Information Sparsification in Visual-Inertial Odometry

Jerry Hsiung, Ming Hsiao, Eric Westman, Rafael Valencia, and Michael Kaess

***Abstract*— In this paper, we present a novel approach to tightly couple visual and inertial measurements in a fixed-lag visual-inertial odometry (VIO) framework using information sparsification. To bound computational complexity, fixed-lag smoothers typically marginalize out variables, but consequently introduce a densely connected linear prior which significantly deteriorates accuracy and efficiency. Current state-of-the-art approaches account for the issue by selectively discarding measurements and marginalizing additional variables. However, such strategies are sub-optimal from an information-theoretic perspective. Instead, our approach performs a dense marginalization step and preserves the information content of the dense prior. Our method sparsifies the dense prior with a nonlinear factor graph by minimizing the information loss. The resulting factor graph maintains information sparsity, structural similarity, and nonlinearity. To validate our approach, we conduct real-time drone tests and perform comparisons to current state-of-the-art fixed-lag VIO methods in the EuRoC visual-inertial dataset. The experimental results show that the proposed method achieves competitive and superior accuracy in almost all trials. We include a detailed run-time analysis to demonstrate that the proposed algorithm is suitable for real-time applications.**

I. Introduction

State estimation is an essential component for autonomous mobile robot operation. For instance, a robust and accurate state estimator is required for agile control and planning in dynamic and challenging scenarios such as indoor and GPS-denied environments [4]. However, designing a real-time state estimator is nontrivial because of limitations such as computational resources and energy capacity. Therefore, a key research focus in the field of localization is to find efficient ways to fuse various sensor information while providing optimal state estimation.

In recent years, much attention has been given to directly combining cameras and inertial sensors due to the complementary nature of their information [28]. While inertial sensors are responsive in short-term dynamics, cameras provide rich exteroceptive information for long-term navigation. In particular, Visual-Inertial Odometry (VIO) has shown effectiveness in challenging scenarios such as indoor and GPS-denied environments compared to previously existing methods. The ability to navigate indoors is particularly important in applications such as search and rescue, damage inspection, and mapping. In densely populated cities and natural environments like forests, VIO could be used to better aid localization when combined with existing methods that utilize GPS and other global reference points.

This work was supported by Autel Robotics under award number A020215.

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {shsiung, mhsiao, westman, rvalenci, kaess}@andrew.cmu.edu

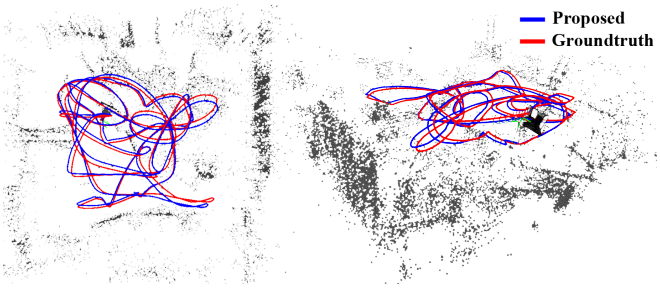


Fig. 1. The trajectories of the proposed method and the groundtruth on EuRoC Vicon Room 2 dataset [3]. The result shows the proposed algorithm achieves highly accurate state estimation in real-time.

In considering efficiency, robustness, and accuracy, typical VIO systems employ either an Extended Kalman Filter (EKF) or a graph-based optimization algorithm to combine inertial information with existing visual odometry methods [4]. While an EKF is known for its efficiency, it is generally less accurate than an optimization approach, which is often computationally expensive. To combine the best of the two, we focus on a fixed-lag smoothing VIO framework, which performs graph optimization on a fixed window of variables in order to bound computational complexity while achieving better accuracy compared to an EKF. However, there are several known drawbacks in a fixed-lag framework. 1) In order to bound computational complexity, a fixed-lag smoother marginalizes out variables, which requires linearizing the system by fixing linearization points. As a consequence, it no longer describes the original nonlinear optimization. 2) This in turn limits the ability of the fixed-lag smoother to converge to the optimal solution at future timesteps because marginalized variables are no longer optimizable. 3) Furthermore, repeating the marginalization process creates a prior that densely connects the remaining variables, which significantly decreases computational efficiency.

To address these shortcomings, we propose a novel information-theoretic approach for a fixed-lag VIO system by utilizing sparsification online. The proposed method maintains the original nonlinear VIO optimization while preserving most of the information and sparse structure. Our main contributions are:

- To the best of our knowledge, this is the first work employing sparsification in the context of fixed-lag VIO to maintain sparsity while minimizing information loss.
- We detail the derivation and design of our sparsification methodology, which retains the sparsity and

nonlinearity of the graphical model in the presence of marginalization.
- We conduct real-world experiments with our software pipeline running on-board an Autel X-Star drone, and provide comparisons of our approach to the current state-of-the-art fixed-lag VIO methods on the EuRoC visual-inertial dataset [3].
- We provide a detailed run-time analysis to demonstrate that the proposed algorithm is suitable for real-time application and suggest ideas for further acceleration of the proposed algorithm.

II. Related Work

VIO algorithms may be roughly categorized into two different types of systems. A *loosely-coupled* system consists of a distinct vision component such as PTAM [25] or DSO [11] to compute visual data as odometry information [43, 12, 26]. The system then combines the odometry data with inertial data to compute the joint solution. In contrast, a *tightly-coupled* system directly incorporates visual and inertial data in a single framework [28, 33, 38, 32], which is shown to be the more accurate approach [28].

Both *loosely-coupled* and *tightly-coupled* systems may also be categorized as either filtering-based [31, 2, 44] or optimization-based [28, 33, 38, 22]. Filtering-based methods are computationally efficient; however, they are known for accumulated linearization errors and inconsistency issues especially in highly nonlinear systems [35]. Huang et al. [18, 19], Li et al. [29], and Hesch et al. [16] propose the First Estimate Jacobian (FEJ) EKF and the Observability Constrained (OC) EKF to limit such issues by enforcing fixed linearization points. Optimization-based methods solve for the optimal estimate by iteratively minimizing the measurement residual. They require more computational resources but achieve higher accuracy. However, the sparse nature of VIO allows existing optimization-based methods to utilize sparsity and apply efficient solvers such as iSAM2 [24], g2o [27], Ceres [1], and SLAM++ [21] to achieve real-time performance.

Our method, like [28, 33], focuses on the tightly-coupled fixed-lag smoothing framework, which combines advantages from both filtering and batch optimization methods [9]. A fixed-lag smoother maintains a bounded computational complexity by fixing the number of target variables in the optimization window while allowing nonlinear optimization to solve for the optimal solution.

To be able to efficiently solve a fixed-lag optimization, existing methods exploit its sparsity in the information form [15, 13, 41]. Sparsity is an important property of a SLAM system [15, 7], which both filtering-based methods and optimization-based methods benefit from. For instance, Eustice et al. [13] and Thrun et al. [37] exploit the sparse structure and develop information filters to efficiently solve the landmark-based SLAM problem. Existing graph solvers [23, 24, 27, 1] exploit sparsity for efficient optimization. In factor graph SLAM, the information matrix specifies the weights and connectivity between variables [7]. However, as the optimization window grows over time, a fixed-lag smoother needs to marginalize variables to maintain a constant computational complexity [45].

Successive marginalizations create "fill-in", additional non-zero entries in the otherwise sparse information matrix, which significantly reduces computational efficiency [15]. To avoid such issue, current state-of-the-art methods such as OKVIS [28] and VINS-MONO [33] 1) selectively discard measurements for sparsity and 2) marginalize additional variables. From an information-theoretic perspective, the information content of the optimization window is reduced and the marginalized variables are no longer optimize-able. The solution to the consecutive optimizations will no longer be optimal with respect to the original problem. As opposed to existing methods, our algorithm addresses the aforementioned issues by incorporating information sparsification to minimize the information loss while maintaining sparsity.

Existing literature in sparsification focuses on the context of large SLAM pose graphs [20, 6, 17, 10]. Wang et al. [42] formulate the sparsification problem by minimizing Kullback-Leibler divergence (KLD) in a laser-based SLAM application. Carlevaris-Bianco et al. [5] propose a generic linear constraint (GLC) which utilizes the Chow-Liu tree to approximate the information of the Markov blanket. Our work follows Mazuran et al.'s Nonlinear Factor Recovery (NFR) [30], which uses specified nonlinear factors to approximate the dense prior by KLD optimization. To our knowledge, our work is the first to demonstrate online sparsification in a fixed-lag VIO framework. We show that our methodology achieves state-of-the-art performance on a public test dataset and is suitable for real-time state estimation.

III. Problem Formulation

At each time w, our fixed-lag smoother optimizes a window of states :

$$\mathcal{X}_w = \{\mathcal{K}_w, \mathcal{F}_w, \mathcal{L}_w\} \tag{1}$$

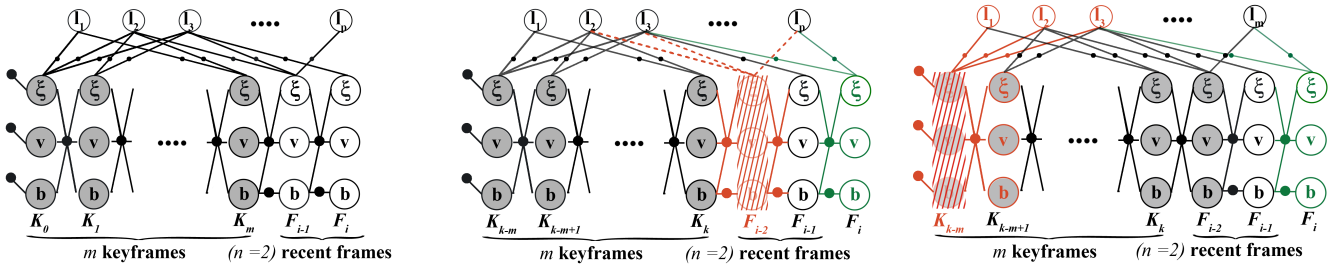
where the set $\mathcal{K} = \{K_1, \ldots, K_m\}$ contains m consecutive keyframes K; $\mathcal{F} = \{F_1, \ldots, F_n\}$ contains n most recent frames F; $\mathcal{L} = \{L_1, \ldots, L_p\}$ contains p landmarks L.

For each frame F_i or keyframe K_i, the IMU state x_i is defined as:

$$x_i = \left[\xi_i^\top, \mathbf{v}_i^\top, \mathbf{b}_i^\top\right]^\top \tag{2}$$

where $\xi \in \mathbb{R}^6$ is the minimum representation of the 3D robot pose, $\mathbf{v} \in \mathbb{R}^3$ the velocity, $\mathbf{b} = \left[\mathbf{b}_a^\top, \mathbf{b}_g^\top\right]^\top \in \mathbb{R}^6$ the IMU accelerometer and gyroscope biases. The measurements $\mathcal{Z}_i$ associated with each F_i or K_i consists of a set of q camera measurements $\mathcal{C}_i = \{c_{i1}, \ldots, c_{iq}\}$ and a *relative* or *marginalized* IMU measurement I_i between two consecutive frames or keyframes respectively. We follow the IMU preintegration method [14] to generate the *relative* IMU measurement. The *marginalized* IMU measurement is detailed in Section IV-A. We define each landmark L_j as a 3D point $\mathbf{l} \in \mathbb{R}^3$ in the world frame.

Using the factor graph formulation, we represent each measurement residual $\mathbf{r}$ as a factor in the graph shown in



(a) (b) (c)

Fig. 2. **(a)**: The proposed fixed-lag VIO factor graph. Each camera frame corresponds to an IMU state $x = [\xi^\top, \mathbf{v}^\top, \mathbf{b}^\top]^\top$, where $\xi \in \mathbb{R}^6$ is the minimum representation of the robot pose. $\mathbf{v} \in \mathbb{R}^3$ is the velocity. $\mathbf{b} \in \mathbb{R}^6$ is the IMU bias. The l's are the visual landmark variables. Measurement factors are represented by solid black circles, including prior factors, IMU factors, and stereo projection factors. **(b)**: Suppose we define the recent frame window size $n = 2$. When a new frame F_i arrives (green), the proposed algorithm looks at frame F_{i-2} being a keyframe (shaded circle), or a regular frame (transparent circle). In this case, F_{i-2} is not a keyframe, the proposed method employs *midframe marginalization* including only the IMU constraints while discarding all visual measurements. **(c)**: If F_{i-2} is a keyframe, the proposed method employs *keyframe marginalization* with sparsification. It includes all variables and measurement information in the Markov blanket (in red).

Fig. 2a. The two main types of factors are IMU preintegration factors and stereo projection factors. The preintegrated IMU factor between x_i and x_{i+1} allows efficient relinearization during optimization. Its residual consists of three terms:

$$\mathbf{r}_{I_i} = \begin{bmatrix} \mathbf{r}_{\Delta\xi_i}^\top & \mathbf{r}_{\Delta\mathbf{v}_i}^\top & \mathbf{r}_{\Delta\mathbf{b}_i}^\top \end{bmatrix}^\top \tag{3}$$

where $\mathbf{r}_{\Delta\xi_i}$ and $\mathbf{r}_{\Delta\mathbf{v}_i}$ and $\mathbf{r}_{\Delta\mathbf{b}_i}$ corresponds to the residuals of pose, velocity, and biases respectively.

Given the states and the measurement residuals, the optimal solution for the factor graph is the *maximum a posteriori* (MAP) estimate according to

$$\begin{aligned} \mathcal{X}_w^\star = \arg\min_{\mathcal{X}_w} \; & \|\mathbf{r}_0\|_{\Sigma_0}^2 + \\ & \sum_{\mathcal{Z}_i \in \{\mathcal{K}_w, \mathcal{F}_w\}} \left(\|\mathbf{r}_{I_i}\|_{\Sigma_{I_i}}^2 + \sum_{c_{ij} \in \mathcal{C}_i} \|\mathbf{r}_{c_{ij}}\|_{\Sigma_{c_{ij}}}^2 \right) \end{aligned} \tag{4}$$

where $\mathbf{r}_0$ represent the prior residual, and the corresponding measurement covariances Σ_0, Σ_{I_i}, and $\Sigma_{c_{ij}}$. To solve the nonlinear SLAM problem, optimizers such as Dogleg and Levenberg-Marquart iterate on the linearized cost of (4) with respect to $\delta\mathcal{X}_w$. At iteration k, the linearized residual of IMU and camera measurements evaluate at the linearization point $\hat{\mathcal{X}}_w^{(k)}$ are in the forms:

$$\begin{aligned} \mathbf{r}_{I_i}(\hat{\mathcal{X}}_\omega^{(k)} + \delta\mathcal{X}_\omega^{(k+1)}) &\approx \mathbf{r}_{I_i}(\hat{\mathcal{X}}_\omega^{(k)}) + H_{I_i}^{(k)} \delta\mathcal{X}_\omega^{(k+1)} \\ \mathbf{r}_{c_{ij}}(\hat{\mathcal{X}}_\omega^{(k)} + \delta\mathcal{X}_\omega^{(k+1)}) &\approx \mathbf{r}_{c_{ij}}(\hat{\mathcal{X}}_\omega^{(k)}) + H_{c_{ij}}^{(k)} \delta\mathcal{X}_\omega^{(k+1)} \end{aligned} \tag{5}$$

where

$$H_{I_i}^{(k)} = \left.\frac{\partial \mathbf{r}_{I_i}}{\partial \mathcal{X}_\omega}\right|_{\mathcal{X}_\omega = \hat{\mathcal{X}}_\omega^{(k)}}, \quad H_{c_{ij}}^{(k)} = \left.\frac{\partial \mathbf{r}_{c_{ij}}}{\partial \mathcal{X}_\omega}\right|_{\mathcal{X}_\omega = \hat{\mathcal{X}}_\omega^{(k)}} \tag{6}$$

are the IMU and camera measurement Jacobians. The optimizer solves for $\delta\mathcal{X}_w^{(k+1)}$ and updates the window as:

$$\hat{\mathcal{X}}_w^{(k+1)} = \hat{\mathcal{X}}_w^{(k)} \oplus \delta\mathcal{X}_w^{(k+1)} \tag{7}$$

The $\oplus$ operator follows vector addition in $\mathbb{R}^n$ and matrix multiplication on Lie manifolds such as SE(3) for poses.

IV. Fixed-Lag VIO With Sparsification

To bound computational complexity, a fixed-lag smoother marginalizes out selected states to maintain a fixed-size optimization window. Marginalization on the Gaussian distribution is typically done by Schur complement on the linearized information matrix $\Lambda_{(\text{MB})}$ of the Markov blanket ($\mathcal{X}_{(\text{MB})}$), which is the collection of state variables incident to the marginalized variables. In Fig. 3a, the red variables and factors show an example of the Markov blanket with respect to the marginalized IMU states of keyframe K_{k-m}.

$\Lambda_{(\text{MB})}$ is constructed by the measurement Jacobian of the factors in the Markov blanket:

$$\Lambda_{(\text{MB})} = H_0^\top \Sigma_0^{-1} H_0^\top + H_{I_i}^\top \Sigma_{I_i}^{-1} H_{I_i} + \sum_{c_{ij} \in \mathcal{C}_i} H_{c_{ij}}^\top \Sigma_{c_{ij}}^{-1} H_{c_{ij}} \tag{8}$$

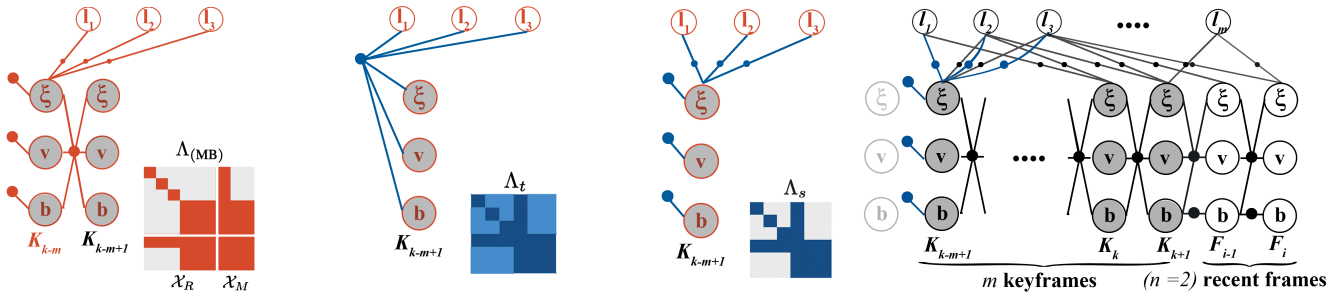
Note that $\Lambda_{(\text{MB})}$ is sparse and its entries correspond to the connectivity in the graph. Define $\mathcal{X}_R \in \mathcal{X}_{(\text{MB})}$ the remaining states, and $\mathcal{X}_M \in \mathcal{X}_{(\text{MB})}$ the marginalized states, we can perform Schur-Complements on $\mathcal{X}_M$:

$$\begin{aligned} \Lambda_{(\text{MB})} &= \begin{bmatrix} \Lambda_{\mathcal{X}_R\mathcal{X}_R} & \Lambda_{\mathcal{X}_M\mathcal{X}_R} \\ \Lambda_{\mathcal{X}_R\mathcal{X}_M} & \Lambda_{\mathcal{X}_M\mathcal{X}_M} \end{bmatrix} \\ \Lambda_{\text{t}} &= \Lambda_{\mathcal{X}_R\mathcal{X}_R} - \Lambda_{\mathcal{X}_R\mathcal{X}_M} \Lambda_{\mathcal{X}_M\mathcal{X}_M}^{-1} \Lambda_{\mathcal{X}_M\mathcal{X}_R} \end{aligned} \tag{9}$$

where Λ_{t} is the target information corresponding to the dense prior. Marginalization degrades the algorithm efficiency as the factor graph loses its sparse structure. To cope with such issue, keyframe-based VIO methods such as OKVIS [28] and VINS-MONO [33] selectively discard measurements to maintain sparsity during marginalization. For landmarks that are not observed by the recent frames, they are marginalized altogether with the marginalized IMU states. It is important to note that such marginalize strategies while maintaining efficiency, potentially lose the capabilities re-estimating the positions of the landmarks and therefore become less accurate. This motivates the main contributions of our work in minimizing information loss during marginalization.

A. Marginalization Strategy

As shown in Fig. 2a, the proposed method maintains n recent frames, and m keyframes. When a new frame



(a) Original Markov blanket (b) Marginalized graph (c) Sparsified graph (d) Re-insertion back to the original graph

Fig. 3. **(a)**: The proposed method first calculates the Markov blanket information $\Lambda_{(\mathrm{MB})}$ from the oldest keyframe. **(b)**: The new target information is then calculated by Schur-Complement. The resulting matrix corresponds to a dense prior factor that connects to every variable in the Markov blanket. **(c)**: Given Λ_t, we employ sparsification with the designed nonlinear factor topology, which we will recover the corresponding information Λ_r for each measurement. **(d)**: The proposed method re-inserts the sparsified topology back to the original fixed-lag window, which retains sparsity and structural similarity.

F_i enters the window, we check whether frame F_{i-n} is a keyframe to select the following *midframe marginalization* or *keyframe marginalization* strategy. In order to enforce consistency, we adopt the method from Dong-Si et al. [9] by using the prior linearization points when corresponding measurement Jacobians are first evaluated.

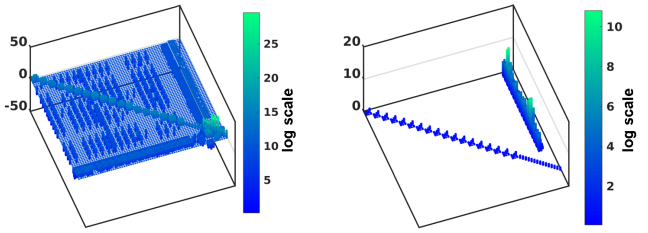
1) Midframe Marginalization: Fig. 2b shows an example of the midframe marginalization strategy. Follow both OKVIS and VINS-MONO on midframe marginalization, we discard all projection factors but only include inertial constraints. This is to keep sparsity but also avoid repeated observations on the landmarks when the robot is stationary. The resulting factor is a *marginalized* IMU measurement that connects to the two corresponding IMU states.

2) Keyframe Marginalization: Fig. 2c shows an example of the keyframe marginalization strategy. If frame F_{i-n} is a keyframe, we perform marginalization on the oldest keyframe at K_{k-m} and landmarks that are only connected to the frame. Unlike existing methods, the rest of the landmarks are preserved during the marginalization step, so that they remain in the optimization window for further nonlinear updates. The result is a dense prior connecting to the next state and all the landmarks defined by the Markov blanket. The blue prior factor in Fig. 3b shows an example connectivity of this prior. In the real system, the associated information Λ_t can be large as shown in Fig. 4a, which significantly reduces computational efficiency. However, keeping landmarks as variables in the optimization window allows further nonlinear updates for subsequent optimizations to reach the optimal solution. To reintroduce sparsity to the graph, our method applies sparsification to the dense prior information.

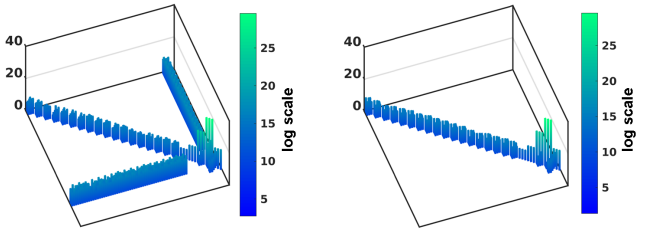
B. Information Sparsification

The dense prior information Λ_t defines a multivariate Gaussian $p(\mathcal{X}_\mathrm{t}) \sim \mathcal{N}(\mu_\mathrm{t}, \Lambda_\mathrm{t})$, with the mean equals to the current linearization point $\mathcal{X}_\mathrm{t}$ of the Markov blanket. We use the global linearization point for the Markov blanket since global priors are included in the marginalization.

Our method first specifies a factor graph topology $\mathcal{T}$ for the Markov blanket, which induces a sparsified distribution $p_\mathrm{s}(\mathcal{X}_\mathrm{t}) \sim \mathcal{N}(\mu_\mathrm{s}, \Lambda_\mathrm{s})$. We follow NFR [30] to recover the



(a) Target information Λ_t (b) Measurement Jacobian H_s by proposed topology



(c) Sparsified information Λ_s (d) Recovered Measurement information Λ_r

Fig. 4. The diagram illustrates the sparsity of the corresponding matrix in pairs of images. **(a)** is the target information from the Markov blanket. **(b)** is the measurement Jacobian matrix corresponding to (13). **(c)** is the sparsified information corresponding to (14). **(d)** is the recovered measurement information corresponding to (13). In each image pair, the left image, the height in the 3D bar graph represent the magnitude of the log absolute value. The corresponding right image shows the informative entries above 10^{-5} threshold.

approximate distribution such that the KLD from $p_\mathrm{s}(\mathcal{X}_\mathrm{t})$ to $p(\mathcal{X}_\mathrm{t})$ is minimized:

$$D_{KL}(p(\mathcal{X}_\mathrm{t}) \| p_\mathrm{s}(\mathcal{X}_\mathrm{t})) = \frac{1}{2}\Big(\langle \Lambda_\mathrm{s}, \Sigma_\mathrm{t} \rangle - \log\det(\Lambda_\mathrm{s}) + \|\Lambda_\mathrm{s}^{\frac{1}{2}}(\mu_\mathrm{s} - \mu_\mathrm{t})\|_2^2 - d\Big) \tag{10}$$

where $\Sigma_\mathrm{t} = \Lambda_\mathrm{t}^{-1}$.

For each factor in $\mathcal{T}$, one must define the sparsified measurement z_s, the sparsified measurement model h_s and the measurement covariance $\Sigma_s = \Lambda_\mathrm{s}^{-1}$, such that $z_s = h_s(\mu_s) + v$, $v \sim \mathcal{N}(0, \Sigma_s)$. First, we set the measurements z_s of each factor to be the expected measurement considering the current state estimate $z_s = h_s(\mu_\mathrm{t})$. This induces the approximate distribution $\mu_s = \mu_t$ which minimizes (10).

TABLE I

ROOT-MEAN-SQUARE ATE (METER) ON THE EUROC DATASET

	MH					V1			V2	
	01_easy	02_easy	03_medium	04_difficult	05_difficult	01_easy	02_medium	03_difficult	01_easy	02_med.
Proposed	**0.059**	**0.060**	**0.099**	0.238	**0.187**	0.060	0.094	0.257	0.080	0.212
OKVIS	0.160	0.106	0.176	**0.208**	0.292	**0.050**	**0.061**	**0.127**	**0.055**	**0.081**
OKVIS (ours)	0.182	0.144	0.278	0.310	0.401	0.272	0.292	0.353	0.153	0.270
VINS-MONO*	0.284	0.237	0.171	0.416	0.308	0.072	0.120	0.159	0.058	0.097
ROVIO	0.354	0.362	0.452	0.919	1.106	0.125	0.160	0.170	0.220	0.392

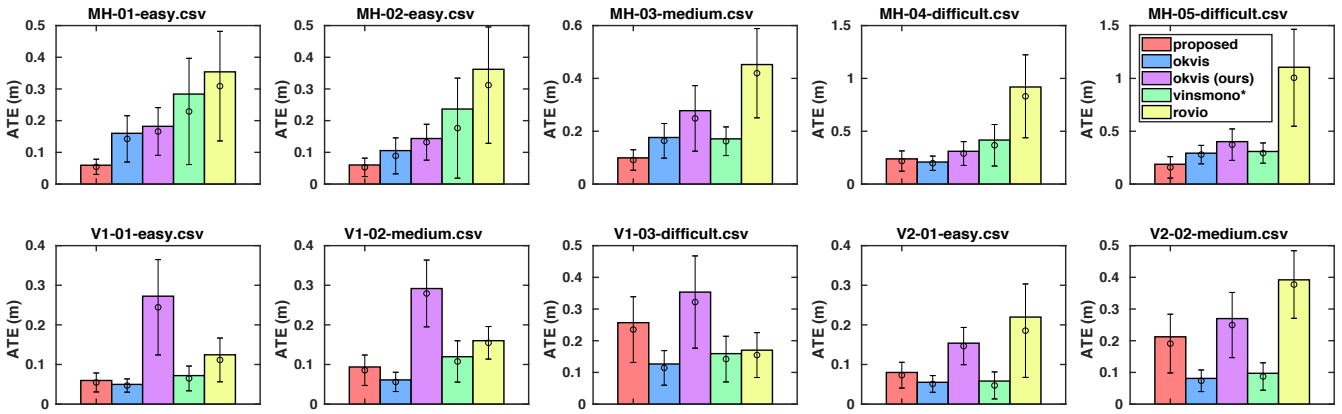


Fig. 5. The diagram shows the comparison of the proposed method against other state-of-the-art algorithms on the EuRoC datasets [3]. Each color represents the result of an algorithm specified by the legend on the top right. The bar value specifies the Root-Mean-Square Error (RMS) of the Absolute Trajectory Error (ATE) metric in meters. Overlaying on each bar, there is an error bar that shows the mean and the standard deviation (std) of the ATE.

Next we look explain the definition of the measurement models. Section IV-C details the method to recover Λ_r for every measurement.

To design the topology, we consider that 1) Λ_s should best approximate Λ_t and 2) $\mathcal{T}$ maintains the sparsity of the graph for future optimizations and 3) $\mathcal{T}$ retains structural similarity to the original graphs. Given the structure of the Markov blanket in our VIO formulation, the most informative entries of Λ_t (see Fig. 4a) are located at the main diagonal blocks and off-diagonal entries corresponding to IMU state and landmarks. Therefore, we have designed the corresponding topology shown in Fig. 3c. The topology consists of independent unary prior factors and binary relative measurement factors. The dense prior information always include the remaining IMU state x_R corresponding to frame K_{k-m+1} and all the landmarks $\mathcal{L}_\mathrm{R} = \{\mathbf{l}_p \in \mathcal{X}_\mathrm{t}\}$.

Denote R_R the rotation, and $\mathbf{p}_\mathrm{R}$ the translation of the pose represented by ξ_R. We design two types of nonlinear topological measurement models to encapsulate the most informative entries in Λ_t. The first is the individual priors for the IMU state x_R,

$$h_\mathrm{r}(\xi_\mathrm{R}) = \xi_\mathrm{R}, \quad h_\mathrm{r}(\mathbf{v}_\mathrm{R}) = \mathbf{v}_\mathrm{R}, \quad h_\mathrm{r}(\mathbf{b}_\mathrm{R}) = \mathbf{b}_\mathrm{R} \tag{11}$$

and the second is the relative pose-to-landmark measurement model

$$h_\mathrm{r}(\xi_\mathrm{R}, \mathbf{l}_p) = R_\mathrm{R}^{-1}(\mathbf{l}_p - \mathbf{p}_\mathrm{R}), \quad \forall \mathbf{l}_p \in \mathcal{X}_\mathrm{t} \tag{12}$$

To construct the sparse information Λ_s of the topology using (11) and (12), we first define H_s and Λ_r as

$$H_\mathrm{s} = \begin{bmatrix} \vdots \\ H_\mathrm{s}^{(j)} \\ \vdots \end{bmatrix}, \quad \Lambda_\mathrm{r} = \begin{bmatrix} \ddots & & \mathbf{0} \\ & \Lambda_\mathrm{r}^{(j)} & \\ \mathbf{0} & & \ddots \end{bmatrix} \tag{13}$$

where $H_\mathrm{s}^{(j)}$ and $\Lambda_\mathrm{r}^{(j)}$ are the Jacobian and the unknown information matrix of the j-th nonlinear topological measurement model. An example of H_s is shown in Fig. 4b. Then Λ_s can be written as

$$\Lambda_\mathrm{s} = H_\mathrm{s}^\top \Lambda_\mathrm{r} H_\mathrm{s} \tag{14}$$

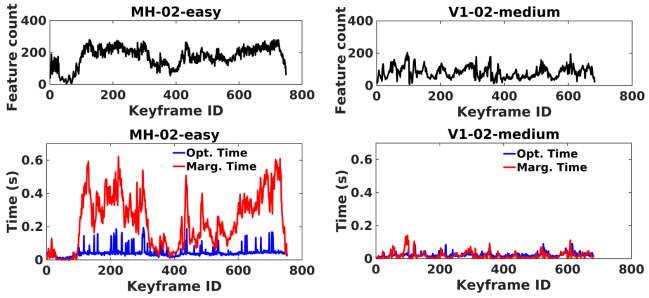
The independent nonlinear topological measurements ensure Λ_r to be block-diagonal, which can be recovered as described in the following section.

C. Topology Measurement Covariances Recovery

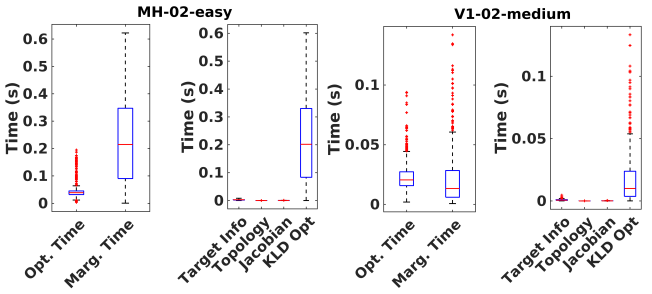
With Λ_t and H_s provided, one can formulate a convex optimization based on KLD to recover the information Λ_r from (13) [30][10]:

$$\begin{aligned} \min \quad & C_{KL} = \langle H_\mathrm{s}^\top \Lambda_\mathrm{r} H_\mathrm{s}, \Sigma_\mathrm{t} \rangle - \log\det(H_\mathrm{s}^\top \Lambda_\mathrm{r} H_\mathrm{s}) \\ \text{s.t.} \quad & \Lambda_\mathrm{r} \succeq 0, \ \Lambda_\mathrm{r} \text{ is block diagonal} \end{aligned} \tag{15}$$

Typically this constrained optimization requires either Interior Point methods (IP) or limited-memory Projected Quasi-Newton (PQN) [34] and the recently proposed Factor Descent Algorithm [39]; however, PQN is only superlinear convergence while IP method requires Hessian. Both methods are costly in terms of computational resources. Because our

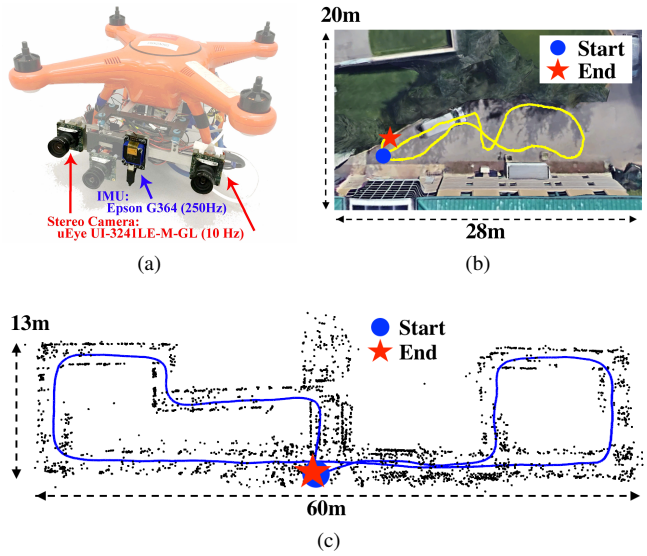


(a) The execution time of proposed algorithm over keyframe ID. The optimization time is in blue, and the marginalization time is in red. **(Left)** MH_02_easy dataset. **(Right)** V1_02_medium dataset



(b) **(Left pair)** The run-time analysis boxplot of the proposed algorithm on MH_02_easy dataset. On the left is the breakdown of total optimization and marginalization time. On the right is the breakdown of time spent on the steps of the marginalization procedure. As shown, the marginalization time is mostly spent on recovering the measurement covariances. **(Right pair)** The same run-time analysis on the V1_02_medium dataset.

Fig. 6. The detailed run-time analysis of EuRoC datasets.



(a) (b) (c)

Fig. 7. **(a)** Our custom built Autel X-Star Premium drone with the visual-inertial payload. It consists of two uEye UI-3241LE-M-GL cameras running at 10Hz and a synchronized Epson G364 IMU running at 250Hz. **(b)** An outdoor data sequence with the proposed algorithm running onboard. The total distance is approximately 48m with the final drift 0.55m (1.11% error). **(c)** A walk through of 4th floor Newel Simon Hall of Carnegie Mellon University. The total distance is approximately 170m, with the final drift about 0.3m (0.17% error)

measurement model always provide a full-rank and invertible Jacobian H_{s}, we are able to solve for (15) in closed-form:

$$\Lambda_{\mathrm{r}}^{(i)} = \left(\{H_{\mathrm{s}}\Sigma_{\mathrm{t}}H_{\mathrm{s}}^{\top}\}^{(i)}\right)^{-1} \tag{16}$$

where $(\cdot)^{(i)}$ denotes the i-th matrix block. The solution from (16) is unique and optimal by the convexity of (15). The proof is shown in [30] by examining the gradient of (15):

$$\begin{aligned}\frac{\partial C_{KL}}{\partial \Lambda_{\mathrm{r}}^{(i)}} &= \left\{H_{\mathrm{s}}[\Sigma_{\mathrm{t}} - (H_{\mathrm{s}}^{\top}\Lambda_{\mathrm{r}}H_{\mathrm{s}})^{-1})]H_{\mathrm{s}}^{T}\right\}^{(i)} \\ &= \left\{H_{\mathrm{s}}\Sigma_{\mathrm{t}}H_{\mathrm{s}}^{T} - H_{\mathrm{s}}H_{\mathrm{s}}^{-1}\Lambda_{\mathrm{r}}^{-1}H_{\mathrm{s}}^{-\top}H_{\mathrm{s}}^{T}\right\}^{(i)} \\ &= \left\{H_{\mathrm{s}}\Sigma_{\mathrm{t}}H_{\mathrm{s}}^{\top} - \Lambda_{\mathrm{r}}^{-1}\right\}^{(i)}\end{aligned} \tag{17}$$

Since (15) is an instance of MAXDET problem [40], the optimal solution is given by the sufficient and necessary condition for (17) to be $\mathbf{0}$.

Fig. 4d shows the recovered sparse information Λ_{r}, where each block corresponds to a nonlinear topological measurement. Our method then replaces the original dense prior Λ_{t} with sparsified topology $\mathcal{T}$ shown in Fig. 3d. The updated smoothing window from (4) now includes the sparsified measurement residuals $\mathbf{r}_{\mathrm{s}}$ with the corresponding covariance $\Sigma_{\mathrm{r}}^{(i)} = \Lambda_{\mathrm{r}}^{(i)-1}$.

V. Experimental Results

A. Implementation

We implement our method in a complete VIO pipeline that includes a visual frontend that matches stereo features, and a backend optimizer. The visual frontend implementation using OpenCV follows the typical pipeline of Shi-Tomashi Corner detector and KLT optical flow tracking for both temporal and stereo images. We implemented a Levenberg-Marquart optimizer and a factor graph based fixed-lag smoother using the GTSAM library [8]. All experiments are run on an Ubuntu desktop with Intel i7-6700 @3.40GHz CPU.

B. Real-time Hardware Test

We have demonstrated our algorithm running real-time onboard using a custom built Autel X-Star Premium drone as shown in Fig. 7a. The visual-inertial payload includes two uEye UI-3241LE-M-GL cameras recording at 10Hz and a synchronized Epson G364 IMU recording at 250Hz. We tested the proposed algorithm outdoors as shown in Fig. 7b. In the outdoor test, we hand-held the drone and traveled a total distance ≈48 meter with the final drift of 0.55 meter (1.11% error). Fig. 7c shows a trajectory walking through the Newell Simon Hall of Carnegie Mellon University. The total length of the sequence is ≈170 meter with the final position drift of 0.3 meter (0.17% error). For visualizations, we have included a link to a video showing the proposed algorithm running onboard with dynamic motions and trajectory plots.

C. Public Test Dataset

We evaluate the proposed method using the EuRoC visual-inertial dataset [3] by the metric of the Absolute Trajectory Error (ATE). ATE indicates the global consistency of the estimated trajectory by comparing the absolute distance to the ground truth [36]. The EuRoC dataset is recorded by a VI sensor with synchronized 20Hz stereo images and

TABLE II
RUN TIME ANALYSIS ON THE EUROC DATASET

		Optimization (unit: s)			Marginalization (unit: s)		
		Mean	RMSE	Std	Mean	RMSE	Std
MH	01_easy	0.054	0.066	0.039	0.248	0.308	0.182
	02_easy	0.043	0.051	0.027	0.230	0.279	0.158
	03_med	0.053	0.064	0.037	0.151	0.211	0.147
	04_diff	0.034	0.042	0.024	0.088	0.129	0.094
	05_diff	0.040	0.048	0.026	0.115	0.163	0.115
V1	01_easy	0.021	0.025	0.016	0.018	0.031	0.026
	02_med	0.023	0.026	0.012	0.020	0.029	0.021
	03_diff	0.021	0.026	0.015	0.015	0.034	0.031
V2	01_easy	0.027	0.033	0.018	0.039	0.062	0.049
	02_med	0.016	0.017	0.006	0.009	0.013	0.010
	03_diff	X	X	X	X	X	X

200Hz IMU data. The dataset consists of three major sets of trajectories, Machine Hall (MH), Vicon Room 1 (V1), and Vicon Room 2 (V2), which vary in smooth and aggressive motions in large and small indoor environment. The V1 and V2 dataset present motion blur and lighting change that produce challenges to the state estimator.

We compare our method against stereo OKVIS [28] and monocular VINS-MONO [33], which are the state-of-the-art fixed-lag VIO systems. The loop-closure and online calibration of VINS-MONO functionality are deactivated to compare pure odometry performance (denote VINS-MONO*). We have included ROVIO [2] to compare fixed-lag VIO approaches with a filtering-based approach. Lastly, to our best knowledge, we implemented OKVIS's marginalization strategy using our GTSAM framework. This is to directly compare the proposed algorithm with OKVIS's marginalization strategy by standardizing the frontend visual module, since OKVIS's frontend module utilizes its backend information for robustness. All results are generated offline in order to ensure the results are the comparison of pure accuracy.

The ATE results are shown in Table. I, and to better visualize we include a bar graph in Fig. 5. The result illustrates our proposed method outperforms the existing methods in four out of five trials in the MH dataset, and achieves comparable results in most of the V1 and V2 datasets. However, in V1_03_difficult and V2_02_medium datasets, the proposed method results in errors significantly larger than those from OKVIS and VINS-MONO. In both cases, the dynamic lighting change has caused the stereo camera images vary in grayscale. Consequently it significantly decreases the performance of our frontend matching algorithm. The sudden loss of visual information has caused a discrepancy in the state estimate. However, prior to the loss of features, our method out performs the existing method. It is important to note that both OKVIS and our proposed algorithm fail in running the V2_03_difficult dataset because of the motion blur. Both our frontend modules have failed to matched stereo features and the state estimate eventually diverges. One idea is to compute both sparse features and direct photometric odometry to handle the blurry images.

D. Run-time Analysis

To demonstrate that our algorithm is appropriate for real-time application, we conduct time profile on the proposed method on the EuRoC dataset with 300 feature cap. The result is shown in Fig. 6 and Table. II. The statistics illustrate that the factor graph optimization maintains around 0.02 to 0.05 second for all the datasets with various difficulties. This is expected as the fixed-lag smoother retains a constant size optimization window. The time spent on sparsification, however, varies across datasets but remain bounded through the sequence. From Fig. 6a, it is shown that the total marginalization time, including sparsification, per frame is correlated with the number of features being optimized. In fact, in the MH_02_easy dataset, our method spends more time in marginalization than in the harder V1_02_medium dataset.

We have also time profiled each step in the sparsification pipeline shown in Fig. 6b. It is important to note that the majority of time is spent on recovering the measurement information using (16). However, there is a significant room for improvement in the implementation, as (16) is clearly highly parallelizable, although we do not take advantage of that here. Furthermore, one can also postpone sparsification step depending on computational resources as shown in [10]. With an improved implementation, more features can be incorporated in the optimization and the required time for marginalization will reduce.

VI. CONCLUSIONS

In this paper, we have introduced a novel fixed-lag smoothing VIO framework with online information sparsification. Compared to the existing methods, the proposed algorithm not only retains sparsity and nonlinearity of the original optimization but also minimizes information loss in the presence of marginalization. Furthermore, we propose a factor graph topology that retains structural similarity of the original fixed-lag window. This allows continuous operation of our algorithm, which is essential for navigating in exploration applications. The proposed method is compared to existing fixed-lag VIO systems and achieves competitive results. Even though the algorithm runs offline, the time analysis has shown its potential for real-time implementation.

For future works, we would like to explore the proposed method with other factor graph topologies. For example, a possible direction is utilizing Chow-Liu Tree by the measure of mutual information between variables in the factor graph.

VII. ACKNOWLEDGMENT

We would like to thank Paloma Sodhi for the extensive theoretical discussions. We would also like to thank the authors of NFR [30] especially Mladen Mazuran, for the discussion, clarification and code base for the NFR.

REFERENCES

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

[2] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[3] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Achtelik, and R. Siegwart, "The EuRoC MAV datasets," *Intl. J. of Robotics Research (IJRR)*, 2015. [Online]. Available: http://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. on Robotics (TRO)*, vol. 32, no. 6, pp. 1309–1332, 2016.

[5] N. Carlevaris-Bianco, M. Kaess, and R. Eustice, "Generic factor-based node removal: Enabling long-term SLAM," *IEEE Trans. on Robotics (TRO)*, vol. 30, no. 6, pp. 1371–1385, Dec. 2014.

[6] S. Choudhary, V. Indelman, H. I. Christensen, and F. Dellaert, "Information-based reduced landmark SLAM," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, vol. 2015-June, no. June, 2015, pp. 4620–4627.

[7] F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, Aug. 2017.

[8] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Tech, Tech. Rep., Sep. 2012. [Online]. Available: https://research.cc.gatech.edu/borg/sites/edu.borg/files/downloads/gtsam.pdf

[9] T. C. Dong-Si and A. I. Mourikis, "Motion tracking with fixed-lag smoothing: Algorithm and consistency analysis," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2011, pp. 5655–5662.

[10] K. Eckenhoff, L. Paull, and G. Huang, "Decoupled, consistent node removal and edge sparsification for graph-based SLAM," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 3275–3282.

[11] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, March 2018.

[12] J. Engel, J. Sturm, and D. Cremers, "Camera-based navigation of a low-cost quadrocopter," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2012, pp. 2815–2821. [Online]. Available: http://ieeexplore.ieee.org/document/6385458/

[13] R. Eustice, M. Walter, and J. Leonard, "Sparse extended information filters: insights into sparsification," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Aug. 2005, pp. 3281–3288.

[14] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Trans. on Robotics (TRO)*, vol. 33, no. 1, pp. 1–21, 2017.

[15] U. Frese, "A proof for the approximate sparsity of SLAM information matrices," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, no. April, 2005, pp. 329–335.

[16] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Consistency analysis and improvement of vision-aided inertial navigation," *IEEE Trans. on Robotics (TRO)*, vol. 30, no. 1, pp. 158–176, 2014.

[17] G. Huang, M. Kaess, and J. Leonard, "Consistent sparsification for graph optimization," in *European Conference on Mobile Robots (ECMR)*, Barcelona, Spain, Sep. 2013, pp. 150–157.

[18] G. P. Huang and S. I. Roumeliotis, "On filter consistency of discrete-time nonlinear systems with partial-state measurements," *Proc. American Control Conference (ACC)*, pp. 5468–5475, 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6580693

[19] G. P. Huang, A. I. Mourikis, and S. I. Roumeliotis, "Observability-based rules for designing consistent EKF SLAM estimators," *Intl. J. of Robotics Research (IJRR)*, vol. 29, no. 5, pp. 502–528, 2010.

[20] V. Ila, J. M. Porta, and J. Andrade-Cetto, "Information-based compact pose SLAM," *IEEE Trans. on Robotics (TRO)*, vol. 26, no. 1, pp. 78–93, 2010.

[21] V. Ila, L. Polok, M. Solony, and P. Svoboda, "Slam++-a highly efficient and temporally scalable incremental slam framework," *Intl. J. of Robotics Research (IJRR)*, vol. 36, no. 2, pp. 210–230, 2017. [Online]. Available: https://doi.org/10.1177/0278364917691110

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

[23] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. on Robotics (TRO)*, vol. 24, no. 6, pp. 1365–1378, Dec. 2008.

[24] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Intl. J. of Robotics Research (IJRR)*, vol. 31, no. 2, pp. 216–235, Feb. 2012.

[25] G. Klein and D. Murray, "Parallel tracking and mapping on a camera phone," *International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 83–86, 2009.

[26] K. Konolige, M. Agrawal, and J. Sola, "Large scale visual odometry for rough terrain," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2007, pp. 201–212.

[27] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[28] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visualinertial odometry using nonlinear optimization," *Intl. J. of Robotics Research (IJRR)*, vol. 34, no. 3, pp. 314–334, 2015. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364914554813

[29] M. Li and A. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *Intl. J. of Robotics Research (IJRR)*, vol. 32, no. 6, pp. 690–711, 2013.

[30] M. Mazuran, W. Burgard, and G. D. Tipaldi, "Nonlinear factor recovery for long-term SLAM," *Intl. J. of Robotics Research (IJRR)*, vol. 35, no. 1-3, pp. 50–72, 2016. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364915581629

[31] A. I. Mourikis and S. I. Roumeliotis, "A multi-state Kalman filter for vision-aided inertial navigation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, no. April, 2007, pp. 10–14.

[32] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters (RA-L)*, vol. 2, no. 2, pp. 796–803, 2017.

[33] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *arXiv preprint arXiv:1708.03852*, 2017.

[34] M. Schmidt, E. van den Berg, M. P. Friedlander, and K. Murphy, "Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm," in *Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida, April 2009, pp. 456–463.

[35] M. A. Skoglund, G. Hendeby, and D. Axehill, "Extended Kalman filter modifications based on an optimization view point," in *Intl. Conf. on Information Fusion (FUSION)*, July 2015, pp. 1856–1861.

[36] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 573–580, 2012.

[37] S. Thrun, Y. Liu, D. Koller, A. Y. Ng, Z. Ghahramani, and H. Durrant-Whyte, "Simultaneous localization and mapping with sparse extended information filters," *Intl. J. of Robotics Research (IJRR)*, vol. 23, no. 7-8, pp. 693–716, 2004. [Online]. Available: https://doi.org/10.1177/0278364904045479

[38] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2016, pp. 1885–1892.

[39] J. Vallv, J. Sol, and J. Andrade-Cetto, "Graph slam sparsification with populated topologies using factor descent optimization," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 2, pp. 1322–1329, April 2018.

[40] L. Vandenberghe, S. Boyd, and S.-P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 2, pp. 499–533, 1998. [Online]. Available: https://doi.org/10.1137/S0895479896303430

[41] M. R. Walter, R. M. Eustice, and J. J. Leonard, "Exactly sparse extended information filters for feature-based SLAM," *Intl. J. of Robotics Research (IJRR)*, vol. 26, no. 4, pp. 335–359, 2007.

[42] Y. Wang, R. Xiong, Q. Li, and S. Huang, "Kullback-Leibler divergence based graph pruning in robotic feature mapping," in *European Conference on Mobile Robots (ECMR)*, Sept 2013, pp. 32–37.

[43] 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 *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2012, pp. 957–964.

[44] K. J. Wu, A. M. Ahmed, G. A. Georgiou, and S. I. Roumeliotis, "A square root inverse filter for efficient vision-aided inertial navigation on mobile devices," *Robotics: Science and Systems (RSS)*, 2015.

[45] Y. Yang, J. Maley, and G. Huang, "Null-space-based marginalization: Analysis and algorithm," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, no. October, 2017, pp. 6749–6755. [Online]. Available: http://ieeexplore.ieee.org/document/8206592/