Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Smartphone Indoor Positioning through Coordinate Alignment and Error State Fusion of Visual Inertial and WiFi Estimates

Jiaxuan Wu
Shenyang Ligong University
Ziyang He
Shenyang Ligong University
Yueqiu Jiang
jiangyueqiu.sylu@gmail.com

Shenyang Ligong University

Research Article

Keywords: indoor positioning, visual-inertial navigation, Wi-Fi fingerprinting, multisensor fusion, error-state Kalman filtering, coordinate alignment, mobile computing

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11154773/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Smartphone Indoor Positioning through Coordinate Alignment and Error State Fusion of Visual Inertial and WiFi Estimates

Jiaxuan Wu[1], Ziyang He[1], Yueqiu Jiang[1*]

[1]School of Information Science and Engineering, Shenyang Ligong University, Shenyang, China.

*Corresponding author(s). E-mail(s): jiangyueqiu.sylu@gmail.com;
Contributing authors: jiaxuanw@sylu.edu.cn; 2503610218@sylu.edu.cn;

Abstract

Integrating visual-inertial and wireless position estimates requires a common coordinate frame and a fusion model that accommodates different update rates. This paper presents a position-level fusion method for smartphone indoor positioning that combines a visual-inertial navigation system (VINS) with Wi-Fi fingerprinting. A least-squares rigid transformation aligns the local VINS trajectory with the Wi-Fi fingerprint coordinate frame, and an error-state extended Kalman filter (ES-EKF) integrates the aligned VINS estimates with intermittent Wi-Fi position observations. Data were collected using a Xiaomi Mi 9 smartphone and processed offline with VINS-Mono. Five repeated trials were conducted in each of two scenarios with trajectories of approximately 50 m and 150 m. In the short-trajectory experiment, VINS+Wi-Fi achieved a root mean square error (RMSE) of 0.3429 $\pm$ 0.038 m, compared with 0.3821 $\pm$ 0.052 m for VINS and 0.7468 $\pm$ 0.088 m for Wi-Fi, a 10.3% reduction relative to VINS. In the long-trajectory experiment, the corresponding RMSE values were 0.5909 $\pm$ 0.079 m, 0.8006 $\pm$ 0.121 m, and 2.1626 $\pm$ 0.345 m, a 26.2% reduction relative to VINS. Values are reported as means and standard deviations across trials. These results support the use of coordinate alignment and asynchronous position-level fusion to reduce positioning error under the tested static, single-floor conditions. The prototype required eight Wi-Fi access points and a preconstructed fingerprint database; real-time on-device processing was not evaluated.

Keywords: indoor positioning, visual-inertial navigation, Wi-Fi fingerprinting, multisensor fusion, error-state Kalman filtering, coordinate alignment, mobile computing

1 Introduction

Indoor positioning supports mobile navigation, asset tracking, and location-based services in environments where global navigation satellite system (GNSS) signals are unavailable or unreliable [1]. For smartphone-based systems, positioning involves estimating motion from heterogeneous sensor streams with different coordinate frames, update rates, and error characteristics. The accuracy of a single sensing modality can degrade because of signal variability, environmental complexity, or accumulated estimation error [2, 3]. Multi-source approaches combine complementary measurements to improve positioning accuracy and continuity [4, 5].

Visual-inertial odometry (VIO) combines image observations with inertial measurements to estimate relative motion in GNSS-denied environments [6–10]. Image features constrain geometric drift, whereas the inertial measurement unit (IMU) provides high-rate motion information between image frames. Nevertheless, VIO and VINS remain locally referenced navigation methods and are susceptible to accumulated drift, particularly with smartphone-grade sensors or sparse visual features [11, 12]. Wi-Fi fingerprinting provides position observations in a predefined building coordinate frame and can therefore supply global corrections [13]. Its accuracy and update rate are affected by received-signal-strength (RSS) variability, multipath propagation, and environmental change [14]. These complementary characteristics motivate a position-level architecture in which aligned VINS estimates provide continuous local motion and Wi-Fi observations provide intermittent global corrections. Their relative contributions are determined by the modeled process and measurement uncertainties.

Several multi-source fusion approaches have been reported. Pritzl et al. [15] combined VIO with LiDAR relative localization using a linear Kalman filter and reported a mean three-dimensional absolute trajectory error of 0.28 m, but the system required a separate LiDAR-equipped aerial platform. Jung et al. [16] proposed U-VIO, which tightly integrated UWB ranging with VIO in a factor graph and reported position errors below approximately 0.2 m on a custom unmanned ground vehicle. Gao and Lian [17] combined visual, inertial, and UWB information through graph optimization with line-feature and non-line-of-sight processing. Kong et al. [14] combined BLE observations with monocular VINS on a smartphone, while Wen and Feng [18] used Wi-Fi fingerprinting for coarse localization with VIO-based refinement. Smartphone approaches that fuse pedestrian dead reckoning, magnetic measurements, and Wi-Fi have also been investigated [5]. Because these studies use different platforms, environments, trajectories, and error metrics, their numerical results provide context rather than a controlled ranking. Table 1 therefore summarizes architectural and deployment differences without treating the reported values as directly comparable.

Table 1: Comparison of representative multi-source indoor positioning methods

Method	Sensors	Fusion strategy	Platform / infrastructure	Reported metric	Main limitation
Pritzl et al. [15]	VIO + LiDAR	Linear KF (loose)	Two UAVs; 3D LiDAR	0.28 m mean 3D ATE	Specialized LiDAR platform
Jung et al. [16]	VIO + UWB	Factor graph (tight)	UGV; UWB anchor	<0.2 m position error	UWB infrastructure
Kong et al. [14]	VINS + BLE	EKF (loose)	Smartphone; BLE beacons	Not directly comparable	BLE infrastructure; different protocol
Wen and Feng [18]	VIO + Wi-Fi	Loosely coupled	Smartphone; Wi-Fi radio map	Not reported	Coarse Wi-Fi localization
Chen et al. [5]	PDR + magnetic + Wi-Fi	EKF (loose)	Smartphone; Wi-Fi radio map	Not directly comparable	PDR and map dependence
Proposed	VINS + Wi-Fi	ES-EKF (position level)	Xiaomi Mi 9; 8 APs; radio map	0.343 / 0.591 m RMSE	2D output; offline processing

This study investigates coordinate alignment and asynchronous position-level fusion of VINS and Wi-Fi estimates for smartphone indoor positioning. The contribution lies in the system formulation, implementation, and experimental evaluation of established estimation techniques. Specifically, the study provides: (1) an explicit error-state filtering formulation with nominal and error states, a state-transition model, an observation model, and covariance matrices; (2) a least-squares rigid-alignment procedure that registers the local VINS trajectory to the Wi-Fi fingerprint frame before horizontal position fusion; and (3) an evaluation using Xiaomi Mi 9 recordings from approximately 50 m and 150 m trajectories, with five repeated trials per scenario and comparisons against standalone VINS and Wi-Fi. A contextual comparison with representative fusion approaches clarifies differences in architecture and deployment requirements. The fusion RMSE values were 0.3429 ± 0.038 m and 0.5909 ± 0.079 m, corresponding to reductions of 10.3% and 26.2%, respectively, relative to VINS.

2 Materials and Methods

2.1 System Architecture

The system comprises four processing stages, as illustrated in Fig. 1. First, the smartphone records images, inertial measurements, and Wi-Fi RSS observations. Second, tightly coupled VIO processes the camera and IMU data offline to generate local position estimates at approximately 10 Hz, while Wi-Fi fingerprinting provides building-frame position estimates at an effective rate of approximately 1 Hz. Third, a least-squares rigid transformation aligns the local VINS trajectory with the Wi-Fi fingerprint frame, after which only the horizontal coordinates are retained. Finally, a position-level ES-EKF integrates the aligned estimates using VINS-rate prediction and Wi-Fi-triggered correction. Sections 2.2–2.4 describe the filtering, Wi-Fi localization, and coordinate-alignment procedures.

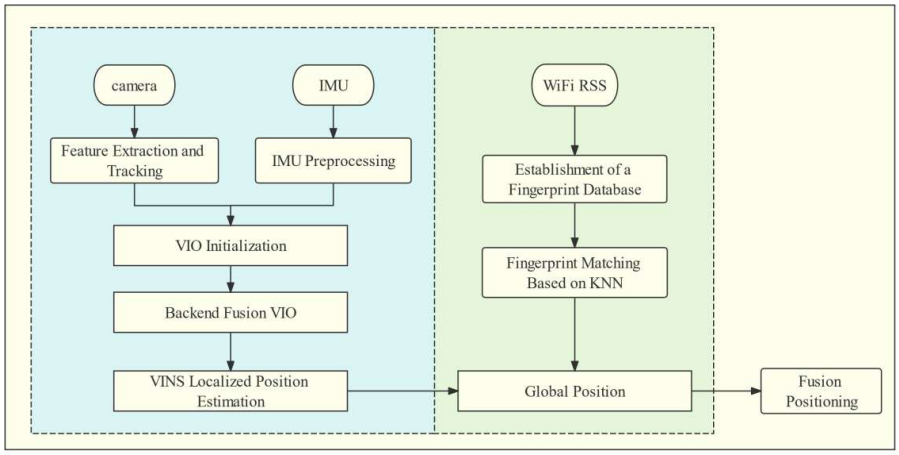


Fig. 1 Overall VINS-Wi-Fi fusion positioning framework

2.2 Error-State Extended Kalman Filter (ES-EKF)

An error-state extended Kalman filter (ES-EKF) represents the estimate as a nonlinear or externally propagated nominal state plus a small error state [19, 20]. The filter propagates the nominal estimate separately and applies Kalman prediction and correction to the linearized error state; the estimated error is then injected into the nominal state and reset. This formulation is common in inertial navigation because small perturbations can be modeled locally without repeatedly estimating the full state in additive coordinates [20]. In the present position-level implementation, both the constant-velocity error dynamics and the Wi-Fi position observation model are linear. The resulting algorithm is therefore equivalent in practice to a linear error-state Kalman filter; the ES-EKF designation is retained to describe the nominal-state injection and reset architecture. Fig. 2 summarizes the workflow.

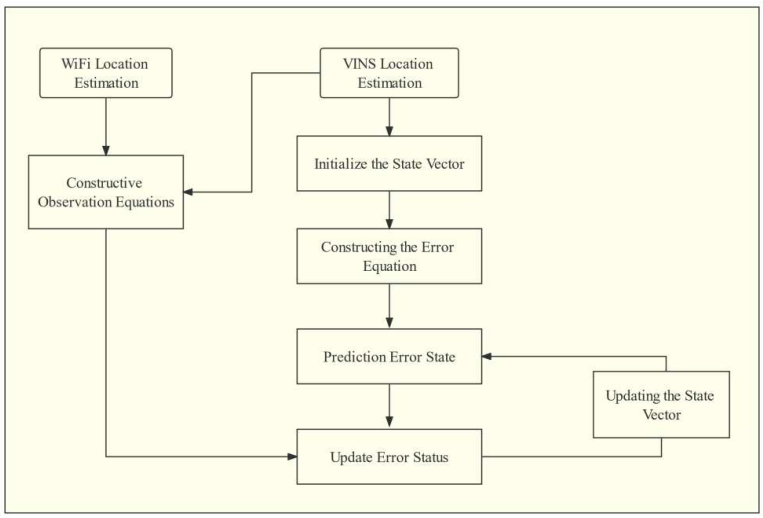


Fig. 2 Error-state extended Kalman filter workflow

2.2.1 Nominal State and Error State

The Wi-Fi fingerprint coordinate frame is used as the global frame. At each update, the system state comprises horizontal position and velocity. The full state is represented as the sum of a nominal state and a small error state:

$$\mathbf{x}_k = \hat{\mathbf{x}}_k + \delta\mathbf{x}_k \tag{1}$$

The nominal state contains the horizontal position and velocity components. The associated error state contains the corresponding position and velocity perturbations:

$$\hat{\mathbf{x}}_k = \begin{bmatrix} p_{x,k} & p_{y,k} & v_{x,k} & v_{y,k} \end{bmatrix}^{\mathsf{T}} \tag{2}$$

$$\delta\mathbf{x}_k = \begin{bmatrix} \delta p_{x,k} & \delta p_{y,k} & \delta v_{x,k} & \delta v_{y,k} \end{bmatrix}^{\mathsf{T}} \tag{3}$$

2.2.2 Error-State Dynamic Model

The nominal state is updated from the aligned VINS position and velocity outputs. A discrete constant-velocity model is used for the position-velocity error state, and unmodeled planar acceleration is represented by process noise:

$$\delta\mathbf{x}_k = F\delta\mathbf{x}_{k-1} + \mathbf{w}_k \tag{4}$$

The 4×4 state-transition matrix is:

$$F = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

Here, Δt is the elapsed time between successive prediction steps and is computed from the VINS timestamps (approximately 0.1 s for the nominal 10 Hz VINS output). Wi-Fi corrections arrive less frequently, at an effective interval of approximately 1 s. The process noise is modeled as zero-mean Gaussian noise with covariance Q_k. With q_a denoting the planar acceleration-noise spectral density and I_2 the 2×2 identity matrix, the discrete constant-velocity process model is:

$$\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, Q_k) \tag{6}$$

$$Q_k = q_a \begin{bmatrix} \frac{\Delta t^4}{4} I_2 & \frac{\Delta t^3}{2} I_2 \\ \frac{\Delta t^3}{2} I_2 & \Delta t^2 I_2 \end{bmatrix} \tag{7}$$

where q_a is an empirically tuned acceleration-noise spectral-density parameter. The implementation used $q_a = 0.5\ \mathrm{m^2 \cdot s^{-3}}$; this filter-tuning value was informed by the observed device noise but was not obtained by directly substituting the Table 3 accelerometer noise density into Equation (7).

2.2.3 Observation Model

Consistent with the definition of the error state as true state minus nominal state, the observation residual is defined as the Wi-Fi fingerprint position minus the aligned VINS position:

$$\mathbf{z}_k = \mathbf{p}_{\mathrm{WiFi},k} - \mathbf{p}_{\mathrm{VINS},k} \tag{8}$$

The observation model relating the observation to the error state is linear:

$$\mathbf{z}_k = H \delta \mathbf{x}_k + \mathbf{v}_k \tag{9}$$

The corresponding 2×4 observation matrix is:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{10}$$

The measurement noise is modeled as zero-mean Gaussian noise with covariance R_k. Assuming identical and uncorrelated horizontal Wi-Fi-position variances, R_k is specified as:

$$\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, R_k) \tag{11}$$

$$R_k = \begin{bmatrix} \sigma^2_{\mathrm{WiFi}} & 0 \\ 0 & \sigma^2_{\mathrm{WiFi}} \end{bmatrix} \tag{12}$$

The implementation uses a fixed nominal horizontal Wi-Fi position variance of $0.25\ \mathrm{m^2}$ per axis, corresponding to a standard deviation of 0.5 m. This value is a filter-tuning parameter rather than an empirical summary of the scenario-level Wi-Fi RMSE. The same covariance is used in both experiments; adaptive or scenario-dependent covariance estimation is outside the scope of this study.

2.2.4 Filter Prediction and Update Steps

Prediction Step: At each time step k, the error state and its covariance are predicted as follows:

$$\delta \mathbf{x}_k^- = F \delta \mathbf{x}_{k-1} \tag{13}$$

$$P_k^- = F P_{k-1} F^\mathsf{T} + Q_k \tag{14}$$

Because the error state is reset after each correction, its prior value is initialized to zero before the next prediction step.

Update Step: When a new Wi-Fi position becomes available, the innovation covariance S_k and Kalman gain K_k are first computed. The innovation residual r_k is then used to update the error state and its covariance:

$$\begin{aligned} S_k &= H P_k^- H^\mathsf{T} + R_k, \\ K_k &= P_k^- H^\mathsf{T} S_k^{-1} \end{aligned} \tag{15}$$

$$\begin{aligned} \mathbf{r}_k &= \mathbf{z}_k - H \delta \mathbf{x}_k^-, \\ \delta \mathbf{x}_k &= \delta \mathbf{x}_k^- + K_k \mathbf{r}_k \end{aligned} \tag{16}$$

$$P_k = (I - K_k H) P_k^- \tag{17}$$

The Kalman gain has dimensions 4×2, and I denotes the 4×4 identity matrix.

Error-State Injection and Reset: After the update, the estimated error state is injected into the nominal state:

$$\hat{\mathbf{x}}_k \leftarrow \hat{\mathbf{x}}_k^- + \delta \mathbf{x}_k \tag{18}$$

After the correction is injected into the nominal state, the error-state estimate is reset to zero while the updated covariance is retained. This injection-reset operation keeps the error state in the neighborhood in which the linearized model is applied.

$$\delta \mathbf{x}_k \leftarrow \mathbf{0} \tag{19}$$

2.2.5 Sensor Synchronization

VINS and Wi-Fi provide position estimates at different effective update rates (approximately 10 Hz and 1 Hz, respectively). The ES-EKF performs a prediction at each VINS update using the elapsed VINS timestamp interval and applies a correction only when a new Wi-Fi position becomes available. Between Wi-Fi updates, the nominal state follows VINS and the error-state covariance grows according to the process model. This asynchronous schedule preserves the higher-rate VINS trajectory while applying Wi-Fi corrections at their measured availability.

2.3 Wi-Fi Fingerprinting Localization

The Wi-Fi fingerprinting subsystem provides building-frame position estimates using weighted K-nearest neighbors (WKNN). Its processing pipeline consists of fingerprint database construction and position estimation from query RSS measurements.

Access Point Deployment: Eight Wi-Fi access points (APs) were deployed in the experimental area to provide signal coverage. The APs were mounted on walls at a height of approximately 2.5 m and arranged with a reported mean inter-AP spacing of approximately 8 m. All APs operated in the 2.4 GHz band using IEEE 802.11n.

The experimental area was discretized into reference points (RPs) on a 1.0 m grid. At each RP, RSS measurements from all detectable APs were recorded for 60 s at an effective rate of approximately 1 Hz, and the sample mean for each AP was stored with the surveyed two-dimensional RP coordinate. With M APs, the i-th fingerprint entry is represented as:

$$D_i = \left\{ (x_i, y_i), \left[\mathrm{RSS}_i^1 \ \cdots \ \mathrm{RSS}_i^M \right] \right\} \tag{20}$$

During the positioning phase, one recorded smartphone RSS vector $\mathrm{r_{online}}$ is used as the query at each Wi-Fi update. Let $\bar{r}_\mathrm{i}$ denote the mean RSS vector stored for the i-th reference point. The WKNN algorithm ranks the fingerprints by their Euclidean distance in RSS space:

$$d_i = \|\mathbf{r}_{\mathrm{online}} - \bar{\mathbf{r}}_i\|_2 \tag{21}$$

The K reference points with the smallest Euclidean distances are selected (K = 5 in this study). The estimated position is then computed as the weighted average of the coordinates of these K nearest neighbors, where the weight for the j-th nearest neighbor is inversely proportional to its RSS distance:

$$\begin{aligned} \mathbf{p}_{\mathrm{WiFi}} &= \frac{\sum_{j=1}^{K} w_j \mathbf{p}_j}{\sum_{j=1}^{K} w_j}, \\ w_j &= \frac{1}{d_j + \varepsilon} \end{aligned} \tag{22}$$

In Equation (22), $\mathrm{p_j}$ is the coordinate of the j-th nearest RP, $\mathrm{d_j}$ is its RSS-space distance, and $\varepsilon = 10^{-6}$ prevents division by zero. The value K = 5 and inverse-distance weighting were used in both experiments.

2.4 Least-Squares Coordinate Alignment

VINS positions are expressed in a local three-dimensional frame established at initialization, whereas the Wi-Fi fingerprint database provides horizontal coordinates in the building frame. The alignment implementation estimates a three-dimensional rigid transformation from synchronized VINS/Wi-Fi position pairs, with the Wi-Fi reference points embedded in a constant-height plane. Only the horizontal components of the transformed VINS positions are supplied to the position-level filter.

The rigid transformation comprises a rotation $\mathrm{R} \in \mathrm{SO(3)}$ and a translation vector $\mathrm{t} \in \mathbb{R}^3$:

$$\mathbf{p}_i^G = R\mathbf{p}_i^{\mathrm{VINS}} + t, \quad R \in \mathrm{SO(3)}, \quad t \in \mathbb{R}^3 \tag{23}$$

The transformation minimizes the sum of squared residuals between the transformed VINS positions and the corresponding plane-embedded Wi-Fi reference positions:

$$\begin{aligned} \mathbf{r}_i(R,t) &= \mathbf{p}_i^G - \left(R\mathbf{p}_i^{\mathrm{VINS}} + t\right), \\ (R^*,t^*) &= \arg\min_{R,t} \sum_{i=1}^{N_{\mathrm{paired}}} \|\mathbf{r}_i\|_2^2 \end{aligned} \tag{24}$$

Equation (24) is solved as a rigid Procrustes problem. The paired point sets are centered, the cross-covariance matrix is formed, and the rotation is obtained by singular value decomposition; the translation is then computed from the two centroids. The implementation uses the 20 most recent synchronized position pairs, corresponding to an approximately 20 s sliding window at the reported Wi-Fi update rate, and re-estimates the transformation when a new Wi-Fi position becomes available. The transformed trajectory is then projected onto the horizontal x-y plane for fusion and evaluation.

2.5 Reference Trajectory Acquisition

The reference trajectory was constructed from AprilTags placed at surveyed positions along the walking path. The two-dimensional coordinates of each marker were measured relative to the building-frame origin using a Leica DISTO D2 laser distance meter (nominal accuracy ± 1.5 mm under favorable conditions). During data collection, the smartphone was held at an approximately consistent height and orientation while the experimenter followed the predefined path at a target speed of approximately 1.0 m/s. Each marker passage was assigned a timestamp, and the reference positions between consecutive markers were obtained by linear interpolation under a constant-velocity assumption. The markers were spaced by approximately 2.0 m. Because this procedure interpolates between discrete surveyed points rather than continuously tracking the device, the resulting trajectory is treated as a reference trajectory rather than metrological ground truth.

3 Results

3.1 Calibration Experiments

A Xiaomi Mi 9 smartphone (Qualcomm Snapdragon 855, Android 10) was used for data collection. The device provided images from its Sony IMX586 main camera and acceleration and angular-velocity measurements from a reported ICM-20690 six-axis IMU. IMU samples were recorded at an effective rate of approximately 200 Hz using the Android Sensor API with SENSOR_DELAY_FASTEST. Camera intrinsic calibration was performed with the Kalibr toolbox [21] using a 6×6 AprilGrid target with a reported tag size of 0.024 m. The estimated parameters comprised focal lengths, principal-point coordinates, and radial and tangential distortion coefficients. Fig. 3 shows the monocular-camera reprojection residuals.

Most monocular-camera reprojection residuals in Fig. 3 lie within approximately 1 pixel, which is consistent with an acceptable calibration for the tested image resolution. The calibrated intrinsic and distortion parameters are listed in Table 2.

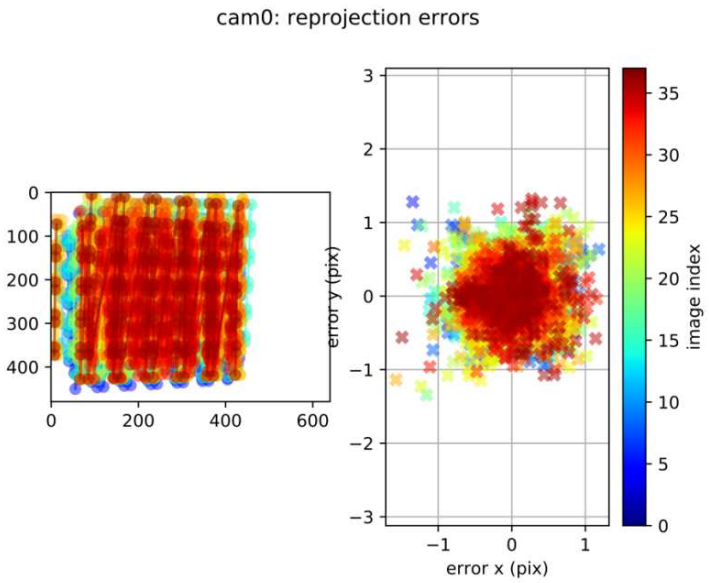


Fig. 3 Monocular-camera reprojection-error distribution

Table 2: Camera intrinsic parameter calibration results

Parameter	**Value**	**Unit**
fx (focal length x)	418.923	pixel
fy (focal length y)	421.084	pixel
cx (principal point x)	314.702	pixel
cy (principal point y)	240.977	pixel
k1 (radial distortion)	-0.2917	unitless
k2 (radial distortion)	0.08228	unitless
p1 (tangential distortion)	0.0005333	unitless
p2 (tangential distortion)	-0.0001578	unitless

IMU noise characterization was performed with the open-source imu_utils toolbox [22]. The imu_utils-derived parameters used by the estimator include accelerometer and gyroscope noise densities, accelerometer and gyroscope random-walk parameters, and an estimated gravity magnitude. The reported values and units are listed in Table 3; the estimated gravity magnitude was 9.7917 $m \cdot s^{-2}$.

Table 3: IMU parameter calibration results

Parameter	Value	Unit
accn (accelerometer noise density)	0.16	$\mathrm{m\,s^{-2}\,Hz^{-1/2}}$
gyrn (gyroscope noise density)	0.024	$\mathrm{rad\,s^{-1}\,Hz^{-1/2}}$
accw (accelerometer random walk)	0.00055	$\mathrm{m\,s^{-3}\,Hz^{-1/2}}$
gyrw (gyroscope random walk)	0.0002	$\mathrm{rad\,s^{-2}\,Hz^{-1/2}}$
gnorm (gravity magnitude)	9.7917	$\mathrm{m\,s^{-2}}$

Joint camera–IMU calibration determines the relative pose (rotation and translation) between the two sensor coordinate frames. This calibration was also performed using the Kalibr toolbox, which requires the camera intrinsic parameters, IMU noise and random walk parameters, and a short calibration recording with sufficient rotational excitation about all three axes. The resulting camera–IMU extrinsic calibration is shown in Table 4. Under the coordinate convention used here, the rotation matrix $R_{C,I}$ maps vectors from the IMU coordinate frame to the camera coordinate frame, and the translation vector $t_{C,I}$ is the position of the IMU origin expressed in the camera coordinate frame (in meters).

Table 4: Camera–IMU extrinsic calibration results (rotation matrix $R_{C,I}$ and translation vector $t_{C,I}$)

Rotation matrix $R_{C,I}$	**Translation vector** $t_{C,I}$ **(m)**
0.014866, -0.999881, 0.004140	-0.021640
0.999557, 0.014967, 0.025716	-0.064677
-0.025774, 0.003756, 0.999661	0.009811

3.2 Experimental Setup

Experiments were conducted on the ground floor of a university building containing corridors, open areas, and office rooms. Two trajectory scales were evaluated. Experiment 1 used an approximately 50 m closed-loop path with multiple corners, narrow corridors, and non-line-of-sight Wi-Fi conditions. Experiment 2 used an approximately 150 m path through several rooms and a long corridor with comparatively sparse visual features and variable illumination. The first scenario is referred to as the short-trajectory experiment, and the second as the long-trajectory experiment.

Table 5 summarizes the experimental equipment and configuration.

Table 5: Experimental equipment and configuration

Equipment / Parameter	**Specification**
Smartphone model	Xiaomi Mi 9 (Snapdragon 855, Android 10)
IMU sensor	ICM-20690 (6-axis)
IMU sampling rate	200 Hz
Camera sensor	Sony IMX586 (48 MP, f/1.75)
Camera resolution / frame rate	640 × 480 pixels, 30 Hz
Wi-Fi module	Qualcomm WCN3998 (802.11 a/b/g/n/ac), reported
Effective Wi-Fi RSS update rate	Approximately 1 Hz in the recorded log
Wi-Fi AP count	8 (2.4 GHz, IEEE 802.11n)
Fingerprint grid interval	1.0 m
WKNN parameter K	5 (inverse-distance weighted)
RSS samples per RP	60 (60 s collection at 1 Hz)
Processing platform	Ubuntu 18.04 (i7-9700K, 32 GB RAM, GTX 1660)
VINS framework	VINS-Mono (offline tightly coupled VIO)
Coordinate alignment	3D rigid Procrustes; horizontal projection (20 pairs)
Fusion filter	Position-level ES-EKF (4D state, 2D observation)

3.3 Positioning Performance

3.3.1 Experiment 1: Short-Trajectory Scenario

Experiment 1 evaluated the method on an approximately 50 m trajectory containing 10 sharp turns and passing through narrow corridors and open areas. Five repeated traversals of the same path were recorded to quantify trial-to-trial variability. Table 6 reports the least-squares alignment parameters for one representative traversal.

Table 6: Experiment 1: Three-dimensional least-squares alignment parameters between the local VINS frame and the building frame (representative trial)

3D Rotation Matrix R	**3D Translation Vector t (m)**
-0.01867, -0.99923, 0.03457	4.06343
0.99946, -0.01771, 0.02785	1.80491
-0.02721, 0.03507, 0.99901	-0.26949

For each trial, let $e_i = \|\hat{\mathbf{p}}_i^{xy} - \mathbf{p}_{\mathrm{ref},i}^{xy}\|_2$ denote the horizontal Euclidean position error at evaluation sample i, where $\hat{\mathbf{p}}_i^{xy}$ and $\mathbf{p}_{\mathrm{ref},i}^{xy}$ are the estimated and reference positions in the same building coordinate frame. With N evaluated samples in a trial, the error metrics are defined as follows:

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}, \\ \mathrm{MAX} &= \max_i e_i, \qquad \mathrm{MIN} = \min_i e_i, \\ \mathrm{MEAN} &= \frac{1}{N}\sum_{i=1}^{N} e_i \end{aligned} \tag{25}$$

Table 7 summarizes the short-trajectory results. Each entry reports the mean and standard deviation of the corresponding trial-level metric across the five trials.

Table 7: Experiment 1: Positioning accuracy comparison (mean and standard deviation over 5 repeated trials). All values in meters

Method	RMSE (m)	MAX (m)	MIN (m)	MEAN (m)
VINS	0.3821 ± 0.052	0.8244 ± 0.105	0.0031 ± 0.001	0.3181 ± 0.048
Wi-Fi	0.7468 ± 0.088	1.3782 ± 0.215	0.1223 ± 0.035	0.6437 ± 0.071
VINS+Wi-Fi	0.3429 ± 0.038	0.7317 ± 0.089	0.0012 ± 0.001	0.2839 ± 0.035

3.3.2 Experiment 2: Long-Trajectory Scenario

Experiment 2 evaluated the method on an approximately 150 m trajectory through multiple rooms and a long corridor. The longer path and lower visual-feature density provide a more demanding test of accumulated VINS drift. Five repeated traversals were recorded, and Table 8 reports the alignment parameters for one representative traversal.

Table 8: Experiment 2: Three-dimensional least-squares alignment parameters (representative trial)

3D Rotation Matrix R	3D Translation Vector t (m)
0.77703, -0.28091, -0.56331	16.17930
0.34370, 0.93906, 0.00582	12.92762
0.52735, -0.19813, 0.82623	-3.13687

Tables 6 and 8 report the 3×3 rotation matrices and three-dimensional translation vectors returned by the alignment implementation. Only the horizontal x-y components of the transformed positions are supplied to the position-level filter and used in the reported two-dimensional error metrics.

Table 9 summarizes the long-trajectory results using the same trial-level metrics and across-trial reporting convention as Table 7.

Table 9: Experiment 2: Positioning accuracy comparison (mean and standard deviation over 5 repeated trials). All values in meters

Method	RMSE (m)	MAX (m)	MIN (m)	MEAN (m)
VINS	0.8006 ± 0.121	3.5338 ± 0.481	0.0171 ± 0.008	0.7931 ± 0.115
Wi-Fi	2.1626 ± 0.345	4.1872 ± 0.564	0.2315 ± 0.052	2.1565 ± 0.338
VINS+Wi-Fi	0.5909 ± 0.079	2.1407 ± 0.283	0.0157 ± 0.006	0.5834 ± 0.074

In the long-trajectory experiment, Wi-Fi, VINS, and VINS+Wi-Fi produced RMSE values of 2.1626 ± 0.345 m, 0.8006 ± 0.121 m, and 0.5909 ± 0.079 m, respectively. Using the reported mean RMSE values, the fusion RMSE was 26.2% lower than that of VINS and 72.7% lower than that of Wi-Fi. The across-trial averages of the maximum errors were 4.1872 m, 3.5338 m, and 2.1407 m, and the across-trial averages of the mean errors were 2.1565 m, 0.7931 m, and 0.5834 m for Wi-Fi, VINS, and VINS+Wi-Fi, respectively. The across-trial RMSE standard deviation was numerically lower for VINS+Wi-Fi (0.079 m) than for VINS (0.121 m); no inferential significance test was performed.

3.3.3 Comparison of Short- and Long-Trajectory Performance

The fusion RMSE increased from 0.3429 m in Experiment 1 to 0.5909 m in Experiment 2, while the relative reduction versus VINS increased from 10.3% to 26.2%. This pattern is consistent with a greater benefit from intermittent global corrections as the duration over which VINS drift can accumulate increases. However, trajectory length, feature density, illumination, and radio propagation were not varied independently; the experiment therefore does not isolate the causal contribution of any single environmental factor.

4 Discussion

4.1 Performance Analysis and Comparison with Existing Methods

The experiments show lower observed positioning errors for VINS+Wi-Fi than for standalone VINS and Wi-Fi. In Experiment 1, VINS+Wi-Fi achieved an RMSE of 0.3429 m compared with 0.3821 m for VINS, a 10.3% reduction. In Experiment 2, the fusion RMSE was 0.5909 m compared with 0.8006 m for VINS, a 26.2% reduction. This pattern is consistent with the complementary roles of the two sources: VINS provides high-rate local trajectory updates, whereas Wi-Fi supplies building-frame position observations. Coordinate alignment allows these estimates to be combined in a shared spatial frame, and the asynchronous update schedule accommodates their different rates. The process- and measurement-noise covariances control their relative influence in the filter.

Cross-study numerical comparison must be interpreted cautiously because the cited studies use different platforms, environments, trajectories, reference systems, and error metrics. Pritzl et al. [15] reported a mean three-dimensional absolute trajectory error of 0.28 m for a cooperative VIO/LiDAR system, while Jung et al. [16] reported position errors below approximately 0.2 m for a custom UWB/VIO platform. These values are not directly comparable with the two-dimensional RMSE reported here. Kong et al. [14] and Wen and Feng [18] address more closely related smartphone scenarios, but their protocols and reported metrics are also not sufficiently matched for a controlled ranking. The present method therefore demonstrates sub-meter RMSE only under the tested conditions; its practical advantage is the use of smartphone sensors and Wi-Fi infrastructure rather than dedicated LiDAR or UWB ranging hardware, with the remaining requirement for deployed AP coverage and a fingerprint database.

Table 10 summarizes these architectural differences and the limitations of cross-study comparisons.

Table 10: Qualitative comparison of the proposed method with representative fusion-based indoor positioning methods

Aspect	**Kong et al. [14]**	**Wen and Feng [18]**	**Pritzl et al. [15]**	**Jung et al. [16]**	**Proposed**
Sensors	VINS + BLE	VIO + Wi-Fi	VIO + LiDAR	VIO + UWB	VINS + Wi-Fi
Infrastructure	BLE beacons	Wi-Fi radio map	3D LiDAR on primary UAV	UWB anchor	8 Wi-Fi APs; radio map
Fusion strategy	EKF (loose)	Loosely coupled	Linear KF	Factor graph	ES-EKF (position level)
Platform	Smartphone	Smartphone	Cooperative UAVs	Custom UGV	Xiaomi Mi 9; offline PC
Reported metric	Different protocol / metric	Not reported	0.28 m mean 3D ATE	<0.2 m position error	0.343 / 0.591 m 2D RMSE
Metric caveat	Not directly comparable	No quantitative comparison	3D ATE; different task	Different task and reference	Two self-collected scenarios
Evaluation repeats	See cited study	See cited study	See cited study	See cited study	5 trials per scenario
Main constraint	BLE deployment	Radio-map dependence	Specialized LiDAR platform	UWB infrastructure	Radio map; 2D output; offline processing

4.2 Trajectory-Scale and Environmental Factors

The higher RMSE in Experiment 2 (0.5909 m) than in Experiment 1 (0.3429 m) is consistent with the greater difficulty of the longer trajectory and its reported visual and radio conditions. In Experiment 2, the maximum errors were 4.1872 m for Wi-Fi, 3.5338 m for VINS, and 2.1407 m for VINS+Wi-Fi. The fusion method therefore reduced the reported maximum error relative to both standalone methods in this scenario. Because the experimental factors were not independently controlled, these results should not be interpreted as proving that trajectory length, multipath, or feature sparsity alone caused the observed difference.

4.3 Limitations and Challenges

Several limitations should be noted. First, the alignment requires a window of temporally corresponding VINS and Wi-Fi positions, producing an initialization period and making the result sensitive to poor early Wi-Fi estimates. Second, the reported output and evaluation are two-dimensional and limited to single-floor trajectories. Third, the method depends on the coverage, quality, and freshness of the Wi-Fi fingerprint database; AP relocation or environmental change can degrade performance. Fourth, the reference trajectory is based on surveyed markers and constant-velocity interpolation rather than continuous external tracking, so its temporal and between-marker uncertainty should be included when interpreting millimeter-level minimum errors or excessive decimal precision. Finally, data processing was performed offline, and neither real-time on-device latency nor energy consumption was measured.

The imu_utils-derived estimator parameters include an accelerometer noise density of 0.16 $m \cdot s^{-2} \cdot Hz^{-1/2}$ and a gyroscope noise density of 0.024 $rad \cdot s^{-1} \cdot Hz^{-1/2}$ (Table 3). These values are estimator inputs obtained from the recorded-device calibration and are not the nominal sensor noise floors stated in a component datasheet. The current filter uses fixed process and measurement covariances. The Wi-Fi tuning variance of 0.25 m^2 per axis corresponds to a per-axis standard deviation of 0.5 m; it is a model parameter, not the measured two-dimensional trajectory RMSE. Its suitability across the two scenarios requires further validation. Adaptive or innovation-based covariance estimation should therefore be evaluated under changing motion and radio conditions [24].

The experiments were conducted in static indoor environments without crowds, dynamic obstacles, deliberate lighting changes, or camera occlusions. References to a complex environment therefore denote spatial characteristics such as corners, narrow corridors, and feature-poor regions, not demonstrated robustness to dynamic environmental change. Evaluation under moving pedestrians, lighting transitions, device-orientation changes, and temporary radio interference remains necessary.

4.4 Practical Implications and Future Directions

The prototype illustrates how visual-inertial estimation and wireless fingerprint matching can be integrated for mobile indoor positioning. Data collection with a Xiaomi Mi 9 demonstrates compatibility with consumer smartphone sensors, while the offline evaluation leaves computational latency and on-device energy demand to

be assessed. Deployment also requires adequate Wi-Fi AP coverage and a fingerprint database for the target environment. Wider application therefore depends on the following practical considerations:

Scalability: Automated construction and maintenance of Wi-Fi fingerprint databases would support deployment across larger buildings. Crowdsourcing-based approaches, where users' smartphones passively collect and upload RSS measurements during normal usage, could help maintain fingerprint databases at scale [23].

Robustness: The system should be extended to handle dynamic environments (moving crowds, temporary obstacles) and sensor anomalies (camera occlusion, IMU saturation, Wi-Fi interference). Incorporating outlier rejection mechanisms in the ES-EKF update step (e.g., Mahalanobis distance gating) and sensor health monitoring could improve robustness.

Energy efficiency: Continuous image acquisition and repeated Wi-Fi scanning can consume substantial energy on mobile devices. Adaptive sensor scheduling, such as reducing image-processing frequency when visual tracking is stable or requesting Wi-Fi observations when VINS uncertainty increases, may reduce energy demand. This trade-off should be quantified in a future real-time implementation.

Future work should: (1) evaluate additional buildings, devices, users, device orientations, and motion patterns; (2) extend the method to multi-floor positioning using barometric and floor-plan constraints; (3) incorporate robust innovation gating and adaptive covariance estimation; (4) quantify on-device latency and energy consumption; and (5) evaluate dynamic conditions including moving pedestrians, illumination changes, camera occlusion, and temporal Wi-Fi variation.

5 Conclusions

This study presented a position-level method that combines locally referenced VINS estimates with building-frame Wi-Fi fingerprint positions. The method includes least-squares rigid alignment followed by horizontal projection, a position-level ES-EKF with explicit transition and observation models, and a WKNN Wi-Fi subsystem. The equations specify the nominal and error states, process and measurement covariances, Kalman update, and injection-reset operation. The alignment is re-estimated over a sliding window of synchronized position pairs.

Across five repeated trials per scenario, VINS+Wi-Fi had lower reported RMSE, mean error, and maximum error than standalone VINS and Wi-Fi. In the approximately 50 m experiment, the fusion RMSE was 0.3429 ± 0.038 m, a 10.3% reduction relative to VINS. In the approximately 150 m experiment, it was 0.5909 ± 0.079 m, a 26.2% reduction relative to VINS. These results support the method under the tested static, single-floor conditions. Generalization requires additional environments, devices, orientations, users, and radio configurations, together with a more accurate continuously tracked reference trajectory.

The results provide a proof of concept for integrating visual-inertial and wireless position estimates through coordinate alignment and asynchronous error-state fusion. For mobile indoor positioning, the remaining implementation question is whether these accuracy gains can be retained during real-time on-device processing within

practical latency and energy budgets. Further evaluation should also address dynamic environments, adaptive noise modeling, and multi-floor positioning.

Declarations

Funding

This research was funded by the Liaoning Provincial Department of Education Basic Research Program (Grant No. LJ222510144004) and the Applied Basic Fund of Liaoning Province (Grant No. 2025JH2/10133076).

Competing interests

The authors declare no competing interests.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

All processed data and summary results available to the authors and supporting the findings of this study are included in this article. The underlying raw sensor recordings are no longer available because they were not retained after data processing.

Author contributions

J.W. conceptualized the study, contributed to the methodology, and revised the manuscript. Z.H. implemented the algorithms, collected and processed the data, conducted the experiments, analyzed the results, prepared the figures and tables, and wrote the initial draft. Y.J. supervised the research, contributed to the study design and interpretation of the results, and critically revised the manuscript. All authors reviewed and approved the final manuscript.

References

[1] Brena RF, Garcia-Vazquez JP, Galvan-Tejada CE, et al. Evolution of Indoor Localization Technologies: A Survey. J. Sensors. 2017;2017:2630413. https://doi.org/10.1155/2017/2630413.

[2] Deng ZL, Yin L, Tang SH, et al. Key Technologies of Indoor Localization: A Review. Journal of Navigation and Positioning. 2018;5:14-23.

[3] Raza A, et al. Comparing and Evaluating Indoor Positioning Techniques. In Proceedings of the 2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Lloret de Mar, Spain, 29 November-2 December 2021; pp. 1-8.

[4] Hailu TG, Guo X, Si H, et al. Theories and Methods for Indoor Positioning Systems: A Comparative Analysis, Challenges, and Prospective Measures. Sensors. 2024;24:6876. https://doi.org/10.3390/s24216876.

[5] Chen J, Song S, Yu H. An Indoor Multi-Source Fusion Positioning Approach Based on PDR/MM/WiFi. AEU-Int. J. Electron. Commun. 2021;135:153733. https://doi.org/10.1016/j.aeue.2021.153733.

[6] Delmerico J, Scaramuzza D. A Benchmark Comparison of Monocular Visual-Inertial Odometry Algorithms for Flying Robots. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21-25 May 2018; pp. 2502-2509. https://doi.org/10.1109/ICRA.2018.8460664.

[7] Huang B, Zhao J, Liu J. A Survey of Simultaneous Localization and Mapping with an Envision in 6G Wireless Networks. arXiv 2020, arXiv:2006.00648.

[8] Sui X, Zhang J, Chen ZJ, et al. A Combined Monocular VIO/UWB Indoor Localization Method. Journal of Navigation and Positioning. 2022;10:1-8.

[9] Yang Z, Shen S. Monocular Visual-Inertial State Estimation with Online Initialization and Camera-IMU Extrinsic Calibration. IEEE Trans. Autom. Sci. Eng. 2017;14:39-51. https://doi.org/10.1109/TASE.2016.2550621.

[10] Dong X, Dong X, Dong J, Zhou H. Monocular Visual-IMU Odometry: A Comparative Evaluation of Detector-Descriptor-Based Methods. IEEE Trans. Intell. Transp. Syst. 2020;21:2471-2484. https://doi.org/10.1109/TITS.2019.2919003.

[11] Aslan MF, Durdu A, Yusefi A, et al. HVIOnet: A Deep Learning Based Hybrid Visual-Inertial Odometry Approach for Unmanned Aerial System Position Estimation. Neural Netw. 2022;155:461-474.

[12] Wang Z, Cheng X. Adaptive Optimization Online IMU Self-Calibration Method for Visual-Inertial Navigation Systems. Measurement. 2021;180:109478. https://doi.org/10.1016/j.measurement.2021.109478.

[13] Bellavista-Parent V, Torres-Sospedra J, Perez-Navarro A. New Trends in Indoor Positioning Based on WiFi and Machine Learning: A Systematic Review. In Proceedings of the 2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Lloret de Mar, Spain, 29 November-2 December 2021; pp. 1-8.

[14] Kong X, Wu C, You Y, Lv Z, Zhao Z. Hybrid Indoor Positioning Method of BLE and Monocular VINS-Based Smartphone. IEEE Trans. Instrum. Meas. 2023;72:1-13. https://doi.org/10.1109/TIM.2023.3295010.

[15] Pritzl V, Vrba M, Stepan P, et al. Fusion of Visual-Inertial Odometry with LiDAR Relative Localization for Cooperative Guidance of a Micro-Scale Aerial Vehicle. arXiv 2023, arXiv:2306.17544.

[16] Jung KY, Shin SJ, Myung H. U-VIO: Tightly Coupled UWB Visual Inertial Odometry for Robust Localization. In Robot Intelligence Technology and Applications 6; Springer: Cham, Switzerland, 2022; pp. 272-283. https://doi.org/10.1007/978-3-030-97672-9_24.

[17] Gao B, Lian BW. Graph-Optimization-Based Vision/Inertial/UWB Fusion Positioning Algorithm for Indoor Environments. Acta Photonica Sin. 2024;53:1012004.

[18] Wen ZT, Feng SL. Design of Indoor Localization System Based on VIO and Wi-Fi Fingerprinting Technology. Telecommunication Engineering. 2019;59(4):449-454. https://doi.org/10.3969/j.issn.1001-893x.2019.04.014.

[19] Im G. Notes on Kalman Filter (KF, EKF, ESKF, IEKF, IESKF). arXiv 2024, arXiv:2406.06427.

[20] Sola J. Quaternion Kinematics for the Error-State Kalman Filter. arXiv 2017, arXiv:1711.02508.

[21] Furgale P, Rehder J, Siegwart R. The Kalibr Visual-Inertial Calibration Toolbox. 2013. Available online: https://github.com/ethz-asl/kalibr

[22] Gao WS. imu_utils: A ROS Package for IMU Calibration. 2018. Available online: https://github.com/gaowenliang/imu_utils

[23] Yang S, Dessai P, Verma M, Gerla M. FreeLoc: Calibration-Free Crowdsourced Indoor Localization. In Proceedings of the IEEE INFOCOM, Turin, Italy, 14-19 April 2013; pp. 2481-2489. https://doi.org/10.1109/INFCOM.2013.6567054.

[24] Mehra R. Approaches to Adaptive Filtering. IEEE Trans. Autom. Control. 1972;17:693-698. https://doi.org/10.1109/TAC.1972.1100100.