State Estimation Method for GNSS/INS/Visual Multi-sensor Fusion Based on Factor Graph Optimization for Unmanned System

ZHU Zekun, YANG Zhong, XUE Bayang, ZHANG Chi, YANG Xin*

College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China

(Received 27 May 2024; revised 13 June 2024; accepted 20 June 2024)

Abstract: With the development of unmanned driving technology, intelligent robots and drones, high-precision localization, navigation and state estimation technologies have also made great progress. Traditional global navigation satellite system / inertial navigation system (GNSS/INS) integrated navigation systems can provide high-precision navigation information continuously. However, when this system is applied to indoor or GNSS-denied environments, such as outdoor substations with strong electromagnetic interference and complex dense spaces, it is often unable to obtain high-precision GNSS positioning data. The positioning and orientation errors will diverge and accumulate rapidly, which cannot meet the high-precision localization requirements in large-scale and long-distance navigation scenarios. This paper proposes a method of high-precision state estimation with fusion of GNSS/INS/Vision using a nonlinear optimizer factor graph optimization as the basis for multi-source optimization. Through the collected experimental data and simulation results, this system shows good performance in the indoor environment and the environment with partial GNSS signal loss.

Key words: state estimation; multi-sensor fusion; combined navigation; factor graph optimization; complex environments

CLC number: V249.3 **Document code:** A **Article ID:** 1005-1120(2024)S-0043-09

0 Introduction

Localization and state estimation are critical technologies enabling intelligent unmanned systems like robots to determine their position and orientation. Visual sensors provide rich visual information at low size and cost, exhibiting high applicability. When integrated with inertial measurement units (IMUs), they offer high-frequency, continuous, jitter-free state estimation and positioning data. Owing to their robustness and accuracy in complex environments, vision-aided inertial navigation systems often achieve high precision in short-range, short-duration experimental measurements[1]. However, cameras and IMUs operate in the local frame, leaving four unobservable degrees of freedom X, Y, Z position and yaw angle. Thus, odometry drift is present when using visual-inertial navigation[2]. Although maintaining good short-term performance, errors can accumulate in long-range environments. Fortunately, global navigation satellite system (GNSS) provides real-time, drift-free global positioning with wide applicability. By concurrently tracking at least four satellites, the receiver obtains precise global geodetic frame coordinates[3].

This paper combines visual-inertial systems and GNSS via factor graph optimization for multi-sensor fusion state estimation. The front-end handles feature extraction, matching, local map alignment, and motion estimation, while the back-end maintains the map and optimizes the sliding window. The workload and innovations of this paper

*Corresponding author, E-mail address: YangZhong@nuaa.edu.cn.

How to cite this article: ZHU Zekun, YANG Zhong, XUE Bayang, et al. State estimation method for GNSS/INS/Visual multi-sensor fusion based on factor graph optimization for unmanned system[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2024, 41(S): 43-51.

http://dx.doi.org/10.16356/j.1005-1120.2024.S.006

are:

(1) A tightly coupled fusion approach based on factor graph optimization combines the visual-inertial odometry and GNSS. The probabilistic framework is shown in Fig.1.

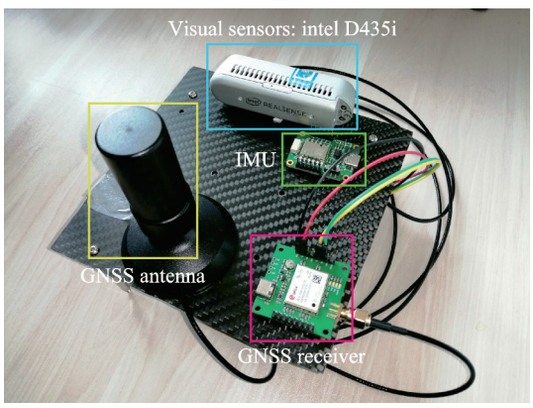


Fig.1 Tightly coupled multi-sensor state estimation platform

(2) The Kanade-Lucas algorithm is tracking multiple feature descriptors without inter-frame matching during corner tracking.

(3) Incorporating spatio-temporal measurement correlations and GNSS error sources into the optimization.

(4) More accurate IMU models are used in this paper, including biases, scale factors, g-dependent terms and skewness, with numerical stability improvements.

1 Related Technical Theory

In visual simultaneous localization and mapping (SLAM), the motion and observation models are fundamental in the probabilistic graphical model.

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_k) + \boldsymbol{w}_k \\ \boldsymbol{z}_{k,j} = \boldsymbol{h}(\boldsymbol{y}_j, \boldsymbol{x}_k) + \boldsymbol{v}_{k,j} \end{cases} \tag{1}$$

The robot pose $\boldsymbol{x}$ is usually parameterized with transformation matrices $\boldsymbol{T}$ in the lie group, with optimization performed via the tangent space using lie algebra

$$\boldsymbol{T} = \exp(\boldsymbol{\xi}) \tag{2}$$

where $\boldsymbol{\xi}$ is the tangent vector of lie algebra. The observation model is determined by the camera projection function π

$$\boldsymbol{z} = \pi(\boldsymbol{x}, \boldsymbol{l}) + \boldsymbol{n} \tag{3}$$

where $\boldsymbol{z}$ is the observed feature location, $\boldsymbol{l}$ the 3D landmark location, $\boldsymbol{n}$ the sensor noise. This projects $\boldsymbol{l}$ onto the image plane based on the camera intrinsics $\boldsymbol{K}$ and camera extrinsics $\boldsymbol{Q}$ relating the body and world frames.

Due to sensor noise, the motion and observation models are not perfect, containing error terms that make them hold only approximately. Accurately estimating the maximum a posteriori states $\hat{\boldsymbol{x}}$ from noisy measurements $\boldsymbol{z}$ is crucial for nonlinear optimization in visual SLAM. Multiple observations of the same landmarks over time and across cameras provide constraints that help resolve the pose ambiguity and noise, thereby recovering the optimal robot trajectory $\hat{\boldsymbol{x}}$ and map $\hat{\boldsymbol{l}}$[4].

Visual odometry (VO) uses visual data for state estimation. Earlier VO methods used extended Kalman filters (EKFs) to iteratively estimate camera poses and landmark positions[1]. However, EKF-based approaches accumulate errors over time. Optimization-based formulations refine poses via bundle adjustment over sliding windows, improving on filtering[5]. Recent VO uses factor graphs and smoothing and mapping (SAM) for incremental nonlinear optimization without continual relinearization[6]. The factor graph and solvers like levenberg-marquardt enable accurate simultaneous trajectory $\hat{\boldsymbol{x}}$ and structure $\hat{\boldsymbol{l}}$ estimation from visual measurements $\boldsymbol{z}$ by maximizing the posterior

$$\hat{\boldsymbol{x}}, \hat{\boldsymbol{l}} = \arg\max P(\boldsymbol{x}, \boldsymbol{l} | \boldsymbol{z}) \tag{4}$$

$$\boldsymbol{x}^*_{\mathrm{MAP}} = \arg\max P(\boldsymbol{x}|\boldsymbol{z}) = \arg\max P(\boldsymbol{z}|\boldsymbol{x}) P(\boldsymbol{x}) \tag{5}$$

Visual-inertial odometry (VIO) which used EKFs for incremental state updates, suffers from linearization errors[1]. Contemporary VIO leverages optimization-based smoothing, better handling nonlinear dynamics and visual constraints[7]. Despite higher costs, optimization-based visual inertial navigation (VIN) shows superior accuracy from joint state optimization over sliding windows[8]. In contrast, GNSS have used filter-based methods for efficient sequential updates. Properly initialized graph optimization also shows potential for improving GNSS accuracy during signal issues[9]. There are

opportunities to integrate filtering and optimization for efficiency and accuracy improvements in multi-sensor state estimation.

2 System Frameworks and Theoretical Methods

2.1 Factor graph optimization framework

In studying localization and state estimation for unmanned systems, determining the current position, attitude comprising translation and rotation in various frames is imperative initially. The unmanned system's pose can be parameterized with spatial points and unit quaternions that avoid singularities in representing rotations[10]. Before constructing the graph optimization framework, related coordinate frames are defined and established per navigation frame conventions[3] in Fig.2. Natural frames include the earth-centered, earth-fixed (ECEF) frame $(*)^{E}$ and the earth-centered inertial (ECI) frame $(*)^{e}$, both centered on the earth and fixed. The transformation between them is given by the earth rotation matrix. Additionally, the natural frames include the local-world navigation frame $(*)^{w}$, which is local-tangent with its Z-axis aligned with gravity towards the geocenter and X, Y axes pointing east and north. Custom frames include the unmanned system body frame $(*)^{b}$ and odometry frame $(*)^{o}$. The transformation relationship of each frame is shown in Fig.3.

The vision and IMU are first jointly initialized, aligning their trajectories by combining the visual structure-from-motion[11] and inertial navigation system (INS) propagation[12]. GNSS initialization follows afterwards. The GNSS/INS optimization constitutes a subset of the full GNSS/INS/Vision optimization. Renders the state estimation question as a nonlinear least square problem by defining a factor graph[13]

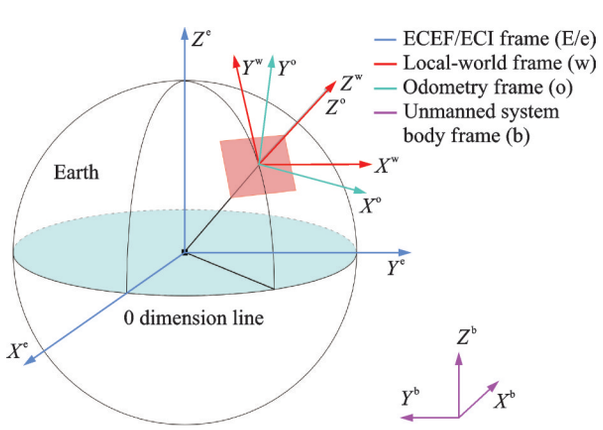


Fig.2 All the frames used in this paper

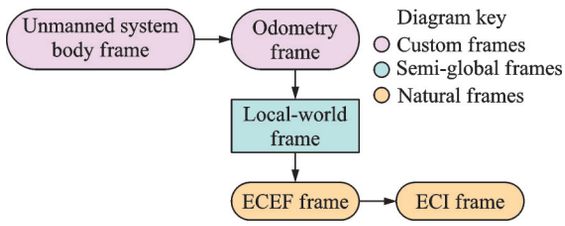


Fig.3 Transform relationships between the frames

$$\hat{\chi}=\arg\min_{\chi}\left\{\|\boldsymbol{\varepsilon}_{\text{pre}}-\boldsymbol{H}_{\text{pre}}\chi\|^2+\|\boldsymbol{\varepsilon}_{\text{c}}-\boldsymbol{h}_{\text{c}}(\chi_{\text{c}},\boldsymbol{z}_{\text{c}})\|^2+\|\boldsymbol{\varepsilon}_{\text{I}}-\boldsymbol{h}_{\text{I}}(\chi_{\text{I}},\boldsymbol{z}_{\text{I}})\|^2+\|\boldsymbol{\varepsilon}_{\text{g}}-\boldsymbol{h}_{\text{g}}(\boldsymbol{z}_{\text{g}})\|^2\right\} \tag{6}$$

where $\hat{\chi}$ represents the state vector to be estimated, including information such as position, velocity, and orientation. χ is the state vector to be optimized, $\boldsymbol{z}$ the observation vector, $\boldsymbol{\varepsilon}$ the error vector, and $\boldsymbol{H}$ the jacobian matrix related to the preintegration error. $\boldsymbol{h}$ is a vector-valued function that associates the state and measurement values. It outputs a vector representing the expected observation. The subscripts c, I, and g represent the visual, IMU, and GNSS components, respectively, while the subscript pre represents the preintegration component.

The multi-sensor state estimation is formulated as a maximum posteriori estimation problem. After pre-processing, the sensor data are incorporated into a factor graph optimization framework as probabilistic constraint factors to constrain the motion state of the unmanned system.

$$\boldsymbol{x}_k=[\boldsymbol{o}_k^{\text{w}},\boldsymbol{v}_k^{\text{w}},\boldsymbol{p}_k^{\text{w}},\boldsymbol{b}_{\text{acc}},\boldsymbol{b}_{\text{gyr}},\delta t,\delta\dot{t}\,]^{\text{T}} \tag{7}$$

where k represents the discrete time; $\boldsymbol{x}_k$ is the state vector of the unmanned system at discrete time k, which includes multiple state information; $\boldsymbol{o}_k^{\text{w}}$ is the orientation of the unmanned system in the world coordinate system, $\boldsymbol{v}_k^{\text{w}}$ the velocity, and $\boldsymbol{p}_k^{\text{w}}$ the position; $\boldsymbol{b}_{\text{acc}}$ and $\boldsymbol{b}_{\text{gyr}}$ represent the IMU's accelerometer bias and gyroscope bias, respectively, and δt, $\delta\dot{t}$ the GNSS time offset and its rate of change.

$$\chi=[\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n,\lambda_1,\lambda_2,\cdots,\lambda_m,\boldsymbol{\psi}\,]^{\text{T}} \tag{8}$$

where n denotes the size of the sliding optimization

window, m the total number of visual features observed within the window. λ_i the inverse depth parameter of the ith visual feature in the current window, and ψ the yaw deviation between the odometry frame and local-world frame.

In contrast to filtering approaches, latest measurements can be incorporated into the factor graph for optimization without requiring perfect time synchronization across sensors. Specifically, new measurement factors can be incrementally added in the factor graph upon availability, regardless of their timestamp. This asynchronous inclusion of multimodal measurements differentiates graph optimization from filtering techniques requiring sequential synchronized updates. The constraint factors can be categorized into GNSS, visual and inertial factors, elaborated below part. The pre-processed measurements from GNSS, camera, and IMU are modeled as factors in a factor graph, which encapsulate the residual and uncertainty as probabilistic constraints between observations and states. The whole framework as shown in Fig.4.

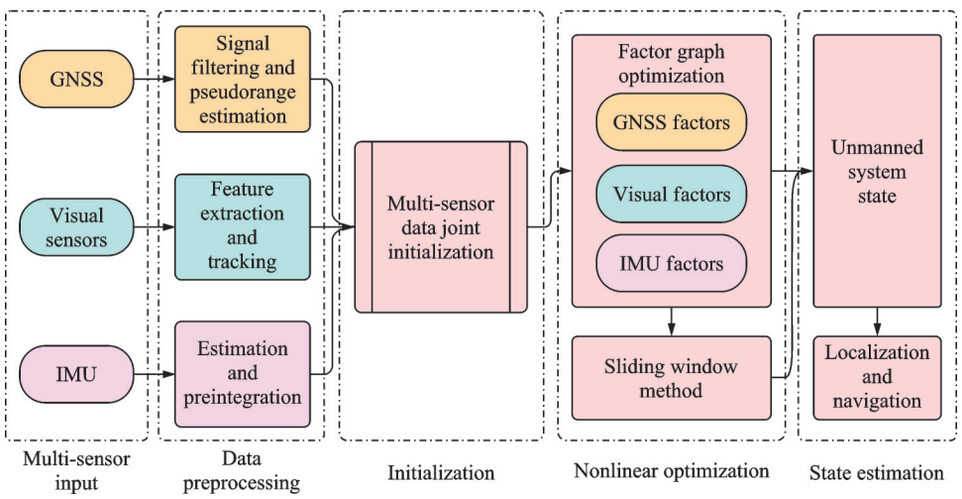


Fig.4 System architecture diagram overview in this paper

2.2 Visual factors

The visual constraint factors come from a set of prominent feature points. These points are detected and tracked across the image frames captured by the visual sensors. A key prerequisite for reliable visual factors is extracting a sufficient number of stable features in the perceived imagery. Thus, the feature point extraction is continually iterated with an adaptive threshold set at 90. If the feature points in a new camera frame are less than 90, additional points will be identified. This ensures robust visual information under varying scene conditions for incorporation into the optimization framework. Only features persisting across multiple views with consistent descriptors are triangulated into 3D landmarks for state estimation. The visual reprojection errors between observed 2D feature coordinates and projected landmark locations provide nonlinear constraints for trajectory estimation. Representing the feature positions in homogeneous form as

$$\tilde{\boldsymbol{p}}_l^{\mathrm{w}} = [\, X_l/Z_l, Y_l/Z_l, 1 \,]^{\mathrm{T}} \tag{9}$$

This provides a compact means to transform the points into the camera view via a linear projection matrix, while avoiding issues with division by zero and reflections implicit in euclidean coordinates. The coordinates of feature points expressed in the local world frame need to be transformed into the image sensor pixel coordinates for further processing. The relationship between the two coordinate systems is given by

$$\tilde{\boldsymbol{p}}_l^{c} = \boldsymbol{K}\boldsymbol{T}_{\mathrm{rec}}^{c}\boldsymbol{T}_{\mathrm{w}}^{\mathrm{rec}}\tilde{\boldsymbol{p}}_l^{\mathrm{w}} + \boldsymbol{n}_c \tag{10}$$

where $\boldsymbol{T}$ is the transformation matrix, $\boldsymbol{n}_c$ the image sensor noise inherent in the camera projection process, and $\boldsymbol{K}$ the camera intrinsic matrix which does not depend on the external camera pose. After reprojecting feature points from the mth frame to the nth frame, the feature coordinates can be expressed as

$$\hat{\tilde{\boldsymbol{p}}}_{l}^{c_n}=\boldsymbol{K}\boldsymbol{T}_{\mathrm{rec}}^{c}\boldsymbol{T}_{\mathrm{w}}^{\mathrm{rec}}\left[\boldsymbol{T}_{\mathrm{rec}}^{\mathrm{w}}\boldsymbol{T}_{c}^{\mathrm{rec}}\boldsymbol{K}^{-1}\left(\boldsymbol{z}_{l}^{c_m}\tilde{\boldsymbol{p}}_{l}^{c_m}\right)\right] \tag{11}$$

The constraint factors are formulated from the deviations between the projected image feature point locations and their actual observed positions after coordinate transformation

$$\boldsymbol{E}_{c}\left(\hat{\boldsymbol{z}}_{l}^{c_n},\boldsymbol{\chi}_{c}\right)=\tilde{\boldsymbol{p}}_{l}^{c_n}-\hat{\tilde{\boldsymbol{p}}}_{l}^{c_n} \tag{12}$$

where $\boldsymbol{E}_c$ is the visual error vector, representing the deviation between the projected position of the feature point and its actual observed position. It is the target value in the formula, meaning the error that needs to be minimized during the optimization process. $\boldsymbol{\chi}_c$ denotes a subvector of the state vector that related with visual information.

2.3 GNSS constraint factors

GNSS measurements typically comprise two components: the legacy code and tracking channel factors. In the graph optimization framework proposed in this study, the constraint factors originating from GNSS incorporate pseudorange factors, Doppler shift factors, and receiver clock bias factors. The pseudorange r_{pse} observation equation and carrier phase φ observation equation can be expressed as

$$\begin{cases} r_{\mathrm{pse}}=r+c\left(\delta t_{\mathrm{rec}}-\delta t_{\mathrm{s}}\right)+L+M+\gamma_{\mathrm{pse}} \\ \varphi=\lambda^{-1}\left[r+c\left(\delta t_{\mathrm{rec}}-\delta t_{\mathrm{s}}\right)+L+M\right]+N+\gamma_{\varphi} \end{cases} \tag{13}$$

where r_{pse} represents the pseudorange measurement, which is the distance from the satellite to the receiver, including various errors and biases; φ is the carrier phase measurement, representing the phase measurement from the satellite to the receiver measured in cycles, which includes integer ambiguity and error terms; N is the integer ambiguity, representing the unknown integer number of cycles in the carrier phase measurement; r is the true geometric distance between the receiver and the satellite and c the speed of light; δt_{rec} and δt_{s} represent the clock biases of the receiver and the satellite, respectively; L and M the ionospheric delay and tropospheric delay, respectively; and γ_{pse} and γ_{φ} the pseudorange measurement noise and carrier phase measurement noise, respectively.

Considering various interference factors, the pseudorange measurement model between a ground receiver and navigation satellite can be expressed as

$$\hat{p}_{\mathrm{rec}}^{\mathrm{s}}=\left\|\boldsymbol{p}_{\mathrm{rec}}^{\mathrm{E}}-\boldsymbol{p}_{\mathrm{s}}^{\mathrm{E}}\right\|+c\left(\delta t_{\mathrm{rec}}+\delta t_{\mathrm{s}}+\Delta L+\Delta M\right)+\gamma \tag{14}$$

where $\boldsymbol{p}_{\mathrm{rec}}^{\mathrm{E}}$ is the position of ground receiver in earth-centered inertial frame, $\boldsymbol{p}_{\mathrm{s}}^{\mathrm{E}}$ the position of navigation satellite in earth-centered inertial frame, $\left\|\boldsymbol{p}_{\mathrm{rec}}^{\mathrm{E}}-\boldsymbol{p}_{\mathrm{s}}^{\mathrm{E}}\right\|$ the measured pseudorange of GNSS signal, and γ the pseudorange carrier phase measurement noise.

The pseudorange constraint factors originate from the measured and true pseudorange data. These pseudorange residuals serve as one part of the GNSS constraint factors incorporated into the factor graph optimization framework

$$E_{\mathrm{pse}}=\left\|\boldsymbol{p}_{\mathrm{rec}}^{\mathrm{E}}-\boldsymbol{p}_{\mathrm{s}}^{\mathrm{E}}\right\|+c\left(\delta t_{\mathrm{rec}}+\delta t_{\mathrm{s}}+\Delta L+\Delta M\right)-\hat{r}_{\mathrm{pse}} \tag{15}$$

where E_{pse} represents the pseudorange residual, which is the error between the measured pseudorange and the predicted pseudorange. The predicted pseudorange $\hat{r}_{\mathrm{pse}}$ is the estimated pseudorange based on the state vector.

The constraint factors originating from GNSS Doppler shift measurements can be formulated as the residuals between the theoretically modeled Doppler shifts and the empirically observed shifts.

$$E_{\mathrm{Dop}}=-\left(\left[\boldsymbol{\theta}_{\mathrm{rec}}^{\mathrm{s}}\left(\boldsymbol{v}_{\mathrm{rec}}^{\mathrm{E}}-\boldsymbol{v}_{\mathrm{s}}^{\mathrm{E}}\right)+c\left(\delta t_{\mathrm{rec}}+\delta t_{\mathrm{s}}\right)\right]-\delta\hat{f}_{\mathrm{rec}}^{\mathrm{s}}\right)/\lambda \tag{16}$$

where E_{Dop} represents the Doppler shift residual, which is the error between the theoretically calculated Doppler shift and the measured Doppler shift; $\boldsymbol{\theta}_{\mathrm{rec}}^{\mathrm{s}}$ the part of the projection matrix or direction cosine matrix, used to translate the velocity difference into the doppler shift effect; $\delta\hat{f}_{\mathrm{rec}}^{\mathrm{s}}$ the measured Doppler shift value; $\boldsymbol{v}_{\mathrm{rec}}^{\mathrm{E}}$ and $\boldsymbol{v}_{\mathrm{s}}^{\mathrm{E}}$ represent the velocity of the receiver and the satellite in the ECEF coordinate system.

2.4 IMU pre-integration and IMU factors

The IMU sensor constraint factors comprise gyroscope measurement residuals and accelerometer measurement residuals. Important parameters are the gyroscope and accelerometer biases. The raw IMU measurements are preintegrated rather than directly incorporated. Specifically, the IMU preintegration includes attitude and position preintegration, integrating the gyroscope and accelerometer read-

ings to propagate the orientation and position increments between sensor updates. Velocity preintegration, integrating the accelerometer-derived specific force to propagate the delta velocity increments.

By preintegrating the IMU data, the nonlinear state optimization is simplified to operate on these delta increments rather than the high-rate raw IMU data. The biases are estimated as part of the state to account for sensor errors. This preintegration approach provides an efficient theoretical framework for fusing IMU measurements in a nonlinear estimator. The IMU preintegration is utilized to provide relative motion constraints between two state nodes. The specific model is as follows

$$\begin{cases} \boldsymbol{x}_{k-1}=[(\boldsymbol{p}_{k-1}^{\mathrm{w}})^{\mathrm{T}}(\boldsymbol{q}_{k-1}^{\mathrm{w}})^{\mathrm{T}}(\boldsymbol{v}_{k-1}^{\mathrm{w}})^{\mathrm{T}}(\boldsymbol{b}_{k-1}^{\mathrm{acc}})^{\mathrm{T}}(\boldsymbol{b}_{k-1}^{\mathrm{gyro}})^{\mathrm{T}}]^{\mathrm{T}} \\ \boldsymbol{x}_k \in \boldsymbol{P}\times \boldsymbol{SO}(3)\times \boldsymbol{B} \end{cases} \tag{17}$$

where $\boldsymbol{q}_{k-1}^{\mathrm{w}}$ represents the orientation of the unmanned system in the world coordinate system at time $k-1$, usually expressed as a quaternion; The state vector $\boldsymbol{x}_k$ is defined within the space that includes position $\boldsymbol{P}$, orientation $\boldsymbol{SO}(3)$, and the velocity along with IMU biases $\boldsymbol{B}$.

3 Experiments

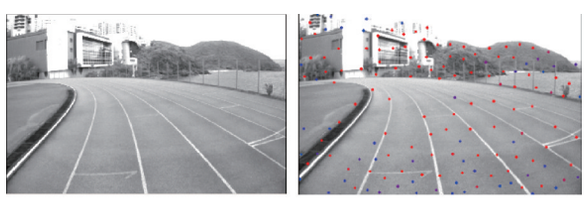
3.1 Simulation experiments based on open-source dataset

The experiments in this section are primarily based on the datasets sport_field.bag and complex_environment.bag. Since the solution proposed in this paper utilizes multi-sensor fusion with multi-channel data input, the open-source dataset should contain raw GNSS data, images captured by visual sensors and related features, and raw IMU recordings. The devices and specific information used in the open-source datasets are shown in Table 1.

Table 1 Devices for open-source dataset

Sensor	Device	Frequency/Hz
Visual sensor	Stereo camera Aptina MT9V034 image sensors	20
IMU sensor	Analog device ADIS 16448 IMU	200
GNSS	u-blox ZED-F9P	10

The dataset utilized in simulation is captured by a customized stereo visual-inertial sensor rig, comprising a calibrated stereo camera system as the primary visual sensor for acquiring images, along with an IMU and GNSS receiver recording inertial and position data, respectively. The stereo visual-inertial data in sport_field.bag is recorded outdoors, providing robust GNSS signals and enabling accurate global position measurements for state estimation. Sample stereo image frames with detected and matched features and state estimation results are visualized in Fig.5. Through the optimization framework of our multi-sensor fusion algorithm, noisy GNSS measurements are effectively suppressed by the visual-inertial data, exploiting their complementary error characteristics. Consequently, the estimated unmanned system trajectory as shown in Fig.6 exhibits negligible drift, with no discernible cumulative errors even after multiple laps along the same trajectory.



(a) Original image (b) Feature tracking image

Fig.5 Extracting features by visual sensor

Fig.6 Outdoor sport field state estimation trajectory

In contrast, the complex_environment.bag dataset presents a more challenging scenario, comprising indoor and outdoor settings with variations in altitude. Specifically, one segment involves ascending indoor stairwells where both GNSS signals and visual odometry drift more significantly. The degraded

visual and inertial measurements in these indoor sections test our approach. While the outdoor performance capitalized on strong GNSS cues, our algorithm's robustness is highlighted by accuracies on par with the sport_field.bag dataset despite compromised sensing in the complex mixed environment. This demonstrates the strengths of our visual-inertial optimization and drift reduction capabilities when GNSS observations are intermittent or highly noisy.

The state estimation trajectory plot indicates the segment corresponding to the indoor stairwell environment, where diminished GNSS signal reception is expected. Despite compromised GNSS measurements in this room, our approach still achieves smooth, unbiased state estimates. This demonstrates the ability of our visual-inertial fusion algorithm to maintain accurate tracking even in GNSS-denied indoor areas by leveraging the complementarity of the visual and inertial cues again likes in Fig.7. As you can see in Fig.8, the robust performance in both outdoor and indoor settings highlights the versatility of our system.

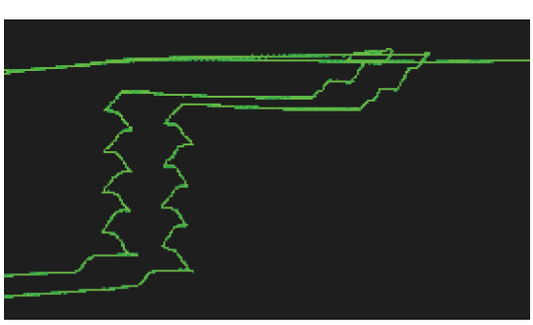


Fig.7 Trajectory of indoor stairs

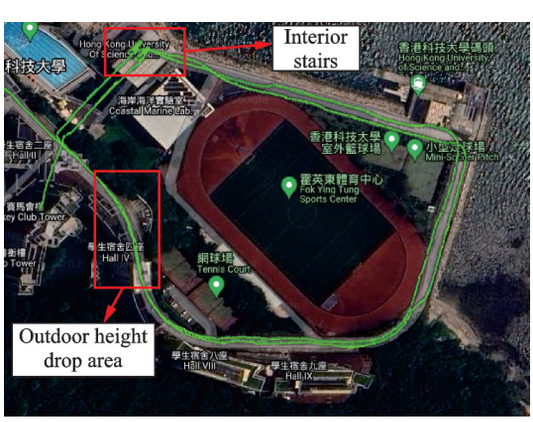
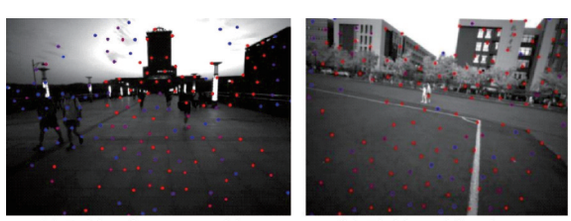


Fig.8 Complex environment state estimation trajectory

3.2 Experiments based on real environment

For real-world experiments, we collect datasets across several locations at Nanjing University of Aeronautics and Astronautics. Due to inherent limitations of onboard sensors, visual-inertial odometry often suffers from drift and accumulation of errors over time. To obtain robust GNSS signals, we select open areas with minimal obstructions. To evaluate performance in complex dynamic environment, we focus on a trajectory from the east playground to Yufeng Park, traversing crowded pedestrian overpasses, tree-lined walkways, and staircases. Prior to data collection, we perform feature extraction testing across scenes with varying lighting conditions and architectural environments as in Fig.9. This is to validate the operational integrity of the visual sensor suite under diverse real-world conditions. Thorough pre-deployment testing ensure robust visual feature detection and matching performance during subsequent experiments. Our devices are shown in Table 2.



(a) Bridge feature tracking (b) Field tracking image

Fig.9 Extracting features by visual sensor in real environment

Table 2 Devices for real environment test

Sensor	Device	Frequency/Hz
Visual sensor	Stereo camera Intel D435i	20
IMU sensor	Microchip MPU9250 IMU	200
GNSS	u-blox ZED-F9P	10

The presence of pedestrians and foliage degrades visual odometry and GNSS reception to an extent. Furthermore, ascending and descending the overpass stairs induce altitude change. During data collection, we minimize erratic sensor motions to avoid confounding the state estimator. Despite these challenges, our approach demonstrates accurate and robust performance. The natural environment stress

tests the limits of visual-inertial navigation, providing insights into real-world viability as shown in Fig.10.

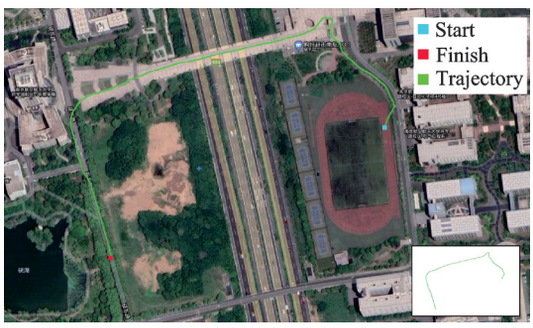


Fig.10 Real environment state estimation trajectory

Due to the four unobservable degrees of freedom inherent in tightly coupled VINS algorithms like VINS-Mono[7], some cumulative drift is inevitable along these directions. The loosely coupled VINS-Fusion[14] mitigates long-term drift by fusing VI and GNSS in a decoupled architecture, bounding the drift to a constant level[15]. Our method provides better error control to both techniques. The visual-inertial constraints in our optimization effectively reduce the GNSS noise while still utilizing its global position information to limit drift. This approach delivers accurate and robust state estimates over extended trajectories. The specific comparison results are shown in Fig.11.

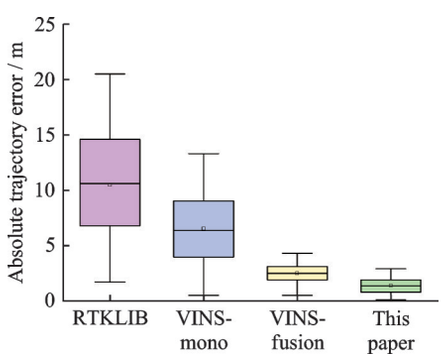


Fig.11 Absolute trajectory error of other methods and this paper method

4 Conclusions

We study on sensor fusion for navigation, developing a multi-sensor state estimation algorithm. Our key contributions encompass the formulation of the core estimation approach, experimental validation, and development of the hardware/software platform. The proposed method achieves robust and accurate state estimation by tightly fusing vision, inertial, and GNSS measurements under a joint graph optimization framework. Extensive experiments demonstrate the effectiveness of our sensor fusion approach in reducing drift and maintaining consistency across diverse environments.

References

[1] MOURIKIS A I, ROUMELIOTIS S I. A multi-state constraint Kalman filter for vision-aided inertial navigation[C]//Proceedings of the IEEE International Conference on Robotics and Automation. Rome, Italy: IEEE, 2007: 3565-3572.

[2] HUANG G P. Visual-inertial navigation: A concise review[C]//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada: IEEE, 2019: 9572-9582.

[3] GROVES P D. Principles of GNSS, inertial, and multisensor integrated navigation systems[M]. Boston: Artech House, 2013.

[4] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age[J]. IEEE Transactions on Robotics, 2016, 32(6): 1309-1332.

[5] STRASDAT H, MONTIEL J M, DAVISON A J. Visual SLAM: Why filter?[J]. Image and Vision Computing, 2010, 30(2): 65-77.

[6] KAESS M, JOHANNSSON H, ROBERTS R, et al. ISAM2: Incremental smoothing and mapping using the Bayes tree[J]. The International Journal of Robotics Research, 2012, 31(2): 217-236.

[7] LEUTENEGGER S, LYNEN S, BOSSE M, et al. Key frame-based visual-inertial odometry using nonlinear optimization[J]. The International Journal of Robotics Research, 2015, 34(3): 314-334.

[8] QIN T, LI P, SHEN S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.

[9] DUNIK J, LUKES Z, SLAVÍK P, et al. Improvement of GNSS positioning accuracy using INS/GNSS integration[J]. Advances in Space Research, 2018, 61(1): 158-169.

[10] SOLÀ J. Quaternion kinematics for the error-state kalman filter[EB/OL]. (2017-11-03). https://arxiv.org/

abs/1711.02508.

[11] IGLHAUT J, CABO C, PULITI S, et al. Structure from motion photogrammetry in forestry: A review[J]. Current Forestry Reports, 2019, 5(3): 155-168.

[12] EL-SHEIMY N, YOUSSEF A. Inertial sensors technologies for navigation applications: state of the art and future trends[J]. Satellite Navigation, 2020, 1(4): 42-69.

[13] FRANK D, MICHAEL K. Factor graphs for robot perception[M]. Boston: Now Foundations and Trends, 2017.

[14] QIN T, SHEN S. Online temporal calibration for monocular visual-inertial systems[C]//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 3662-3669.

[15] GENEVA P, ECKENHOFF K, LEE W, et al. LIPS: Lidar-inertial 3D plane SLAM[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3143-3150.

Acknowledgement The work was supported in part by the Guangxi Power Grid Company's 2023 Science and Technology Innovation Project (No.GXKJXM20230169).

Authors Mr. **ZHU Zekun** received the B.S. degree in electrical engineering and intelligent control from Shanghai Maritime University, Shanghai, China, in 2021. He is currently pursuing the M.S. degree in College of Automation Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. He has currently published two international conference papers and presented at the conference. He has participated in multiple research projects and holds several technical patents. His research interests are computer vision localization and robotic environment perception.

Prof. **YANG Zhong** received the B.S., M.S., and Ph.D. degrees from NUAA, Nanjing, China, in 1991, 1994, and 1998, respectively. In 2005, he completed his postdoctoral research at the Control Science and Engineering Postdoctoral Research Station of NUAA. He is a professor at NUAA. He has received more than ten awards at various levels, published dozens of high-quality papers, and holds numerous technical patents. His research is focused on UAV flight control, intelligent robot motion control, and computer vision.

Author contributions Mr. **ZHU Zekun** formulated the overall technical direction, designed the multi-sensor fusion model, developed the data fusion and state estimation methods, and performed the overall experimental design and result analysis, culminating in the writing of the paper. Prof. **YANG Zhong** provided guidance on the proposed technical methods and feasibility, as well as oversight of the overall completeness of the paper. Dr. **XUE Bayang** enhanced the reliability of the algorithms and provided useful feedback on specific programs. Dr. **ZHANG Chi** offered relevant guidance on the writing style of the paper, and Mr. **YANG Xin** assisted in improving the writing of the research background and significance. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editors: WANG Jie, XU Chengting)

基于因子图优化的无人系统GNSS/INS/视觉多传感器融合状态估计方法

朱泽堃，杨　忠，薛八阳，张　驰，杨　欣

(南京航空航天大学自动化学院，南京 211106，中国)

摘要：随着无人驾驶技术、智能机器人和无人机的发展，高精度定位、导航和状态估计技术也取得了很大进步。传统的全球导航卫星/惯性(Global navigation satellite system / inertial navigation system, GNSS/INS)集成导航系统可以持续提供高精度的导航信息。然而，当该系统应用于室内或GNSS受限环境(如具有强电磁干扰和复杂密集空间的户外变电站)时，通常无法获得高精度的GNSS定位数据。定位和定向误差会迅速发散和积累，无法满足大规模和长距离导航场景中的高精度定位要求。本文提出了一种基于非线性因子图优化的GNSS/INS/视觉融合的高精度状态估计方法。通过收集的实验数据和仿真结果，该系统在室内环境和部分GNSS信号丢失的环境中表现良好。

关键词：状态估计；多传感器融合；组合导航；因子图优化；复杂环境