Visual-Inertial Mapping with Non-Linear Factor Recovery

Vladyslav Usenko[1] Nikolaus Demmel[1] David Schubert[1] Jörg Stückler[2] Daniel Cremers[1]
[1]Technical University of Munich [2]MPI for Intelligent Systems Tübingen

Abstract

Cameras and inertial measurement units are complementary sensors for ego-motion estimation and environment mapping. Their combination makes visual-inertial odometry (VIO) systems more accurate and robust. For globally consistent mapping, however, combining visual and inertial information is not straightforward. To estimate the motion and geometry with a set of images large baselines are required. Because of that, most systems operate on keyframes that have large time intervals between each other. Inertial data on the other hand quickly degrades with the duration of the intervals and after several seconds of integration, it typically contains only little useful information.

In this paper, we propose to extract relevant information for visual-inertial mapping from visual-inertial odometry using non-linear factor recovery. We reconstruct a set of non-linear factors that make an optimal approximation of the information on the trajectory accumulated by VIO. To obtain a globally consistent map we combine these factors with loop-closing constraints using bundle adjustment. The VIO factors make the roll and pitch angles of the global map observable, and improve the robustness and the accuracy of the mapping. In experiments on a public benchmark, we demonstrate superior performance of our method over the state-of-the-art approaches.

1. Introduction

Visual-inertial odometry (VIO) is a popular approach for tracking the motion of a camera in application domains such as robotics or augmented reality. By combining visual and IMU measurements, one can exploit the complementary strengths of both sensors and thereby increase accuracy and robustness. Commonly, the optimization of camera trajectory and map is performed locally on a small window of recent camera frames and IMU measurements. This approach, however, is inevitably prone to drift in the estimates.

Globally consistent optimization for visual-inertial mapping is less explored in the computer vision community. While in principle the optimization could be formulated as bundle adjustment with additional IMU measurements, this approach would quickly become computationally infeasible due to the high number of frames which would lead to a large number of optimization parameters in a naive formulation. To keep the computational burden in bounds, bundle adjustment subsamples the high-frame rate images of the camera to a smaller set of keyframes. The common choice in VIO is to preintegrate IMU measurements between consecutive frames. If we select keyframes temporally far apart to make the optimization efficient, the preintegrated IMU

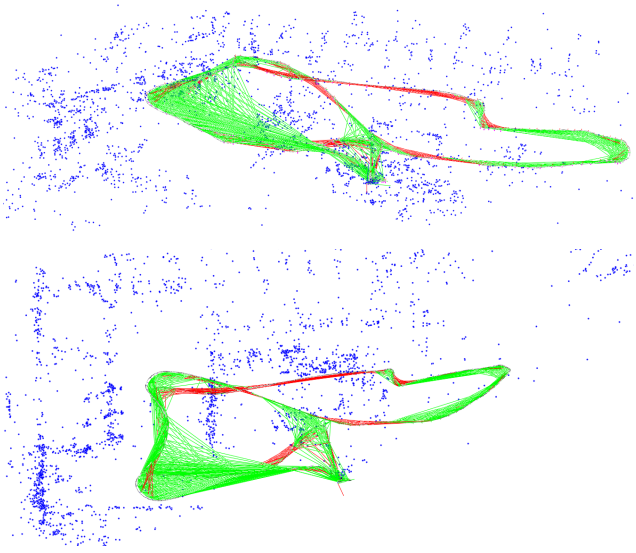


Figure 1: Demonstration of the proposed mapping system on the MH_05 sequence of the EuRoC dataset [3]. Side view (top) and top-down orthographic projection (bottom). Non-linear factors are recovered from the marginalization prior of the VIO and combined with keypoint-based bundle adjustment to achieve a globally consistent, gravity-aligned map. Green lines visualize keyframe connections resulting from bundle adjustment factors and red lines connections from the recovered relative pose factors. Additionally each keyframe has a recovered factor that penalizes deviation from the gravity direction observed in VIO. This approach results in better trajectory estimates compared to approaches that use preintegrated IMU measurements between keyframes.

measurements provide only little information to constrain the trajectory due to the accumulated sensor noise. The small frame rate also affects the quality of the estimated velocities and biases from visual and inertial cues which are required for pose prediction using preintegrated IMU measurements.

We propose a novel approach that formulates visual-inertial mapping as bundle adjustment on a high-frame rate set of visual and inertial measurements. Instead of directly optimizing the camera trajectory for all frames, we propose a hierarchical approach which first recovers a local VIO estimate at the frame rate of the camera. Once keyframes are removed and marginalized from the current local VIO optimization window, we extract non-linear factors [12] that approximate the accumulated visual-inertial information about the camera motion between keyframes. The keyframes and non-linear factors are subsequently used on the global bundle-adjustment layer.

On the VIO layer, our method uses image features designed for fast and accurate tracking, while for the mapping layer we employ distinctive but lighting and viewpoint invariant keypoints that are suitable for loop closing. By this our approach can leverage information from the IMU and short-term visual tracking at high frame rates together with keypoint matching and loop-closing on low frame rates for globally consistent mapping. The factors also help to keep the map gravity-aligned, bridge between frames that do not have enough visual information. Our approach also makes the optimization problem smaller, since we do not have to estimate velocities and biases.

In summary, our contributions are:

- We propose a novel two-layered visual-inertial mapping approach that integrates keypoint-based bundle-adjustment with inertial and short-term visual tracking through non-linear factor recovery.
- As the first layer of our mapping approach we propose a VIO system which outperforms the state-of-the-art methods in terms of trajectory accuracy on the majority of the evaluated sequences.
- Unlike other state-of-the-art systems that use preintegrated IMU measurements, we subsume high-frame rate visual-inertial information in non-linear factors extracted from the marginalization prior of the VIO layer. This results not only in a smaller optimization problem but also in better pose estimates in the resulting gravity aligned map.

We encourage the reader to watch the demonstration video and inspect the open-source implementation of the system, which is available at:

https://vision.in.tum.de/research/vslam/basalt

2. Related Work

Visual-inertial odometry: Early methods for visual-inertial odometry are primarily filter-based [8, 15]. In tightly integrated filters, the prediction step typically propagates the current camera state estimate using the IMU measurements. The state is recursively corrected based on the camera images. A significant drawback in filters is that the linearization point for the non-linear measurement and state transition models cannot be changed, once a measurement is integrated. Fixed-lag smoothers (a.k.a. optimization-based approaches) such as [10, 24, 9] relinearize at the current states in a local optimization window of recent frames. The visual-inertial state estimation is formulated as a full bundle adjustment (BA) over keyframes and IMU measurements. The problem is reduced to a computationally manageable size by marginalization of old frames up to the recent set in the optimization window. The continuous relinearization, windowed optimization and maintenance of the marginalization prior increase the accuracy of the methods. The above methods need to discard keypoints and observations that are observed in marginalized keyframes in order to maintain the sparse structure of the marginalization prior. Hsiung et al. [6] apply non-linear factor recovery to achieve a sparse marginalization prior without discarding information about observed keypoints. This way, the approach can further refine the keypoints and achieve higher accuracy, but in contrast to our work it is limited to local BA.

Visual-inertial mapping: Only few works have tackled globally consistent mapping from visual and inertial measurements. In [16], the authors add inertial measurements to a keyframe-based SLAM system through IMU preintegration. The IMU measurements are preintegrated into a set of pseudo-measurements between keyframes. They notice that the accuracy of preintegrated measurements degrades over time and restrict the time between keyframes to 0.5 seconds in local BA and 3 seconds in global BA. A further shortcoming of the method is its requirement of estimating the camera velocity and IMU biases at each keyframe which is less well constrained through visual measurements than in our approach due to the strong temporal subsampling into keyframes. Schneider et al. [21] follow a similar approach in which preintegrated IMU measurements are inserted into the optimization. The approach in [17] proposes a combination of VIO and 4 degree-of-freedom (DoF) pose optimization for visual-inertial mapping. They fix 2 DoF (roll and pitch) and optimize only for the others. We also constrain roll and pitch from visual-inertial measurements. However, we extract non-linear factors in a probabilistic formulation which account for uncertainties in those values and are traded off with other information in the probabilistic optimization.

3. Preliminaries

In this paper, we write matrices as bold capital letters (e.g. $\mathbf{R}$) and vectors as bold lowercase letters (e.g. $\boldsymbol{\xi}$). Rigid-body poses are represented as $(\mathbf{R}, \mathbf{p}) \in \mathrm{SO}(3) \times \mathbb{R}^3$ or as transformation matrices $\mathbf{T} \in \mathrm{SE}(3)$ when needed. Incrementing a rotation $\mathbf{R}$ by an increment $\boldsymbol{\xi} \in \mathbb{R}^3$ is defined as $\mathbf{R} \oplus \boldsymbol{\xi} = \mathrm{Exp}(\boldsymbol{\xi})\mathbf{R}$. The difference between two rotations $\mathbf{R}_1$ and $\mathbf{R}_2$ is calculated as $\mathbf{R}_1 \ominus \mathbf{R}_2 = \mathrm{Log}(\mathbf{R}_1\mathbf{R}_2^{-1})$ such that $(\mathbf{R} \oplus \boldsymbol{\xi}) \ominus \mathbf{R} = \boldsymbol{\xi}$. Here we use $\mathrm{Exp}\colon \mathbb{R}^3 \to \mathrm{SO}(3)$, which is a composition of the hat operator ($\mathbb{R}^3 \to \mathfrak{so}(3)$) and the matrix exponential ($\mathfrak{so}(3) \to \mathrm{SO}(3)$) and maps rotation vectors to their corresponding rotation matrices, and its inverse $\mathrm{Log}\colon \mathrm{SO}(3) \to \mathbb{R}^3$. For all other variables, such as translation, velocity and biases, we define $\oplus$ and $\ominus$ as regular addition and subtraction. as regular addition.

In the following we will use a state $\mathbf{s}$ that is defined as a tuple of several rotation and vector variables, and a vector-valued function $\mathbf{r}(\mathbf{s})$ that depends on it and can also produce rotations and vectors as the result. An increment $\boldsymbol{\xi} \in \mathbb{R}^n$ is a stacked vector with all the increments of the variables in $\mathbf{s}$. Then, the Jacobian of the function with respect to the increment is defined as

$$\mathbf{J}_{\mathbf{r}(\mathbf{s})} = \lim_{\boldsymbol{\xi} \to \mathbf{0}} \frac{\mathbf{r}(\mathbf{s} \oplus \boldsymbol{\xi}) \ominus \mathbf{r}(\mathbf{s})}{\boldsymbol{\xi}} \,. \tag{1}$$

Here, $\mathbf{s} \oplus \boldsymbol{\xi}$ denotes that each component in $\mathbf{s}$ is incremented with the corresponding segment in $\boldsymbol{\xi}$ using the appropriate definition of the $\oplus$ operator, and similarly for $\ominus$. The limit is done component-wise, such that the Jacobian is a matrix. For Euclidean quantities, this definition is just a normal derivative, with an extension for rotations, both as function value and as function argument.

In non-linear least squares problems, we minimize functions of the form

$$E(\mathbf{s}) = \frac{1}{2}\mathbf{r}(\mathbf{s})^\top \mathbf{W}\mathbf{r}(\mathbf{s}) \,, \tag{2}$$

which is a squared norm of the sum of residuals with block-diagonal weight matrix $\mathbf{W}$. In this case, $\mathbf{r}(\mathbf{s})$ is purely vector-valued. Near the current state $\mathbf{s}$ we can use a linear approximation of the residual, which leads to

$$E(\mathbf{s} \oplus \boldsymbol{\xi}) = E(\mathbf{s}) + \boldsymbol{\xi}^\top \mathbf{J}_{\mathbf{r}(\mathbf{s})}^\top \mathbf{W}\mathbf{r}(\mathbf{s}) + \frac{1}{2}\boldsymbol{\xi}^\top \mathbf{J}_{\mathbf{r}(\mathbf{s})}^\top \mathbf{W}\mathbf{J}_{\mathbf{r}(\mathbf{s})}\boldsymbol{\xi} \,. \tag{3}$$

The optimum of this approximated energy can be attained using the Gauss-Newton increment

$$\boldsymbol{\xi}^* = -(\mathbf{J}_{\mathbf{r}(\mathbf{s})}^\top \mathbf{W}\mathbf{J}_{\mathbf{r}(\mathbf{s})})^{-1}\mathbf{J}_{\mathbf{r}(\mathbf{s})}^\top \mathbf{W}\mathbf{r}(\mathbf{s}) \,. \tag{4}$$

With this, we can iteratively update the state $\mathbf{s}_{i+1} = \mathbf{s}_i \oplus \boldsymbol{\xi}^*$ until convergence.

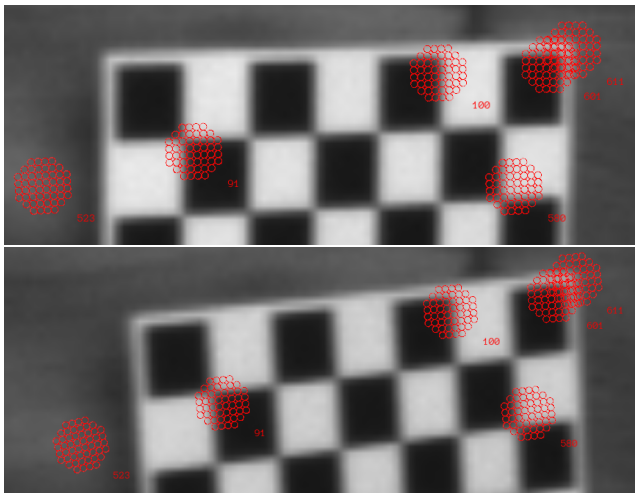


Figure 2: Example of sparse optical flow estimated by our system. Despite changes in exposure time the proposed method is able to estimate the warp in SE(2) between the patches in the images.

4. Visual-Inertial Odometry

We formulate the incremental motion tracking of the camera-IMU setup over time as fixed-lag smoothing. First, we use optical flow to track a sparse set of points in the 2D image plane between consecutive frames. This information is then used in a bundle-adjustment framework which for every frame minimizes an error that consists of point reprojection and IMU propagation terms. To maintain a fixed parameter size of the optimization problem we marginalize out old states. In the remainder of this section we will discuss these stages in more detail.

4.1. Sparse Optical Flow

As a first step of our algorithm we detect a sparse set of keypoints in the frame using the FAST [19] corner detector. To track the motion of these points over a series of consecutive frames we use a sparse optical flow framework [11]. To achieve fast, accurate and robust tracking we combine the inverse-compositional approach as described in [1] with a patch dissimilarity norm that is invariant to intensity scaling. Several authors suggested zero-normalized cross-correlation (ZNCC) for illumination-invariant optical flow [14, 22], but we use locally-scaled sum of squared differences (LSSD) defined in [18] which is computationally less expensive than alternatives.

We formulate the patch tracking problem as estimating the transform $\mathbf{T} \in \mathrm{SE}(2)$ between two corresponding patches in two consecutive frames that minimizes the differences between the patches according to the selected norm. Essentially, we minimize a sum of squared residuals, where

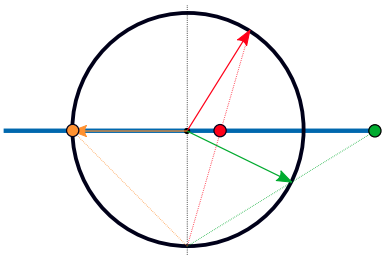


Figure 3: Geometric interpretation of stereographic projection used to represent unit vectors. The two parameters define a point in the XY-plane of the coordinate system shown in blue. To obtain the corresponding 3D unit vector we cast a ray from $(0 \;\; 0 \;\; -1)^\top$ and find an intersection with the unit sphere shown in black. Three example points are visualized in red, green and yellow, with dashed lines representing the rays intersecting with the sphere and arrows showing the resulting unit vectors.

every residual is defined as

$$r_i(\boldsymbol{\xi}) = \frac{I_{t+1}(\mathbf{T}\mathbf{x}_i)}{\overline{I_{t+1}}} - \frac{I_t(\mathbf{x}_i)}{\overline{I_t}} \quad \forall \mathbf{x}_i \in \Omega. \tag{5}$$

Here, $I_t(\mathbf{x})$ and $I_{t+1}(\mathbf{x})$ are the image intensities of images t and $t+1$ at pixel location $\mathbf{x}$. The set of image coordinates that defines the patch is denoted Ω, and the mean intensity of the patch in image t and $t+1$ is $\overline{I_{t+1}}$ and $\overline{I_t}$, respectively. A visualization of the patch and tracking results is shown in Fig. 2.

To achieve robustness to large displacements in the image we use a pyramidal approach, where the patch is first tracked on the coarsest level and then on increasingly finer levels. For outlier filtering, instead of an absolute threshold on the error, we track the patches from the current frame to the target frame and back to check consistency. Points that do not return to the initial location with the second tracking are considered as outliers and discarded.

4.2. Visual-Inertial Bundle Adjustment

To estimate the motion of the camera we combine error terms based on tracked feature locations from sparse optical flow with IMU error terms based on preintegrated IMU measurements [5].

We use the following coordinate frames throughout the paper: W is the world frame, I is the IMU frame and C_i is the frame of camera i, where i is the index of the camera in a stereo setup where applicable. We estimate transformations $\mathbf{T}_{\mathrm{WI}} \in \mathrm{SE}(3)$ from IMU to world coordinate frame. The transformations $\mathbf{T}_{\mathrm{IC}_i}$ from camera frame i to IMU frame and the projection functions π_i are assumed to be static and known from calibration. For the formulation of reprojection errors we denote the transformations from camera i to world by $\mathbf{T}_{\mathrm{WC}_i}$. Those do not constitute additional optimization variables and are calculated using $\mathbf{T}_{\mathrm{WI}}$ and $\mathbf{T}_{\mathrm{IC}_i}$ in practice.

At different points in time, we optimize a state

$$\mathbf{s} = \{\mathbf{s}_\mathrm{k}, \mathbf{s}_\mathrm{f}, \mathbf{s}_\mathrm{l}\}, \tag{6}$$

where $\mathbf{s}_\mathrm{k}$ contains IMU poses for n older keyframes, $\mathbf{s}_\mathrm{f}$ contains IMU poses, velocities and biases of the m most recent frames, which possibly are also keyframes if they host landmarks, and $\mathbf{s}_\mathrm{l}$ contains landmarks. A graphical representation of the problem is shown in Fig. 4 (a). Landmarks are stored relative to the keyframe where they were observed for the first time [13] and defined by a unit-length direction vector in the coordinate frame of the camera and an inverse distance to the landmark [4].

4.2.1 Representation of Unit Vectors in 3D

In order to avoid the necessity of additional constraints for the optimization and to keep the number of optimiziation variables small, we parametrize the bearing vector in 3D space using a minimal representation, which is two-dimensional. In [2] the authors provide an extensive review of possible parametrizations and suggest a new parametrization based on $\mathrm{SO}(3)$ rotations that yields simple derivatives with respect to 2D increments.

In this work we use a parametrization based on stereographic projection that given 2D coordinates $(u, v)^\top$ generates a unit-length bearing vector

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \eta u \\ \eta v \\ \eta - 1 \end{pmatrix}, \quad \eta = \frac{2}{1 + u^2 + v^2}. \tag{7}$$

This parametrization is efficient as it only uses simple operations such as multiplication and division and is defined for all u and v. A geometric interpretation is shown in Fig. 3. The only direction vector that cannot be represented with finite u, v is the negative Z-direction $\begin{pmatrix} 0 & 0 & -1 \end{pmatrix}^\top$, however this is not a drawback in practice, as cameras usually have a limited field of view and cannot see points behind them.

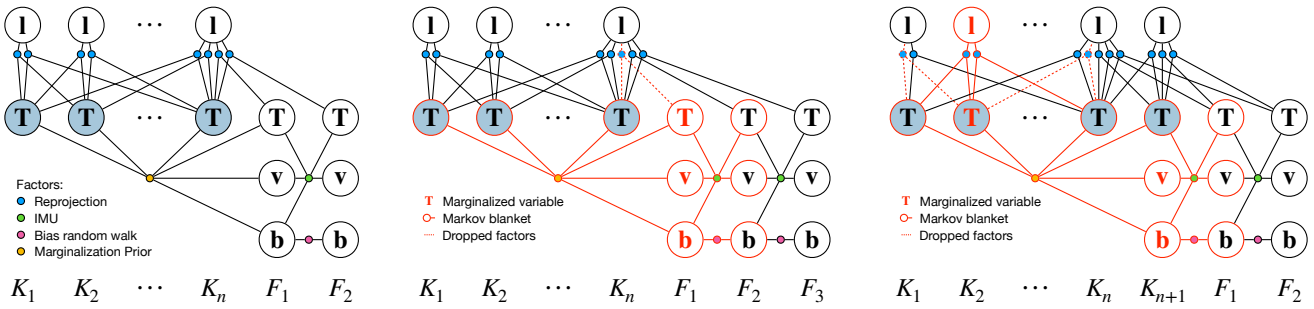
4.2.2 Reprojection Error

The first cue that we can use for motion estimation is the reprojection error. When point i that is hosted in frame $h(i)$ is detected in target frame t at image coordinates $\mathbf{z}_{it}$, the residual is defined as

$$\mathbf{r}_{it} = \mathbf{z}_{it} - \pi_{c(t)}(\mathbf{T}_t^{-1}\mathbf{T}_{h(i)}\mathbf{q}_i(u, v, d)), \tag{8}$$

$$\mathbf{q}_i(u, v, d) = \begin{pmatrix} x(u, v) & y(u, v) & z(u, v) & d \end{pmatrix}^\top, \tag{9}$$

where $c(t)$ is the index of the camera used to take frame t. The pose $\mathbf{T}_t$ denotes $\mathbf{T}_{\mathrm{WC}_{c(t)}}$ at the time when frame t has



(a) (b) (c)

Figure 4: Factor graphs. (a) After marginalizing a frame, the system consists of n older keyframes $K_1 \dots K_n$ and the $m-1$ most recent frames F_1 and F_2 (which could potentially also host landmarks and hence be keyframes). After a new frame has been added, the oldest velocity $\mathbf{v}$ and the oldest bias $\mathbf{b}$ are marginalized. If they do not belong to a keyframe (b), the whole frame including its pose $\mathbf{T}$ is marginalized. If they belong to a keyframe (c), another keyframe is selected for marginalization, including the landmarks hosted in it and its pose. In both cases, reprojection factors where the target frame is the marginalized frame are dropped. In the latter case, reprojection factors from the marginalized frame to F_2 are dropped to allow relinearization. Note that not all possible combinations of host and target frames for reprojection factors are shown.

been taken, and similarly for $\mathbf{T}_{h(i)}$. The first three entries of the homogeneous point coordinates $\mathbf{q}_i(u,v,d)$ are computed from the minimal representation (u,v) as described in Sec. 4.2.1, with an additional fourth entry d, the inverse distance. Since the projection function is independent of scale we do not have to normalize $\mathbf{q}_i$, which makes this formulation numerically stable even when d is close or equal to zero.

4.2.3 IMU Error

The second cue for motion estimation is the IMU data. To deal with the high frequency of IMU measurements we preintegrate several consecutive IMU measurements into a pseudo-measurement. When adding an IMU factor between frame i and frame j, we compute pseudo-measurement $\Delta\mathbf{s} = (\Delta\mathbf{R}, \Delta\mathbf{v}, \Delta\mathbf{p})$ similar to [5]. For this, we compute bias-corrected accelerations $\mathbf{a}_t = \mathbf{a}_t^{\text{raw}} - \bar{\mathbf{b}}_i^{\text{a}}$ and rotational velocities $\boldsymbol{\omega}_t = \boldsymbol{\omega}_t^{\text{raw}} - \bar{\mathbf{b}}_i^{\text{g}}$ using the raw accelerometer $\mathbf{a}_t^{\text{raw}}$ and gyroscope $\boldsymbol{\omega}_t^{\text{raw}}$ measurements. We fix the corresponding biases $\bar{\mathbf{b}}_i^{\text{a}}$ and $\bar{\mathbf{b}}_i^{\text{g}}$ for the entire preintegration time and use linear approximation to account for changes in these variables.

For the timestamp t_i of frame i, we assign the initial state delta $\Delta\mathbf{s}_{t_i} = (\mathbf{I}, \mathbf{0}, \mathbf{0})$. Then, for each IMU timestamp t satisfying $t_i < t \leq t_j$ the following updates are calculated.

$$\Delta\mathbf{R}_{t+1} = \Delta\mathbf{R}_t \text{Exp}(\boldsymbol{\omega}_{t+1}\Delta t)\,, \tag{10}$$
$$\Delta\mathbf{v}_{t+1} = \Delta\mathbf{v}_t + \Delta\mathbf{R}_t\mathbf{a}_{t+1}\Delta t\,, \tag{11}$$
$$\Delta\mathbf{p}_{t+1} = \Delta\mathbf{p}_t + \Delta\mathbf{v}_t\Delta t\,. \tag{12}$$

This defines $\Delta\mathbf{s}_{t+1}$ as a function of $\Delta\mathbf{s}_t$, $\mathbf{a}_{t+1}$, and $\boldsymbol{\omega}_{t+1}$,

$$\Delta\mathbf{s}_{t+1} = f(\Delta\mathbf{s}_t, \mathbf{a}_{t+1}, \boldsymbol{\omega}_{t+1})\,, \tag{13}$$

with corresponding Jacobian $\mathbf{J}_f = [\mathbf{J}_f^{\text{s}}, \mathbf{J}_f^{\text{a}}, \mathbf{J}_f^{\text{g}}]$. Furthermore, all previous iterations of f up to $t+1$ define $\Delta\mathbf{s}_{t+1}$ as a function of the biases,

$$\Delta\mathbf{s}_{t+1} = g_{t+1}(\mathbf{b}_i^{\text{a}}, \mathbf{b}_i^{\text{g}})\,. \tag{14}$$

Starting with zero-initialization, the corresponding Jacobian $\mathbf{J}_{g_{t+1}} = [\mathbf{J}_{g_{t+1}}^{\text{a}}, \mathbf{J}_{g_{t+1}}^{\text{g}}]$ can be computed recursively using $\mathbf{J}_f$,

$$\mathbf{J}_{g_{t+1}}^{\text{a}} = \mathbf{J}_f^{\text{s}}\mathbf{J}_{g_t}^{\text{a}} - \mathbf{J}_f^{\text{a}}\,, \tag{15}$$
$$\mathbf{J}_{g_{t+1}}^{\text{g}} = \mathbf{J}_f^{\text{s}}\mathbf{J}_{g_t}^{\text{g}} - \mathbf{J}_f^{\text{g}}\,, \tag{16}$$

which results from the chain rule. Eventually, the Jacobians of g_{t_j} are denoted $\mathbf{J}^{\text{g}}$ and $\mathbf{J}^{\text{a}}$. Small changes in biases can be represented as increments to the linearization point $\mathbf{b}_i^{\text{a}} = \bar{\mathbf{b}}_i^{\text{a}} + \boldsymbol{\epsilon}^{\text{a}}$ and $\mathbf{b}_i^{\text{g}} = \bar{\mathbf{b}}_i^{\text{g}} + \boldsymbol{\epsilon}^{\text{g}}$. Then, $\Delta\mathbf{s}$ is approximated as

$$\Delta\tilde{\mathbf{s}}(\mathbf{b}_i^{\text{a}}, \mathbf{b}_i^{\text{g}}) = \Delta\mathbf{s}(\bar{\mathbf{b}}_i^{\text{a}}, \bar{\mathbf{b}}_i^{\text{g}}) \oplus (\mathbf{J}^{\text{a}}\boldsymbol{\epsilon}^{\text{a}} + \mathbf{J}^{\text{g}}\boldsymbol{\epsilon}^{\text{g}})\,, \tag{17}$$

with components $\Delta\tilde{\mathbf{s}} = (\Delta\tilde{\mathbf{R}}, \Delta\tilde{\mathbf{v}}, \Delta\tilde{\mathbf{p}})$. The residuals are then calculated as

$$\mathbf{r}_{\Delta\mathbf{R}} = \text{Log}\left(\Delta\tilde{\mathbf{R}}\mathbf{R}_j^\top\mathbf{R}_i\right)\,, \tag{18}$$
$$\mathbf{r}_{\Delta\mathbf{v}} = \mathbf{R}_i^\top(\mathbf{v}_j - \mathbf{v}_i - \mathbf{g}\Delta t) - \Delta\tilde{\mathbf{v}}\,, \tag{19}$$
$$\mathbf{r}_{\Delta\mathbf{p}} = \mathbf{R}_i^\top(\mathbf{p}_j - \mathbf{p}_i - \frac{1}{2}\mathbf{g}\Delta t^2) - \Delta\tilde{\mathbf{p}}\,, \tag{20}$$

where $\mathbf{g}$ is the gravity vector and $\mathbf{R}$ and $\mathbf{p}$ denote the rotation and translation components of $\mathbf{T}_{\text{WI}}$, respectively. These residuals have to be weighted with an appropriate covariance matrix, which can be also calculated recursively. Starting from $\boldsymbol{\Sigma}_{t_i} = \mathbf{0}$, updates are calculated as

$$\boldsymbol{\Sigma}_{t+1} = \mathbf{J}_f^{\text{s}}\boldsymbol{\Sigma}_t{\mathbf{J}_f^{\text{s}}}^\top + \mathbf{J}_f^{\text{a}}\boldsymbol{\Sigma}^{\text{a}}{\mathbf{J}_f^{\text{a}}}^\top + \mathbf{J}_f^{\text{g}}\boldsymbol{\Sigma}^{\text{g}}{\mathbf{J}_f^{\text{g}}}^\top\,, \tag{21}$$

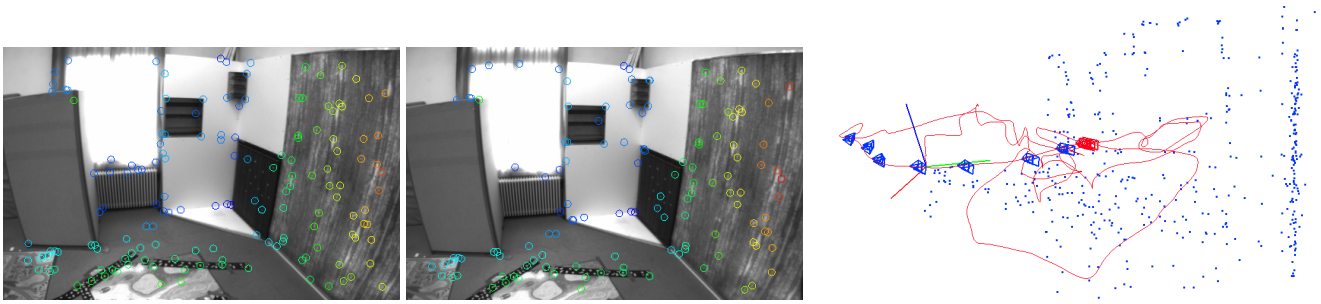


Figure 5: Visual-inertial odometry subsystem proposed in Section 4. Projections of the landmarks with color-coded inverse distance used for estimating the position of the current frame are shown on the left. The results of local visual-inertial bundle adjustment are shown on the right. Keyframe poses with the associated landmarks are visualized in blue, current states and the estimated trajectory are visualized in red. Information about the keyframe poses in the local window is approximated using a set of non-linear factors as described in Section 5 and reused for global mapping.

with diagonal matrices $\mathbf{\Sigma}^{\mathrm{a}}$ and $\mathbf{\Sigma}^{\mathrm{g}}$ that contain the hardware-specific IMU noise parameters for accelerometer and gyroscope. For more detailed information about the underlying physical model of the IMU and preintegration theory we refer the reader to [5].

4.2.4 Optimization and Partial Marginalization

For each new frame we minimize a non-linear energy that consists of reprojection terms, IMU terms and a marginalization prior E_{m}

$$E = \sum_{\substack{i \in \mathcal{P} \\ t \in \mathrm{obs}(i)}} \mathbf{r}_{it}^{\top} \mathbf{\Sigma}_{it}^{-1} \mathbf{r}_{it} + \sum_{(i,j) \in \mathcal{C}} \mathbf{r}_{ij}^{\top} \mathbf{\Sigma}_{ij}^{-1} \mathbf{r}_{ij} + E_{\mathrm{m}}. \tag{22}$$

The reprojection errors are summed over the set of points $\mathcal{P}$ and for each point i over the set $\mathrm{obs}(i)$ of frames where the point is observed, including its host frame. The set $\mathcal{C}$ contains pairs of frames which are connected by IMU factors.

The energy E is optimized using the Gauss-Newton algorithm. To constrain the problem size we fix the number of keyframe poses and consecutive states that we optimize at every iteration. When a new frame is added, there are n pose-only keyframes in $\mathbf{s}_{\mathrm{k}}$ and the m newest frames including the newly added one in $\mathbf{s}_{\mathrm{f}}$. After optimizing, we perform a partial marginalization of the state to prevent the problem size from growing.

Two possible scenarios for marginalization are shown in Fig. 4. In the first one we marginalize out the latest frame that we have in the state. In this case we drop the landmark factors that have this frame as a target to maintain the sparsity of the problem. In the second case we have a new keyframe as the last state, so we marginalize out velocity and biases for this frame and one old keyframe with corresponding landmarks.

In both cases the marginalization is done on the linearized Markov blanket of the variables we want to remove, where the Markov blanket is a collection of incident states to those variables. The linearization $\mathbf{H}$ and $\mathbf{b}$ represent a distribution of the estimated state in the vector space of the increment $\boldsymbol{\xi}$. If we split the increment $\boldsymbol{\xi} = [\boldsymbol{\xi}_{\alpha}^{\top}, \boldsymbol{\xi}_{\beta}^{\top}]^{\top}$ into variables $\boldsymbol{\xi}_{\alpha}$ to stay in the system and variables $\boldsymbol{\xi}_{\beta}$ to be marginalized, we can compute the parameters of the new distribution using the Schur complement,

$$\mathbf{H}_{\alpha\alpha}^{\mathrm{m}} = \mathbf{H}_{\alpha\alpha} - \mathbf{H}_{\alpha\beta} \mathbf{H}_{\beta\beta}^{-1} \mathbf{H}_{\beta\alpha}\,, \tag{23}$$

$$\mathbf{b}_{\alpha}^{\mathrm{m}} = \mathbf{b}_{\alpha} - \mathbf{H}_{\alpha\beta} \mathbf{H}_{\beta\beta}^{-1} \mathbf{b}_{\beta}\,, \tag{24}$$

where we have split the original $\mathbf{H}$ and $\mathbf{b}$ into

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\alpha\alpha} & \mathbf{H}_{\alpha\beta} \\ \mathbf{H}_{\beta\alpha} & \mathbf{H}_{\beta\beta} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_{\alpha} \\ \mathbf{b}_{\beta} \end{bmatrix}. \tag{25}$$

$\mathbf{H}_{\alpha\alpha}^{\mathrm{m}}$ and $\mathbf{b}_{\alpha}^{\mathrm{m}}$ now define an energy term that only depends on $\boldsymbol{\xi}_{\alpha}$ and can be added to the total energy at the next iteration.

We use first-estimate Jacobians [7] to maintain the nullspace properties of the linearized marginalization prior. As soon as a variable becomes a part of the marginalization prior, its linearization point is fixed, and the Jacobian used to calculate $\mathbf{H}$ and $\mathbf{b}$ is evaluated at this linearization point, while the residuals are calculated at the current state estimate. Residuals already in the marginalization term have to be linearly approximated, thus not $\mathbf{b}_{\alpha}^{\mathrm{m}}$, but $\mathbf{b}_{\alpha}^{\mathrm{m}} + \mathbf{H}_{\alpha\alpha}^{\mathrm{m}} \boldsymbol{\delta}_{\alpha}$ is added to the Gauss-Newton optimization once $\boldsymbol{\xi}_{\alpha}$ deviates by $\boldsymbol{\delta}_{\alpha}$ from the state used to calculate the residuals in $\mathbf{b}_{\alpha}^{\mathrm{m}}$.

5. Visual-Inertial Mapping

The fixed-lag smoothing method for visual-inertial odometry (Fig. 5) presented in the previous section accumulates drift in the estimate due to the fixed linearization points outside the optimization window. A typical approach to eliminate such drift is to detect loop closures and incorporate loop-closing constraints into the optimization. We propose a two-layered approach which runs our visual-inertial

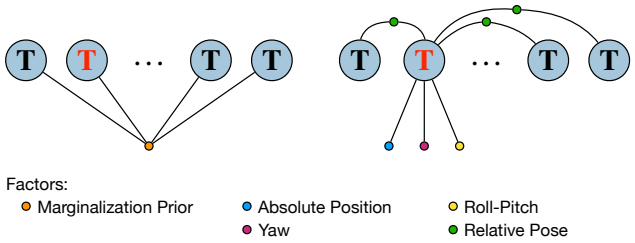


Figure 6: Visualization of non-linear factor recovery. Left: Densely connected factor from marginalization saved from the VIO before removing a keyframe pose. Right: Extracted non-linear factors that approximate the distribution stored in the original factor.

odometry on the lower layer. On the visual-inertial mapping layer, we use non-linear factors that summarize the keyframe pose information from the odometry layer in a bundle-adjustment framework. The BA optimizes the camera poses of keyframes and positions of keypoints. We detect loop closures using the keypoints and add reprojection constraints to the optimization to achieve globally consistent mapping.

5.1. Global Map Optimization

To get statistically independent observations we detect and match ORB [20] features between the keyframes in the global map optimization. This allows us to use the reprojection error function as defined in (8). Combining this reprojection error with the error terms from the recovered non-linear factors yields the objective function:

$$E^{\mathrm{G}}(\mathbf{s}) = \sum_{\substack{i \in \mathcal{P} \\ t \in \mathrm{obs}(i)}} \mathbf{r}_{it}^{\top} \mathbf{\Sigma}_{it}^{-1} \mathbf{r}_{it} + E_{\mathrm{nfr}}(\mathbf{s}), \tag{26}$$

where $E_{\mathrm{nfr}}(\mathbf{s})$ collects the error terms by the recovered non-linear factors. These factors and their recovery are detailed in the following. The state $\mathbf{s}$ that we optimize on this global optimization layer includes the keyframe poses and the positions of the new landmarks (parametrized as in Sec. 4.2.1).

We interface the global map optimization with the VIO layer at the keyframe poses. When a keyframe is marginalized out from the VIO we save the linearization of the Markov blanket (Fig. 4 (c)) and marginalize all other variables except of keyframe poses. From this marginalization prior, we recover a set of non-linear factors on the keyframe poses that approximate the distribution stored in it.

5.2. Non-Linear Factor Recovery

Non-linear factor recovery (NFR [12]) approximates a dense distribution stored in the linearized Markov blanket of the original factor graph with a different set of non-linear factors that yield a sparse factor graph topology. While the initial aim of NFR is to keep the computational complexity of SLAM optimization bounded, we use it to transfer information accumulated during VIO to our globally consistent visual-inertial map optimization.

By linearization of the residual function of a non-linear least squares problem (2), we obtain a multivariate Gaussian distribution $p(\mathbf{s}) \sim N(\boldsymbol{\mu}_{\mathrm{o}}, \mathbf{H}_{\mathrm{o}}^{-1})$ in which the mean $\boldsymbol{\mu}_{\mathrm{o}}$ equals the state estimate. We want to construct another distribution $p_{\mathrm{a}}(\mathbf{s}) \sim N(\boldsymbol{\mu}_{\mathrm{a}}, \mathbf{H}_{\mathrm{a}}^{-1})$ that well approximates the original distribution with a sparser factor graph topology.

We follow NFR [12] and minimize the Kullback-Leibler divergence (KLD) between the recovered distribution and the original distribution. More formally, we minimize

$$D_{\mathrm{KL}}(p(\mathbf{s}) || p_{\mathrm{a}}(\mathbf{s})) = \frac{1}{2} \left(\langle \mathbf{H}_{\mathrm{a}}, \mathbf{\Sigma}_{\mathrm{o}} \rangle - \log\det(\mathbf{H}_{\mathrm{a}} \mathbf{\Sigma}_{\mathrm{o}}) + || \mathbf{H}_{\mathrm{a}}^{\frac{1}{2}} (\boldsymbol{\mu}_{\mathrm{a}} - \boldsymbol{\mu}_{\mathrm{o}}) ||^2 - d \right), \tag{27}$$

where $\mathbf{\Sigma}_{\mathrm{o}} = \mathbf{H}_{\mathrm{o}}^{-1}$ and d is constant.

For the ith non-linear factor that we want to recover, we need to define a residual function such that $\mathbf{r}_i(\mathbf{s}, \mathbf{z}_i) = \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{H}_i^{-1})$. NFR estimates the pseudo measurements $\mathbf{z}_i$ and information matrices $\mathbf{H}_i$ for the factors. Choosing $\mathbf{z}_i$ such that $\mathbf{r}_i(\boldsymbol{\mu}_{\mathrm{o}}, \mathbf{z}_i) = \mathbf{0}$ induces $\boldsymbol{\mu}_{\mathrm{a}} = \boldsymbol{\mu}_{\mathrm{o}}$ which makes the third term of (27) vanish. To estimate $\mathbf{H}_i$ we define

$$\mathbf{J}_{\mathrm{r}} = \begin{bmatrix} \vdots \\ \mathbf{J}_i \\ \vdots \end{bmatrix} \quad \mathbf{H}_{\mathrm{r}} = \begin{bmatrix} \ddots & & 0 \\ & \mathbf{H}_i & \\ 0 & & \ddots \end{bmatrix}, \tag{28}$$

where $\mathbf{J}_{\mathrm{r}}$ stacks the Jacobians of the defined residual functions with respect to the state, and $\mathbf{H}_{\mathrm{r}}$ is a block diagonal matrix that consists of the $\mathbf{H}_i$ for the corresponding residual functions. This allows us to write $\mathbf{H}_{\mathrm{a}} = \mathbf{J}_{\mathrm{r}}^{\top} \mathbf{H}_{\mathrm{r}} \mathbf{J}_{\mathrm{r}}$, and consequently, we can recover the information matrices $\mathbf{H}_i$ by minimizing

$$D_{\mathrm{KL}}(\mathbf{H}_{\mathrm{r}}) = \langle \mathbf{J}_{\mathrm{r}}^{\top} \mathbf{H}_{\mathrm{r}} \mathbf{J}_{\mathrm{r}}, \mathbf{\Sigma}_{\mathrm{o}} \rangle - \log\det(\mathbf{J}_{\mathrm{r}}^{\top} \mathbf{H}_{\mathrm{r}} \mathbf{J}_{\mathrm{r}}). \tag{29}$$

This is an instance of the convex MAXDET problem [25]. For full-rank and invertible $\mathbf{J}_{\mathrm{r}}$, [12, 6] showed that the following closed-form solution exists,

$$\mathbf{H}_i = (\{ \mathbf{J}_{\mathrm{r}} \mathbf{\Sigma}_{\mathrm{o}} \mathbf{J}_{\mathrm{r}}^{\top} \}_i)^{-1}, \tag{30}$$

where $\{\}_i$ denotes the corresponding diagonal block.

5.3. Non-Linear Factors for Distribution Approximation

When we need to marginalize out a keyframe as shown in Fig. 4 (c), we save the current linearization and marginalize out everything except the keyframe poses. This gives us

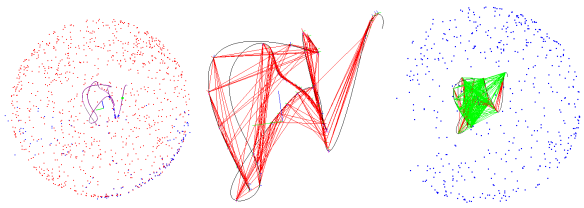


Figure 7: Left: Simulation environment used for the evaluation. Continuous b-spline (black) provides a ground-truth trajectory and IMU data. Ground-truth points (red) are used to obtain point projections. Middle: Measurement with added noise are used for running the VIO system and extract non-linear factors. Right: Extracted factors are combined with a reprojection error for the points (blue) detected in keyframes to get a consistent gravity-aligned map.

a factor that densely connects all keyframe poses in the optimization window. We use it to recover non-linear factors between the marginalized keyframe and all other keyframes as shown in Fig. 6. We define the following residual functions:

$$\mathbf{r}_{\text{rel}}(\mathbf{s}, \mathbf{z}_{\text{rel}}) = \text{Log}(\mathbf{z}_{\text{rel}} \mathbf{T}_j^{-1} \mathbf{T}_i), \tag{31}$$

$$\mathbf{r}_{\text{rp}}(\mathbf{s}, \mathbf{z}_{\text{rp}}) = \lfloor \mathbf{z}_{\text{rp}} \mathbf{R}_i^{-1} (0,0,-1)^\top \rfloor_{xy}, \tag{32}$$

$$\mathbf{r}_{\text{pos}}(\mathbf{s}, \mathbf{z}_{\text{pos}}) = \mathbf{z}_{\text{pos}} - \mathbf{p}_i, \tag{33}$$

$$\mathbf{r}_{\text{yaw}}(\mathbf{s}, \mathbf{z}_{\text{yaw}}) = \lfloor \mathbf{R}_i \mathbf{z}_{\text{yaw}} \rfloor_y, \tag{34}$$

where with $\lfloor \rfloor_{xy}$ we denote x and y components of the vector and with $\mathbf{z}$ we denote the recovered measurements from the estimated state at the time of linearization. In our case $\mathbf{z}_{\text{rel}} = \mathbf{T}_i^{-1}\mathbf{T}_j \in \text{SE}(3)$, $\mathbf{z}_{\text{rp}} = \mathbf{R}_i \in \text{SO}(3)$, $\mathbf{z}_{\text{pos}} = \mathbf{p}_i \in \mathbb{R}^3$ and $\mathbf{z}_{\text{yaw}} = \mathbf{R}_i^{-1} \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^\top \in \mathbb{R}^3$.

We recover pairwise relative-pose factors between the keyframe we will remove and all other current VIO keyframes. For that keyframe we also recover roll-pitch, absolute position and yaw factors (Fig. 6). This gives us a full-rank invertable Jacobian $\mathbf{J}_{\text{r}}$ which means that we can use (30) for recovering information matrices for the factors.

Since yaw and absolute position are 4 unobservable states of the VIO, the only information we have there comes from the initial prior on the start pose. As we do not need this information for the global map we drop yaw and absolute position factors, and only take relative pose and roll-pitch factors for the map optimization. With these factors, the energy terms $E_{\text{nfr}}^{\text{G}}$ become

$$E_{\text{nfr}}^{\text{G}}(\mathbf{s}) = \sum_{(i,j)\in\mathcal{R}} \mathbf{r}_{ij}^\top \mathbf{H}_{ij} \mathbf{r}_{ij} + \sum_{i \in \mathcal{P}} \mathbf{r}_i^\top \mathbf{H}_i \mathbf{r}_i, \tag{35}$$

where $\mathcal{R}$ is a set of all relative pose factors and $\mathcal{P}$ is the set of all roll-pitch factors.

N	VIO	BA + NFR (ours)	BA + IMU preint.
1000	0.0242	**0.0060**	0.0094
500	0.0575	**0.0072**	0.0095
200	0.0538	**0.0168**	0.0203

Table 1: RMS ATE of the estimated trajectory in simulation with N sampled landmarks for visual-inertial odometry, bundle adjustment for all keyframes with recovered factors as proposed in this paper and bundle adjustment with preintegrated IMU measurements.

6. Evaluation

To evaluate the presented approach we conduct evaluation on synthetic data and the EuRoC dataset [3]. We present the evaluation for both our VIO subsystem and our full visual-inertial mapping approach. Our VIO runs the optimization in a local window of frames and provides a pose for every tracked frame, while the mapping system performs global map optimization for keyframes that were selected by the VIO. To measure the accuracy of the evaluated systems, we use the root mean square (RMS) of the absolute trajectory error (ATE) after aligning the estimates with ground truth.

6.1. Simulation

For testing the system in ideal conditions we generate a simulated environment that consists of N random keypoints sampled on a sphere with radius 10 meters and a smooth B-spline that describes the ground-truth motion of the IMU over time (see Fig. 7). From the spline we generate ground-truth poses, velocities, accelerations and artificially add white noise and bias to simulate the IMU measurements. For the vision part, we project the keypoints into the image planes of the cameras and add white noise to simulate the results of the optical flow estimation.

Table 1 summarizes the results of our system on the simulated data. We report RMS ATE for visual-inertial odometry and two global mapping strategies. In the first one, we recover a set of non-linear factors and combine them with bundle-adjustment factors from the points that are detected in the keyframes as described in Section 5. For the second approach we use the method proposed in [16] where instead of non-linear factors, preintegrated IMU measurements are used. In this second setting we estimate not only the pose for each keyframe, but also velocity and biases, which makes the optimization problem 2.5 times larger (15 states instead of 6 for each keyframe). The results suggest that our way of including the IMU information is more accurate in RMS ATE than the competing approach.

Sequence	MH_01	MH_02	MH_03	MH_04	MH_05	V1_01	V1_02	V1_03	V2_01	V2_02
VI DSO [23], mono	**0.06**	**0.04**	0.12	0.13	**0.12**	0.06	0.07	**0.10**	**0.04**	0.06
OKVIS [10] mono	0.34	0.36	0.30	0.48	0.47	0.12	0.16	0.24	0.12	0.22
OKVIS [10] stereo	0.23	0.15	0.23	0.32	0.36	**0.04**	0.08	0.13	0.10	0.17
VINS FUSION [17] mono	0.18	0.09	0.17	0.21	0.25	0.06	0.09	0.18	0.06	0.11
VINS FUSION [17] stereo	0.24	0.18	0.23	0.39	0.19	0.10	0.10	0.11	0.12	0.10
IS VIO [6] stereo	**0.06**	0.06	0.10	0.24	0.19	0.06	0.10	0.26	0.08	0.21
Proposed VIO, stereo	0.07	0.05	**0.06**	**0.12**	**0.12**	0.05	**0.05**	**0.10**	**0.04**	**0.05**
VI SLAM [9] mono, KF	0.25	0.18	0.21	0.30	0.35	0.11	0.13	0.20	0.12	0.20
VI SLAM [9] stereo, KF	0.11	0.09	0.19	0.27	0.23	0.04	0.05	0.11	0.10	0.18
VI ORB-SLAM [16], mono, KF	**0.07**	0.08	0.09	0.22	**0.08**	**0.03**	**0.03**	inf	**0.03**	0.04
Proposed VI Mapping, stereo, KF	0.08	**0.05**	**0.05**	**0.10**	**0.08**	0.04	**0.03**	**0.03**	**0.03**	**0.03**

Table 2: RMS ATE of the estimated trajectory in meters on the EuRoC dataset for several different methods. In the upper part we summarize the results for the VIO methods that run optimization in a local window and estimate the pose of every camera frame. In the lower part we evaluate mapping methods that operate on all keyframes and perform global map optimization. In both evaluations the proposed system shows the lowest error on the majority of the sequences and outperforms the competitors. Note: The V2_03 sequence is excluded from the comparison because it has more than 400 missing frames for one of the cameras.

6.2. EuRoC Dataset

We also evaluate our system on the EuRoC dataset [3] and compare it to other state-of-the-art systems for visual-inertial odometry and mapping. Note that in our definition of a mapping system, the focus is on the global consistency and the absolute accuracy of the recovered keyframe poses in a common world frame. The results of the evaluation are summarized in Table 2. When considering visual-inertial odometry methods our system shows the best performance on seven out of ten sequences while the closest competitor (VI DSO [23]) shows the best results on five.

To evaluate the mapping part we compare it to the visual-inertial version of ORB-SLAM [16], where the vision subsystem is very similar to the one proposed in our mapping layer (ORB keypoints). The main difference lies in the inertial part where ORB-SLAM uses preintegrated measurements between keyframes, while we use recovered non-linear factors that summarize IMU and visual tracking on the VIO layer.

The proposed system clearly outperform ORB-SLAM on the "machine hall" sequences where the large scale of the environment results in large time intervals between keyframes. On the "Vicon room" sequences the difference is smaller, since the rapid motion of the MAV that carries the camera in a small room results in many keyframes with small time intervals between them.

Qualitative results of reconstructed maps are shown in Fig. 1. With the proposed system we are able to reconstruct globally consistent gravity-aligned maps and recover keyframe poses even for segments where no matches between detected ORB features can be estimated.

7. Conclusions

In this paper we present a novel approach for visual-inertial mapping that combines the strengths of highly accurate visual-inertial odometry with globally consistent keyframe-based bundle adjustment. We achieve this in a hierarchical framework that successively recovers non-linear factors from the VIO estimate that summarize the accumulated inertial and visual information between keyframes. VIO is formulated as fixed-lag smoothing which optimizes a set of active recent frames in a sliding window and keeps all previous information in marginalization priors. The accumulated VIO information between keyframes is extracted and retained for the visual-inertial mapping when a keyframe falls outside the window and is marginalized.

Compared to alternative approaches that use preintegrated IMU measurements between keyframes our system shows better trajectory estimates on a public benchmark and in simulation. This formulation has the potential to reduce the computational cost of optimization by reducing the dimensionality of the state space and enable large-scale visual-inertial mapping. Integrating information from other sensor modalities or extending the system for multi-camera setting are the interesting directions for future research.

Acknowledgment

This work was partially supported by the ERC Consolidator Grant "3D Reloaded" and the grant "For3D" by the Bavarian Research Foundation.

References

[1] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Comput. Soc, 2001.

[2] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart. Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *The International Journal of Robotics Research*, 36(10):1053–1072, sep 2017.

[3] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, jan 2016.

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

[5] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. In *Robotics: Science and Systems XI*. Robotics: Science and Systems Foundation, jul 2015.

[6] J. Hsiung, M. Hsiao, E. Westman, R. Valencia, and M. Kaess. Information sparsification in visual-inertial odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, oct 2018.

[7] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis. A first-estimates jacobian EKF for improving SLAM consistency. In *Experimental Robotics*, pages 373–382. Springer Berlin Heidelberg, 2009.

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

[9] A. Kasyanov, F. Engelmann, J. Stückler, and B. Leibe. Keyframe-based visual-inertial online SLAM with relocalization. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, sep 2017.

[10] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, dec 2014.

[11] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[12] M. Mazuran, W. Burgard, and G. D. Tipaldi. Nonlinear factor recovery for long-term SLAM. *The International Journal of Robotics Research*, 35(1-3):50–72, jun 2015.

[13] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid. RSLAM: A system for large-scale mapping in constant-time using stereo. *International Journal of Computer Vision*, 94(2):198–214, jun 2010.

[14] J. Molnár, D. Chetverikov, and S. Fazekas. Illumination-robust variational optical flow using cross-correlation. *Computer Vision and Image Understanding*, 114(10):1104–1114, oct 2010.

[15] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, apr 2007.

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

[17] T. Qin, J. Pan, S. Cao, and S. Shen. A general optimization-based framework for local odometry estimation with multiple sensors. *CoRR*, abs/1901.03638, 2019.

[18] N. Roma, J. Santos-Victor, and J. Tomé. A comparative analysis of cross-correlation matching algorithms using a pyramidal resolution approach. In *Series in Machine Perception and Artificial Intelligence*, pages 117–142. World Scientific, may 2002.

[19] E. Rosten, R. Porter, and T. Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119, jan 2010.

[20] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society.

[21] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart. Maplab: An open framework for research in visual-inertial mapping and localization. *IEEE Robotics and Automation Letters*, 3(3):1418–1425, jul 2018.

[22] F. Steinbrücker, T. Pock, and D. Cremers. Advanced data terms for variational optic flow estimation. In *Proceedings of the Vision, Modeling, and Visualization Workshop (VMV)*, Braunschweig, Germany, 2009.

[23] L. V. Stumberg, V. Usenko, and D. Cremers. Direct sparse visual-inertial odometry using dynamic marginalization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2018.

[24] V. Usenko, J. Engel, J. Stückler, and D. Cremers. Direct visual-inertial odometry with stereo cameras. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2016.

[25] L. Vandenberghe, S. Boyd, and S.-P. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533, apr 1998.