

# An Efficient Schmidt-EKF for 3D Visual-Inertial SLAM

Patrick Geneva, James Maley, Guoquan Huang  
University of Delaware

pgeneva@udel.edu, james.m.maley2@udel.edu, ghuang@udel.edu

## Abstract

*It holds great implications for practical applications to enable centimeter-accuracy positioning for mobile and wearable sensor systems. In this paper, we propose a novel, high-precision, efficient visual-inertial (VI)-SLAM algorithm, termed Schmidt-EKF VI-SLAM (SEVIS), which optimally fuses IMU measurements and monocular images in a tightly-coupled manner to provide 3D motion tracking with bounded error. In particular, we adapt the Schmidt Kalman filter formulation to selectively include informative features in the state vector while treating them as nuisance parameters (or Schmidt states) once they become matured. This change in modeling allows for significant computational savings by no longer needing to constantly update the Schmidt states (or their covariance), while still allowing the EKF to correctly account for their cross-correlations with the active states. As a result, we achieve linear computational complexity in terms of map size, instead of quadratic as in the standard SLAM systems. In order to fully exploit the map information to bound navigation drifts, we advocate efficient keyframe-aided 2D-to-2D feature matching to find reliable correspondences between current 2D visual measurements and 3D map features. The proposed SEVIS is extensively validated in both simulations and experiments.*

## 1. Introduction

Enabling centimeter-accuracy positioning for mobile and wearable devices such as smart phones and micro air vehicles (MAVs), holds potentially huge implications for practical applications. One of the most promising methods providing precision navigation in 3D is through the fusion of visual and inertial sensor measurements (i.e., visual-inertial navigation systems or VINS) [30, 13, 22, 21, 17, 15]. This localization solution has the advantages of being both cheap and ubiquitous, and has the potential to provide position and orientation (pose) estimates which are on-par in terms of accuracy with more expensive sensors such as LiDAR. To date, various algorithms are available for VINS problems including visual-inertial (VI)-SLAM [19, 45] and

visual-inertial odometry (VIO) [30, 29, 22], such as the extended Kalman filter (EKF) [30, 20, 14, 22, 17, 16, 50, 37], unscented Kalman filter (UKF) [10, 4], and batch or sliding-window optimization methods [46, 18, 21, 33, 52, 45, 40], among which the EKF-based approaches remain arguably the most popular for resource constrained devices because of their efficiency. While current approaches can perform well over a short period of time in a small-scale environment (e.g., see [13, 22, 15]), they are not robust and accurate enough for long-term, large-scale deployments in challenging environments, due to their limited available resources of sensing, memory and computation, which, if not properly addressed, often result in short mission duration or intractable real-time estimator performance.

In this paper, we will primarily focus on EKF-based VI-SLAM rather than VIO. VI-SLAM has the advantage of building a map of the surrounding environment, which enables “loop closing” to bound long-term navigation drift. VIO systems do not build a map and therefore cannot leverage information from prior observations to help improve estimator performance. However, one of the largest downsides of SLAM is that its computational complexity grows quadratically with the number of landmarks in the map, which commonly makes it computationally intractable without simplifying assumptions to allow for them to run on resource constrained sensor platforms such as mobile devices. To address this complexity issue, we leverage the computationally-efficient multi-state constraint Kalman filter (MSCKF) [30] and selectively keep a number of features (say  $n$ ) in the state vector as a map of the environment, enabling the system to use them for a long period of time and thus allowing for (implicit) loop closures to bound drifts. This, however, would still exhibit  $O(n^2)$  computational complexity as in the standard EKF-based SLAM. By observing that features’ estimates do not have significant updates if they approach their steady state (i.e., becoming matured/converged), we could gain substantial computational savings by avoiding performing EKF updates for those matured map features while still taking into account their uncertainty. To this end, we adapt the Schmidt Kalman filter (SKF) [44] and treat map features as nuisance parameters

which will no longer be updated but whose covariance and cross-correlations to other states are still utilized in the EKF update. As a result, this renders only  $O(n)$  computational complexity, making our proposed Schmidt-EKF Visual-Inertial SLAM (SEVIS) significantly more amenable to running on resource-constrained sensor platforms.

In particular, the main contributions of the paper include:

- We design a high-precision, efficient Schmidt-EKF based VI-SLAM (i.e., SEVIS) algorithm which leverages the Schmidt-KF formulation to allow for concurrent estimation of an environmental map used for long-term loop closures to bound navigation drifts with linear computational complexity.
- We propose a keyframe-aided 2D-to-2D matching scheme for the challenging data association problem of matching 2D visual measurements to 3D map features, without performing 3D-to-2D matching (which may not be applicable to sparse 3D environmental maps). This 2D-to-2D matching is not effected by estimation performance, allowing for long-term loop closures and recovery from extreme drifts.
- We validate the proposed SEVIS algorithm extensively in both Monte-Carlo simulations and real-world experiments, showing the applicability and performance gains offered by our system. The experimental study of computation requirements further shows that the proposed SEVIS remains real-time while building and maintaining a 3D feature-based map.

## 2. Related Work

While SLAM estimators – by jointly estimating the location of the sensor platform and the features in the surrounding environment – are able to easily incorporate loop closure constraints to bound localization errors and have attracted much research attention in the past three decades [8, 1, 6, 3], there are also significant research efforts devoted to open-loop VIO systems (e.g., [30, 13, 14, 22, 17, 50, 37, 49, 53, 5, 2, 15, 40]). For example, a hybrid MSCKF/SLAM estimator was developed for VIO [23], which retains features that can be continuously tracked beyond the sliding window in the state as SLAM features while removing them when they get lost.

It is challenging to achieve accurate localization by performing large-scale VI-SLAM due to the inability to remain computationally efficient without simplifying assumptions such as treating keyframe poses and/or map features to be perfect (i.e., zero uncertainty). Many methods use feature observations from different keyframes to limit drift over the trajectory (e.g., [34, 21]), and with most leveraging a two-thread architecture that optimizes a small window of local keyframes and features, while a background thread solves a long-term sparse pose graph containing loop closure constraints [11, 32, 24, 40, 39]. For example, VINS-

Mono [40, 39] uses loop closure constraints in both the local sliding window and in the global batch optimization. During the local optimization, feature observations from keyframes provide implicit loop closure constraints, while the problem size remains small by assuming the keyframe poses are perfect (thus removing them from optimization), while their global batch process optimizes a relative pose graph. In [31] a dual-layer estimator uses the MSCKF to perform real-time motion tracking and triggers the global bundle adjustment (BA) on loop closure detection. This allows for the relinearization and inclusion of loop closure constraints in a consistent manner, while requiring substantial additional overhead time where the filter waits for the BA to finish. A large-scale map-based VINS [25] assumes a compressed prior map containing feature positions and their uncertainty and uses matches to features in the prior map to constrain the localization globally. The recent Cholesky-Schmidt-KF [9] however explicitly considers the uncertainty of the prior map, by employing the sparse Cholesky factor of the map’s information matrix and further relaxing it by reducing the map size with more sub-maps for efficiency. In contrast, in this work, we formulate a single-threaded Schmidt-EKF for VI-SLAM, allowing for full probabilistic fusion of measurements without sacrificing real-time performance and permitting the construction and leverage of an environmental map to bound long-term navigation drift indefinitely.

## 3. Visual-Inertial SLAM

The process of VI-SLAM optimally fuses camera images and IMU (gyroscope and accelerometer) measurements to provide 6DOF pose estimates of the sensor platform as well as reconstruct 3D positions of environmental features (map). In this section, we briefly describe VI-SLAM within the EKF framework, which serves as the basis for our proposed SEVIS algorithm.

The state vector of VI-SLAM contains the IMU navigation state  $x_I$  and a sliding window of cloned past IMU (or camera) poses  $x_C$  as in the MSCKF [30], as well as the map features’ positions  $x_S$  expressed in the global frame:<sup>1</sup>

$$x_k = [x_I \quad x_C \quad x_S]^T =: [x_A \quad x_S]^T \quad (1)$$

$$x_I = [{}^I_k \bar{q} \quad b_k \quad {}^G v_{I_k} \quad b_{a_k} \quad {}^G p_{I_k}]^T \quad (2)$$

$$x_C = [{}^{I_{k-1}}_G \bar{q} \quad {}^G p_{I_{k-1}} \quad \dots \quad {}^{I_{k-m}}_G \bar{q} \quad {}^G p_{I_{k-m}}]^T \quad (3)$$

$$x_S = [{}^G p_{f_1} \quad \dots \quad {}^G p_{f_n}]^T \quad (4)$$

<sup>1</sup>Throughout this paper the subscript  $|j$  refers to the estimate of a quantity at time-step  $j$ , after all measurements up to time-step  $j$  have been processed.  $\hat{x}$  is used to denote the estimate of a random variable  $x$ , while  $\tilde{x} = x - \hat{x}$  is the error in this estimate.  $I_{n \times m}$  and  $O_{n \times m}$  are the  $n \times m$  identity and zero matrices, respectively. Finally, the left superscript denotes the frame of reference the vector is expressed with respect to.

where  ${}^l_k \bar{q}$  is the unit quaternion parameterizing the rotation  $C({}^l_k \bar{q}) = {}^l_k C$  from the global frame of reference  $\{G\}$  to the IMU local frame  $\{I_k\}$  at time  $k$  [47],  $b_g$  and  $b_a$  are the gyroscope and accelerometer biases, and  ${}^G v_{I_k}$  and  ${}^G p_{I_k}$  are the velocity and position of the IMU expressed in the global frame, respectively. The clone state  $x_C$  contains  $m$  historical IMU poses in a sliding window, while the map state  $x_S$  has  $n$  features. With the state decomposition (1), the corresponding covariance matrix can be partitioned as:

$$P_k = \begin{bmatrix} P_{AA_k} & P_{AS_k} \\ P_{SA_k} & P_{SS_k} \end{bmatrix} \quad (5)$$

### 3.1. IMU Propagation

The inertial state  $x_I$  is propagated forward using incoming IMU measurements of linear accelerations ( $a_m$ ) and angular velocities ( $\omega_m$ ) based on the following generic nonlinear IMU kinematics [7]:

$$x_{k+1} = f(x_k, a_{m_k} - n_{a_k}, \omega_{m_k} - n_{\omega_k}) \quad (6)$$

where  $n_a$  and  $n_{\omega}$  are the zero-mean white Gaussian noise of the IMU measurements. We linearize this nonlinear model at the current estimate, and then propagate the state covariance matrix forward in time:

$$P_{k|k-1} = \begin{bmatrix} {}^{k-1}P_{AA_{k-1}|k-1} & {}^{k-1}P_{AS_{k-1}|k-1} \\ P_{SA_{k-1}|k-1} & P_{SS_{k-1}|k-1} \end{bmatrix} + \begin{bmatrix} Q_{k-1} & 0 \\ 0 & 0 \end{bmatrix} \quad (7)$$

where  ${}^{k-1}P_{AA_{k-1}|k-1}$  and  $Q_{k-1}$  are respectively the system Jacobian and discrete noise covariance matrices for the active state [30]. Since the repeated computation of the above covariance propagation can become computationally intractable as the size of the covariance or rate of the IMU (e.g., > 200Hz) grows, we instead compound the state transition matrix and noise covariance as follows:

$${}^{(i+1)}P = {}^{k-1}P_{AA_{k-1}|k-1} \quad (i) \quad (8)$$

$$Q(i+1) = {}^{k-1}Q(i) + Q_{k-1} \quad (9)$$

with the initial conditions of  ${}^{(i=0)}P = I$  and  $Q(i=0) = 0$ . After compounding  ${}^{(i+1)}P$  and  $Q(i+1)$ , we directly apply them to propagate  $P_{k-1|k-1}$  based on (7).

### 3.2. Camera Measurement Update

Assuming a calibrated perspective camera, the measurement of a corner feature at time-step  $k$  is the perspective projection of the 3D point,  ${}^C p_{f_i}$ , expressed in the current camera frame  $\{C_k\}$ , onto the image plane, i.e.,

$$z_k = \frac{1}{z_k} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + n_{f_k} \quad (10)$$

$$\begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} = {}^C p_{f_i} = C({}^C \bar{q}) C({}^l_k \bar{q}) {}^G p_{f_i} - {}^G p_{I_k} + {}^C p_I \quad (11)$$

where  $n_{f_k}$  is the zero-mean, white Gaussian measurement noise with covariance  $R_k$ . In (11),  $\{{}^C \bar{q}, {}^C p_I\}$  is the extrinsic rotation and translation between the camera and IMU. This transformation can be obtained, e.g., by performing camera-IMU extrinsic calibration offline [28]. For the use of EKF, linearization of (10) yields the following residual:

$$r_{f_k} = H_k x_{k|k-1} + n_{f_k} \quad (12)$$

$$= H_{I_k} x_{I_{k|k-1}} + H_{f_k} {}^G p_{f_i, k|k-1} + n_{f_k} \quad (13)$$

where  $H_k$  is computed by (for simplicity assuming  $i = 1$ ):

$$H_k = \begin{bmatrix} H_{I_k} & 0_{3 \times 6m} & H_{f_k} & 0_{3 \times (3n-3)} \end{bmatrix} = \quad (14)$$

$$H_{proj} C({}^C \bar{q}) \begin{bmatrix} H_{I_k} & 0_{3 \times 9} & H_{p_k} & 0_{3 \times 6m} & C({}^l_k \bar{q}) & 0_{3 \times (3n-3)} \end{bmatrix} \quad (15)$$

$$H_{proj} = \begin{bmatrix} 1 & \hat{z}_k & 0 & -\hat{x}_k \\ \hat{z}_k & 0 & \hat{z}_k & -\hat{y}_k \end{bmatrix} \quad (15)$$

$$H_{I_k} = C({}^l_k \bar{q}) {}^G \hat{p}_{f_i} - {}^G \hat{p}_{I_k} \times, \quad H_{p_k} = -C({}^l_k \bar{q}) \quad (16)$$

Once the measurement Jacobian and residual are computed, we can apply the standard EKF update equations to update the state estimates and error covariance [26].

## 4. Schmidt-EKF based VI-SLAM

It is known that the EKF update of state estimates and covariance has quadratic complexity in terms of the number of map features [38], making naive implementations of VI-SLAM too expensive to run in real-time. Leveraging the SKF [44], we propose a novel Schmidt-EKF for VI-SLAM (SEVIS) algorithm which mitigates this quadratic complexity. The key idea is to selectively treat map features as nuisance parameters in the state vector [i.e., Schmidt state  $x_S$  (1)] whose mean and covariance will no longer be updated, while their cross-correlations with the active state  $x_A$  are still utilized and updated.

In particular, the IMU propagation of the proposed SEVIS is identical to that of the standard EKF in Section 3.1. In what follows we primarily focus on the update with monocular images, which is at the core of our SEVIS, but the approach is easily extendable to stereo systems. As the camera-IMU sensor pair moves through the environment, features are tracked using descriptor-based tracking. FAST features are first detected [41] and ORB descriptors [43] are extracted for each. The OpenCV [36] ‘‘BruteForce-Hamming’’ KNN descriptor matcher is used to find correspondences, after which we perform both a ratio test between the top two returns to ensure valid matches and 8-point RANSAC to reject any additional outliers. Once visual tracks are found, three types of tracked features are used to efficiently update state estimates and covariance: (i) VIO features that are opportunistic and can only be tracked for a short period time, (ii) SLAM features that are more stable than the above one and can be tracked beyond the current sliding window, and (iii) map features that are the

matured and informative SLAM features which are kept in the Schmidt state for an indefinite period of time.

#### 4.1. VIO Features: MSCKF Update

For those features that have lost active track in the current window (termed VIO features), we perform the standard MSCKF update [30]. In particular, we first perform BA to triangulate these features for computing the feature Jacobians  $H_f$  [see (14)], and then project  $r_k$  [see (13)] onto the left nullspace of  $H_f$  (i.e.,  $N H_f = 0$ ) to yield the measurement residual independent of features:

$$N r_f = N H_x \tilde{x}_{A_{k|k-1}} + N H_f^G \tilde{p}_{f_i} + N n_f \quad (17)$$

$$r_f = H_x \tilde{x}_{A_{k|k-1}} + n_f \quad (18)$$

where  $H_x$  is the stacked measurement Jacobians with respect to the navigation states in the current sliding window,  $R_f = N R_f N$  is the inferred noise covariance [30].

#### 4.2. SLAM Features: EKF Update

For those features that can be reliably tracked longer than the current sliding window, we will initialize them into the active state and perform EKF updates as in the standard EKF-based VI-SLAM (see Section 3.2). However, it should be noted that SLAM features will not remain active forever, instead they will either be moved to the Schmidt state as nuisance parameters (see Section 4.3) or marginalized out for computational savings as in [23].

#### 4.3. Map Features: Schmidt-EKF Update

If we perform VIO by linearly marginalizing out features [51] as in the MSCKF [30], the navigation errors may grow unbounded albeit achieving efficiency; on the other hand, if performing full VI-SLAM by continuously maintaining features (map) in the state, the computational cost may become prohibitive albeit gaining accuracy. In particular, two challenges arise in SLAM that must be tackled: (i) the increase in computational complexity due to number of map features included, and (ii) the data association of detecting whether actively tracked features match previously mapped features in the state vector. This motivates us to design our SEVIS algorithm that builds a sparse feature-based map of the environment which can then be leveraged to prevent long-term drift while still preserving necessary efficiency via the SKF.

##### 4.3.1 Keyframe-aided 2D-to-2D Matching

To overcome the data association challenge, given 3D positions of map features already included in the state vector, one straightforward approach might be through 3D-to-2D projection (i.e., projecting the 3D map feature onto the current frame) to find the correspondence of current visual

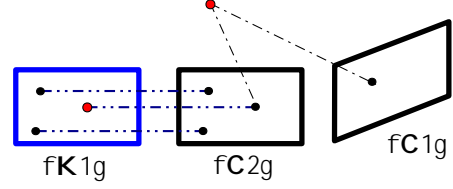


Figure 1: Illustration of the proposed keyframe-aided 2D-to-2D matching for data association. Assuming a cloned frame {C2} matches to a keyframe {K1} with all actively tracked features, and among these positive matches, one feature (red) corresponds to a map feature, the measurements in {C2} and {C1} will be used to update the active state by performing Schmidt-EKF update.

measurements to the mapped feature, which is often used in the literature (e.g., [9, 25]). However, in a typical SLAM scenario, estimating a map of 3D point features and matching them to current features is often sparse; for example, we found that it was common for a multi-floor indoor environment with up to 600 map features to only have about 10 features that can successfully project back into the active frame. Moreover, if there is any non-negligible drift in the state estimates (which is inevitable in practice), then projected features are likely to not correspond to the same spatial area as the current image is observing, thus preventing utilization of map information to reduce navigation errors.

For these reasons, we advocate 2D-to-2D matching for data association with the aid of “keyframes” that observe previous areas in the environment, due to its ability to provide high quality estimates and not be effected by estimation drift. Each keyframe contains a subset of the extracted features that correspond to map features in the state vector, and thus, if we match active feature tracks to previous keyframes we can find the correspondence between the newly tracked features and the previously mapped features that reside in our state.

Specifically, we first query the keyframe database to retrieve the closest keyframe to the current frame. To this end, different place recognition approaches such as DBow2 [12] and CALC [27] can be used to find the best candidate. After retrieval, we perform an additional geometric check by ensuring that the fundamental matrix can be calculated between the current frame and the proposed keyframe match, which we found provided extremely good matches to the best keyframe in the database. After retrieving a matching keyframe, we perform descriptor-based matching from features in the current frame to the keyframe with *all* extracted features from both frames followed by 8-point RANSAC to reject outliers. We now have the correspondences between the current frame feature tracks and keyframe map features. Fig. 1 visualizes this process.



### 4.3.2 Schmidt-EKF Update

To gain significant computational savings while still performing SLAM and exploiting map constraints to bound navigation errors, we adapt the SKF methodology [44] and treat the map features as nuisance parameters by only tracking their cross-correlations with the active states while still allowing for probabilistic inclusion of them during update. In particular, we compute the gain matrix of the Schmidt-EKF as follows:

$$\begin{aligned} K_{A_k} &= P_{AA_{k|k-1}} H_{A_k} + P_{AS_{k|k-1}} H_{S_k} S_k^{-1} \\ K_{S_k} &= P_{SA_{k|k-1}} H_{A_k} + P_{SS_{k|k-1}} H_{S_k} \end{aligned} \quad (19)$$

where  $H_A$  and  $H_S$  are respectively the measurement Jacobians with respect to the active and Schmidt states features [see (13)], and  $S_k$  is the residual covariance given by:

$$S_k = \begin{bmatrix} H_{A_k} & H_{S_k} \end{bmatrix} P_{k|k-1} \begin{bmatrix} H_{A_k}^T & H_{S_k}^T \end{bmatrix} + R_k \quad (20)$$

To reduce the computational complexity, we do *not* update the map feature nuisance parameters (Schmidt state), and thus, as in the SKF, we set the gain corresponding to the Schmidt state to zero, i.e.,  $K_{S_k} = 0$ . As a result, the state estimate is updated as follows:

$$\hat{x}_{A_{k|k}} = \hat{x}_{A_{k|k-1}} + K_{A_k} r_{f_k}, \quad \hat{x}_{S_{k|k}} = \hat{x}_{S_{k|k-1}} \quad (21)$$

The covariance is efficiently updated in its partitioned form:

$$\begin{aligned} P_{AA_{k|k}} &= P_{AA_{k|k-1}} \\ &\quad - K_{A_k} (H_{A_k} P_{AA_{k|k-1}} + H_{S_k} P_{AS_{k|k-1}}) \end{aligned} \quad (22)$$

$$\begin{aligned} P_{AS_{k|k}} &= P_{AS_{k|k-1}} \\ &\quad - K_{A_k} (H_{A_k} P_{AS_{k|k-1}} + H_{S_k} P_{SS_{k|k-1}}) \end{aligned} \quad (23)$$

$$P_{SS_{k|k}} = P_{SS_{k|k-1}} \quad (24)$$

Up to this point, we have fully utilized the current camera measurement information to update the SEVIS state estimates and covariance [see (1) and (5)]. The main steps of the proposed SEVIS are outlined in Algorithm 1.

### 4.4. Computational Complexity Analysis

Here we demonstrate the computational efficiency of the proposed SEVIS by providing detailed analysis that shows the complexity is *linear* with respect to the number of map features. This efficiency will also be demonstrated with experimental data in Section 6.

**Propagation:** The main computational cost of propagation comes from the matrix multiplication of  $_{k-1} P_{AS_{k-1|k-1}}$  [see (7)], where  $_{k-1}$  is a square matrix of  $\dim(x_A)$  size and  $P_{AS_{k-1|k-1}}$  is a fat matrix with size of  $O(n)$ . This incurs a total cost of  $O(n)$  because the number of map features far exceeds the size of the active state.

#### Algorithm 1 Schimdt-EKF Visual-Inertial SLAM (SEVIS)

**Propagation:** Propagate the IMU navigation state estimate  $\hat{x}_{I_{k|k-1}}$  based on (6), the active state's covariance  $P_{AA_{k|k-1}}$  and cross-correlation  $P_{AS_{k|k-1}}$  based on (7).

**Update:** For an incoming image,

- Perform stochastic cloning [42] of current state.
- Track features into the newest frame.
- Perform keyframe-aided 2D-to-2D matching to find map feature correspondences:
  - Query keyframe database for a keyframe visually similar to current frame.
  - Match currently active features to the features in the keyframe.
  - Associate those active features with mapped features in the keyframe.
- Perform MSCKF update for VIO features (i.e., those that have lost their tracks) as in Section 4.1.
- Initialize new SLAM features if needed and perform EKF update as in Section 4.2.
- Perform Schmidt-EKF update for map features as in Section 4.3.2.

**Management of Features and Keyframes:**

- Active SLAM features that have lost track are moved to the Schmidt state or marginalized out.
- Marginalize the oldest cloned pose from the sliding window state.
- Marginalize map features if exceeding the maximum map size.
- Insert a new keyframe into database if we have many map features in the current view.
- Remove keyframes without map features in view.

**Update:** After propagation, we augment the state by appending the propagated state to the active clone state  $x_C$ . This is an  $O(n)$  computation as we simply need to append a new row and column on the active covariance  $P_{AA_k}$  and then a row on the Schmidt cross-correlation terms  $P_{AS_k}$  yielding an  $O(n)$  operation. SLAM feature initialization follows the same logic and is an  $O(n)$  operation. During update, naively, the operation allows for the computation cost to be on order  $O(n^2)$  in the case that the current frame matches to *all* features in the map at the same time instance. A close inspection of (20) reveals that if the size of  $H_{S_k}$  is order  $n$ , the calculation of  $S_k$  will be of  $O(n^2)$ . However, this is *not* common in practice due to large environments and limited viewpoints. Therefore, we limit the number of features that can be used in one update to be far lower than the order of  $n$  and additional features can be processed at future instances to spread the computation over a period of time allowing for  $O(n)$  complexity at every time step.

**Management:** We manage the matrices  $P_{AA_k}$ ,  $P_{AS_k}$ , and  $P_{SS_k}$  as separate entities and pre-allocate  $P_{SS_k}$  to the max-

(a) Orientation RSSE

(b) Position RSSE

Figure 2: Monte-Carlo simulation averaged RSSE of pose (position and orientation) estimates for the three considered VIO and VI-SLAM algorithms.

imum number of allowed features to prevent overhead from memory allocation operations. When moving a state from the active state  $x_{A_k}$  to the Schmidt state  $x_{S_k}$ , special care is taken such that this operation remains on order  $O(n)$ . In particular, we first copy the associated block column from  $P_{AA_k}$  onto the last column of  $P_{AS_k}$ , after which we copy the associated block row in  $P_{AS_k}$  to the last row and column in the pre-allocated  $P_{SS_k}$ , thus yielding an total cost of  $O(n)$ . Marginalizing states in the active state is of  $O(n)$  as it requires removal of a row from the  $P_{AS_k}$  matrix which is achieved through copying all rows after the to-be-removed upwards overwriting the to-be-removed entries. During marginalization of map features from the Schmidt state  $P_{SS_k}$ , we overwrite the rows and columns corresponding with the to-be-removed state with the last inserted map feature, allowing for an  $O(n)$  operation.

## 5. Monte-Carlo Simulation Results

To validate the back-end estimation engine of the proposed SEVIS, we first perform Monte-Carlo simulations of visual-inertial SLAM with known measurement-feature correspondences, where a monocular-visual-inertial sensor platform is moving on a circular trajectory within a cylinder arena observing a series of environmental features. The simulation parameters about the sensors and the trajectory are listed in Table 1.

In particular, we compare three VINS algorithms to reveal the benefits of the proposed SEVIS: (i) The baseline VIO approach, which consists of the MSCKF augmented with 6 SLAM features (see [23]). These SLAM features are explicitly marginalized out when they leave the field of view. (ii) The baseline SLAM method, which uses the same MSCKF window but is augmented with 90 SLAM features. Different from the above VIO, in this case the SLAM features are never marginalized so that they can be used for (implicit) loop closures. (iii) The proposed SEVIS algo-

Table 1: Monte-Carlo Simulation Parameters

Parameter	Value	Units
IMU Angle Random Walk Coeff.	0.4	deg/ $\sqrt{\text{Hr}}$
IMU Rate Random Walk Coeff.	0.02	deg/sec/ $\sqrt{\text{Hr}}$
IMU Velocity Random Walk Coeff.	0.03	m/sec/ $\sqrt{\text{Hr}}$
IMU Acceleration Random Walk Coeff.	0.25	milli-G/ $\sqrt{\text{Hr}}$
IMU Sample Rate	100	Hz
Image Processing Rate	5	Hz
Feature Point Error $\sigma$	0.17	deg
Number of MSCKF Poses	15	
Approximate Loop Period	32	sec

rithm, which consists of the same MSCKF window and 6 SLAM features as in the baseline VIO, while being augmented with a bank of 90 map features that are modeled as nuisance parameters. When the SLAM features leave the field of view, they are moved into the Schmidt states, becoming the map features as described in Algorithm 1.

The average root sum squared error (RSSE) performance of 50 Monte-Carlo simulation runs are shown in Fig. 2. As expected, the baseline VIO accumulates drift in both orientation and position over time while the baseline SLAM provides bounded error performance without long term drift. It is interesting to point out that the position RSSE oscillates slightly depending on the location relative to the initial loop closure. This is because that the EKF has limited ability to correct these errors as it cannot re-linearize past measurements unlike optimization-based approaches [48]. More importantly, it is clear that the proposed SEVIS algorithm also does not accumulate long-term drift, although it is slightly less accurate than the baseline SLAM. However, this degradation in accuracy is a small price to pay considering that the SEVIS is of *linear* computational complexity with respect to the number of map features, while the baseline SLAM has *quadratic* complexity.

## 6. Real-World Experimental Results

We further evaluated the baseline MSCKF-based VIO (without map features), the baseline full VI-SLAM, and the proposed SEVIS on real-world datasets. In what follows, we first examine the estimator accuracy and computational overhead, after which the systems are evaluated on a challenging nighttime multi-floor dataset, showing that the proposed SEVIS can robustly be extended to realistic applications.

### 6.1. Vicon Loops Dataset

We first validated the proposed system on the Vicon loops dataset [21] that spans 1.2km in a single room over a 13 minute collection period. A hand-held VI-sensor [35] provides grayscale stereo image pairs and inertial informa-

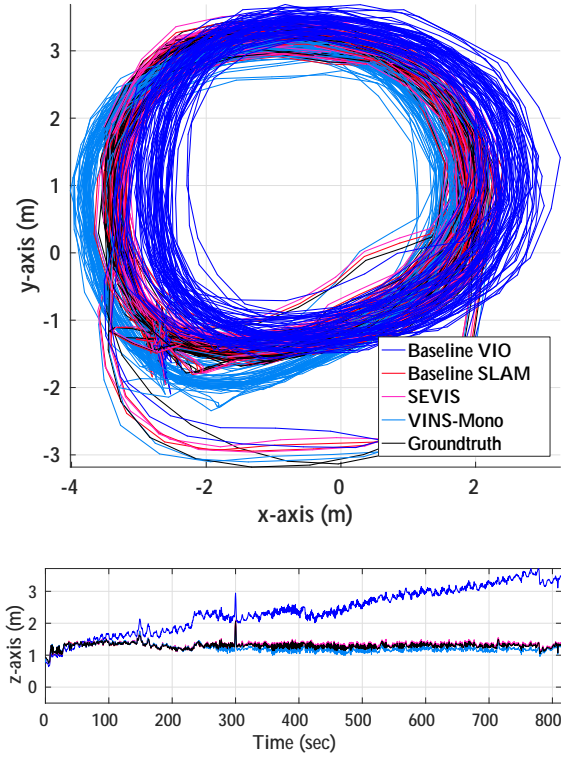


Figure 3: Trajectory of the baseline VIO, baseline SLAM with map features, proposed SEVIS with Schmidt covariance update, and VINS-Mono [40, 39]. Clearly the inclusion of map features has limited the drift and allows for high accuracy.

Table 2: Relative trajectory error for different segment lengths along with the overall absolute trajectory error. Values were computed using Zhang and Scaramuzza’s open sourced utility [54].

Segment Length	Baseline VIO	Baseline SLAM	SEVIS	VINS-Mono
123m	0.383	0.102	0.111	0.184
247m	0.645	0.099	0.108	0.238
370m	0.874	0.104	0.123	0.325
494m	1.023	0.095	0.121	0.381
618m	1.173	0.107	0.139	0.425
ATE	0.779	0.121	0.128	0.323

tion, while full 6DOF groundtruth is captured using a Vicon motion tracking system at 200 Hz. The maximum number of map features was set to 600 points to ensure real-time performance over the entire trajectory with images inserted into the query keyframe database at 0.5 Hz and a max of 5 SLAM features in the active state at a time. The results presented show three different configurations: (i) the baseline VIO augmented with 5 SLAM features, (ii) the base-

Figure 4: Boxplot of the relative trajectory error statistics. The middle box spans the first and third quartiles, while the whiskers are the upper and lower limits. Plot best seen in color.

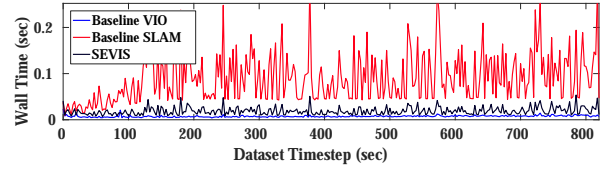


Figure 5: The wall clock execution time in seconds comparing the three methods can be seen. Plot best seen in color.

line VI-SLAM with 600 SLAM/map features, and (iii) the proposed SEVIS with 600 map features that leverages the Schmidt formulation for computational gains.

We evaluated the proposed method using two different error metrics: Absolute Trajectory Error (ATE) and Relative Error (RE). We point the reader to [54] for detailed definitions of these error metrics. Alongside our baseline and proposed methods, we additionally evaluated VINS-Mono [40, 39] to provide a comparison to a current state-of-the-art method that leverages loop closure information. Shown in Table 2 and Fig. 4, the proposed SEVIS is able to localize with high accuracy and perform on the level of the full baseline VI-SLAM system. Looking at the RE it is clear that the inclusion of map features prevents long-term drift and offers a greater accuracy shown by the almost constant RE as the trajectory segment length grows. The proposed SEVIS provides a computationally feasible filter that has similar accuracy as full baseline VI-SLAM with competitive performance to that of VINS-Mono (although the VINS-Mono leverages batch optimization).

The primary advantage of the proposed SEVIS algorithm over full-covariance SLAM is a decrease in computational complexity. The practical utility of this is evident in the run-times of the different algorithms. As shown in Fig. 5, we evaluated the three systems and collected timing statistics of our implementation.<sup>2</sup> The proposed SEVIS is able

<sup>2</sup>Single thread on an Intel(R) Xeon(R) E3-1505Mv6 @ 3.00GHz

to remain real-time (20 Hz camera means we need to be under 0.05 seconds total computation), while the full VI-SLAM method with 600 map features, has update spikes that reach magnitudes greater than four times the computational limit. This is due to the full covariance update being of order  $O(n^2)$ . Note that there is an additional overhead in the propagation stage as symmetry of the covariance matrix needs to be enforced for the entire matrix instead of just the active elements to ensure numerical stability.

## 6.2. Nighttime Multi-Floor Dataset

We further challenged the proposed system on a difficult indoor nighttime multi-floor dataset, which has multiple challenges including low light environments, long exposure times, and low contrast images with motion blur unsuitable for proper feature extraction (see Fig. 6). If features can be extracted, the resulting descriptor matching is poor due to the high noise and small gradients, and as compared to the Vicon Loops Dataset, more outliers are used during update, causing large estimator jumps and incorrect corrections. We stress that the proposed SEVIS can recover in these scenarios due to keyframe-aided 2D-to-2D matches which are invariant to poor estimator performance or drift and map feature updates correct and prevent incorrect drift.

A Realsense ZR300 sensor<sup>3</sup> was used to collect 20 minutes of grayscale monocular fisheye images with inertial readings, with the **1.5km** trajectory spanning two floors. We additionally performed online calibration of the camera to IMU extrinsic to further refine the transform provided by the manufacture’s driver. A max of 700 map-points allowed for sufficient coverage of the mapping area, keyframes were inserted into the query database at 4Hz to ensure sufficient coverage of all map features, and 2 SLAM features in the active state at a time. The trajectory generated by the baseline VIO and the proposed SEVIS are shown in Fig. 7. Clearly, the inclusion of map features prevent long-term drift experienced by the baseline VIO which exhibits large errors in both the yaw and z-axis direction. Since no groundtruth was available for this dataset, as a common practice, we computed the start-end error of the trajectory which should ideally be equal to zero as the sensor platform was returned to the starting location. The baseline VIO had an error of **4.67m (0.31%** of trajectory distance) while the proposed SEVIS had an error of only **0.37m (0.02%** of trajectory distance).

## 7. Conclusions and Future Work

In this paper, we have developed the high-precision, efficient SEVIS algorithm that adapts the SKF formulation for long-term visual-inertial SLAM. In particular, the probabilistic inclusion of map features within SEVIS allows for

Figure 6: Selected views during the night multi-floor trajectory show the high noise, poor lighting conditions, and motion blur that greatly challenge visual feature tracking.

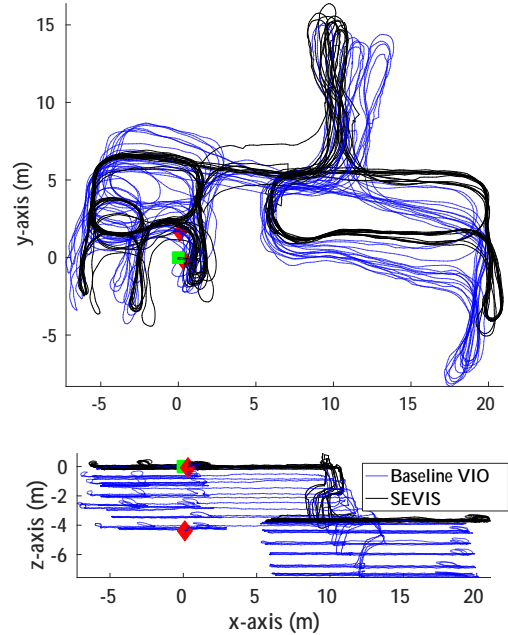


Figure 7: Estimated trajectories of the baseline VIO (blue) and SEVIS (black) show the improved performance due to inclusion of map features. The start-end positions are denoted with a green square and red diamond respectively.

bounded navigation drifts while retaining linear computational complexity. To achieve this, the keyframe-aided 2D-to-2D feature matching of current visual measurements to 3D map features greatly facilitates the full utilization of the map information. We then performed extensive Monte-Carlo simulations and real-world experiments whose results showed that the inclusion of map features greatly impact the long-term accuracy while the proposed SEVIS still allows for real-time performance without effecting estimator performance. In the future, we will investigate how to refine the quality of map features added for long-term localization and further evaluate our system on resource-constrained mobile sensor systems.

## 8. Acknowledgment

This work was partially supported by the University of Delaware (UD) College of Engineering, Google Daydream, and by the U.S. Army Research Lab.

<sup>3</sup><https://software.intel.com/en-us/realsense/zr300>



## References

- [1] T. Bailey and H. Durrant-Whyte. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robotics Automation Magazine*, 13(3):108–117, 2006. **2**
- [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, 2017. **2**
- [3] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220, Sept 2017. **2**
- [4] M. Brossard, S. Bonnabel, and A. Barrau. Invariant kalman filtering for visual inertial slam. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 2021–2028. IEEE, 2018. **1**
- [5] M. Brossard, S. Bonnabel, and J. Condomines. Unscented kalman filtering on lie groups. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2485–2491, Sept 2017. **2**
- [6] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016. **2**
- [7] A. B. Chatfield. *Fundamentals of High Accuracy Inertial Navigation*. AIAA, 1997. **3**
- [8] H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: Part I. *IEEE Robotics Automation Magazine*, 13(2):99–110, June 2006. **2**
- [9] R. C. DuToit, J. A. Hesch, E. D. Nerurkar, and S. I. Roumeliotis. Consistent map-based 3d localization on mobile devices. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6253–6260, May 2017. **2, 4**
- [10] S. Ebcin and M. Veth. Tightly-coupled image-aided inertial navigation using the unscented Kalman filter. Technical report, Air Force Institute of Technology, Dayton, OH, 2007. **1**
- [11] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Proc. European Conference on Computer Vision*, Zurich, Switzerland, Sept. 6–12, 2014. **2**
- [12] D. Gálvez-López and J. D. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, October 2012. **4**
- [13] J. Hesch, D. Kottas, S. Bowman, and S. Roumeliotis. Consistency analysis and improvement of vision-aided inertial navigation. *IEEE Transactions on Robotics*, 30(1):158–176, 2013. **1, 2**
- [14] J. Hesch, D. Kottas, S. Bowman, and S. Roumeliotis. Camera-IMU-based localization: Observability analysis and consistency improvement. *International Journal of Robotics Research*, 33:182–201, 2014. **1, 2**
- [15] Z. Huai and G. Huang. Robocentric visual-inertial odometry. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, Oct. 1-5, 2018. (to appear). **1, 2**
- [16] G. Huang, K. Eickenhoff, and J. Leonard. Optimal-state-constraint EKF for visual-inertial navigation. In *Proc. of the International Symposium on Robotics Research*, Sestri Levante, Italy, Sept. 12-15 2015. **1**
- [17] G. Huang, M. Kaess, and J. Leonard. Towards consistent visual-inertial navigation. In *Proc. of the IEEE International Conference on Robotics and Automation*, pages 4926–4933, Hong Kong, China, May 31-June 7 2014. **1, 2**
- [18] V. Indelman, S. Williams, M. Kaess, and F. Dellaert. Information fusion in navigation systems via factor graph based incremental smoothing. *Robotics and Autonomous Systems*, 61(8):721–738, 2013. **1**
- [19] J. Kim and S. Sukkarieh. Real-time implementation of airborne inertial-SLAM. *Robotics and Autonomous Systems*, 55(1):62–71, Jan. 2007. **1**
- [20] D. G. Kottas, J. A. Hesch, S. L. Bowman, and S. I. Roumeliotis. On the consistency of vision-aided inertial navigation. In *Proc. of the 13th International Symposium on Experimental Robotics*, Quebec City, Canada, June 17–20, 2012. **1**
- [21] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *International Journal of Robotics Research*, 34(3):314–334, 2015. **1, 2, 6**
- [22] M. Li and A. Mourikis. High-precision, consistent EKF-based visual-inertial odometry. *International Journal of Robotics Research*, 32(6):690–711, 2013. **1, 2**
- [23] M. Li and A. I. Mourikis. Optimization-based estimator design for vision-aided inertial navigation. In *Robotics: Science and Systems*, pages 241–248, Sydney, Australia, June 2012. **2, 4, 6**
- [24] H. Liu, M. Chen, G. Zhang, H. Bao, and Y. Bao. Iceba: Incremental, consistent and efficient bundle adjustment for visual-inertial slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1974–1982, 2018. **2**
- [25] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart. Get out of my lab: Large-scale,

- real-time visual-inertial localization. In *Robotics: Science and Systems*, 2015. 2, 4
- [26] P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, London, 1979. 3
- [27] N. Merrill and G. Huang. Lightweight unsupervised deep loop closure. In *Proc. of Robotics: Science and Systems (RSS)*, Pittsburgh, PA, June 26–30, 2018. 4
- [28] F. M. Mirzaei and S. I. Roumeliotis. A Kalman filter-based algorithm for IMU-camera calibration: Observability analysis and performance evaluation. *IEEE Transactions on Robotics*, 24(5):1143–1156, Oct. 2008. 3
- [29] A. Mourikis, N. Trawny, S. Roumeliotis, A. Johnson, A. Ansar, and L. Matthies. Vision-aided inertial navigation for spacecraft entry, descent, and landing. *IEEE Transactions on Robotics*, 25(2):264–280, 2009. 1
- [30] A. I. Mourikis and S. I. Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3565–3572, Rome, Italy, Apr. 10–14, 2007. 1, 2, 3, 4
- [31] A. I. Mourikis and S. I. Roumeliotis. A dual-layer estimator architecture for long-term localization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008.*, pages 1–8. IEEE, 2008. 2
- [32] R. Mur-Artal and J. D. Tardós. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017. 2
- [33] R. Mur-Artal and J. D. Tardós. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, April 2017. 1
- [34] E. Nerurkar, K. Wu, and S. Roumeliotis. C-klam: Constrained keyframe-based localization and mapping. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3638–3643, May 2014. 2
- [35] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. T. Furgale, and R. Siegwart. A synchronized visual-inertial sensor system with fpga pre-processing for accurate real-time slam. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 431–437. IEEE, 2014. 6
- [36] OpenCV Developers Team. Open source computer vision (OpenCV) library. Available: <http://opencv.org>. 3
- [37] M. K. Paul, K. Wu, J. A. Hesch, E. D. Nerurkar, and S. I. Roumeliotis. A comparative analysis of tightly-coupled monocular, binocular, and stereo VINS. In *Proc. of the IEEE International Conference on Robotics and Automation*, pages 165–172, Singapore, July 2017. 1, 2
- [38] L. M. Paz, J. D. Tardós, and J. Neira. Divide and conquer: EKF slam in  $O(n)$ . *IEEE Transactions on Robotics*, 24(5):1107–1120, Oct 2008. 3
- [39] T. Qin, P. Li, and S. Shen. Relocalization, global optimization and map merging for monocular visual-inertial slam. *arXiv preprint arXiv:1803.01549*, 2018. 2, 7
- [40] T. Qin, P. Li, and S. Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1, 2, 7
- [41] 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, 2010. 3
- [42] S. I. Roumeliotis and J. W. Burdick. Stochastic cloning: A generalized framework for processing relative state measurements. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1788–1795, Washington, DC, May 11–15, 2002. 5
- [43] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 IEEE international conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011. 3
- [44] S. F. Schmidt. Application of state-space methods to navigation problems. In C. LEONDES, editor, *Advances in Control Systems*, volume 3, pages 293–340. Elsevier, 1966. 1, 3, 5
- [45] S. Shen, N. Michael, and V. Kumar. Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft mavs. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5303–5310, May 2015. 1
- [46] D. Strelow. *Motion Estimation From Image and Inertial Measurements*. PhD thesis, CMU, 2004. 1
- [47] N. Trawny and S. I. Roumeliotis. Indirect Kalman filter for 3D attitude estimation. Technical report, University of Minnesota, Dept. of Comp. Sci. & Eng., Mar. 2005. 3
- [48] B. Triggs, P. McLauchlan, R. Hartley, and Andrew Fitzgibbon. Bundle adjustment – A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000. 6
- [49] K. Wu, T. Zhang, D. Su, S. Huang, and G. Dissanayake. An invariant-ekf vins algorithm for improving consistency. In *Proc. of the IEEE/RSJ Interna-*

*tional Conference on Intelligent Robots and Systems*, pages 1578–1585, Sept 2017. [2](#)

- [50] 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. In *Robotics: Science and Systems Conference (RSS)*, 2015. [1](#), [2](#)
- [51] Y. Yang, J. Maley, and G. Huang. Null-space-based marginalization: Analysis and algorithm. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6749–6755, Vancouver, Canada, Sept. 24-28, 2017. [4](#)
- [52] Z. Yang and S. Shen. Monocular visualinertial state estimation with online initialization and cameraimu extrinsic calibration. *IEEE Transactions on Automation Science and Engineering*, 14(1):39–51, Jan 2017. [1](#)
- [53] T. Zhang, K. Wu, J. Song, S. Huang, and G. Dissanayake. Convergence and consistency analysis for a 3D invariant-ekf slam. *IEEE Robotics and Automation Letters*, 2(2):733–740, April 2017. [2](#)
- [54] Z. Zhang and D. Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018. [7](#)