2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

# Tightly-Coupled Single-Anchor Ultra-wideband-Aided Monocular Visual Odometry System

Thien Hoang Nguyen, Thien-Minh Nguyen*, and Lihua Xie, *Fellow, IEEE*

***Abstract*— In this work, we propose a tightly-coupled odometry framework, which combines monocular visual feature observations with distance measurements provided by a single ultra-wideband (UWB) anchor with an initial guess for its location. Firstly, the scale factor and the anchor position in the vision frame will be simultaneously estimated using a variant of Levenberg-Marquardt non-linear least squares optimization scheme. Once the scale factor is obtained, the map of visual features is updated with the new scale. Subsequent ranging errors in a sliding window are continuously monitored and the estimation procedure will be reinitialized to refine the estimates. Lastly, range measurements and anchor position estimates are fused when needed into a pose-graph optimization scheme to minimize both the landmark reprojection errors and ranging errors, thus reducing the visual drift and improving the system robustness. The proposed method is implemented in Robot Operating System (ROS) and can function in real-time. The performance is validated on both public datasets and real-life experiments and compared with state-of-the-art methods.**

## I. Introduction

In recent years, it can be seen that vision-based localization methods such as visual odometry (VO) and simultaneous localization and mapping (VSLAM) have become an integral part of robotics research. As technology progresses, many lightweight, energy-efficient but high performance cameras have come into the field and offer strong advantages over more expensive and heavier alternative sensors for SLAM, such as laser and LiDAR. While a wide variety of sensors have been studied in various VO and SLAM systems, e.g., stereo camera, infrared (IR) and thermal cameras, laser scanner LiDAR, etc., traditional monocular camera systems are still attracting great interests from the community due to its high flexibility and easy integration with many mobile platforms. A detailed comparison of monocular, stereo and multi-camera visual odometry pipelines [1] shows that among these approaches, monocular setup would be the most preferred solution for many real-world applications where constraints on size and weight put hard restriction on the kind of sensors and computational resources that can be carried by the robot. More specifically, monocular VO setup does not require sufficiently large baseline between cameras like stereo/multi-camera setups, and while visual-inertial alternatives can employ the ubiquitous IMU sensor on Micro Aerial Vehicle (MAV), the low-cost nature of the sensor often makes good performance unattainable. An additional high-quality IMU sensor is still required, which adds a major cost and extra layer of complexity to the system.

In exchange for the flexibility and low demand on computational resources, two particular challenges need to be addressed when using a monocular VO/SLAM system. The first is scale ambiguity, whereby from a sequence of images provided by one camera, one can only obtain the knowledge of relative scale about 3D distances in the perceived environment and not metric scale. For monocular systems, scale ambiguity is inherent since depth information of 3D scene is lost when projected onto 2D frame. All estimates of camera positions and a 3D representation of the environment are therefore calculated "up to a scale". The second issue is "scale drift" [2] where the scale factor has to be adjusted in different regions of the map. In principle, initial scale estimate in monocular odometry can be corrected by adjusting the values of parameters. However, since scale is arbitrary in each run even with the same algorithm, this is not a viable solution for fast deployable platform like MAV.

To overcome these challenges, this work aims to leverage point-to-point range measurements provided by UWB sensor such that not only a metric scale factor can be obtained, but also the scale drift would be corrected since scale correction is performed continuously along the MAV's trajector.. Among various wireless technologies such as ZigBee, BLE or Wifi, UWB is chosen thanks to its properties of strong multi-path resistance and accurate ranging measurement in indoor, GPS-denied and cluttered environment [3], [4].

## II. Literature Review

### A. Methods for metric scale correction

Many methods have been proposed to address the scale problems in the literature. For metric scale recovery, the common solution is to introduce at least one additional sensor that provides metric measurements of some kind into the monocular system. Through bundle adjustment (BA) in the back-end task [5]–[7], one can find an optimal solution for the whole trajectory, then the drifts can be corrected, at the expense of high computational cost. However, if no loops are encountered in the trajectory of the robot, scale drift would still be an issue for the front-end odometry.

In stereo [8] or multi-camera [9] setups, since fixed baseline length is provided, depth estimates can be directly computed and metric scale can thus be obtained. However, depth estimation range is highly dependent on baseline length, distances from objects to baseline, and calibration accuracy between cameras. When the scene is far away or objects are too close to one of the cameras, stereo vision will deteriorate

The authors are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, 50 Nanyang Avenue.
*Email of corresponding author: thienminh.npn@ieee.org

A video summarizing the main concepts and experimental results of this paper can be viewed at https://youtu.be/Z8hDW6Zf8io.

to monocular case. Significant computational resources are also required to perform various image processing tasks.

A popular approach is incorporating sensors that can provide range measurements such as 1D/2D/3D LiDAR [10]–[12], RGB-D camera [13], ultrasonic altimeter, radar etc., each of which effectively provides distance measurements in different dimensions. By exploiting the right properties in each configuration, the scale can be accurately recovered. On the other hand, considering that each sensor comes with its own limitation, the applicable scenario might be restricted after the fusion of odometry and range data [10], [14]. For example, when 1D LiDAR is used, the camera is assumed to face a flat surface. Most 2D and 3D LiDAR sensors are relatively heavy, power hungry and require high computational power, which is not available on a lightweight MAV. Depth resolution and range of RGB-D camera is often limited and deteriorate in outdoor scenes.

It can be seen that the most successful method to acquire metric-scaled odometry data is the visual-inertial odometry (VIO), where IMU measurements and VO are fused together to estimate both ego-motion and map coordinates in true scale. Many state-of-the-art approaches [15]–[17] can demonstrate very high accuracy and robustness in a wide range of environments. Nonetheless, coupling of IMU and camera often requires carefully supervised initial states and accurate multi-sensor calibration (including intrinsic and extrinsic calibration parameters, IMU biases characterization, time-synchronization between sensors), so that local minima [18] or divergence can be avoided. Notably, VINS-Mono [15] proposed an in-flight initialization routine to solve the problem of obtaining precise camera-IMU calibrations. Even then, the performance is heavily dependent on the quality of the IMU sensor, of which the size and cost are often major considerations in practical applications.

Other approaches to recover scale include learning the scene depth, dimensions of objects [19]–[21] or applying an adaptive control strategy [22]. Depth learning-based methods [19], [20] are applicable in a wide range of scenarios, but requires large amount of data and high computing power platform for training and deployment. If a known target is exploited as a priori information for initialization [21], once the robot moves away from the original scene the scale drift problem will not be addressed. An interesting solution is controlling the robot's movements to recover metric scale through observations on control gains [22], in which neither additional sensors nor high computation resources are needed. However, an accurate motion model of the robot in the vertical axis is necessary, and the robot will have to spend its limited flight time on correcting the scale with vertical motions multiple times throughout the mission.

### B. UWB-based and UWB-Aided Localization

Range-based localization such as UWB has the capability to overcome the shortcomings of vision-based methods in reflective or featureless environments [23]. Other methods of localization can also employ UWB data to improve estimation accuracy, such as VO [23], [24], LiDAR-based [25] or RGB-D-based [26]. Furthermore, when fused with other sensors, UWB is able to provide centimeter level of localization accuracy while being robust to multipath and non-line-of-sight (NLOS) effects [3], [27]. However, these approaches assume prior knowledge of anchor positions and require at least four UWB anchors for full 3D localization which might not be practical in cluttered, dynamic environments like industrial facilities and warehouses.

Without a setup of sufficient number of anchors for full 3D localization, UWB distance measurements between robots in a formation have been used for cooperative relative localization and control problem in [28]. The use of single UWB anchor placed at an arbitrary unknown position was studied in [29], [30] for a distance-based docking problem of MAVs without the need of visual information. In [31], an optimization-based approach was proposed to perform scale and orientation correction of different trajectories with 1D UWB distance measurements between points on those trajectories. However, this method only considers movement in 2D case and required a height sensor. Furthermore, the whole trajectory was used to reach a desirable solution, thus the performance of the system was only verified offline.

### C. Main Contributions

In summary, the main contributions of this work include:

- a variant of the Levenberg-Marquardt method in which given an initial guess of the UWB anchor position, the original up-to-scale position from monocular odometry is tightly combined with distance measurements to simultaneously estimate the scale factor and anchor position, effectively addressing the scale ambiguity problem;
- a real-time monitoring process of ranging errors in a sliding window to trigger the estimation procedure to continuously refine the scale and anchor position estimates, hence the scale drift problem is diminished;
- an extension of the ORB-SLAM [5] pose-graph optimization scheme that aims to minimize both the landmark reprojection errors and ranging errors, which aims to reduce the visual drift and improve the overall robustness of the system.

The advantages of such system are: 1) a monocular camera would greatly alleviate the complexity of mechanical and software design, compared to stereo/multi-camera setup that requires large baseline or good IMU-camera intrinsic/extrinsic calibrations for VIO setup, 2) the proposed approach requires the least number of UWB sensors and with unknown positions as opposed to UWB-only localization methods. However, it is noted that the solution would be limited by the maximum range of the UWB sensor and line-of-sight (LOS) conditions of the environment. Additionally, movements on a sphere centered at the anchor position, in theory, would render the scale unobservable. As such, the system is most effective if the environment permits LOS from the anchor to the robot, and the robot's movement is established on multiple axes.

## III. Problem Formulation

### A. System Overview

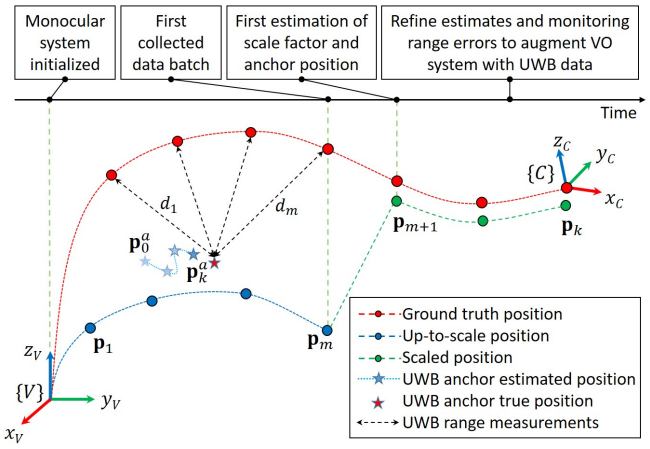



Fig. 1: Overview of the coordinate frames and the operating phases of the system.

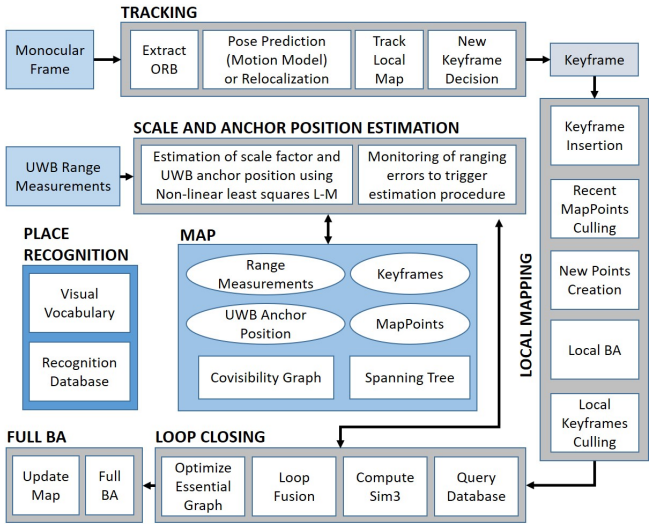



Fig. 2: The architecture of the proposed system, based on the monocular ORB-SLAM framework [5].

*1) Basic definitions:* As depicted in Fig. 1, let $\{C\}$ and $\{V\}$ be the camera and vision coordinate frame, respectively. The origin of $\{V\}$ corresponds to $\{C\}$ at the camera's initial pose, i.e. $\{V\} \widehat{=} \{C\}_{t=0}$. At time instance $k$, the odometry output for position of our system is $\mathbf{p}_k = [p_k^x,\ p_k^y,\ p_k^z]^\top \in \mathbb{R}^3$ in $\{V\}$ frame, the nearest associated range measurement from the UWB anchor to UWB sensor on the MAV is $d_k \in \mathbb{R}$, and $\mathbf{p}_k^a = [p_k^{ax},\ p_k^{ay},\ p_k^{az}]^\top \in \mathbb{R}^3$ is the estimated UWB anchor position in $\{V\}$ frame, with an initial guess $\mathbf{p}_0^a$. The range measurement $d_k$ is obtained by multiplication of light speed $c$ and time of flight, which is measured from the time stamps $T_{M1}^{Rx}$ and $T_{M0}^{Tx}$ of when the UWB ranging signal is sent and received. Taking into account the processing time delay $\sigma_k$ of the UWB sensors, we have

$$d_k = c\frac{T_{M1}^{Rx} - T_{M0}^{Tx} - \sigma_k}{2} + \eta_k \tag{1}$$

with $\eta_k \sim \mathcal{N}(0, \Omega_k)$ is assumed to be a zero mean Gaussian noise [23]. $d_k$ is directly referred as distance from UWB anchor to camera since the camera and UWB sensor are attached on a rigid body and the translational offset between UWB antenna and camera is considered neglectable. In this work, two-way TOF UWB sensor is used for it does not require clock synchronization between anchors.

*2) Operating phases:* The overall architecture of the proposed system is illustrated in Fig. 2. A new thread for estimating the scale factor and the anchor position as well as monitoring ranging errors is augmented on top of the existing ORB-SLAM [6] framework. The operation consists of the following three phases:

1) After the monocular odometry pipeline is initialized, a data $\mathcal{N}_m$ is collected, which composes of $m$ arbitrarily scaled odometry positions and their nearest range measurements, i.e. $\mathcal{N}_m = \{(\mathbf{p}_i{}^\top, d_i)\}_{i=1}^{i=m}$.
2) Once $\mathcal{N}_m$ is filled, an estimation procedure as described in III-B is carried out to estimate both $s$ and $\mathbf{p}^a$ concurrently. An initial guess $\mathbf{p}_0^a$ is provided for the anchor position but not for the scale factor ($s_0 = 1$). After each successful run, the visual map is updated with the new scale factor, with the position of all the keyframes and map points rescaled.
3) The sum of ranging errors of the data in $\mathcal{N}_m$ is monitored to reinitialize phase 2 with the current anchor position estimates if it rises above a certain threshold. As discussed in III-E, if the anchor position is stable over a pre-defined number of runs, the position is added as a vertex with subsequent ranges as edges to the current pose in the pose-graph optimization scheme.

### B. Non-linear least squares regression

Let $s \in \mathbb{R}^1$ be the scale factor, such that metric position can be recovered with $|s|\,\mathbf{p}_k$. The absolute operator is used to restrain sign flip of scale estimates without an explicit constraint $s > 0$, as explained in III-D. We have

$$\begin{aligned} d_k = \|\mathbf{p}_k^a - |s_k|\,\mathbf{p}_k\| &= \left\| \begin{bmatrix} p_k^{ax} \\ p_k^{ay} \\ p_k^{az} \end{bmatrix} - |s_k| \begin{bmatrix} p_k^x \\ p_k^y \\ p_k^z \end{bmatrix} \right\| \\ &= \left\| \begin{bmatrix} -p_k^x & 1 & 0 & 0 \\ -p_k^y & 0 & 1 & 0 \\ -p_k^z & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} |s_k| \\ p_k^{ax} \\ p_k^{ay} \\ p_k^{az} \end{bmatrix} \right\| \\ &= \|A_k \boldsymbol{\beta}_k\|, \end{aligned} \tag{2}$$

with $\|*\|$ denotes Euclidean norm of the argument vector, all the unknown parameters are grouped into a vector $\boldsymbol{\beta}_k = (s_k, p_k^{ax}, p_k^{ay}, p_k^{az})^\top$. Given a set of $m$ data $\mathcal{N}_m = \{(\mathbf{p}_i{}^\top, d_i)\}_{i=1}^{i=m}$, our aim is to find the vector of parameters

$$\boldsymbol{\beta}_k = (s_k, p_k^{ax}, p_k^{ay}, p_k^{az})^\top, \tag{3}$$

where $k$ is the time index of the latest data added to the window, that minimizes the sum of squared errors

$$E_k^r = \sum_{i=k-m}^{k} {r_i}^2(\boldsymbol{\beta}_k). \tag{4}$$

Let be $y_i = d_i$, $f(\boldsymbol{\beta}_k) = \|A_k \boldsymbol{\beta}_k\|$, the residuals are

$$r_i(\boldsymbol{\beta}_k) = {y_i}^2 - f(\boldsymbol{\beta}_k)^2. \tag{5}$$

With the cost function (4), the optimal values for $\mathbf{p}^a$ and $s$ can be obtained through the minimization of

$$\boldsymbol{\beta}_k^* = \arg\min_{\boldsymbol{\beta}_k} E_k^r. \tag{6}$$

**Remark III.1.** *The following formulas for the residuals $r_i$ have been validated in experiments:*

$$r_i = d_i - \|\mathbf{p}_i^a - |s|\,\mathbf{p}_i\|,$$
$$r_i = {d_i}^2 - \|\mathbf{p}_i^a - |s|\,\mathbf{p}_i\|^2,$$

*neither of which showed significant and consistent improvements compare to the other based on experimental results of RMSE errors. However, the calculation of the Jacobian matrix from the former involves a square root in the denominator, which might lead to the division by zero issue.*

### C. Existence of solution

For UWB sensors that use time-of-flight technology, range measurement $d_k$ will be the shortest when there is direct LOS path between the antennas. Taking multipath and non-LOS effects into account, the actual measurements would tend to be higher than ideal cases due to propagation delay. With that insight, we can then rewrite equation (2) as

$$d_k \geq \|A_k \boldsymbol{\beta}_k\|, \text{or } {\boldsymbol{\beta}_k}^\top {A_k}^\top A_k \boldsymbol{\beta}_k \leq {d_k}^2. \tag{7}$$

The problem can be viewed as finding a solution within the intersection of the convex regions created by each range measurement constraints since ${A_k}^\top A_k$ is positive semi-definite. With $d_k > 0$, a feasible solution can be found since the point $\boldsymbol{\beta}_k = 0$ is contained in each of the convex regions. It has been shown that the optimization problem of similar structure will have a convex hull defined by convex constraints, and the global optima must lie on the convex hull [31].

### D. Modified Levenberg-Marquardt algorithm

In this section we describe our variant and modifications of the Levenberg-Marquardt (L-M) algorithm [32] to solve the optimization problem (6).

*1) Recursive LM algorithm:* At time instance $k$, starting with an initial guess $\boldsymbol{\beta}_k^{(0)}$, the parameter vector $\boldsymbol{\beta}_k$ is refined after each iteration $l$ by applying:

$$\boldsymbol{\beta}_k^{(l+1)} = \boldsymbol{\beta}_k^{(l)} + \boldsymbol{\gamma}\boldsymbol{\Delta\beta}, \tag{8}$$

where $\boldsymbol{\Delta\beta}$ is the shift vector and $\boldsymbol{\gamma}$ is the dynamic learning rate. $\boldsymbol{\Delta\beta}$ can be found via solving the equation:

$$(\mathbf{J}^\top \mathbf{J} + \lambda \mathbf{I})\boldsymbol{\Delta\beta} = \mathbf{J}^\top [\mathbf{y} - \boldsymbol{f}(\boldsymbol{\beta}_k)], \tag{9}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{J} = [J_{ij}]$ with $\mathbf{J}_{ij} = -\partial r_i / \partial \beta_j$ being the Jacobian matrix, $\lambda$ is a non-negative damping parameter, adjusted at each iteration $l$.

*2) Dynamic learning rate:* To incorporate the static nature of the anchor, we apply a time-varying learning rate $\boldsymbol{\gamma} = \mathrm{diag}(1, \gamma(k), \gamma(k), \gamma(k))$ to the components of the shift vector $\boldsymbol{\Delta\beta}$ corresponding to the anchor position $\mathbf{p}^a$, with

$$\gamma(k) = \gamma_0 e^{-(k/\tau)}, \tag{10}$$

where $\gamma_0$ is an initial value, $\tau$ is a time constant and $k$ is the number of run times. In essense, $\mathbf{p}^a$ should be initially estimated with large gradient and fine-tuned with smaller and smaller gradient as $\mathbf{p}^a$ becomes more precisely determined.

*3) Sign-bounded cost function:* All distance measurements would be unchanged if the trajectory and the position of the anchor are mirrored on one plane or axis of $\{V\}$. In such cases, signs of scale estimates $s$ can be flipped but the cost function value will still be the same, resulting in local minimas. To address this issue, one can keep the sign of $s$ positive or fix the position of anchor as known parameters, with the former being the only viable option as the true anchor position in $\{V\}$ is unknown. Since the theory of LM algorithm does not define a way to handle explicit bound constraints, the scale-corrected position is formulated as $|s|\,\mathbf{p}_k$ so that final position output never changes sign.

*4) Start and stop conditions:* Except for the first batch of data, a condition is checked when new odometry data $\tilde{\mathbf{p}}$ arrives. If the new position data is outside a pre-defined diameter $\rho$ from the last point in the window $\mathcal{N}_m$, i.e.

$$\|\tilde{\mathbf{p}} - \mathbf{p}_m\| > \rho, \tag{11}$$

then it will be added to the window. Otherwise, $\tilde{\mathbf{p}}$ is discarded and previous estimations are carried on until the next data is received. For the first run or whenever $E_k^r$ rises above a threshold, the estimation procedure is triggered and will run iteratively until the number of iterations $l$ has exceeded a pre-defined limit $L$ or a convergence criterion defined as

$$\left| \left[ E_k^{r(l-1)} - E_k^{r(l)} \right] / E_k^{r(l-1)} \right| < \zeta, \tag{12}$$

is met, where $E_k^{r(l)}$ is the sum of squared errors at data sample $k$ and iteration $l$, $\zeta = 0.0001$ is a numerical constant.

### E. Joint optimization for UWB-aided visual odometry

Updating scale factor would not address the inevitable drift that exists in vision-only estimator. In this work, we propose reducing the visual drift by utilizing camera and UWB measurements in a tightly coupled manner. In a loosely coupled scheme [33], image and range measurements are processed separately to produce a up-to-scale position and a scale factor, which are then multiplied to obtain the final position. The proposed tightly coupled approach would compute the odometry output directly from the images and range measurements, which would improve the robustness of the system since range data is a reliable source of constraint to facilitate position tracking as well as reducing visual drift.

Every new $j$-th keyframe is augmented by the nearest range measurement $d_j$ and estimated anchor position $\mathbf{p}_j^a$. The

cost function for the optimization scheme in the local bundle adjustment thread would include the range error constraints:

$$\arg\min_{\mathbf{p}_i} E^v + \sum_{i=0}^{P}\sum_{j=0}^{K}\rho(\left\|\mathbf{p}_j^a - \mathbf{p}_i\right\| - d_j), \qquad (13)$$

where $P$ is the number of optimized camera poses, $K$ is the number of keyframes associated with range constraints. $E^v$ is the cost of visual measurements which uses the weighted norm of the reprojection error, as originally described in [5], and $\rho(.)$ is the Huber loss function that ensures the effects of outliers and noise can be diminished.

## IV. Experimental Results

In this section, we present results performed on EuRoC MAV datasets [34] and real-life experiments. The proposed system was implemented in ROS[1] with loop-closure thread disabled. The results are validated with Root Mean Square Error (RMSE) of absolute translation error (ATE) and relative pose error (RPE) [35], which indicates the global consistency and drift of the estimated trajectory, respectively. The system runs in real-time for all of the experiments.

### A. Public datasets

Since UWB distance measurements are not available in the EuRoC datasets, UWB anchor is simulated as a virtual anchor placed at the origin of the *ground truth frame*. Range data is the Euclidean norm of ground truth positions added Gaussian noise with standard deviation $\Omega = 0.05$. The simulated UWB runs at the same frequency as the ground truth position data (20Hz). Initial guess for the UWB anchor position is always set at $\mathbf{p}_0^a = (0.5, 0.5, 0)^\top$ in all experiments, which would gradually approach to the true location of the simulated anchor in $\{V\}$ frame. The initial guess for the scale factor is always set as $s_0 = 1$, since the original scale is arbitrary. The algorithm generally performs better with increasing size of sliding window, but in our evaluations the size of $\mathcal{N}_m$ is fixed to $m = 100$ data points, which typically spans over 4s of image streams. It is noted that given the same initial parameters and over the same dataset, the final RMSE value might vary depends on how far the initial arbitrary scale is from the true value. Thus, 5 trials are conducted on each dataset and the reported RMSE results are the average of all the RMSEs obtained. Fig. 3 shows the simulated UWB range measurements, position output on each axis and overall 3D trajectory in one of the experiments performed on MH_01 dataset. Table I shows that the proposed method can achieve the same level of accuracy for RMSE, with results of state-of-the-art methods from [36].

### B. Indoor and outdoor real-life experiments

The proposed system was further validated with real-life experiments where results are summarized in Table II. Hardware setup consists of an Intel NUC i7, a stereo camera[2] and P440 UWB sensor[3], which provide images at 30Hz and ranges at 40Hz. While only the left image stream and UWB data are used by the proposed system, stereo images are recorded and tested offline with ORB-SLAM2 Stereo [6].

The indoor testing area is a 6m x 6m room equipped with Vicon system. All experiments have different trajectories and anchor positions. Potential outliers in range data are rejected and a smoothing filter is applied before fusion. Size of $\mathcal{N}_m$ is set at $m = 100$, which corresponds to $3.5s$ of data. Initial guess for scale is identity, while initial guess for anchor position is measured from the starting point. Other

[1] https://www.ros.org/
[2] https://www.mynteye.com/
[3] https://www.humatics.com/products/scholar/

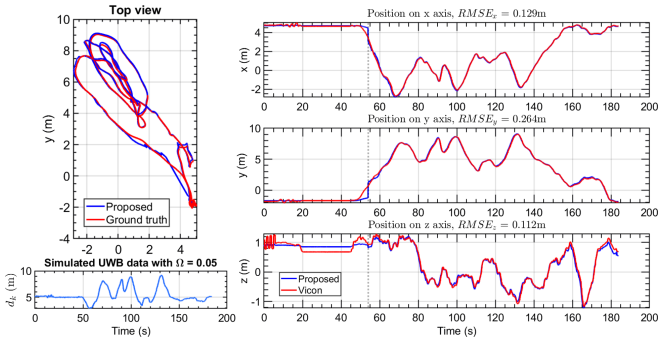



Fig. 3: Results with MH_01 dataset: proposed (blue) and ground truth (red) trajectories. The first batch of data is collected at t=54s, which is 4s after the MAV starts moving.

| Sequence | Proposed | OKVIS | ROVIO | VINS-Mono |
|---|---|---|---|---|
| MH_01 | **0.16** | **0.16** | 0.21 | 0.27 |
| MH_02 | **0.11** | 0.22 | 0.25 | 0.12 |
| MH_03 | 0.15 | 0.24 | 0.25 | **0.13** |
| MH_04 | 0.25 | 0.34 | 0.49 | **0.23** |
| MH_05 | **0.24** | 0.47 | 0.52 | 0.35 |
| V1_01 | 0.18 | 0.09 | 0.10 | **0.07** |
| V1_02 | 0.25 | 0.20 | **0.10** | **0.10** |
| V1_03 | 0.32 | 0.24 | 0.14 | **0.13** |

TABLE I: Comparison of RMSE values of ATE (m) on EuRoC datasets. The best results are highlighted in **bold**.

| Sequence | ATE (m) | | RPE (m/s) | |
|---|---|---|---|---|
| | Proposed | ORB-Stereo | Proposed | ORB-Stereo |
| HH_01 | **0.23** | 0.25 | 0.04 | **0.2** |
| HH_02 | **0.20** | 0.24 | 0.04 | **0.03** |
| HH_03 | **0.09** | 0.24 | **0.02** | 0.03 |
| HH_04 | 0.24 | **0.22** | **0.02** | 0.03 |
| HH_05 | 0.28 | **0.21** | 0.06 | **0.02** |
| HH_06 | **0.18** | 0.23 | 0.04 | **0.02** |

TABLE II: Comparison of RMSE of ATE (m) and RPE (m/s) on indoor experiments. The best results are in **bold**.

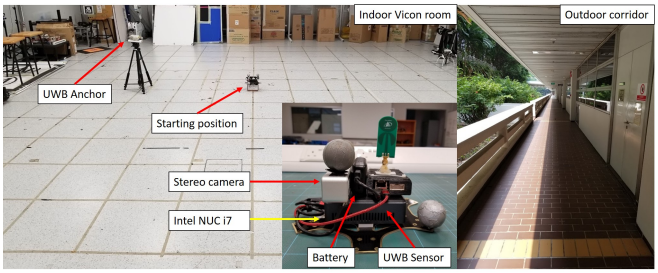



Fig. 4: Hardware setup and experimental environments: 6m × 6m Vicon room (left) and 170m long corridor (right).

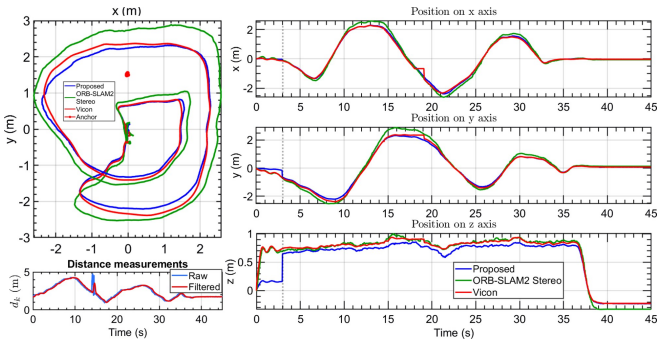



Fig. 5: Results with HH_06 experiment: proposed (blue), ORB-SLAM2 Stereo (green), ground truth trajectories (red).

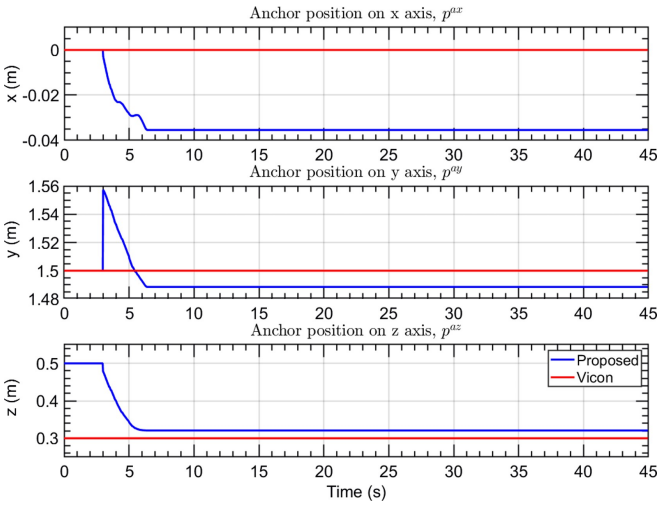



Fig. 6: Estimates of the UWB anchor position in HH_06 experiment, which is visible along the trajectory in Fig. 5.

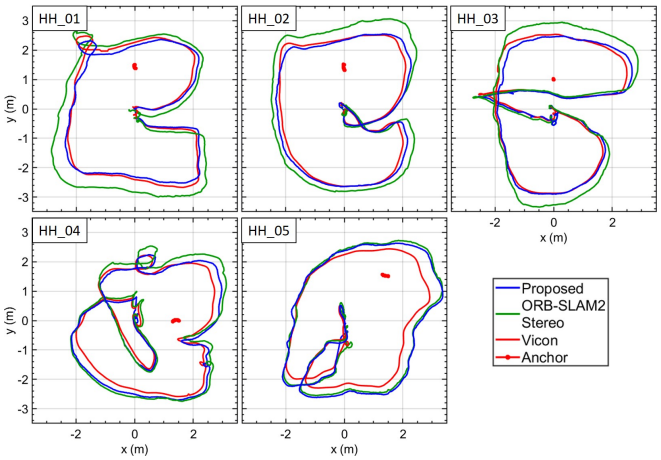
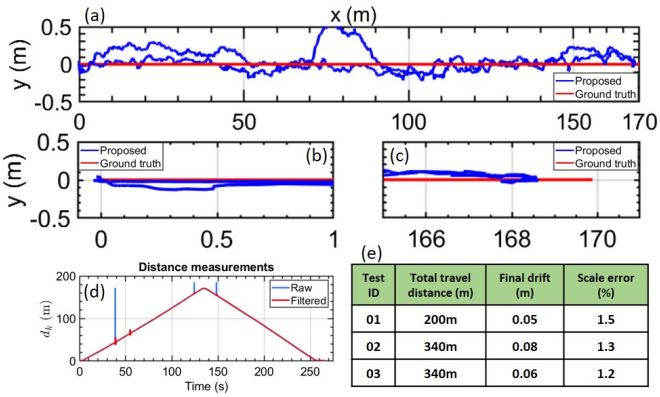



Fig. 7: Trajectories and results of the indoor experiments.

parameters are tuned selectively. Fig. 5 depicts the position in HH_06 experiments. After the first batch of data is collected at $t = 3.5s$, the odometry position on each axis as well as the anchor position can be seen to quickly approach the true values, oscillate for a short period before stabilizing with little change thereafter. Fig. 6 shows that with the initial guess $\mathbf{p}_0^a = (0, 1.5, 0.5)^\top$, $\mathbf{p}^a$ converges to the true position in $3s$. Since first movement of the MAV is usually in z




| Test ID | Total travel distance (m) | Final drift (m) | Scale error (%) |
|---|---|---|---|
| 01 | 200m | 0.05 | 1.5 |
| 02 | 340m | 0.08 | 1.3 |
| 03 | 340m | 0.06 | 1.2 |

Fig. 8: Results with outdoor experiments: for test 03, (a) shows the full trajectory, (b)-(c) are close-up views of the two ends of the trajectory, (d) is UWB distance measurements, (e) shows the results of multiple trials on the same corridor.

axis, initial guess for $p^{az}$ often has a larger error to the true value than for the other axes. Both methods work well since the visual features in the environment and the experiments are relatively small-scale. In Table II, it is evident that the proposed method can achieve the same level of accuracy as ORB-SLAM2 Stereo.

Outdoor experiments were carried out in a $170\text{m}$ long corridor to showcase the ability to reduce drift and the robustness of the system. In these experiments, the anchor was placed at roughly $[-1.3, 0.5, 1.5]^\top$. The system was initialized at one end of the corridor, then walked to the other end and back. Since the movements were mostly on a straight line, UWB range measurements are sufficient as ground truth for scale evaluation. The proposed system was able to consistently follow $340\text{m}$ long trajectories with less than $0.1\text{m}$ drift at the end, while the scale error between the estimated and ground truth trajectories is less than $1.5\%$, with the results can be seen in Fig. 8. ORB-SLAM2 Stereo failed to complete any of the experiments, due to losing track at some instances with challenging movements and illumination condition during the trials, and thus was not available for comparison.

## V. Conclusions and Future Works

In this work, a tightly-coupled fusion scheme is proposed to use both up-to-scale monocular visual odometry and UWB ranging measurements to estimate and refine metric scale and anchor position. Range errors are monitored to incorporate range measurements in the pose graph optimization scheme to reduce visual drift in a tightly-coupled manner. The accuracy is on par with state-of-the-art stereo and VIO systems with datasets and indoor experiments, while drifts in large-scale outdoor experiments are notably diminished.

However, the proposed method still requires an initial guess for the anchor positions and excludes the spatial offsets between the sensors. Eliminating the necessity for this initial guess, taking into account the extrinsic and temporal offsets between sensors, and extending to multiple anchors cases are the main directions for future development.

## REFERENCES


[1] K. Mohta, M. Watterson, Y. Mulgaonkar, S. Liu, C. Qu, A. Makineni, K. Saulnier, K. Sun, A. Zhu, J. Delmerico *et al.*, "Fast, autonomous flight in gps-denied and cluttered environments," *Journal of Field Robotics*, vol. 35, no. 1, pp. 101–120, 2018.

[2] H. Strasdat, J. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular slam," *Robotics: Science and Systems VI*, vol. 2, 2010.

[3] T. M. Nguyen, A. H. Zaini, K. Guo, and L. Xie, "An ultra-wideband-based multi-uav localization system in gps-denied environments," in *2016 International Micro Air Vehicles Conference*, 2016.

[4] T.-M. Nguyen, A. H. Zaini, C. Wang, K. Guo, and L. Xie, "Robust target-relative localization with ultra-wideband ranging and communication," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2312–2319.

[5] M. J. M. M. Mur-Artal, Raúl and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] 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.

[7] G. Dubbelman and B. Browning, "Cop-slam: closed-form online pose-chain optimization for visual slam," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1194–1213, 2015.

[8] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1935–1942.

[9] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, "Towards robust visual odometry with a multi-camera system," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1154–1161.

[10] Z. Zhang, R. Zhao, E. Liu, K. Yan, and Y. Ma, "Scale estimation and correction of the monocular simultaneous localization and mapping (slam) based on fusion of 1d laser range finder and vision data," *Sensors*, vol. 18, no. 6, p. 1948, 2018.

[11] Q. Lv, J. Ma, G. Wang, and H. Lin, "Absolute scale estimation of orb-slam algorithm based on laser ranging," in *2016 35th Chinese Control Conference (CCC)*. IEEE, 2016, pp. 10 279–10 283.

[12] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular camera localization in 3d lidar maps," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1926–1931.

[13] C. Kerl, J. Stuckler, and D. Cremers, "Dense continuous-time tracking and mapping with rolling shutter rgb-d cameras," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2264–2272.

[14] R. Giubilato, S. Chiodini, M. Pertile, and S. Debei, "Scale correct monocular visual odometry using a lidar altimeter," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3694–3700.

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

[16] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *Robotics: Science and Systems*, 2015.

[17] 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.

[18] J. Kaiser, A. Martinelli, F. Fontana, and D. Scaramuzza, "Simultaneous state initialization and gyroscope bias calibration in visual inertial aided navigation," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 18–25, 2017.

[19] K. Tateno, F. Tombari, I. Laina, and N. Navab, "Cnn-slam: Real-time dense monocular slam with learned depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017.

[20] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2043–2050.

[21] D. Eberli, D. Scaramuzza, S. Weiss, and R. Siegwart, "Vision based position control for mavs using one single circular landmark," *Journal of Intelligent & Robotic Systems*, vol. 61, no. 1-4, pp. 495–512, 2011.

[22] S. H. Lee and G. de Croon, "Stability-based scale estimation for monocular slam," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 780–787, 2018.

[23] C. Wang, H. Zhang, T.-M. Nguyen, and L. Xie, "Ultra-Wideband Aided Fast Localization and Mapping System," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017.

[24] J. Tiemann, A. Ramsey, and C. Wietfeld, "Enhanced uav indoor navigation through slam-augmented uwb localization," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[25] Y. Song, M. Guan, W. P. Tay, C. L. Law, and C. Wen, "Uwb/lidar fusion for cooperative range-only slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6568–6574.

[26] F. J. Perez-Grau, F. Caballero, L. Merino, and A. Viguria, "Multi-modal mapping and localization of unmanned aerial robots based on ultra-wideband and rgb-d sensing," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3495–3502.

[27] T. H. Nguyen, M. Cao, T.-M. Nguyen, and L. Xie, "Post-mission autonomous return and precision landing of uav," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2018, pp. 1747–1752.

[28] T.-M. Nguyen, Z. Qiu, T. H. Nguyen, M. Cao, and L. Xie, "Distance-based cooperative relative localization for leader-following control of mavs," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3641–3648, 2019.

[29] T.-M. Nguyen, Z. Qiu, M. Cao, T. H. Nguyen, and L. Xie, "An integrated localization-navigation scheme for distance-based docking of uavs," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5245–5250.

[30] T. M. Nguyen, Z. Qiu, M. Cao, T. H. Nguyen, and L. Xie, "Single landmark distance-based navigation," *IEEE Transactions on Control Systems Technology*, 2019.

[31] A. Shariati, K. Mohta, and C. J. Taylor, "Recovering relative orientation and scale from visual odometry and ranging radio measurements," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 3627–3633.

[32] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of applied mathematics*, vol. 2, no. 2, pp. 164–168, 1944.

[33] T. H. Nguyen, T.-M. Nguyen, M. Cao, and L. Xie, "Loosely-coupled ultra-wideband-aided scale correction for monocular visual odometry," *Unmanned Systems*, 2020.

[34] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[35] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 573–580.

[36] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2502–2509.