymbol{\chi}}_\mathrm{R}^{k+1}}^T\right)^{-1}. \tag{15}$$

The Jacobian $\frac{\partial\mathbf{e}_\mathrm{s}^k}{\partial\delta\hat{\boldsymbol{\chi}}_\mathrm{R}^{k+1}}$ is straightforward to obtain but non-trivial, since the orientation error will be nonzero in general.

Finally, the Jacobians with respect to $\delta\boldsymbol{\chi}_\mathrm{R}^k$ and $\delta\boldsymbol{\chi}_\mathrm{R}^{k+1}$ will be needed for efficient solution of the optimization problem. While differentiating with respect to $\delta\boldsymbol{\chi}_\mathrm{R}^{k+1}$ is straightforward (but non-trivial), some attention is given to the other Jacobian. Recall that the IMU error term (14) is calculated by iteratively applying the prediction. Differentiation with respect to the state $\delta\boldsymbol{\chi}_\mathrm{R}^k$ thus leads to application of the chain rule, yielding

$$\begin{aligned}\frac{\partial\mathbf{e}_\mathrm{s}^k}{\partial\delta\boldsymbol{\chi}_\mathrm{R}^k} =& \mathbf{F}_\mathrm{d}(\mathbf{x}_\mathrm{R}^k, t(p^k) - t(k))\left(\prod_{p=p^k}^{p^{k+1}-1}\mathbf{F}_\mathrm{d}(\hat{\mathbf{x}}_\mathrm{R}^p, \Delta t)\right) \\ & \mathbf{F}_\mathrm{d}(\hat{\mathbf{x}}_\mathrm{R}^{p^{k+1}-1}, t(k+1) - t(p^{k+1}-1))\frac{\partial\mathbf{e}_\mathrm{s}^k}{\partial\delta\hat{\boldsymbol{\chi}}_\mathrm{R}^{k+1}}.\end{aligned} \tag{16}$$

Hereby, $t(.)$ denotes the timestamp of a specific discrete step, and p^k stands for the first IMU sample index after the acquisition of camera frame k.

C. Keypoint Matching and Keyframe Selection

Our processing pipeline employs a customized multi-scale SSE-optimized Harris corner detector combined with BRISK descriptor extraction [12]. The detector enforces uniform keypoint distribution in the image by gradually suppressing corners with weaker score as they are detected at a small distance to a stronger corner. Descriptors are extracted oriented along the gravity direction (projected into the image) which is observable thanks to tight IMU fusion.

Initially, keypoints are stereo-triangulated and inserted into a local map. We perform brute-force matching against all of the map landmarks; outlier rejection is simply performed by applying a chi-square test in image coordinates by using the (uncertain) pose predictions obtained by IMU-integration. There is no costly RANSAC step involved—another advantage of tight IMU involvement. For the subsequent optimization, a bounded set of camera frames is maintained, i.e. poses with associated images taken at that time instant; all landmarks visible in these images are kept in the local map. As illustrated in Figure 5, we distinguish two kinds of frames: we introduce a temporal window of the S most recent frames including the current frame; and we use a number of N keyframes that may have been taken far in the past. For keyframe selection, we use a simple heuristic: if the ratio between the image area spanned by matched points versus the area spanned by all detected points falls below 50 to 60%, the frame is labeled keyframe.

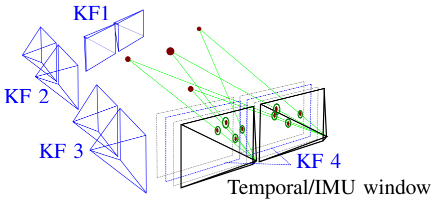

Fig. 5. Frames kept for matching and subsequent optimization.

D. Partial Marginalization

It is not obvious how nonlinear temporal constraints can reside in a bounded optimization window containing keyframes that may be arbitrarily far spaced in time. In the following, we first provide the mathematical foundations for marginalization, i.e. elimination of states in nonlinear optimization, and apply them to visual-inertial SLAM.

1) Mathematical Formulation of Marginalization in Nonlinear Optimization: The Gauss-Newton system of equations is constructed from all the error terms, Jacobians and information: it takes the form $\mathbf{H}\delta\boldsymbol{\chi} = \mathbf{b}$. Let us consider a set of states to be marginalized out, $\mathbf{x}_\mu$, the set of all states related to those by error terms, $\mathbf{x}_\lambda$, and the set of remaining states, $\mathbf{x}_\rho$. Due to conditional independence, we can simplify the marginalization step and only apply it to a sub-problem:

$$\begin{bmatrix} \mathbf{H}_{\mu\mu} & \mathbf{H}_{\mu\lambda_1} \\ \mathbf{H}_{\lambda_1\mu} & \mathbf{H}_{\lambda_1\lambda_1} \end{bmatrix} \begin{bmatrix} \delta\boldsymbol{\chi}_\mu \\ \delta\boldsymbol{\chi}_\lambda \end{bmatrix} = \begin{bmatrix} \mathbf{b}_\mu \\ \mathbf{b}_{\lambda_1} \end{bmatrix} \tag{17}$$

Application of the Schur-Complement operation yields:

$$\mathbf{H}^*_{\lambda_1\lambda_1} := \mathbf{H}_{\lambda_1\lambda_1} - \mathbf{H}_{\lambda_1\mu}\mathbf{H}^{-1}_{\mu\mu}\mathbf{H}_{\mu\lambda_1}, \tag{18a}$$

$$\mathbf{b}^*_{\lambda_1} := \mathbf{b}_{\lambda_1} - \mathbf{H}_{\lambda_1\mu}\mathbf{H}^{-1}_{\mu\mu}\mathbf{b}_\mu, \tag{18b}$$

where $\mathbf{b}^*_{\lambda_1}$ and $\mathbf{H}^*_{\lambda_1\lambda_1}$ are nonlinear functions of $\mathbf{x}_\lambda$ and $\mathbf{x}_\mu$.

The equations in (18) describe a single step of marginalization. In our keyframe-based approach, must apply the marginalization step repeatedly and incorporate the resulting information as a prior in our optimization as our state estimate continues to change. Hence, we fix the linearization point around $\mathbf{x}_0$, the value of $\mathbf{x}$ at the time of marginalization. The finite deviation $\Delta\boldsymbol{\chi} := \Phi^{-1}(\log(\bar{\mathbf{x}} \boxplus \mathbf{x}_0^{-1})))$ represents state updates that occur after marginalization, where $\bar{\mathbf{x}}$ is our current estimate for $\mathbf{x}$. In other words, $\mathbf{x}$ is composed as

$$\mathbf{x} = \exp\left(\Phi(\delta\boldsymbol{\chi})\right) \boxplus \underbrace{\exp\left(\Phi(\Delta\boldsymbol{\chi})\right) \boxplus \mathbf{x}_0}_{=\bar{\mathbf{x}}}. \tag{19}$$

This generic formulation allows us to apply prior information on minimal coordinates to any of our state variables—including unit length quaternions. Introducing $\Delta\boldsymbol{\chi}$ allows the right hand side to be approximated (to first order) as

$$\mathbf{b} + \left.\frac{\partial \mathbf{b}}{\partial \Delta\boldsymbol{\chi}}\right|_{\mathbf{x}_0} \Delta\boldsymbol{\chi} = \mathbf{b} - \mathbf{H}\Delta\boldsymbol{\chi}. \tag{20}$$

Now we can represent the Gauss-Newton system (17) as:

$$\begin{bmatrix} \mathbf{b}_\mu \\ \mathbf{b}_{\lambda_1} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{\mu,0} \\ \mathbf{b}_{\lambda_1,0} \end{bmatrix} - \begin{bmatrix} \mathbf{H}_{\mu\mu} & \mathbf{H}_{\mu\lambda_1} \\ \mathbf{H}_{\lambda_1\mu} & \mathbf{H}_{\lambda_1\lambda_1} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\chi}_\mu \\ \Delta\boldsymbol{\chi}_\lambda \end{bmatrix}. \tag{21}$$

In this form, the right-hand side (18) becomes

$$\mathbf{b}^*_{\lambda_1} = \underbrace{\mathbf{b}_{\lambda_1,0} - \mathbf{H}^T_{\lambda_1\mu}\mathbf{H}^{-1}_{\mu\mu}\mathbf{b}_{\mu,0}}_{\mathbf{b}^*_{\lambda_1,0}} - \mathbf{H}^*_{\lambda_1\lambda_1}\Delta\boldsymbol{\chi}_{\lambda_1}. \tag{22}$$

In the case where marginalized nodes comprise landmarks at infinity (or sufficiently close to infinity), or landmarks visible only in one camera from a single pose, the Hessian blocks associated with those landmarks will be (numerically) rank-deficient. We thus employ the pseudo-inverse $\mathbf{H}^+_{\mu\mu}$, which provides a solution for $\delta\boldsymbol{\chi}_\mu$ given $\delta\boldsymbol{\chi}_\lambda$ with a zero-component into nullspace direction.

The formulation described above introduces a fixed linearization point for both the states that are marginalized $\mathbf{x}_\mu$, as well as the remaining states $\mathbf{x}_\lambda$. This will also be used as as point of reference for all future linearizations of terms involving these states. After application of (18), we can remove the nonlinear terms consumed and add the marginalized $\mathbf{H}^{*,N}_{\lambda_1\lambda_1}$ and $\mathbf{b}^{*,N}_{\lambda_1}$ as summands to construct the overall Gauss-Newton system. The contribution to the chi-square error may be written as $\chi^2_{\lambda_1} = \mathbf{b}^{*T}_{\lambda_1}\mathbf{H}^{*+}_{\lambda_1\lambda_1}\mathbf{b}^*_{\lambda_1}$.

2) Marginalization Applied to Keyframe-Based Visual-Inertial SLAM: The initially marginalized error term is constructed from the first $N+1$ frames $\mathbf{x}^k_{\mathrm{T}}, k = 1, \ldots, N+1$ with respective speed and bias states as visualized graphically in Figure 6. The N first frames will all be interpreted as keyframes and the marginalization step consists of eliminating the corresponding speed and bias states.

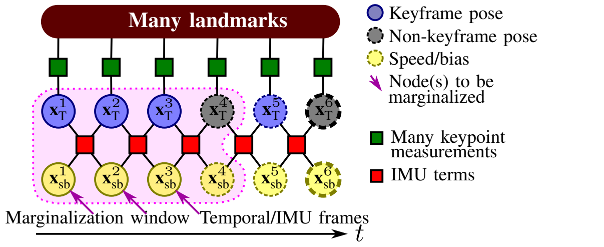
Many landmarks
$\mathbf{x}_{\mathrm{T}}^{1}$
$\mathbf{x}_{\mathrm{T}}^{2}$
$\mathbf{x}_{\mathrm{T}}^{3}$
$\mathbf{x}_{\mathrm{T}}^{4}$
$\mathbf{x}_{\mathrm{T}}^{5}$
$\mathbf{x}_{\mathrm{T}}^{6}$
$\mathbf{x}_{\mathrm{sb}}^{1}$
$\mathbf{x}_{\mathrm{sb}}^{2}$
$\mathbf{x}_{\mathrm{sb}}^{3}$
$\mathbf{x}_{\mathrm{sb}}^{4}$
$\mathbf{x}_{\mathrm{sb}}^{5}$
$\mathbf{x}_{\mathrm{sb}}^{6}$
Marginalization window
Temporal/IMU frames
t
Keyframe pose
Non-keyframe pose
Speed/bias
Node(s) to be marginalized
Many keypoint measurements
IMU terms

Fig. 6. Graph showing the initial marginalization on the first $N+1$ frames.

When a new frame $\mathbf{x}_{\mathrm{T}}^{c}$ (current frame, index c) is inserted into the optimization window, we apply a marginalization operation. In the case where the oldest frame in the temporal window ($\mathbf{x}_{\mathrm{T}}^{c-S}$) is not a keyframe, we will drop all its landmark measurements and then marginalize it out together with the oldest speed and bias states. Figure 7 illustrates this process. Dropping landmark measurements is suboptimal; however, it keeps the problem sparse for fast solution. Visual SLAM with keyframes successfully proceeds analogously, dropping entire frames with their landmark measurements.

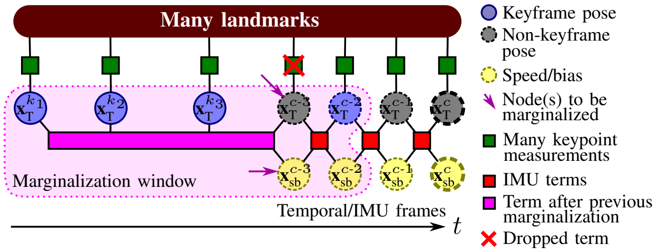
Many landmarks
$\mathbf{x}_{\mathrm{T}}^{k_1}$
$\mathbf{x}_{\mathrm{T}}^{k_2}$
$\mathbf{x}_{\mathrm{T}}^{k_3}$
$\mathbf{x}_{\mathrm{T}}^{c-3}$
$\mathbf{x}_{\mathrm{T}}^{c-2}$
$\mathbf{x}_{\mathrm{T}}^{c-1}$
$\mathbf{x}_{\mathrm{T}}^{c}$
$\mathbf{x}_{\mathrm{sb}}^{c-3}$
$\mathbf{x}_{\mathrm{sb}}^{c-2}$
$\mathbf{x}_{\mathrm{sb}}^{c-1}$
$\mathbf{x}_{\mathrm{sb}}^{c}$
Marginalization window
Temporal/IMU frames
t
Keyframe pose
Non-keyframe pose
Speed/bias
Node(s) to be marginalized
Many keypoint measurements
IMU terms
Term after previous marginalization
Dropped term

Fig. 7. Graph illustration with $N=3$ keyframes and an IMU/temporal node size $S=3$. A regular frame is slipping out of the temporal window.

In the case of $\mathbf{x}_{\mathrm{T}}^{c-S}$ being a keyframe, the information loss of simply dropping all keypoint measurements would be more significant: all relative pose information between the oldest two keyframes encoded in the common landmark observations would be lost. Therefore, we additionally marginalize out the landmarks that are visible in $\mathbf{x}_{\mathrm{T}}^{k_1}$ but not in the most recent keyframe. Figure 8 depicts this procedure graphically. The sparsity of the problem is again preserved.

III. Results

We present experimental results using a custom-built sensor prototype as shown in Figure 1, which provides WVGA stereo images with 14 cm baseline synchronized to the IMU (ADIS16488) measurements. The proposed method runs in *real-time* for all experiments on a standard laptop (2.2 GHz Quad-Core Intel Core i7, 8 Gb RAM). We use g^{2}o [11] as an optimization framework. A precise intrinsic and extrinsic calibration of the camera with respect to the IMU using [4] was available beforehand. The IMU characteristics used (Table I) are slightly more conservative than specified.

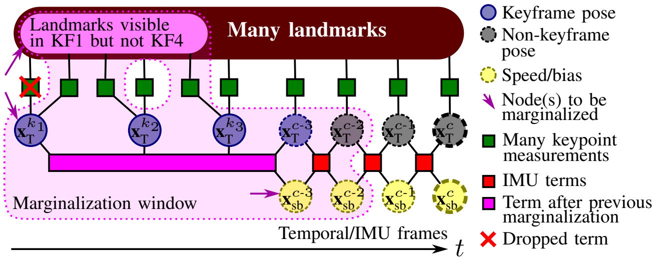
Landmarks visible in KF1 but not KF4
Many landmarks
$\mathbf{x}_{\mathrm{T}}^{k_1}$
$\mathbf{x}_{\mathrm{T}}^{k_2}$
$\mathbf{x}_{\mathrm{T}}^{k_3}$
$\mathbf{x}_{\mathrm{T}}^{c-3}$
$\mathbf{x}_{\mathrm{T}}^{c-2}$
$\mathbf{x}_{\mathrm{T}}^{c-1}$
$\mathbf{x}_{\mathrm{T}}^{c}$
$\mathbf{x}_{\mathrm{sb}}^{c-3}$
$\mathbf{x}_{\mathrm{sb}}^{c-2}$
$\mathbf{x}_{\mathrm{sb}}^{c-1}$
$\mathbf{x}_{\mathrm{sb}}^{c}$
Marginalization window
Temporal/IMU frames
t
Keyframe pose
Non-keyframe pose
Speed/bias
Node(s) to be marginalized
Many keypoint measurements
IMU terms
Term after previous marginalization
Dropped term

Fig. 8. Graph for marginalization of $\mathbf{x}_{\mathrm{T}}^{c-S}$ being a keyframe: the first (oldest) keyframe ($\mathbf{x}_{\mathrm{T}}^{k_1}$) will be marginalized out.

TABLE I
IMU Characteristics

Rate gyros			Accelerometers		
σ_{g}	4.0e-4	$\mathrm{rad}/(\mathrm{s}\sqrt{\mathrm{Hz}})$	σ_{a}	2.0e-3	$\mathrm{m}/(\mathrm{s}^2\sqrt{\mathrm{Hz}})$
$\sigma_{b_{\mathrm{g}}}$	3.0e-3	$\mathrm{rad}/(\mathrm{s}^2\sqrt{\mathrm{Hz}})$	$\sigma_{b_{\mathrm{a}}}$	8.0e-5	$\mathrm{m}/(\mathrm{s}^3\sqrt{\mathrm{Hz}})$
			τ	3600	s

We adopt the evaluation scheme of [5]: for many starting times, the ground truth and estimated trajectories are aligned and the error is evaluated for increasing distances travelled from there. Our *tightly-coupled* algorithm is evaluated against ground truth, *vision-only* and a *loosely-coupled* approach. To ensure that only the estimation algorithms are being compared, we fix the feature correspondences for all algorithms to the ones derived from the *tightly-coupled* approach. The estimates of the *vision-only* approach are then used as input to the *loosely-coupled* baseline algorithm of [20] (with fixed scale and inter-sensor calibration).

A. Vicon: Walking in Circles

The vision-IMU sensor is hand-held while walking in circular loops in a room equipped with a Vicon[1] providing accurate 6D poses at 200 Hz. No loop closures are enforced, yielding exploratory motion of 90 m. Figure 9 illustrates the position and orientation errors in this sequence. The *loosely-coupled* approach mostly helps limiting the orientation error with respect to gravity, which is extremely important for control of aerial systems which the method was designed for. The proposed *tightly-coupled* method produces the smallest error of all, most significantly concerning position.

B. Car: Long Outdoor Trajectory

The sensor was mounted on a car rooftop, simultaneously capturing 6D GPS-INS ground truth using an Applanix POS

[1] http://www.vicon.com/

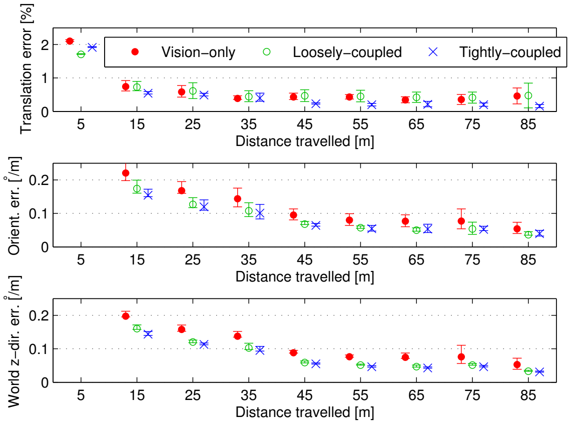

Fig. 9. Comparison with respect to Vicon ground truth. The same keypoint measurements and associations were used in all cases. The 5th and 95th percentiles as well as the means within 10 m bins are shown.

LV at 100 Hz on a trajectory of about 8 km. Figure 10 shows the top view comparison of estimated trajectories with ground truth. Figure 11 provides a quantitative comparison of translation and orientation errors, revealing the clear improvement when using tight fusion of visual and IMU measurements. As expected, the *loosely-coupled* approach exhibits roughly the same performance as the *vision-only* method. This is due to the fact that the former has not been designed to improve the pose estimates over such a long time horizon other than aligning the gravity direction.

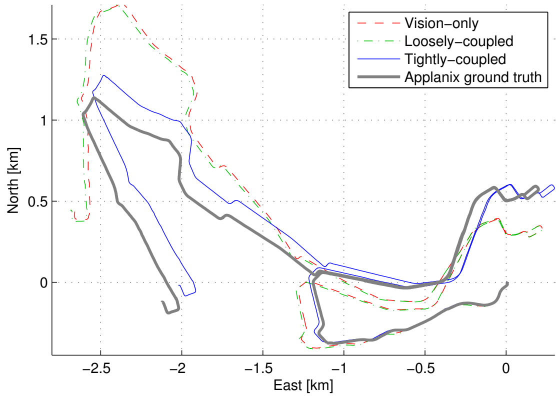

Fig. 10. Car trajectory reconstructions versus Applanix ground truth.

C. *Building: Long Indoor Loop*

As a final experiment, the sensor is hand-held while walking on a long indoor loop spanning 5 floors. As no ground truth is available, we present a qualitative evaluation of the 3D reconstruction of the interior of the building as computed by our method, superimposing the *vision-only* trajectory for comparison. This sequence exhibits challenging lighting and texture conditions while walking through corridors and staircases. The top view plot in Figure 12 demonstrates the applicability of the proposed method in such scenarios with a loop-closure error of 0.6 m, while the error of the *vision-only* baseline reaches 2.2 m.

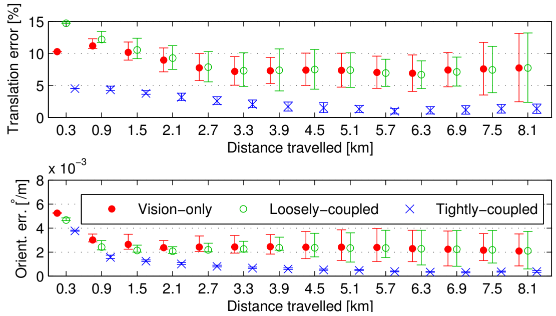

Fig. 11. Quantitative performance evaluation of the estimation approaches with respect to 6D Applanix ground truth.

IV. Conclusion

This paper presents a method of tightly integrating inertial measurements into keyframe-based visual SLAM. The combination of error terms in the non-linear optimization is motivated by error statistics available for both keypoint detection and IMU readings—thus superseding the need for any tuning parameters. Using the proposed approach, we obtain global consistency of the gravity direction and robust outlier rejection employing the IMU kinematics motion model. At the same time, all the benefits of keyframe-based nonlinear optimization are obtained, such as pose keeping in stand-still. Results obtained using a stereo-camera and IMU sensor demonstrate real-time operation of the proposed framework while exhibiting increased accuracy and robustness over vision-only or a loosely coupled approach.

Acknowledgments

The research leading to these results has received funding from the European Commission's Sevenths Framework Programme (FP7/2007-2013) under grant agreement n°285417 (ICARUS), as well as n°600958 (SHERPA) and n°269916 (V-charge). Furthermore, the work was sponsored by Willow Garage and the Swiss CTI project no. 13394.1 PFFLE-NM (Visual-Inertial 3D Navigation and Mapping Sensor). The authors would like to thank Janosch Nikolic, Michael Burri, Jérôme Maye, Simon Lynen and Jörn Rehder from ASL/ETH for their support with hardware, dataset recording, evaluation and calibration.

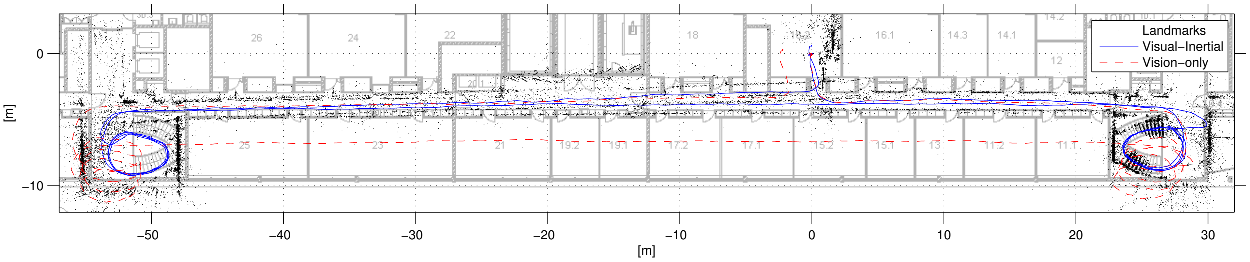

Fig. 12. Ortho-normal top view of the building paths as computed by the different approaches. These are manually aligned with an architectural plan.

References

[1] T. Barfoot, J. R. Forbes, and P. T. Furgale. Pose estimation using linearized rotations and quaternion algebra. *Acta Astronautica*, 68(12):101 – 112, 2011.

[2] T-C. Dong-Si and A. I. Mourikis. Motion tracking with fixed-lag smoothing: Algorithm and consistency analysis. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[3] P. T. Furgale. *Extensions to the Visual Odometry Pipeline for the Exploration of Planetary Surfaces*. PhD thesis, University of Toronto, 2011.

[4] P. T. Furgale, J. Rehder, and R. Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013. To appear.

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis. Towards consistent vision-aided inertial navigation. In *Proc. of the Int'l Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2012.

[7] V. Indelman, S. Williams, M. Kaess, and F. Dellaert. Factor graph based incremental smoothing in inertial navigation systems. In *Information Fusion (FUSION), International Conference on*, 2012.

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

[9] J. Kelly and G. S. Sukhatme. Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. *International Journal of Robotics Research (IJRR)*, 30(1):56–79, 2011.

[10] K. Konolige, M. Agrawal, and J. Sola. Large-scale visual odometry for rough terrain. In *Robotics Research*, pages 201–212. Springer, 2011.

[11] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g^2o: A general framework for graph optimization. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[12] S. Leutenegger, M. Chli, and R.Y. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[13] C. Mei, G. Sibley, M. Cummins, P. M. Newman, and I. D. Reid. Rslam: A system for large-scale mapping in constant-time using stereo. *International Journal of Computer Vision*, pages 198–214, 2011.

[14] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2007.

[15] A. Ranganathan, M. Kaess, and F. Dellaert. Fast 3d pose estimation with out-of-sequence measurements. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007.

[16] G. Sibley, L. Matthies, and G. Sukhatme. Sliding window filter with application to planetary landing. *Journal of Field Robotics*, 27(5):587–608, 2010.

[17] D. Sterlow and S. Singh. Motion estimation from image and intertial measurements. *International Journal of Robotics Research (IJRR)*, 23(12):1157–1195, 2004.

[18] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Real-time monocular SLAM: Why filter? In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[19] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice, September, 1999*, pages 298–372. Springer-Verlag, 1999.

[20] S. Weiss, M.W. Achtelik, S. Lynen, M. Chli, and R. Siegwart. Real-time onboard visual-inertial state estimation and self-calibration of MAVs in unknown environments. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2012.