Range-Visual-Inertial Odometry: Scale Observability Without Excitation

Jeff Delaune, David S. Bayard and Roland Brockers

***Abstract*—Traveling at constant velocity is the most efficient trajectory for most robotics applications. Unfortunately without accelerometer excitation, monocular Visual-Inertial Odometry (VIO) cannot observe scale and suffers severe error drift. This was the main motivation for incorporating a 1D laser range finder in the navigation system for NASA's *Ingenuity* Mars Helicopter. However, *Ingenuity*'s simplified approach was limited to flat terrains. The current paper introduces a novel range measurement update model based on using facet constraints. The resulting range-VIO approach is no longer limited to flat scenes, but extends to any arbitrary structure for generic robotic applications. An important theoretical result shows that scale is no longer in the right nullspace of the observability matrix for zero or constant acceleration motion. In practical terms, this means that scale becomes observable under constant-velocity motion, which enables simple and robust autonomous operations over arbitrary terrain. Due to the small range finder footprint, range-VIO retains the minimal size, weight, and power attributes of VIO, with similar runtime. The benefits are evaluated on real flight data representative of common aerial robotics scenarios. Robustness is demonstrated using indoor stress data and full-state ground truth. We release our software framework, called xVIO, as open source.**

***Index Terms*—Visual-Inertial SLAM, Aerial Systems: Perception and Autonomy, Observability, Inertial Excitation, Mars Helicopter.**

I. Introduction

MONOCULAR Visual-Inertial Odometry (VIO) is a popular approach in robotics to obtain accurate metric state estimates close to a scene, or in GPS-denied conditions. Indeed, a camera and an Inertial Measurement Unit (IMU) form a minimal sensor suite in terms of size, weight and power, which is readily available on most robots.

However, monocular VIO can only observe the motion scale when the acceleration is not constant. This leads to severe error drift under zero or constant-velocity trajectories, which are very common in robotics. This problem is critical problem for applications which must rely on accurate VIO scale estimates for control. Our work is motivated by Mars helicopters [1], [2], but it is applicable to planetary, military, and urban robots in general; as well as indoor or underground traverses along a straight corridor or tunnel.

Our novel range-visual-inertial odometry algorithm can observe scale even under zero or constant-acceleration trajectories. It uses a 1D Laser Range Finder (LRF) that keeps the sensor suite lightweight, while efficiently leveraging VIO sparse structure estimates. Our main contributions are:

- a range measurement model which prevents VIO scale drift and adapts to any scene structure,
- a linearized range-VIO observability analysis, showing scale is observable without excitation,
- outdoor demonstration on a realistic dataset,
- indoor stress case analysis with full-state ground truth,
- an open-source C++ implementation.

In [1], a range-VIO method was presented that navigates over relatively flat terrain while supporting a stable motionless hover needed for demonstrating NASA's *Ingenuity* Mars Helicopter. The current paper extends these range-VIO results with a new method that makes scale observable for 3D terrain without requiring any inertial excitation. This generalization addresses an important need in the field of robotics as well as for future Mars helicopters. The current paper is a journal extension of a previous conference paper [2], which focused specifically on the Mars helicopter application. This included a real-time demonstration with candidate spaceflight hardware operating over Mars-like terrain. The conference paper treatment was non-theoretical and focused on obtaining proof-of-concept empirical results. The current journal paper derives and analyzes its theoretical observability properties. The error drift reduction is evaluated on urban aerial robotics data, which is significantly more complex and 3D than a Mars environment. The robustness of the facet-scene assumption is demonstrated with an indoor stress test supported by a full-state ground truth comparison. Finally, we make the source code publicly available[1].

We refer to monocular VIO as VIO in the rest of this paper. Section II reviews the VIO literature, observability limitations, and drift mitigation techniques using additional sensors. Section III introduces our range-VIO framework. Section IV analyzes the observability benefits of our approach. Sections V and VI present our tests and results. A video and a technical report [3] supplement this paper.

Manuscript received: October, 15, 2020; Revised January, 12, 2021; Accepted February, 5, 2021.

This paper was recommended for publication by Javier Civera upon evaluation of the Associate Editor and Reviewers' comments. The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

The authors are with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA. {jeff.h.delaune, david.s.bayard, roland.brockers}@jpl.nasa.gov

Digital Object Identifier (DOI): see top of this page.

[1] https://github.com/jpl-x

II. Literature Review

A. Visual-Inertial Odometry

One branch of VIO is based on loosely-coupled visual-inertial sensing. In these approaches, a vision-only algorithm estimates position and velocity up to scale, and orientation up to gravity, before fusion with the IMU [4]. The visual odometry module can be swapped between any of the modern algorithms developed in the computer vision community, such as PTAM [5], SVO [6], ORB-SLAM [7] or DSO [8].

The most accurate and robust VIO methods come from tightly-coupled approaches, in which visual measurements consisting of feature tracks or image patch intensities directly constrain the inertial state integration in one single estimator. These approaches require a larger state vector, which leads to a higher computational cost. But they gain in accuracy through the cross-correlations between the inertial and visual states [9], and in robustness with the ability to propagate the state even when no or few image primitives can be tracked. Recent approaches include both filter-based [10], [11] and nonlinear optimization-based methods [12], [13], [14]. Some solutions use image feature coordinates for measurements [10], [12], [13] while others use image intensity values [11], [14]. With good excitation, typical position errors can be under 1% of distance travelled [15].

B. VIO Observability Analysis

VIO observability with unknown IMU bias has been studied at length in the robotics literature. Under generic excitation, the VIO states were found to be observable except for the global position and the rotation about the gravity vector. [16], [17], [18] demonstrated this for the nonlinear system; while [19], [20] proved it for the linearized system and improved its consistency. These unobservable quantities imply that VIO position and heading estimates drift under any conditions with noise. In practice, this drift is acceptable in many robotics scenarios operating at small scale.

[21] further analyzed the unobservable directions under two specific motions for the linearized system with unknown bias. First, they showed that all three global rotations become unobservable if the system has no rotational motion of its own. Second, they showed that under constant acceleration, the scale of motion is unobservable. [22] derived results in line with these for the specific case of hovering. The complete absence of rotation is unlikely in most real applications, and even if it happens, the relative orientation of the camera with respect to the scene structure is still preserved. Constant or zero-acceleration is likely along straight traverses though, and the consequences of scale errors in terms of position and velocity drift can be catastrophic for the planning and control of a robot's trajectory.

C. VIO Scale Drift Mitigation

Range sensors and their equivalents can be used in addition to, or in replacement of, a monocular camera, in order to eliminate the scale observability issue of VIO. Most approaches leverage either lidar or radar scans [23], RGBD cameras [24], or stereo visual measurements [25]. Unlike VIO, these options suffer either from range limitation or integration costs[2] that limit their use in robotics applications.

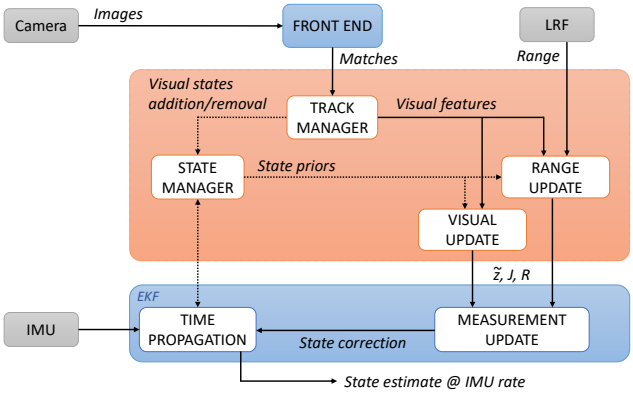


Fig. 1: Range-visual-inertial odometry architecture. Range and visual measurement innovations $\tilde{\boldsymbol{z}}$, Jacobians $\boldsymbol{J}$ and covariance matrices $\boldsymbol{R}$ are used to correct the inertial navigation errors in an EKF. The track manager sorts image features matches into tracks, while the state manager adds and removes or vision states dynamically.

1D Laser Range Finders (LRF) are an underrepresented sensing option in the SLAM literature. Modern units can sense over tens of meters with centimeter resolution. They fit within a small, lightweight and power-efficient package that can be accommodated even on resource-constrained robots. In our previous work for NASA's *Ingenuity* Mars Helicopter [1], we implemented a range-visual-inertial odometry algorithm that integrates LRF measurements to make the scale observable. The low dimension of the resulting estimator, only 21 states, comes at the cost of assuming the scene is flat and level, which is not compatible with most robotics scenarios. [26] solve a similar problem over 3D scenes by initializing the depth of some VIO features with ultrasonic range measurements. This relaxes the scene assumption from globally-flat to locally-flat, but it also assumes the local terrain slope is perpendicular to the range sensor axis within the ranged area. This is problematic over 3D scenes, given the large beam width of ultra sonic sensors.

In this paper, we eliminate VIO scale drift over any scene structure using a novel LRF measurement model. The accuracy and narrow beam width of the LRF create a strong range constraint with the depth of the visual features estimated by VIO in an Extended Kalman Filter (EKF). This constraint assumes the scene can be partitioned into triangular facets with the visual image features as vertices.

III. Range-Visual-Inertial Odometry

The architecture of our framework in Figure 1 is based on an Extended Kalman Filter (EKF). It tightly couples visual and range updates with inertial state propagation. We provide complete derivation details in our technical report [3].

[2]E.g. lidar scanner weight, or stereo baseline.

A. Inertial State Propagation

The EKF state vector $\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_I{}^T & \boldsymbol{x}_V{}^T\end{bmatrix}^T$ is divided between the states related to the IMU $\boldsymbol{x}_I$, and those related to vision $\boldsymbol{x}_V$. The inertial states

$$\boldsymbol{x}_I = \begin{bmatrix}\boldsymbol{p}_w^i{}^T & \boldsymbol{v}_w^i{}^T & \boldsymbol{q}_w^i{}^T & \boldsymbol{b}_g{}^T & \boldsymbol{b}_a{}^T\end{bmatrix}^T \tag{1}$$

include the position, velocity and orientation of the IMU frame $\{i\}$ with respect to the world frame $\{w\}$, the gyroscope biases $\boldsymbol{b}_g$ and the accelerometer biases $\boldsymbol{b}_a$. Rotation quaternions are used to model orientations.

IMU measurements are used to propagate the state estimate and the corresponding subblocks of the error covariance matrix to first order, using [27] and [28].

B. Visual Update

We perform visual updates using the hybrid SLAM-MSCKF paradigm [29]. This requires additional vision states

$$\boldsymbol{x}_V = \begin{bmatrix}\boldsymbol{p}_w^{c_1}{}^T & \dots & \boldsymbol{p}_w^{c_M}{}^T & \boldsymbol{q}_w^{c_1}{}^T & \dots & \boldsymbol{q}_w^{c_M}{}^T & \boldsymbol{f}_1{}^T & \dots & \boldsymbol{f}_N{}^T\end{bmatrix}^T \tag{2}$$

which includes a sliding window with the orientations $\{\boldsymbol{q}_w^{c_i}\}_i$ and positions $\{\boldsymbol{p}_w^{c_i}\}_i$ of the camera frame at the last M image time instances, along with the 3D coordinates of N visual features $\{\boldsymbol{f}_j\}_j$. Each feature state $\boldsymbol{f}_j = \begin{bmatrix}\alpha_j & \beta_j & \rho_j\end{bmatrix}^T$ represents the inverse-depth parametrization of world feature point $\boldsymbol{F}_j$ with respect to an anchor pose $\{c_{i_j}\}$ selected from the sliding window of pose states. Inverse depth improves feature depth convergence properties [30].

The visual measurement is the pinhole projection of terrain feature $\boldsymbol{F}_j$ over the normalized image plane $z = 1$ of the camera frame $\{c_i\}$ at time i

$${}^{i,j}\boldsymbol{z}_{v,m} = \frac{1}{{}^{c_i}z_j}\begin{bmatrix}{}^{c_i}x_j \\ {}^{c_i}y_j\end{bmatrix} + \boldsymbol{n}_v \ , \tag{3}$$

where $\boldsymbol{n}_v$ is a zero-mean white Gaussian feature measurement noise in image space. Equation 3 can be related to the state if we express the Cartesian coordinates of feature $\boldsymbol{F}_j$ in camera frame $\{c_i\}$ as

$$\boldsymbol{p}_{c_i}^{F_j} = \begin{bmatrix}{}^{c_i}x_j & {}^{c_i}y_j & {}^{c_i}z_j\end{bmatrix}^T \tag{4}$$

$$= \boldsymbol{C}(\boldsymbol{q}_w^{c_i})\left(\boldsymbol{p}_w^{c_{i_j}} + \frac{1}{\rho_j}\boldsymbol{C}(\boldsymbol{q}_w^{c_{i_j}})^T\begin{bmatrix}\alpha_j \\ \beta_j \\ 1\end{bmatrix} - \boldsymbol{p}_w^{c_i}\right) \tag{5}$$

This enables SLAM updates for features which are included in the state vector [31]. Features which are not included in the state vector are processed using MSCKF [10]. MSCKF updates have a linear cost per feature, as opposed to a cubic cost for SLAM. However MSCKF requires translational motion since the feature has to be triangulated, which is not always satisfied in practice. Hence, we always perform SLAM updates, and only use MSCKF when the translational motion allows for it. This hybrid approach is also the most computationally-efficient [29]. SLAM features are either initialized with semi-infinite depth uncertainty [30], or using a MSCKF prior if possible [29].

Visual corner features are detected in the image using the FAST algorithm [32], and tracked with the pyramidal implementation of the Kanade-Lucas-Tomasi algorithm [33], [34]. Outlier features are detected at two levels: first at the image level with RANSAC [35], and then at the filter level with a Mahalanobis distance test. The track manager module in Figure 1 assigns each feature to either the SLAM or MSCKF paradigm based on the track length, detection score, and image coordinates. We use image tiles to ensure SLAM features are distributed throughout the field of view, and ensure strong pose constraints.

C. Ranged Facet Update

Our main contribution is a novel range measurement model to constrain the VIO scale drift. Like VIO, it is designed to work over arbitrary unknown 2D or 3D scenes.

1) Measurement Model: A range measurement depends on both the pose of the range sensor and the structure of the scene. The associated measurement model in a Bayesian estimator should account for uncertainty on both. Since structure uncertainty is included in the SLAM feature states, we leverage these states to construct the new range update model.

Our key assumption is that the structure is locally flat between three SLAM features surrounding the intersection point of the LRF beam with the scene. This assumption derives from the observation that visual features are often located at depth discontinuities, and that the structure of the scene between features is often smooth. The impact of this assumption in real-world sequences is discussed in the results section. For simplification purposes in this paper, we also assume zero translation between the optical center of the camera and the origin of the LRF[3]. Figure 2 illustrates the geometry of the scene. $\boldsymbol{u}_{r_i}$ is the unit vector oriented along

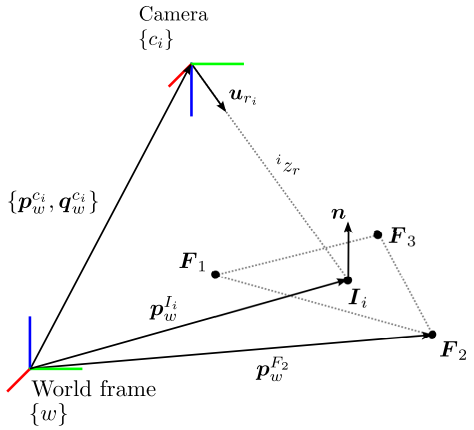


Fig. 2: Geometry of the range measurement ${}^i z_r$ at time i. The scene is assumed to be locally flat within a facet formed by visual features $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$ to build the range constraint.

the optical axis of the LRF at time i. $\boldsymbol{I_i}$ is the intersection of this axis with the terrain. $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$ are SLAM features forming a triangle around $\boldsymbol{I_i}$ in image space. $\boldsymbol{n}$ is a normal vector to the plane containing $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, $\boldsymbol{F}_3$ and $\boldsymbol{I_i}$.

[3]This offset can be measured and introduced in the model if needed.

If the dot product $\boldsymbol{u}_{r_i} \cdot \boldsymbol{n} \neq 0$, we can express the range measurement at time i as

$$
\begin{aligned}
{}^i z_r &= {}^i z_r \frac{\boldsymbol{u}_{r_i} \cdot \boldsymbol{n}}{\boldsymbol{u}_{r_i} \cdot \boldsymbol{n}} && (6)\\
&= \frac{(\boldsymbol{p}_w^{I_i} - \boldsymbol{p}_w^{c_i}) \cdot \boldsymbol{n}}{\boldsymbol{u}_{r_i} \cdot \boldsymbol{n}} && (7)\\
&= \frac{(\boldsymbol{p}_w^{I_i} - \boldsymbol{p}_w^{F_2} + \boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{c_i}) \cdot \boldsymbol{n}}{\boldsymbol{u}_{r_i} \cdot \boldsymbol{n}} && (8)\\
&= \frac{(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{c_i}) \cdot \boldsymbol{n}}{\boldsymbol{u}_{r_i} \cdot \boldsymbol{n}} && (9)
\end{aligned}
$$

since $(\boldsymbol{p}_w^{I_i} - \boldsymbol{p}_w^{F_2}) \cdot \boldsymbol{n} = 0$, where

$$\boldsymbol{n} = (\boldsymbol{p}_w^{F_1} - \boldsymbol{p}_w^{F_2}) \times (\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2}) \tag{10}$$

Here, $\{c_i\}$ is the camera frame at time i, and $\boldsymbol{p}_w^*$ represents the position of an object in the world frame $\{w\}$. Note that $\boldsymbol{n}$ is not necessarily a unit vector in this analysis.

Equations (9) and (10) demonstrate that range can be expressed as a nonlinear function h_r of the state vector $\boldsymbol{x}$ in Equation (11), without requiring any additional state beyond those of VIO. For use in our EKF estimation framework, we assume that the LRF measurements are disturbed by additive zero-mean white Gaussian noise n_r.

$${}^i z_{r,m} = {}^i z_r + n_r = h_r(\boldsymbol{x}) + n_r \tag{11}$$

We refer the reader to [3] for the linearization of Equation 11, providing the expressions of the measurement Jacobians.

2) Delaunay triangulation: To construct the range update in practice, we perform a Delaunay triangulation in image space over the SLAM features, and select the triangle in which the intersection of the LRF beam with the scene is located. We opted for the Delaunay triangulation since it maximizes the smallest angle of all possible triangulations [36]. This property avoids "long and skinny" triangles that do not provide strong local planar constraints.

Figure 3 shows the Delaunay triangulation, and the triangle selected as a ranged facet, over a sample image from our outdoor test sequence. It also illustrates the partitioning of the scene into triangular facets, with SLAM features at their corners. Note that if the state estimator uses only 3 SLAM features in a lightweight fashion, this is equivalent to a globally-flat world assumption. Conversely, if the density of SLAM features increases, the areas of the facets tend to zero and the facet scene assumption virtually disappears.

3) Range outlier rejection: Before being used in the filter, the range measurements go through a Mahalanobis distance test to detect outliers. This gating compares the range measurement to a prior built from the coordinates of the three visual features in the facet. It rejects violations of the facet assumption that cannot be explained by the prior uncertainty model derived from the error covariance matrix.

IV. Observability Analysis

We perform the observability analysis of the linearized range-VIO system, since it is based on an EKF. Although the observability of the nonlinear system would be required for completeness, it is out of the scope of this paper.

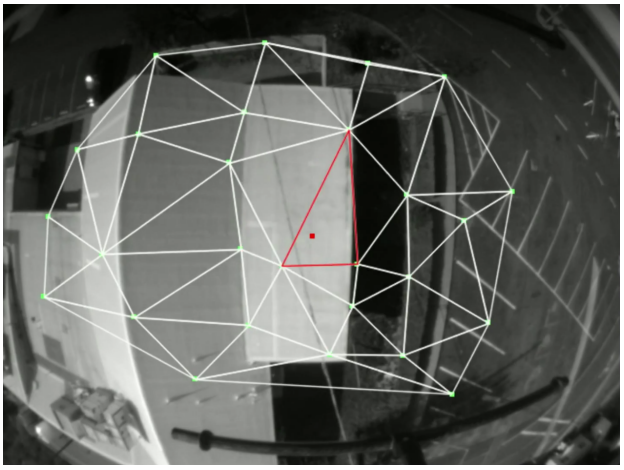


Fig. 3: Delaunay triangulation between SLAM image features tracked in the outdoor flight dataset. The red dot represents the intersection point of the LRF beam and the surface. The surrounding red triangle is the ranged facet.

To simplify the equations, our analysis assumes a state vector $\boldsymbol{x}^0 = \begin{bmatrix} \boldsymbol{x}_I{}^T & \boldsymbol{x}_P{}^T \end{bmatrix}^T$, where $\boldsymbol{x}_I$ was defined in Equation (1) and $\boldsymbol{x}_P = \begin{bmatrix} {}^w\boldsymbol{p}_1{}^T & \dots & {}^w\boldsymbol{p}_N{}^T \end{bmatrix}^T$ includes the Cartesian coordinates of the N SLAM features, $N \geq 3$. [19] proved that observability analysis for the linearized system based on $\boldsymbol{x}^0$ is equivalent to observability analysis for the linearized system defined with $\boldsymbol{x}$ in the previous section.

A. Observability Matrix

For $k \geq 1$, $\boldsymbol{M}_k = \boldsymbol{H}_k \boldsymbol{\Phi}_{k,1}$ is the k-th block row of observability matrix $\boldsymbol{M}$. $\boldsymbol{H}_k$ is the Jacobian of the range measurement in Equation (9) at time k with respect to $\boldsymbol{x}^0$, which is derived in Equations (42-47) in [3]. $\boldsymbol{\Phi}_{k,1}$ is the state transition matrix from time 1 to time k [37].

Without loss of generality, we can assume the ranged facet is constructed from the first 3 features in $\boldsymbol{x}_P$[4]. Then we derive the following expression for $\boldsymbol{M}_k$ in [3].

$$
\begin{aligned}
\boldsymbol{M}_k = \frac{1}{b}\big[\ & \boldsymbol{M}_{k,p} \mid \boldsymbol{M}_{k,v} \mid \boldsymbol{M}_{k,q} \mid \boldsymbol{M}_{k,b_g} \mid \boldsymbol{M}_{k,b_a} \mid \\
& \boldsymbol{M}_{k,p_1} \mid \boldsymbol{M}_{k,p_2} \mid \boldsymbol{M}_{k,p_3} \mid \boldsymbol{0}_{1\times 3(N-3)} \ \big]
\end{aligned} \tag{12}
$$

where

$$\boldsymbol{M}_{k,p} = -{}^w\boldsymbol{n}^T \tag{13}$$

$$\boldsymbol{M}_{k,v} = -(k-1)\delta t\, {}^w\boldsymbol{n}^T \tag{14}$$

$$
\begin{aligned}
\boldsymbol{M}_{k,\theta} = {}^w\boldsymbol{n}\Big(& -\frac{a}{b}\boldsymbol{C}\left(\boldsymbol{q}_w^{c_k}\right)^T \lfloor {}^c\boldsymbol{u}_r \times \rfloor\, \boldsymbol{C}\left(\boldsymbol{q}_w^{i_k}\right) \\
& - \Big\lfloor \boldsymbol{p}_w^{i_1} - \boldsymbol{v}_w^{i_1}(k-1)\delta t - \frac{1}{2}{}^w\boldsymbol{g}(k-1)^2\delta t^2 \\
& - \boldsymbol{p}_w^{i_k} \times \Big\rfloor \Big)\boldsymbol{C}\left(\boldsymbol{q}_{i_1}^w\right)
\end{aligned} \tag{15}
$$

$$\boldsymbol{M}_{k,b_g} = -\frac{a}{b}{}^w\boldsymbol{n}^T \boldsymbol{C}\left(\boldsymbol{q}_w^{c_k}\right)^T \lfloor {}^c\boldsymbol{u}_r \times \rfloor \boldsymbol{\phi}_{12} - {}^w\boldsymbol{n}\boldsymbol{\phi}_{52} \tag{16}$$

[4]This ordering of the states can be obtained at any timestep using permutation matrices. Permutation matrices are full rank and hence do not affect the rank of the observability matrix.

$$\boldsymbol{M}_{k,b_a} = -{}^w\boldsymbol{n}^T\boldsymbol{\phi}_{54} \tag{17}$$

$$\boldsymbol{M}_{k,p_1} = \left(\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)\right)^T \tag{18}$$

$$\boldsymbol{M}_{k,p_2} = \left({}^w\boldsymbol{n} + \left\lfloor(\boldsymbol{p}_w^{F_1} - \boldsymbol{p}_w^{F_3})\times\right\rfloor\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)\right)^T \tag{19}$$

$$\boldsymbol{M}_{k,p_3} = \left(\left\lfloor(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{F_1})\times\right\rfloor\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)\right)^T \tag{20}$$

$$a = (\boldsymbol{p}_w^{f_{j_2}} - \boldsymbol{p}_w^{c_i})^T{}^w\boldsymbol{n} \tag{21}$$

$$b = {}^w\boldsymbol{u}_{r_i}{}^T{}^w\boldsymbol{n} \tag{22}$$

and $\boldsymbol{\phi}_*$ are integral terms defined in [37]. For a generic vector $\boldsymbol{u} \in \mathbb{R}^3$, let ${}^w\boldsymbol{u}$ represent its coordinates in the world frame $\{w\}$, and $\lfloor\boldsymbol{u}\times\rfloor = \begin{bmatrix} 0 & -u_z & u_y \\ u_z & 0 & -u_x \\ -u_y & u_x & 0 \end{bmatrix}$ the skew-symmetric matrix associated with it. c_{i_1}, c_{i_2} and c_{i_3} are the anchor poses associated to $\boldsymbol{F}_1$, $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$, respectively, for their inverse-depth coordinates at time i.

B. Unobservable Directions

1) Generic motion: One can verify that the vectors spanning a global position or a rotation about the gravity vector still belong to the right nullspace of $\boldsymbol{M}_k$. Thus, the ranged facet update does not improve the observabily over VIO under generic motion [19], which is intuitive. Likewise, in the absence of rotation, the global orientation is still not observable [21].

2) Constant acceleration: In this subsection, we demonstrate that in the constant acceleration case, unlike VIO [21], the vector spanning the scale dimension

$$\boldsymbol{N}_s = \begin{bmatrix} \boldsymbol{p}_w^{i_1}{}^T & \boldsymbol{v}_w^{i_1}{}^T & \boldsymbol{0}_{6\times1}{}^T & -{}^i\boldsymbol{a}_w^i{}^T & \boldsymbol{p}_w^{F_1}{}^T & \dots & \boldsymbol{p}_w^{F_N}{}^T \end{bmatrix}^T \tag{23}$$

does not belong the right nullspace of $\boldsymbol{M}_k$, i.e. $\boldsymbol{M}_k\boldsymbol{N}_s \neq \boldsymbol{0}$. ${}^i\boldsymbol{a}_w^i$ is the zero or constant acceleration of the IMU frame in world frame, resolved in the IMU frame. One can write

$$\begin{aligned}\boldsymbol{M}_k\boldsymbol{N}_s = &-{}^w\boldsymbol{n}^T\boldsymbol{p}_w^{i_1} - (k-1)\delta t{}^w\boldsymbol{n}^T\boldsymbol{v}_w^{i_1} + {}^w\boldsymbol{n}^T\boldsymbol{\phi}_{54}{}^i\boldsymbol{a}_w^i \\ &+ \left(\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)\right)^T\boldsymbol{p}_w^{F_1} \\ &+ \left({}^w\boldsymbol{n} + \left\lfloor(\boldsymbol{p}_w^{F_1} - \boldsymbol{p}_w^{F_3})\times\right\rfloor\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)\right)^T\boldsymbol{p}_w^{F_2} \\ &+ \left(\left\lfloor(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{F_1})\times\right\rfloor\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)\right)^T\boldsymbol{p}_w^{F_3}\end{aligned} \tag{24}$$

Reference [21] shows that, under constant acceleration,

$$\boldsymbol{\phi}_{54}{}^i\boldsymbol{a}_w^i = -\left(\boldsymbol{p}_w^{i_k} - \boldsymbol{p}_w^{i_1} - (k-1)\delta t\boldsymbol{v}_w^{i_1}\right) \tag{25}$$

so

$$\begin{aligned}\boldsymbol{M}_k\boldsymbol{N}_s = &{}^w\boldsymbol{n}^T\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{i_k}\right) \\ &+ \left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)^T\Big(\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor^T\boldsymbol{p}_w^{F_1} \\ &+ \left\lfloor(\boldsymbol{p}_w^{F_1} - \boldsymbol{p}_w^{F_3})\times\right\rfloor^T\boldsymbol{p}_w^{F_2} \\ &+ \left\lfloor(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{F_1})\times\right\rfloor^T\boldsymbol{p}_w^{F_3}\Big)\end{aligned} \tag{26}$$

The cross product of the first term can be modified such that

$$\begin{aligned}&\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)^T\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor^T\boldsymbol{p}_w^{F_1} \\ &= \left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)^T\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor^T\left(\boldsymbol{p}_w^{F_1} - \boldsymbol{p}_w^{I_k} + \boldsymbol{p}_w^{I_k}\right)\end{aligned} \tag{27}$$

$$= \left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)^T\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor^T\boldsymbol{p}_w^{I_k} \tag{28}$$

By definition of the cross product,

$$\exists\,\lambda\in\mathbb{R},\ \left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2})\times\right\rfloor^T\left(\boldsymbol{p}_w^{F_1} - \boldsymbol{p}_w^{I_k}\right) = \lambda{}^w\boldsymbol{n} \tag{29}$$

and

$$\lambda\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)^T{}^w\boldsymbol{n} = 0 \tag{30}$$

Thus, by applying this to all cross-product terms,

$$\begin{aligned}\boldsymbol{M}_k\boldsymbol{N}_s = &{}^w\boldsymbol{n}^T\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{i_k}\right) \\ &+ \left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{I_k}\right)^T\left\lfloor(\boldsymbol{p}_w^{F_3} - \boldsymbol{p}_w^{F_2} + \boldsymbol{p}_w^{F_1}\right. \\ &\left. - \boldsymbol{p}_w^{F_3} + \boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{F_1})\times\right\rfloor^T\boldsymbol{p}_w^{I_k}\end{aligned} \tag{31}$$

$$= {}^w\boldsymbol{n}^T\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{i_k}\right) \tag{32}$$

By definition, if the three features of the facet are not aligned in the image, ${}^w\boldsymbol{n}^T\left(\boldsymbol{p}_w^{F_2} - \boldsymbol{p}_w^{i_k}\right) \neq \boldsymbol{0}$. End of proof.

Unlike VIO, range-VIO thus enables scale convergence even in the absence of acceleration excitation.

3) Zero-velocity: Note that when the velocity is null, i.e. in hover, $\boldsymbol{v}_w^{i_1} = \boldsymbol{0}$, $\boldsymbol{p}_w^{i_1} = \boldsymbol{p}_w^{i_k}$ in the previous demonstration, and the following unobservable direction appears instead.

$$\boldsymbol{N}_h = \begin{bmatrix} \boldsymbol{0}_{24\times1}{}^T & \boldsymbol{p}_w^{F_4}{}^T & \dots & \boldsymbol{p}_w^{F_N}{}^T \end{bmatrix}^T \tag{33}$$

It corresponds to the depth of the SLAM features not included in the facet. This result means that in the absence of translation motion, when feature depths are uncorrelated to each other, the ranged facet provides no constraint on the features outside the facet. As soon as the platform starts moving, visual measurements begin to correlate all feature depths, and the depths of all features become observable from a single ranged facet.

V. Experimental Setup

We recorded two datasets to characterize the performance of our approach. The first one is an outdoor flight in an urban environment, which is a common scenario in aerial robotics. The second sequence is aimed at stressing the facet assumption in an indoor environment, where full-state ground truth based on motion capture is available. Both tests consist of a uniform-velocity traverse, i.e. a straight line at constant speed, since this is the most limiting unobservable direction of VIO. Processing was done offline for this paper since the focus is on analysis, but we previously demonstrated the real-time performance of the range-visual-inertial odometry algorithm at 30 frames per seconds on a Snapdragon 820 processor [2], which is the space hardware baseline for the next Mars helicopter mission concept.

A. Sensors

Range data was provided by a Garmin Lidar Lite V3 single-point static laser range finder. This can range up to 40 m with a 2.5-cm accuracy, weighs only 22 g, and is less than 5 cm-long. The monocular navigation camera was a global-shutter Omnivision OV7251, providing 640×480 8-bit grayscale images in auto-exposure mode. Inertial data was delivered by a STIM300 tactical-grade IMU. The camera was collecting data at 30 Hz, the LRF at 25 Hz, and the IMU at 250 Hz. The sensor

suite was mounted on a rigid platform. The camera intrinsics and extrinsincs were calibrated, including the angles between the camera and LRF optical axes. The distance between the camera and the LRF was neglected, as the two sensors were mounted side by side.

B. Outdoor Test

1) Flight Sequence: For this test, the sensor platform was mounted on a GPS-controlled hexacopter. After take-off, the rotorcraft ascended to a cruise altitude of 11 m, and initiated a 150 m straight horizontal traverse at a constant speed of 2 m/s. The traverse was controlled within the performance limits of the on-board Pixhawk 2.1 Cube autopilot set up with the ArduCopter APM firmware. No inertial excitation was provided before take-off.

The flight path was chosen so that the first half of the trajectory covers flat ground, where the facet assumption is likely to be respected; and the second half is over buildings with non-flat roofs, including structure discontinuities where the facet assumption may be challenged. Our video supplement includes the full sequence, and a height profile to illustrate structure variations along the flight path. Figure 3 shows the image facets at the transition between flat ground and rooftops.

2) GPS Ground Truth: Position ground truth was provided by a RTK-GPS system composed of a Trimble BD930-UHF receiver with a BX982 base station. This system provides centimeter accuracy in clear outdoor environments. It also served as time server for our logging computer, so sensor data was readily time-synchronized with ground truth.

We attempted to run a GPS-IMU filter to get attitude ground truth, in addition to position only. However the horizontal accelerations were too small to observe the heading component of attitude [38], which ended up drifting.

C. Indoor Test

1) Hand-Held Sequence: Our second data set was recorded in an indoor environment, with the intention to create a stress test for the facet assumption. We arranged boxes of different heights next to each other in straight line under the LRF path, to create a structure with multiple 90° drop-offs, which result in severe relative height changes when the sensors are in close proximity to the top of the boxes. We refer the reader our video material to observe the violations of the facet assumption, which happens frequently in this sequence.

To highlight the benefit of the ranged facet model, we also made it a stress case for VIO: the sensor platform was hand-held to smooth the motion and eliminate the residual accelerations coming from the hexacopter controls[5]; the environment had surfaces with little texture to limit long feature tracks and increase visual scale drift; and once again, no inertial excitation was provided before the horizontal traverse.

2) Motion Capture Ground Truth: Another motivation for an indoor dataset was to obtain complete and accurate ground truth to fully characterize range-VIO performance in a stress case. Our test arena was equipped with 10 Vicon Vero motion capture cameras, which typically provide millimeter accuracy in position and sub-degree in orientation. For velocity ground truth, we filtered Vicon pose measurement with the on-board IMU [27]. We should note that this ground truth is not fully independent from the state estimates since the same IMU was used for both.

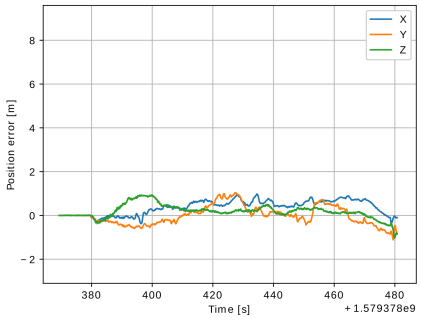
[5]Residual accelerations from hand motion are still present though.

VI. Results

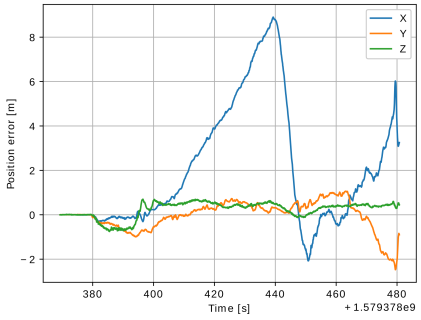
This section discusses the performance of our range-VIO algorithm on the sequences presented in the previous section. The visual state was set to accommodate $M = 4$ poses in the sliding window, and $N = 27$ SLAM features. The Mahalanobis distance test to capture outliers was set to 2σ, with σ the estimated range standard deviation. VIO was run with the exact same settings as range-VIO in all our comparison tests. The only difference was the additional processing of the range measurement with the ranged facet model in range-VIO.

A. Outdoor Flight Tests

Figure 4 compares the position errors of range-VIO and VIO during the outdoor traverse. Range-VIO maximum errors remain below 1 m on each axis, which is under 0.6% of the distance travelled. This performance is similar to state-of-the-art VIO under excitation [15]. Conversely, the VIO error rises along the direction of the traverse (X axis) from the time it is initiated, and up to values 9 times larger compared to range-VIO.

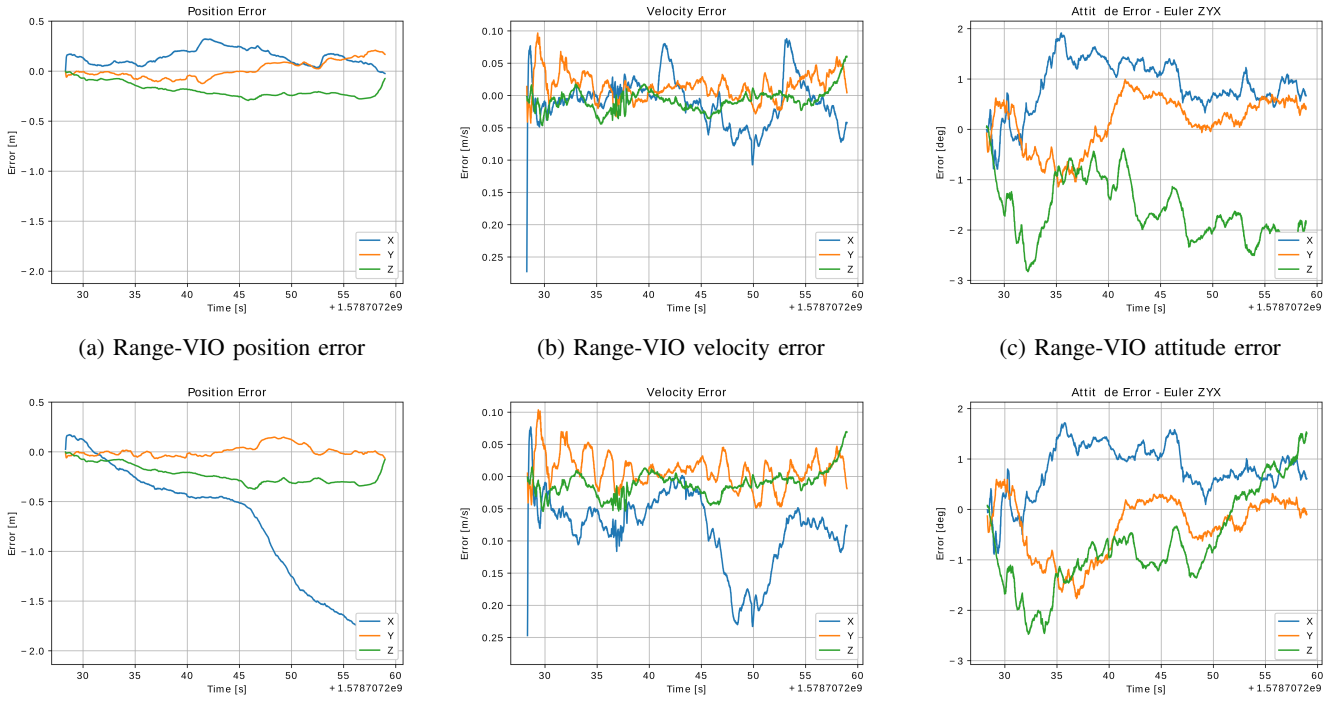


(a) Range-VIO position error



(b) VIO position error

Fig. 4: Position errors for range-VIO (top) and VIO (bottom) on the outdoor dataset. The X and Y axes are horizontal, Z is up. X was aligned with the direction of the traverse.



(a) Range-VIO position error (b) Range-VIO velocity error (c) Range-VIO attitude error

(d) VIO position error (e) VIO velocity error (f) VIO attitude error

Fig. 5: Position (left), velocity (center) and attitude (right) errors for range-VIO (top) and VIO (bottom) on the indoor stress dataset. The X and Y axes are horizontal, Z is up. X was aligned with the direction of the traverse.

We note that the VIO error is consistent with a scale error, which is not observable for VIO under the constant-acceleration traverse. This is a clear illustration of the observability benefit of range-VIO on a trajectory commonly used in robotics. We also note that the range-VIO errors in Figure 4(a) do not suffer from the transition between a flat terrain and a 3D structure, that occurs at $t = 425$ s. This is a good indication that the facets constructed with real-world visual features efficiently capture the structure of the scene. Additionally a 7-m ranged facet outlier occurred at $t = 410$ s, as the LRF hits a street light. This can be observed in the range profile shown in our video material. However it does not affect the range-VIO estimates in Figure 4(a), showing the efficiency of our range outlier rejection scheme.

B. Indoor Stress Tests

To further assess the robustness of the facet model, the indoor sequence discussed in Subsection V-C was used as a stress case. Figure 5 compares range-VIO and VIO errors in position, velocity and orientation, since ground truth was available for all these states indoors. The scale drift reduction is clearly visible along the direction of travel in Figures 5(a) and 5(d). Range-VIO has a maximum position error of 30 cm, or 2.5% while VIO errors grow to 2 m, or 17% in these challenging visual conditions and without excitation.

The velocity and orientation plots are a good illustration of how the facet assumption can work over challenging environments. Figure 5(b) and 5(e) show the velocity errors benefit from scale observability in range-VIO, since they are up to twice lower than VIO. Likewise, range-VIO orientation errors in Figure 5(c) slightly differ from that of VIO in Figure 5(f), especially in the Z (yaw) axis. We interpret these differences as error accumulation cause by ranged facet assumption violations too small to be caught by the Mahalanobis range outlier rejection. This only happens around the global yaw axis, which is unobservable to both methods. However, even in this extreme stress case, the yaw error has the same order of magnitude between range-VIO and VIO, while range-VIO clearly outperforms VIO in terms of position and velocity drift reduction.

Finally, we refer the reader to our video material for additional comparison results in a sequence with large excitation and good visual texture. Under these optimal conditions, VIO performs on par with range-VIO, having a maximum position error of 40 cm. This confirms that VIO was not detuned previously, but only suffered from the lack of excitation. It also confirms that range-VIO does not degrade VIO performance in the presence of good excitation and visual conditions.

VII. Conclusion

VIO-based robotic applications are limited by the inability to observe scale without excitation. In aerial robotics, scale observability without excitation is critical for even the most basic hovering and straight-line trajectories. Our main interest is for aerial exploration of distant worlds, like Mars [1]. Common terrestrial applications include conditions where GPS is unavailable (defense, underground), degraded (tall buildings, canyons), or not accurate enough (indoors).

Using a simple 1D laser range finder, our range-VIO approach eliminates scale drift in the absence of excitation while retaining the minimal size, weight and power requirements of VIO. A theoretical analysis demonstrated the observability of scale in such conditions. Results on constant-velocity real flight data showed error reduction by a factor of 9 compared to VIO.

The novel range update is based on a facet scene assumption that efficiently leverages VIO feature depth estimates to handle unknown structures. The facets can scale from a flat world assumption, to virtually no structure assumption at all based on visual feature density. This paper and supplement report [3] provide a full derivation of the range-VIO model. Range-VIO does not require additional states with respect to VIO, and does not add significant computational cost. We demonstrated the robustness of our facet assumption in a stress case.

Future extensions include increasing visual feature density around the LRF impact point on the scene to improve accuracy further. We also investigate the use of magnetometers and sun sensors to address the next major unobservable direction: orientation about the gravity vector.

REFERENCES

[1] D. S. Bayard, D. T. Conway, R. Brockers, J. Delaune, L. Matthies, H. Grip, G. Merewether, T. Brown, and A. San Martin. Vision-Based Navigation for the NASA Mars Helicopter. In *AIAA Scitech Forum*, 2019.

[2] J. Delaune, R. Brockers, D. S. Bayard, H. Dor, R. Hewitt, J. Sawoniewicz, G. Kubiak, T. Tzanetos, L. Matthies, and J. Balaram. Extended Navigation Capabilities for a Future Mars Science Helicopter Concept. In *IEEE Aerospace Conference*, 2020.

[3] J. Delaune, D.S. Bayard, and R. Brockers. xVIO: A Range-Visual-Inertial Odometry Framework. Technical report, Jet Propulsion Laboratory, 2020, arXiv:2010.06677.

[4] 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 *International Conference on Robotics and Automation (ICRA)*, pages 957–964. IEEE, 2012.

[5] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, 2007.

[6] C. Forster, M. Pizzoli, and D. Scaramuzza. SVO: Fast Semi-Direct Monocular Visual Odometry. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2014.

[7] R. Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[8] J. Engel, V. Koltun, and D. Cremers. Direct Sparse Odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 2018.

[9] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-Based Visual-Inertial Odometry Using Nonlinear Optimization. *The International Journal of Robotics Research*, 34, 02 2014.

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

[11] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust Visual Inertial Odometry Using a Direct EKF-Based Approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[12] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2017.

[13] 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.

[14] L. von Stumberg, V. C. Usenko, and D. Cremers. Direct Sparse Visual-Inertial Odometry Using Dynamic Marginalization. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2510–2517, 2018.

[15] J. A. Delmerico and D. Scaramuzza. A Benchmark Comparison of Monocular Visual-Inertial Odometry Algorithms for Flying Robots. *International Conference on Robotics and Automation (ICRA)*, pages 2502–2509, 2018.

[16] A. Martinelli. Closed-Form Solutions for Attitude, Speed, Absolute Scale and Bias Determination by Fusing Vision and Inertial Measurements. Technical report, INRIA, 2011.

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

[18] J. Kelly and G. S. Sukhatme. Visual-Inertial Sensor Fusion: Localization, Mapping and Sensor-to-Sensor Self-calibration. *The International Journal of Robotics Research*, 30(1):56–79, 2011.

[19] M. Li and A. I. Mourikis. High-precision, consistent EKF-based visual-inertial odometry. *The International Journal of Robotics Research*, 32(6):690–711, 2013.

[20] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis. Consistency Analysis and Improvement of Vision-aided Inertial Navigation. *IEEE Transactions on Robotics*, 30(1):158–176, 2014.

[21] K. J. Wu and S. I. Roumeliotis. Unobservable Directions of VINS Under Special Motions. Technical report, University of Minnesota, 2016.

[22] D. G. Kottas, K. J. Wu, and S. I. Roumeliotis. Detecting and dealing with hovering maneuvers in vision-aided inertial navigation systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3172–3179, 2013.

[23] S. Mohamed, M. Haghbayan, T. Westerlund, J. Heikkonen, H. Tenhunen, and J. Plosila. A Survey on Odometry for Autonomous Navigation Systems. *IEEE Access*, PP:1–1, 07 2019.

[24] V. Angladon, S. Gasparini, V. Charvillat, T. Pribanic, T. Petković, M. Donlic, B. Ahsan, and F. Bruel. An evaluation of real-time RGB-D visual odometry algorithms on mobile devices. *Journal of Real-Time Image Processing*, 02 2017.

[25] T. Taketomi, H. Uchiyama, and S. Ikeda. Visual SLAM algorithms: a survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 9:1–11, 2017.

[26] S. Urzua, R. Munguía, and A. Grau. Vision-based SLAM system for MAVs in GPS-denied environments. *International Journal of Micro Air Vehicles*, 9(4):283–296, 2017.

[27] S. Weiss and R. Siegwart. Real-Time Metric State Estimation for Modular Vision-Inertial Systems. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2012.

[28] N. Trawny and S. I. Roumeliotis. Indirect Kalman Filter for 3D Attitude Estimation. Technical report, University of Minnesota, 2005.

[29] M. Li and A. Mourikis. Optimization-Based Estimator Design for Vision-Aided Inertial Navigation. In *Robotics: Science and Systems conference*, 2012.

[30] J. Civera, A. Davison, and J. Montiel. Inverse Depth Parametrization for Monocular SLAM. *IEEE Transactions on Robotics*, 24(5):932–945, 2008.

[31] J. Delaune, R. Hewitt, L. Lytle, C. Sorice, R. Thakker, and L. Matthies. Thermal-Inertial Odometry for Autonomous Flight Throughout the Night. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2019.

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

[33] J.-Y. Bouguet. Pyramidal implementation of the Lucas Kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.

[34] J. Shi and C. Tomasi. Good Features to Track. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 594–600. IEEE, 1994.

[35] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[36] B. Gärtner and M. Hoffmann. Computational Geometry Lecture Notes HS 2013. Technical report, ETH Zürich, 2013.

[37] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis. Observability-constrained Vision-aided Inertial Navigation. Technical report, University of Minnesota, 2012.

[38] S. Weiss. *Vision Based Navigation for Micro Helicopters*. PhD thesis, ETH Zürich, 2012.