Conquering Textureless with RF-referenced Monocular Vision for MAV State Estimation

Shengkai Zhang[†], Sheyang Tang[†], Wei Wang[*†], Tao Jiang[†], and Qian Zhang[§]

***Abstract*— The versatile nature of agile micro aerial vehicles (MAVs) poses fundamental challenges to the design of robust state estimation in various complex environments. Achieving high-quality performance in textureless scenes is one of the missing pieces in the puzzle. Previously proposed solutions either seek a remedy with visual loop closure or leverage RF localizability with inferior accuracy. None of them support accurate MAV state estimation in textureless scenes. This paper presents RFSift, a new state estimator that conquers the textureless challenge with RF-referenced monocular vision, achieving centimeter-level accuracy in textureless scenes. Our key observation is that RF and visual measurements are tied up with pose constraints. Mapping RF to feature quality and sift well-matched ones significantly improves accuracy. RFSift consists of 1) an RF-sifting algorithm that maps 3D UWB measurements to 2D visual features for sifting the best features; 2) an RF-visual-inertial sensor fusion algorithm that enables robust state estimation by leveraging multiple sensors with complementary advantages. We implement the prototype with off-the-shelf products and conduct large-scale experiments. The results demonstrate that RFSift is robust in textureless scenes, 10× more accurate than the state-of-the-art monocular vision system. The code of RFSift is available at `https://github.com/weisgroup/RFSift`.**

I. Introduction

Micro aerial vehicles (MAVs) are constituting a fast-paced emerging technology that has the profound potential to decrease the risks to human life, *e.g.*, assisting firefighters in searching survivors [1], decrease the execution time and increase the efficiency of the overall process, *e.g.*, increasing inventory efficiency by 30× in warehouses with MAV auto-scanning [2]. Currently, the mainstream for MAV state estimation uses a monocular camera thanks to its small size, lightweight, and low cost [3]–[6]. However, it requires well lighting and rich texture to capture enough visual cues for state estimation via projective geometry, *e.g.*, optical flow needs texture for matching. Therefore, it is very challenging to work in textureless scenes that are commonplace, *e.g.*, a room with white walls, a solid color floor, mirrors on the wall, and large windows [7].

One workaround is to manually add visual markers on textureless surfaces, which is intrusive and labor-intensive. Another avoids textureless areas by assessing the perception of quality [8]. More conventional approaches [3], [4] use loop closure detection as a remedy to correct the accumulated error in textureless areas. While this may be sufficient to cope with temporary textureless flight, the vehicle cannot be stabilized for a long-term flight in such scenes. Moreover, there may not even be any loop at all to correct the error. One may consider deep-learning egomotion estimators, which are insensitive in texture [9]–[12]. Unfortunately, their stochastic results are not suitable for deterministic state estimation, failing to stabilize a MAV over a long-term run. Recently, efforts have been made to use RF signals, *e.g.*, LoRa [13], Ultra-WideBand (UWB) [14], [15], and WiFi [16], [17], as an alternative for MAV state estimation. They are highly resilient to visual conditions. However, due to the large wavelength of RF signals and environmental interference, their accuracy is decimeter-level, an order of magnitude worse than the monocular vision sensing in vision-friendly venues.

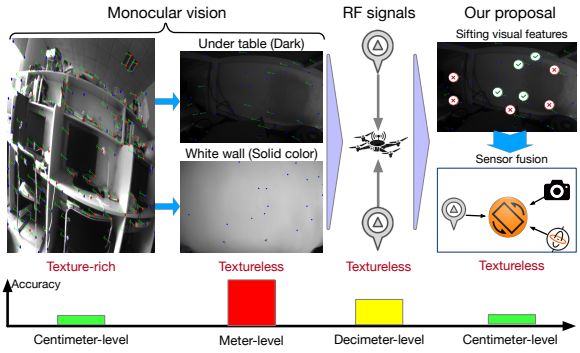


Fig. 1. RFSift conquers the textureless challenge, allowing MAV state estimation in dark or solid color scenes at a centimeter-level accuracy.

In this paper, we present RFSift, a robust state estimator that conquers the textureless challenge as shown in Figure 1, achieving centimeter-level accuracy with a monocular camera, a UWB module, and an inertial measurement unit (IMU). RFSift is free of the monocular camera's adverse effect in textureless scenes with few (≥ 1) UWB sources, *e.g.*, iPhone 11, without the prior knowledge of their positions. The fundamental problem of textureless that cripples computer vision is two-fold: fewer features are detected and more erroneous feature matching via optical flow algorithm [18]. Conventional approaches take all the features to optimize the state. Surprisingly, we find that sifting few "good" visual features that are better matched, rather than taking all of them, significantly improves the accuracy. To sift such "good" features, the conventional scheme, RANSAC [19], is ineffective due to the lack of features in textureless scenes. Our observation is that RF measurements, which are immune to visual conditions, encode the information of a vehicle's pose, providing a reference to distinguish the quality of features.

We realize the above high-level idea by designing two

Authors[†] are with School of Electronic Information and Communications, HUST, Wuhan, China. {szhangk, sheyangtang, weiwangw, taojiang}@hust.edu.cn

Author[§] is with Department of Computer Science and Engineering, HKUST, Hong Kong, SAR China. qianzh@cse.ust.hk

[*]The corresponding author is Wei Wang (weiwangw@hust.edu.cn).

components:

Visual feature sifting. Quantifying the quality of visual features requires the mapping function from UWB to visual measurements. UWB ranges and angles are 3D cues of locations and orientations, while visual features are the projection of points in the 3D world onto 2D image planes. It is nontrivial to establish the mapping from 3D UWB measurements to 2D visual features. We propose an RF-sifting algorithm that leverages the nature of micro-motion between two consecutive frames to compute a score for each feature that indicates the quality of feature matching.

RF-visual-inertial sensor fusion. Sifted visual features provide more reliable information of the vehicle's pose. To further improve the robustness and accuracy of state estimation, we fuse RF measurements (range and angle), sifted visual features, and IMU measurements into a new bundle adjustment (BA) paradigm and formulate a large optimization problem. Since RF measurements provide drift-free localizability, RFSift no longer requires the loop closure detection for monocular vision systems.

Highlights of our original contributions are as follows. *First*, we verify the observation by data-driven simulations and design an RF-sifting algorithm to sift better visual features. *Second*, we fuse RF, visual, and IMU measurements into a new BA paradigm to enable robust state estimation. *Finally*, we demonstrate the effectiveness of RFSift by a prototype implementation with extensive real-world experiments.

II. Related Work

Vision-based state estimation. Monocular vision-based odometry/SLAM/state estimation has been extensively studied [4]–[6], [20]–[23]. They are lightweight and highly accurate as long as application scenarios are well-lighted and texture-rich. On the contrary, in textureless venues, a camera cannot capture enough visual features to estimate MAV states via epipolar geometry. One may say deep-learning based solutions are insensitive to image texture [9]–[12]. But they require prior site survey for training, not accurate when working in unknown venues. Direct methods take all the raw pixel information in images to mitigate the effect of textureless scenes [24], [25]. However, they require high computing power (GPUs) to achieve real-time processing, which is unavailable for payload-limited MAVs.

RF-based state estimation. Recent years have witnessed advances in state estimation/navigation using RF signals, *e.g.*, WiFi [16], [17], LoRa [13], and UWB [15], [26]. They provide complementary sensing modalities to optical sensors in that RF signals can penetrate, reflect, or diffract from objects, being resilient to visual conditions. Zhang *et al.* designed WINS [16] that consists of advanced algorithms to solve multipath and rotation estimation problems, making WiFi-based state estimation into indoor scenarios. Although WiFi is ubiquitously available in modern cities, its narrow bandwidth still limits the accuracy to decimeter-level. To allow MAVs flying in scenarios where WiFi is not available, Zhang *et al.* proposed Marvel [13], a LoRa backscatter assisted state estimator that works in emergency scenarios, *e.g.*, firefighting operations. However, this system requires a dedicated device to produce chirp signals, and its accuracy is still limited to decimeter-level due to the narrow bandwidth so that it cannot support precise control for MAVs. To overcome the limitation of narrow bandwidth, UWB, an off-the-shelf product that transmits ultra-wideband signals, has been attractive in MAV state estimation [14], [15]. The wide bandwidth is highly robust to the multipath problem of RF signals. However, due to the lack of clock synchronization and environmental interference, the accuracy is typical > 10 cm while requiring multiple anchors' support.

RF-visual fusion. Combining RF and visual sensing modalities will conceivably improve MAV state estimation's robustness in general use cases [27]–[29]. Nyqvist *et al.* [27] uses an extended Kalman filter to fuse UWB and visual measurements for better robustness. Wang *et al.* [28] takes UWB into a graph-based optimization framework to correct monocular vision drift. Xu *et al.* [29] takes advantage of UWB's omnidirectional sensing range and fuses it with visual measurements for relative state estimation of an aerial swarm. The design of these approaches is an if-else paradigm that enables an automatic switch on the two sensing modalities and chooses the appropriate one based on their measurement uncertainties. In textureless scenes, such systems let RF sensing take over from visual sensing, and thus it reduces to an RF-based state estimation eventually. Therefore, they still stuck the accuracy at the decimeter level.

III. Feasibility Study

Conventional vision-based approaches take visual features as many as possible in the BA. It works fine in texture-rich regions because ill-matched features are outliers. Majority inliers contribute reliable information to achieve high accuracy. However, in textureless or dark scenarios, fewer corners in images result in less visual features [30], and the camera is harder to focus, leading to more erroneous feature matching [18]. Well-matched features no longer dominate the pose estimation. Taking all the observed features will probably hurt the accuracy.

To verify our claim, we write a data-driven simulator using monocular vision and run it with the EuRoC dataset [31] for MAV state estimation. We artificially blur random parts of the images in the dataset to simulate a textureless condition. The simulation works in two steps: 1) ranking the tracked visual features in terms of the re-projection distance of features; 2) testing the performance with a different number of features in descending order of the rank.

Specifically, Figure 2 shows the feature ranking method. We take two consecutive frames, for instance. Two features tracked by optical flow correspond to a 3D point in the world. A pair of tracked features encode the camera's rotation and translation via epipolar geometry [32]. Due to the distortion of a camera lens and imaging settings, the feature tracking can be noisy. It is conceivable that better tracking will give the relative pose closer to the ground truth, provided by the dataset [31]. To qualify the tracking quality, we reproject the visual features from the previous frame to the current

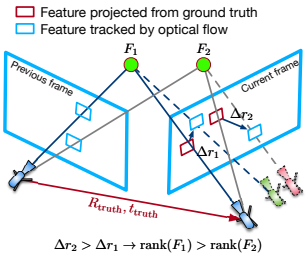


Fig. 2. The feature ranking method using the ground-truth pose.

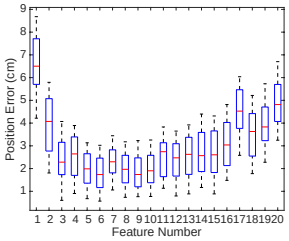


Fig. 3. Multiple trials with best features.

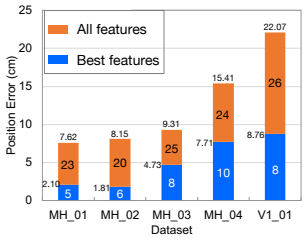


Fig. 4. Statistical results on multiple datasets.

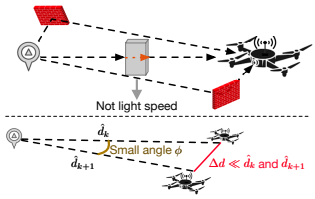


Fig. 5. The RF bias and our approximation.

frame based on the ground-truth camera pose. Then we calculate the distance Δr between the reprojected feature and its corresponding tracked feature on the image plane. The smaller the distance the higher rank it will be.

After the ranking, we evaluate the localization performance with different numbers of best features. Figure 3 shows the result of the "MH_02_easy" data by testing different numbers of best features and repeating 20 times, each of which randomly blurs parts of the images. "k features" means the system only takes the top-k best features into the BA. The data in our tests last for 23 seconds. 20 features are tracked at most. It shows that the best accuracy appears at "6 features", which is 1.81 cm. If incorporating all observed features, the accuracy degenerates to be 4.98 cm. We use top-30% of the observed features to improve accuracy 2.75×.

To prove our observation's generality, Figure 4 shows the statistical results on four Machine Hall datasets and one Vicon Room dataset. The inner bar of a column denotes the best accuracy while the upper bar is the accuracy of taking all the observed features. The number below the inner bar and upper bar denotes the number of best features used and the total number of observed features. It shows that there is 3× accuracy gain when taking less than top-30% best features. Therefore, sifting the best features is very promising to bolster state estimation accuracy in textureless settings.

IV. System Design of RFSift

A. Visual Feature Sifting

Our high-level idea is that if we can find reliable relative poses from other sensors, we can similarly rank the visual features as in § III. A strawman option is to use the onboard IMU to infer the pose. However, IMU is an interoceptive sensor that only measures the vehicle's internal state without any connection to environmental data like image features, making IMU measurements completely uncorrelated with image features. For example, a MAV inevitably flies at a constant speed over a long-term run. IMU gives zero accelerations, *i.e.*, zero translations.

We choose RF range and angle via UWB [33] to find relative poses. They are immune to visual limitations and highly resilient to multipath fading. RF measurements and visual features are correlated with the hidden vehicle's pose. Our idea is to estimate the relative pose of the vehicle from RF measurements and then reproject the visual features based on the pose. Unfortunately, RF range $\hat{d}$ and angle $\hat{\theta}$ are scalars. $\hat{(\cdot)}$ denotes the estimated or measured variable throughout the paper. Notice that we only measure azimuth angle as IMU only drifts in azimuth direction for rotation [4]. It is infeasible to recover the relative translation and rotation as vectors using RF measurements at two consecutive timestamps.

Our solution resorts to a combination of RF and IMU measurements. Range $\hat{d}$ and angle $\hat{\theta}$ are drift-free in pose and encode the information of 3D positions. They can combat the IMU drift in four degree-of-freedom (DoF) [4]. However, we cannot trivially use the raw measurements of these heterogeneous sensors, there are three issues: 1) the RF and IMU measurements are temporally misaligned in that their data rates are different; 2) the RF measurements are biased due to environmental interference; 3) there is also additive Gaussian noise for RF measurements.

To align the RF and IMU measurements in time, we use the IMU preintegration technique [34], [35] to preintegrate buffered IMU readings as one measurement at the time when receiving an RF measurement. Typically, the IMU data rate (100 Hz) is higher than the RF data rate (20 Hz). Given two time instants $[k, k+1]$ when obtaining two RF measurements, we preintegrate acceleration $\hat{\mathbf{a}}_t$ and angular rate $\hat{\omega}_t$ as [34]

$$\begin{aligned} \hat{\boldsymbol{\alpha}}^{i_k}_{i_{k+1}} &= \iint_{t\in[k,k+1]} \mathbf{R}^{i_k}_t \hat{\mathbf{a}}_t \, \mathrm{d}t^2, \; \hat{\boldsymbol{\beta}}^{i_k}_{i_{k+1}} = \int_{t\in[k,k+1]} \mathbf{R}^{i_k}_t \hat{\mathbf{a}}_t \, \mathrm{d}t, \\ \hat{\boldsymbol{\gamma}}^{i_k}_{i_{k+1}} &= \int_{t\in[k,k+1]} \boldsymbol{\gamma}^{i_k}_t \otimes \begin{bmatrix} 0 & \frac{1}{2}\hat{\omega}_t \end{bmatrix}^\top \mathrm{d}t, \end{aligned} \tag{1}$$

where $\otimes$ is the quaternion multiplication operation, $(\cdot)^i$ denotes a measurement in IMU frame, $(\cdot)^{i_k}_{i_{k+1}}$ the measurement in IMU frame when receiving $(k+1)^{\text{th}}$RF measurement with respect to the IMU frame at k^{th}. We use quaternion and rotation matrix interchangeably to represent rotation throughout this paper. $\mathbf{R}^{i_k}_t \in \text{SO}(3)$ is the rotation matrix from i_k to current time t. $\boldsymbol{\gamma}^{i_k}_t$ is the quaternion representation of a incremental rotation from i_k to t, which is available through short-term integration of gyroscope measurements. $\hat{\boldsymbol{\alpha}}^{i_k}_{i_{k+1}}$, $\hat{\boldsymbol{\beta}}^{i_k}_{i_{k+1}}$, and $\hat{\boldsymbol{\gamma}}^{i_k}_{i_{k+1}}$ are the preintegrated terms from IMU and they are temporally aligned with RF measurements.

To address the RF bias issue, we exploit the micro translation between two consecutive ranges. As shown in Figure 5, the bias comes from multipath and traversing through objects. Although UWB signals are highly resilient to multipath compare with narrow-band signals such as WiFi, it still can be interfered with objects nearby transmitter or receiver. Meanwhile, when the signal traversing through objects, the propagation speed within objects is different from the speed of light in a vacuum. Thus, the measured range with respect l^{th}UWB node $\hat{d}_{lk} = d_{lk} + b_l + n_{r_l}$ where

d_{lk} is the real range, b_l the bias, and n_{r_l} additive Gaussian noise of ranges. Since the UWB data rate is 20 Hz, 0.05 s between two ranges is within the channel coherence time, we can assume $\hat{d}_{lk}$ and $\hat{d}_{l(k+1)}$ share the same bias. Moreover, the speed of an indoor MAV is typically less than 2 m/s, then the translation between two RF ranges is about 0.1 m. In most cases, the distance between a UWB node and a vehicle can be greater than 1 m. From the cosine rule, the cosine value of the apex angle ϕ is > 0.995. Thus, we can make an approximation $\cos\phi \approx 1$. From the cosine rule, we have

$$\Delta d_l^2 = \hat{d}_{lk}^2 + \hat{d}_{l(k+1)}^2 - 2\hat{d}_{lk}\hat{d}_{l(k+1)}\cos\phi \approx \left(\hat{d}_{lk} - \hat{d}_{l(k+1)}\right)^2. \quad (2)$$

The above operation cancels the bias. On the other hand, since the azimuth angle $\hat{\theta}_{lk}$ to l^{th}UWB node directly reflects the geometric relationship, there is no bias effect. We only need to suppress its noise n_{a_l}.

To suppress the Gaussian noise $n_{r_l} \sim \mathcal{N}(0, \sigma_{r_l}^2)$ and $n_{a_l} \sim \mathcal{N}(0, \sigma_{a_l}^2)$ ($\sigma_{r_l}^2 = 5$ cm and $\sigma_{a_l}^2 = 5^\circ$ for our UWB node), we employ the Kalman filter (KF) to fuse the RF ranges and angles with the short-term integration of IMU. We use the acceleration and angular velocity provided by IMU to predict the next position and orientation. From the dynamics and Eqn. (1), the IMU-derived relative translation $\mathbf{p}_{i_{k+1}}^{i_k}$ can be expressed as

$$\begin{aligned} \mathbf{p}_{i_{k+1}}^{i_k} &= \mathbf{v}_{i_k}^{i_k}\Delta t_k - \mathbf{g}^{i_k}\frac{\Delta t_k^2}{2} + \hat{\boldsymbol{\alpha}}_{i_{k+1}}^{i_k}, \\ \mathbf{v}_{i_{k+1}}^{i_{k+1}} &= \hat{\boldsymbol{\gamma}}_{i_k}^{i_{k+1}}\left(\mathbf{v}_{i_k}^{i_k} - \mathbf{g}^{i_k}\Delta t_k + \hat{\boldsymbol{\beta}}_{i_{k+1}}^{i_k}\right), \ \mathbf{g}^{i_{k+1}} = \hat{\boldsymbol{\gamma}}_{i_k}^{i_{k+1}}\mathbf{g}^{i_k}. \end{aligned} \quad (3)$$

Δt_k is the time interval of two RF measurements. $\mathbf{g}^{i_0}$ is the initial gravity representation in the IMU frame, which can be set through a sampling method [4]. The predicted relative orientation is $\hat{\boldsymbol{\gamma}}_{i_k}^{i_{k+1}}$.

Then we take the RF measurements to update the predicted position and orientation. We first consider the measurements from l^{th}UWB node. Taking the translational direction from IMU and rendering the RF-based quantity Δd_l observes the relative translation $\mathbf{t}_{i_{k+1}}^{i_k} = \frac{\mathbf{p}_{i_{k+1}}^{i_k}}{\left|\mathbf{p}_{i_{k+1}}^{i_k}\right|}\Delta d_l$.

On the other hand, we convert the azimuth angle $\hat{\theta}_{lk}$ to be a quaternion $\mathbf{q}_{u_k}^{u_l}$. We mark $(\cdot)^{u_l}$ as the measurement in the UWB frame with respect to node l. $(\cdot)_{u_k}^{u_l}$ is the k^{th}measurement in frame u_l. We can use it to correct the IMU-derived rotation by minimizing the following cost function:

$$\mathbf{r}_{i_{k+1}}^{i_0} = \min_{\mathbf{q}_{i_{k+1}}^{i_0}} \left[\left(\hat{\mathbf{q}}_{u_l}^{i} \otimes \hat{\mathbf{q}}_{u_k}^{u_l}\right)^{-1} \otimes \mathbf{q}_{i_{k+1}}^{i_0} \otimes \hat{\boldsymbol{\gamma}}_{i_k}^{i_{k+1}} \right], \quad (4)$$

where $\hat{\mathbf{q}}_{u_l}^{i}$ is the relative rotation from UWB frame u_l to IMU frame, which can be calibrated with the first observed $\tilde{\mathbf{q}}_{u_k}^{u_l}$ and the orientation at that moment $\tilde{\mathbf{q}}_{i_k}^{i_0}$, $\hat{\mathbf{q}}_{u_l}^{i} = \tilde{\mathbf{q}}_{i_k}^{i_0} \otimes \left(\tilde{\mathbf{q}}_{u_k}^{u_l}\right)^{-1}$. Then the observed relative rotation $\mathbf{r}_{i_{k+1}}^{i_k} = \left(\mathbf{r}_{i_k}^{i_0}\right)^{-1} \otimes \mathbf{r}_{i_{k+1}}^{i_0}$.

With the observation $\mathbf{t}_{i_{k+1}}^{i_k}$ and $\mathbf{r}_{i_{k+1}}^{i_k}$, we can update the predicted $\mathbf{p}_{i_{k+1}}^{i_k}$ and $\hat{\boldsymbol{\gamma}}_{i_k}^{i_{k+1}}$ via KF. We omit the detailed derivation of KF since it is a standard tool for combating Gaussian noise. When multiple UWB nodes connect to the onboard UWB tag, each can estimate a relative pose.

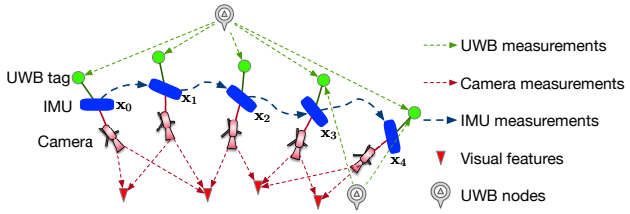


Fig. 6. An illustration of the bundle adjustment formulation with UWB, visual, and IMU measurements.

We choose the pose produced by the node transmitting the strongest signal.

Based on the RF-based relative pose, we reproject the visual features from k^{th}frame to $(k+1)^{\text{th}}$frame via epipolar geometry (refer to Figure 2). This operation forms a set of RF-matched features, each of which associates with an optical-matched feature. Finally, we rank the features concerning the distance between the RF-matched and optical-matched results. The top-rank feature should have a minimum difference. Each feature associates with a score, *i.e.*, the normalized image distance between its optical-matched and RF-matched results. Based on our abundant tests, we empirically define a threshold $\epsilon = 0.4$ and allow features whose scores less than ϵ to participate in the state estimation.

B. RF-Visual-Inertial Sensor Fusion

The sifted features in § IV-A encodes more reliable pose information. We now fuse such features with RF and IMU measurements for better state estimation. The IMU has already been used to assist the feature sifting (Eqn. (3)) by taking a short-term integration of IMU to compute relative translations. The temporal drift of such an integration is negligible [36]. In the long-term, we need to continuously fuse IMU measurements to bring metric information to visual odometry and improve the robustness when visual cues are lost occasionally. The IMU temporal drift can be corrected by adequately fusing the RF and visual measurements.

An illustration of our RF-visual-inertial sensor fusion is shown in Figure 6. We formulate a new BA paradigm that aims to find a configuration of state parameters that best match all measurement constraints. During the long-term flight, the system requires to track not only position and velocity but also the vehicle's orientation. Moreover, we can recover the depth of sifted visual features for sparse mapping. The full state vector in a bundle can be defined as

$$\begin{aligned} \boldsymbol{\mathcal{X}} &= [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m, \lambda_1, \lambda_2, \cdots, \lambda_o]^\top \\ \mathbf{x}_k &= \left[\mathbf{p}_{i_k}^{i_0}, \mathbf{v}_{i_k}^{i_k}, \mathbf{q}_{i_k}^{i_0}\right]^\top, \ k \in [1, n], \end{aligned} \quad (5)$$

where $\mathbf{x}_k \in \mathbb{R}^{10}$ denotes the state in the bundle when obtaining k^{th}keyframe, which includes position $\mathbf{p}_{i_k}^{i_0} \in \mathbb{R}^3$, velocity $\mathbf{v}_{i_k}^{i_k} \in \mathbb{R}^3$, and orientation $\mathbf{q}_{i_k}^{i_0} \in \mathbb{R}^4$. $\mathbf{u}_l$ denotes the position of l^{th}UWB node. m is the total number of the UWB nodes connected by the states in the bundle. λ_η is the depth of η^{th}visual feature from its first observation. There are o features tracked in the bundle. We fix the size of the bundle to ensure enough multi-view constraints and bound the computation complexity.

Since we optimize the orientation $\mathbf{q}_{i_k}^{i_0}$ during the long-term flight, the BA problem becomes highly nonlinear. We solve the state vector by minimizing the Mahalanobis norm of all measurement residuals:

$$\min_{\mathcal{X}} \left\{ \sum_{(l,j)\in\mathcal{U}} \left\| \mathbf{e}_{\mathcal{U}} \left(\hat{\mathbf{z}}_l^{u_j}, \mathcal{X} \right) \right\|^2_{\mathbf{P}_l^{u_j}} + \sum_{(\eta,j)\in C} \left\| \mathbf{e}_C \left(\hat{\mathbf{z}}_\eta^{c_j}, \mathcal{X} \right) \right\|^2_{\mathbf{P}_\eta^{c_j}} + \sum_{k\in\mathcal{I}} \left\| \mathbf{e}_{\mathcal{I}} \left(\hat{\mathbf{z}}_{i_{k+1}}^{i_k}, \mathcal{X} \right) \right\|^2_{\mathbf{P}_{i_{k+1}}^{i_k}} \right\}, \tag{6}$$

where $\mathcal{U}$ denotes the set of UWB measurements in the bundle, C the set of observed visual features, $\mathcal{I}$ the set of IMU measurements. $\mathbf{e}_{\mathcal{U}} \left(\hat{\mathbf{z}}_l^{u_j}, \mathcal{X} \right)$, $\mathbf{e}_C \left(\hat{\mathbf{z}}_\eta^{c_j}, \mathcal{X} \right)$, and $\mathbf{e}_{\mathcal{I}} \left(\hat{\mathbf{z}}_{i_{k+1}}^{i_k}, \mathcal{X} \right)$ are measurement residuals of UWB, camera, and IMU, respectively. We choose the Mahalanobis norm as the objective because it rescales measurements by their covariance, enabling fair correlations of parameters in different scales. These correlations are key for any high-precision system [37].

We solve this problem using the Gauss-Newton algorithm implemented by Ceres Solver [38], which is an open-source C++ library for solving complicated optimization problems. To use this tool, we need to 1) provide an initialization point of the state vector to bootstrap the iterative solution; 2) linearize the nonlinear system and derive the first-order Jacobian matrix to error state to define a templated functor that computes the residuals. We adopt the initialization method from [29]. Next, we derive the residuals, their Jacobians, and covariance matrices. We operate on the error state representation with the above definition to linearize the system (6).

UWB measurement model. The UWB tag [33] provides range $\hat{d}_{lj}$ and azimuth angle $\hat{\theta}_{lj}$. We first convert the azimuth angle into an orientation representation $\hat{\mathbf{q}}_{u_j}^{u_l}$ in the UWB frame with respect to node l using roll and pitch angles provided by IMU. They are accurate as IMU only drifts in the azimuth direction for rotation [4]. The residual $\mathbf{e}_{\mathcal{U}} \left(\hat{\mathbf{z}}_l^{u_j}, \mathcal{X} \right)$ (briefly denote $\mathbf{e}_l^{u_j}$) can be expressed as

$$\mathbf{e}_l^{u_j} = \begin{bmatrix} \hat{d}_{lj}^2 - \left(\mathbf{u}_l - \mathbf{p}_{i_j}^{i_0} - \mathbf{R}_{i_j}^{i_0} \hat{\mathbf{p}}_u^i \right)^\top \left(\mathbf{u}_l - \mathbf{p}_{i_j}^{i_0} - \mathbf{R}_{i_j}^{i_0} \hat{\mathbf{p}}_u^i \right) \\ 2 \left[\left(\hat{\mathbf{q}}_{u_l}^i \otimes \hat{\mathbf{q}}_{u_j}^{u_l} \right)^{-1} \otimes \mathbf{q}_{i_j}^{i_0} \right]_{xyz} \end{bmatrix}, \tag{7}$$

where $\mathbf{u}_l$ is the unknown position of l^{th}UWB node and $\hat{\mathbf{q}}_{u_l}^i$ obtained in § IV-A. $\hat{\mathbf{p}}_u^i$ is the relative position between the onboard UWB tag and the IMU, which can be manually calibrated once the UWB tag and the IMU are installed on the platform. $\mathbf{q}_{xyz}$ extracts the vector part of the quaternion.

Its Jacobian is $\mathbf{J}_l^{u_j} = \begin{bmatrix} \frac{\partial \mathbf{e}_l^{u_j}}{\partial \delta \mathbf{x}_j} & \frac{\partial \mathbf{e}_l^{u_j}}{\partial \delta \mathbf{u}_l} \end{bmatrix}$. The residual covariance $\mathbf{P}_l^{u_j} \in \mathbb{R}^{4\times4}$ is the diagonal matrix whose entries are the noise of UWB measurements, which can be determined by a statistical analysis of measurements.

Camera measurement model. The pinhole camera model has been studied by [34]. The difference in our context is that we consider the frame transformation between IMU and camera using calibrated extrinsic parameters, relative translation $\hat{\mathbf{p}}_i^c$ as well as rotation $\hat{\mathbf{R}}_i^c$, making the model more realistic.

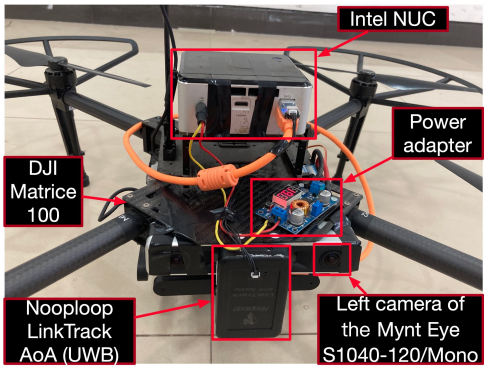


Fig. 7. The prototype of RFSift.

Given η^{th}sifted visual feature observed in j^{th}camera frame $\hat{\mathbf{z}}_\eta^{c_j} = [\hat{x}_\eta^{c_j}, \hat{y}_\eta^{c_j}]^\top$. Suppose it is first observed in h^{th}frame, the residual of this feature in j^{th}frame $\mathbf{e}_C \left(\hat{\mathbf{z}}_\eta^{c_j}, \mathcal{X} \right)$ (briefly denote $\mathbf{e}_\eta^{c_j}$) is its reprojection error. The corresponding 3D point of this feature $\mathbf{f}_\eta^{c_j}$ is $\mathbf{f}_\eta^{c_j} = [x_{f_\eta}^{c_j}, y_{f_\eta}^{c_j}, z_{f_\eta}^{c_j}]^\top = \mathbf{R}_{c_0}^{c_j} \left(\mathbf{p}_{c_h}^{c_0} - \mathbf{p}_{c_j}^{c_0} + \lambda_\eta \mathbf{R}_{c_h}^{c_0} \hat{\mathbf{w}}_\eta^{c_h} \right)$, where $\hat{\mathbf{w}}_\eta^{c_h} = [\hat{x}_\eta^{c_h}, \hat{y}_\eta^{c_h}, 1]^\top$ is the homogeneous coordinate of the feature. $\mathbf{p}_{c_x}^{c_0} = \hat{\mathbf{R}}_i^c \left(\mathbf{p}_{i_x}^{i_0} - \hat{\mathbf{p}}_c^i \right) - \hat{\mathbf{R}}_i^c \mathbf{R}_{i_x}^{i_0} \hat{\mathbf{R}}_c^i \hat{\mathbf{p}}_i^c$ for $\mathbf{p}_{c_h}^{c_0}$ and $\mathbf{p}_{c_j}^{c_0}$. $\mathbf{R}_{c_x}^{c_0} = \hat{\mathbf{R}}_i^c \mathbf{R}_{i_x}^{i_0} \hat{\mathbf{R}}_c^i$ for $\mathbf{R}_{c_h}^{c_0}$ and $\mathbf{R}_{c_j}^{c_0}$. Its reprojection point is $\boldsymbol{\zeta}_\eta^{c_j} = \left[x_{f_\eta}^{c_j} / z_{f_\eta}^{c_j}, y_{f_\eta}^{c_j} / z_{f_\eta}^{c_j} \right]$. Then the residual can be defined as

$$\mathbf{e}_\eta^{c_j} = \boldsymbol{\zeta}_\eta^{c_j} - \hat{\mathbf{z}}_\eta^{c_j}. \tag{8}$$

The Jacobian can be computed by taking standard partial derivatives. The residual covariance $\mathbf{P}_\eta^{c_j} \in \mathbb{R}^{2\times2}$ is the diagonal matrix whose entries are the noise of visual feature measurements.

The **IMU measurement model** has been studied by [4], [13], [34]. We omit the details here for brevity.

V. System Implementation and Evaluation

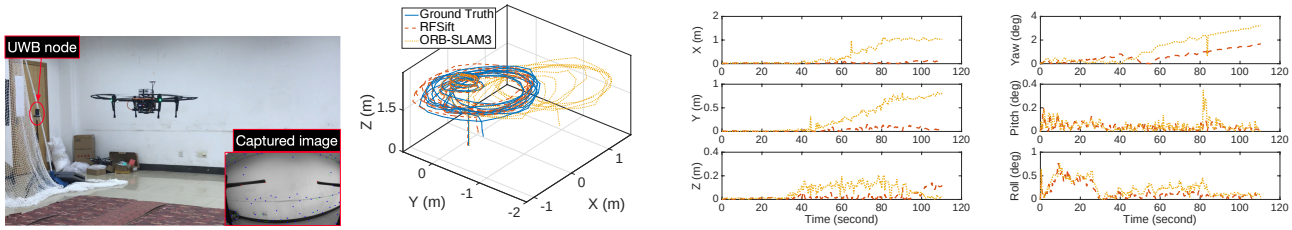
A. System Implementation and Experiment Setup

We implement RFSift on an Intel NUC with a 1.8 GHz Core i5 processor with 4 cores, running Ubuntu 16.04 LTS. A Mynt Eye S1040-120/Mono camera and a Nooploop LinkTrack AoA UWB node are attached to the NUC. An IMU has been integrated into the Mynt Eye camera. The NUC is equipped on a DJI M100 platform, and the vehicle's battery powers it. All the sensors of our system are commercially available. The prototype is shown in Figure 7.

We conduct experiments in two venues. We first test RFSift in an 8×6 m^2 indoor drone test site of our lab for indoor experiments. OptiTrack [39] provides the ground truth of the MAV odometry in the lab. Then, we carry out a large-scale experiment through a mixed indoor and outdoor setting to demonstrate RFSift's long-term practicality.

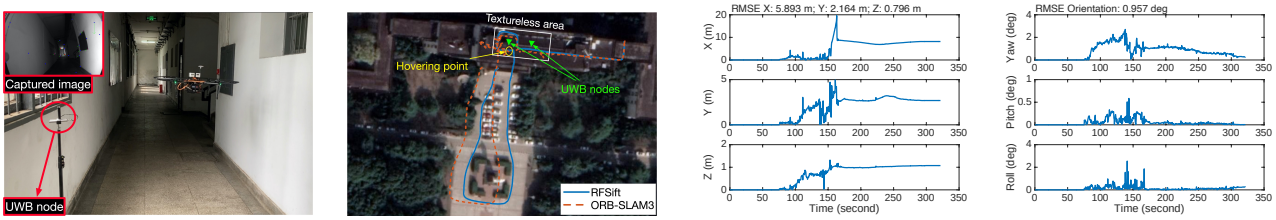
B. Experiment Result and Comparison

1) Experiments in a textureless indoor site: In the indoor test site, we deploy one UWB node on the site's edge to assist the navigation. The Mynt Eye camera is facing the white wall of the site to ensure the textureless challenge. We fly 10 rounds to obtain statistical data and summarize the results in Table I. It shows that RFSift achieves the best



(a) Flight facing the white wall (b) Circular trajectory (c) Position error (d) Orientation error

Fig. 8. Experiment in the textureless indoor site.



(a) Flight through dim corridors (b) Trajectory aligned with Google Maps (c) Difference in position (d) Difference in orientation

Fig. 9. Experiment inside and outside our academic building.

TABLE I

RMS ATE COMPARISON WITH DIFFERENT APPROACHES IN A TRAJECTORY OF 55.915 M.

	VINS-FUSION	ORB-SLAM3	**RFSift**	W/o UWB	W/o sifted features
$\mathbf{p}_{i_k}^{i_0}(x)$	70.5 cm	59.3 cm	**4.8 cm**	15.5 cm	44.3 cm
$\mathbf{p}_{i_k}^{i_0}(y)$	48.9 cm	41.1 cm	**5.3 cm**	19.2 cm	18.8 cm
$\mathbf{p}_{i_k}^{i_0}(z)$	11.1 cm	8.7 cm	**4.0 cm**	9.4 cm	13.7 cm
Orient.	2.142°	1.684°	**0.803°**	1.515°	3.014°

performance in textureless scenes in terms of the root mean square (RMS) of absolute trajectory error (ATE), compared with ORB-SLAM3 [6] and VINS-FUSION [40]. It also shows that the sifted visual features (without UWB) can improve the position accuracy about 2.75× than ORB-SLAM3. On the other hand, if we only use UWB measurements (without sifted visual), the system reduces to conventional RF-based solutions with decimeter-level (50.04 cm) accuracy.

Figure 8 shows the performance of state estimation over time. The trajectory lasts 111 seconds with length 55.915 m. The final position error over the trajectory for RFSift is 9.18 cm, and the final orientation error is 1.69°. We use the rosbag package of ROS to record all sensor data during the flight for running ORB-SLAM3 for comparison. The final errors of ORB-SLAM3 are 78.93 cm and 3.23°.

2) Experiments in a large-scale environment: In this experiment, we go out of the lab and test RFSift through the academic building at dusk. Some corridors are dim and textureless. We apply RFSift for feedback control of the vehicle. The MAV starts from a landing on the first floor of the building. It is initialized in the texture-rich landing where there are stairs and doors with proper lighting. Then the MAV flies through a dim corridor. We deploy two UWB nodes (only one is visible in Figure 9 (a)) on the two ends of the corridor to assist the navigation. Then, the MAV flies out the building into a texture-rich open field. Finally, the MAV flies around a statue in front of the building without UWB nodes. The trajectory length is more than 300.255 meters and the flight lasts about 320 seconds.

Figure 9 shows the state estimation results of ORB-SLAM3 and RFSift. Although we do not have the ground truth in such a large-scale experiment, the performance can still be visually inspected. We can see that the trajectory is smooth and can be appropriately aligned with Google's satellite map. We again use the rosbag package to record the sensor data over the trajectory and run ORB-SLAM3 by the record. ORB-SLAM3 works fine in texture-rich areas but drifts in the textureless area. Since the trajectory does not have any loop, the drift cannot be corrected. The results in Figure 9 (c) and (d) show that the drift happens from 73 to 164 seconds, corresponding to the flight in the textureless area. In particular, during 150 – 160 seconds, the situation worsens because of the hover of the MAV. The final drift of ORB-SLAM3 is $[8.16, 2.69, -1.07]$ m.

VI. CONCLUSION

This paper presents RFSift, a novel state estimation system towards robust MAV navigations in general use cases. It fills the gap of accurate state estimation in textureless regions by two new designs. One is the RF-sifting algorithm that quantifies and sifts well-matched visual features to prevent ill-matched ones' harm. The second is the RF-visual-inertial sensor fusion method that enables drift-free accurate state estimation. We implement RFSift on a DJI Matrice 100 platform with a UWB node, a monocular camera, and an IMU. The experiments demonstrate the effectiveness of working in textureless regions. Our future work will extend RFSift to a self-contained system that can work in the wild without external support.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under Grant 2019YFB180003400, 2020YFB1806606, National Science Foundation of China with Grant 62071194, 61729101, 91738202, Young Elite Scientists Sponsorship Program by CAST under Grant 2018QNRC001.

References

[1] A. Imdoukh, A. Shaker, A. Al-Toukhy, D. Kablaoui, and M. El-Abd, "Semi-autonomous indoor firefighting uav," in *Proc. IEEE ICRA*, 2017, pp. 310–315.

[2] M. Power, "Walmart testing warehouse drones to manage inventory," https://www.supplypro.ca/wal-mart-testing-drones-warehouses-manage-inventory/, 2018, online; accessed 16 June 2020.

[3] Y. Lin, F. Gao, T. Qin, W. Gao, T. Liu, W. Wu, Z. Yang, and S. Shen, "Autonomous aerial navigation using monocular visual-inertial fusion," *J. Field Robot.*, vol. 35, no. 1, pp. 23–51, 2018.

[4] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular slam system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] C. Campos, R. Elvira, J. J. Gómez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *arXiv preprint arXiv:2007.11898*, 2020.

[7] G. De Croon and C. De Wagter, "Challenges of autonomous flight in indoor environments," in *Proc. IEEE/RSJ IROS*, 2018.

[8] Z. Zhang and D. Scaramuzza, "Perception-aware receding horizon navigation for mavs," in *Proc. IEEE ICRA*, 2018, pp. 2534–2541.

[9] A. Loquercio, A. I. Maqueda, C. R. Del-Blanco, and D. Scaramuzza, "Dronet: Learning to fly by driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1088–1095, 2018.

[10] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE CVPR*, 2018, pp. 1983–1992.

[11] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE CVPR*, 2017, pp. 1851–1858.

[12] Y. Meng, Y. Lu, A. Raj, S. Sunarjo, R. Guo, T. Javidi, G. Bansal, and D. Bharadia, "Signet: Semantic instance aided unsupervised 3d geometry perception," in *Proc IEEE CVPR*, 2019, pp. 9810–9820.

[13] S. Zhang, W. Wang, N. Zhang, and T. Jiang, "Rf backscatter-based state estimation for micro aerial vehicles," in *Proc. IEEE INFOCOM*, 2020.

[14] R. Liu, C. Yuen, T.-N. Do, D. Jiao, X. Liu, and U.-X. Tan, "Cooperative relative positioning of mobile users by fusing imu inertial and uwb ranging information," in *Proc. IEEE ICRA*, 2017, pp. 5623–5629.

[15] J. Li, Y. Bi, K. Li, K. Wang, F. Lin, and B. M. Chen, "Accurate 3d localization for mav swarms by uwb and imu fusion," in *Proc. IEEE ICCA*, 2018, pp. 100–105.

[16] S. Zhang, W. Wang, and T. Jiang, "Wifi-inertial indoor pose estimation for micro aerial vehicles," *IEEE Transactions on Industrial Electronics*, 2020.

[17] B. Li, S. Zhang, and S. Shen, "CSI-based WiFi-inertial state estimation," in *Proc. IEEE MFI*, 2016.

[18] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.

[19] 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*, vol. 24, no. 6, pp. 381–395, 1981.

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

[21] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 422–429, 2019.

[22] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[23] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[24] M. Pizzoli, C. Forster, and D. Scaramuzza, "Remode: Probabilistic, monocular dense reconstruction in real time," in *Proc. IEEE ICRA*, 2014, pp. 2609–2616.

[25] S. Maity, A. Saha, and B. Bhowmick, "Edge slam: Edge points based monocular visual slam," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2408–2417.

[26] J. Li, C. Li, and R. Zheng, "3d indoor localization with commercial-o-the-shelf ultra-wide band radios," 2018.

[27] H. E. Nyqvist, M. A. Skoglund, G. Hendeby, and F. Gustafsson, "Pose estimation using monocular vision and inertial sensors aided with ultra wide band," in *Proc. IEEE IPIN*, 2015, pp. 1–10.

[28] C. Wang, H. Zhang, T.-M. Nguyen, and L. Xie, "Ultra-wideband aided fast localization and mapping system," in *Proc. IEEE IROS*, 2017, pp. 1602–1609.

[29] H. Xu, L. Wang, Y. Zhang, K. Qiu, and S. Shen, "Decentralized visual-inertial-uwb fusion for relative state estimation of aerial swarm," in *Proc. IEEE ICRA*, 2020.

[30] J. Shi *et al.*, "Good features to track," in *Proc. IEEE CVPR*, 1994, pp. 593–600.

[31] "The euroc mav dataset," https://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets, Online; Accessed: 18 June, 2020.

[32] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[33] "Nooploop: Linktrack aoa," https://www.nooploop.com/en/linktrack-aoa/, Online; Accessed: 22 June, 2020.

[34] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs," in *Proc. IEEE ICRA*, 2015.

[35] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.

[36] H. Nyqvist and F. Gustafsson, "A high-performance tracking system based on camera and imu," in *Proc. FUSION*. IEEE, 2013.

[37] 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*, vol. 34, no. 3, pp. 314–334, 2015.

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

[39] "Optitrack – motion capture systems," https://optitrack.com/, Online; Accessed: 01 July, 2020.

[40] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," *arXiv preprint arXiv:1901.03638*, 2019.