J. Shanghai Jiao Tong Univ. (Sci.), 2023, 28(1): 70-76
https://doi.org/10.1007/s12204-023-2570-6

Lidar-Visual-Inertial Odometry with Online Extrinsic Calibration

MAO Tianyang (茅天阳), *ZHAO Wentao* (赵文韬),
*WANG Jingchuan** (王景川), *CHEN Weidong* (陈卫东)
(Department of Automation, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China;
Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai 200240, China;
Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai 200240, China)

Abstract: To achieve precise localization, autonomous vehicles usually rely on a multi-sensor perception system surrounding the mobile platform. Calibration is a time-consuming process, and mechanical distortion will cause extrinsic calibration errors. Therefore, we propose a lidar-visual-inertial odometry, which is combined with an adapted sliding window mechanism and allows for online nonlinear optimization and extrinsic calibration. In the adapted sliding window mechanism, spatial-temporal alignment is performed to manage measurements arriving at different frequencies. In nonlinear optimization with online calibration, visual features, cloud features, and inertial measurement unit (IMU) measurements are used to estimate the ego-motion and perform extrinsic calibration. Extensive experiments were carried out on both public datasets and real-world scenarios. Results indicate that the proposed system outperforms state-of-the-art open-source methods when facing challenging sensor-degenerating conditions.

Key words: autonomous vehicles, multi-sensor fusion, online calibration
CLC number: TP 242.6, TP 391 **Document code:** A

0 Introduction

Nowadays, autonomous vehicles are being equipped with multi-sensor systems to perform diverse perception tasks, including map construction[1], vehicle tracking[2], and six degrees of freedom (6-DoF) ego-motion estimation[3]. With the advantages of robustness and accuracy, multi-sensor fusion solutions, including visual-inertial odometry (VIO)[4-6] and lidar-inertial odometry (LIO)[7-9] are becoming popular research trends to address the ego-motion estimation problem. Although lidar-based methods can extract detailed structural measurements at a large scale, their performance tends to degrade when facing circumstances without sufficient structural features. Vision-based methods are effective when operating in a textured environment, but tend to fail under low light or textureless conditions. Both methods can cover only limited situations; thus increasing attention has been paid to the combination of lidar, camera, and inertial measurement unit (IMU) sensors to obtain better localization performance. To fuse heterogeneous sensors and produce precise odometry information, extrinsic calibration between sensors should be performed beforehand. There are many calibration methods applied to autonomous vehicles that achieve outstanding performance in terms of calibration accuracy[10-16]. However, they usually require a calibration target to be observed by all sensors, which is difficult for autonomous vehicles to achieve, because sensors are distributed in various places on the self-driving platform, and most of them do not share overlapping views. Therefore, online calibration methods are receiving increased attention by researchers.

In this paper, a lidar-visual-inertial system based on nonlinear optimization is proposed to estimate poses and perform extrinsic calibration online robustly and accurately. The main contributions of this study are: ① An adapted sliding window mechanism is proposed to manage measurements at different frequencies from heterogeneous sensors including 3D lidar, cameras, and IMUs. ② A nonlinear optimization formulation is proposed to simultaneously perform online calibration and motion estimation using visual features, cloud features, and IMU measurements. ③ A series of extensive experiments are conducted to validate the performance of the proposed system. Results show that our method

Received: 2021-12-10 **Accepted:** 2021-12-29
Foundation item: the National Key R&D Program of China (No. 2020YFC2007500), and the National Natural Science Foundation of China (No. U2013203)
***E-mail:** jchwang@sjtu.edu.cn

outperforms state-of-the-art (SOTA) methods.

1 Related Work

In this section, we review multi-sensor calibration and multi-sensor fusion odometry, which are the most relevant to our proposed concept.

1.1 Multi-Sensor Calibration

Regarding multi-sensor calibration methods, an offline algorithm was presented to enable extrinsic calibration of a camera and lidar sharing a common view[14]. Corner features in images and planar features in point clouds are extracted from a chessboard and are matched to estimate the extrinsic views. However, the method requires an object (the chessboard) that can be observed by both sensors, which is impractical for calibrating the sensors on self-driving platforms. The multi-state constrained extended Kalman filter (MSCKF) framework estimates motion and online calibrated extrinsic parameters for a camera-IMU system[4]. The algorithm requires low computational resources, but achieves relatively low accuracy because a linear approximation is performed at every update step. Recently, the M-LOAM[17] online extrinsic initialization method based on hand-eye calibration has been proposed. It requires no extrinsic prior, but it estimates rotational and translation parts separately, which reduces its accuracy, especially in the translation estimates.

1.2 Multi-Sensor Fusion Odometry

LIC-fusion[18] fuses asynchronous lidar, camera, and IMU measurements within an MSCKF framework to estimate poses and calibrate extrinsic parameters online. However, as mentioned previously, all MSCKF-based methods have theoretical defects because of the linear approximations used. R^{2}LIVE[19] uses a high-frequency lidar with embedded IMU and a global shutter camera to realize high-rate iEKF-based odometry and factor graph optimization to refine poses and visual features within a sliding window. However, the method combines filtering and optimization directly, which are two different theoretical frameworks resulting in reduced accuracy. LVI-SAM[20] is the integration of VINS-Mono[6] and LIO-SAM[8]. The VIO subsystem leverages point cloud information to perform depth registration, which normally provides a good guess in the VIO initialization stage. The LIO subsystem leverages the odometry information obtained from the VIO to provide an initial guess for scan-matching. However, the combination of two subsystems means that each subsystem cannot fully utilize the correlation among the three sensors, which degrades the overall performance of the system.

Our proposed method can be classified as a tightly-coupled method. The most similar concept to ours is LVI-SAM[20]. The main difference is that our method handles IMU measurements, visual features, and cloud features simultaneously within a sliding window as a whole to achieve better localization accuracy, whereas LVI-SAM constructs lidar-inertial and visual-inertial subsystems separately.

2 Optimization-Based Lidar-Visual-Inertial System

2.1 System Overview

The framework of the proposed system is shown in Fig. 1, where $\boldsymbol{T}_{B_{t_l}}^{B_{t_c}}$ is the propagated transformation, and $\boldsymbol{b}_a$ and $\boldsymbol{b}_\omega$ are the IMU accelerometer and gyroscope biases respectively. In the measurement preprocessing step, visual features are detected and tracked between successive images based on the Kanade-Lucas-Tomasi (KLT) algorithm[21]. Cloud features are extracted from point clouds based on curvature and tracked within the local map by 3D lidar, similar to

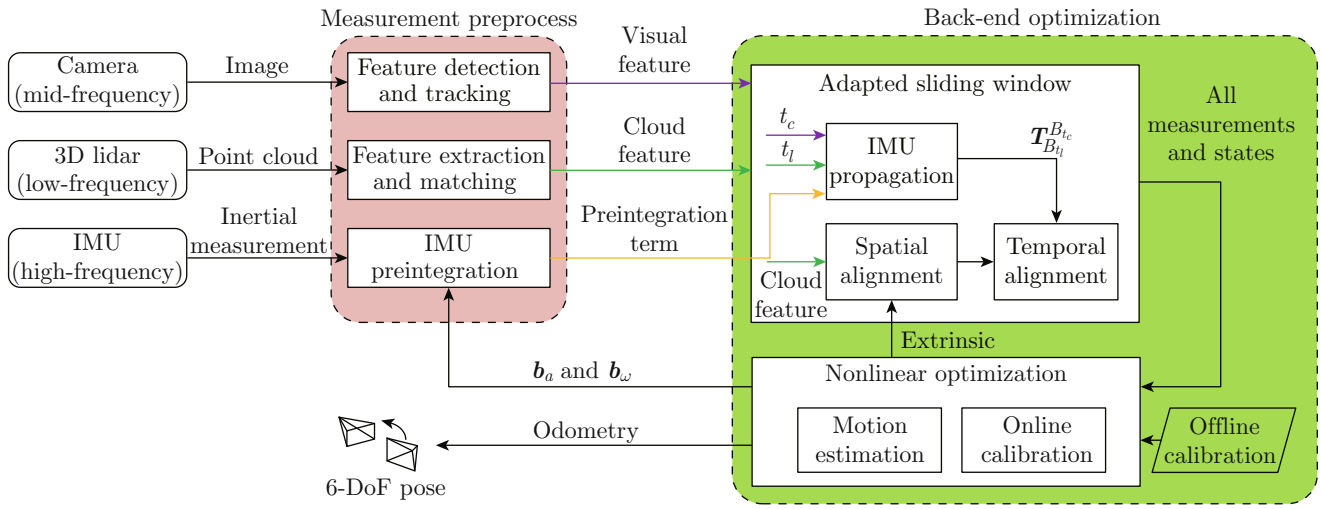


Fig. 1 Overview of the proposed system

LIO-SAM[8]. IMU pre-integration terms are calculated using all IMU measurements between successive visual keyframes.

All preprocessed measurements are managed by an adapted sliding window. When a cloud feature (lowest frequency feature) arrives, it triggers the spatial-temporal alignment step. The cloud feature is first spatially aligned to the body frame using the extrinsic. Denoting the cloud feature arrival time as t_l, after finding the most recent visual feature time t_c, we perform temporal alignment using IMU measurements between t_c and t_l.

A nonlinear optimization problem is constructed involving all system states and measurements, which refines the system states (motion and extrinsic) within the sliding window. If we denote W, B, C, and L as the frames of world, IMU, camera, and 3D lidar respectively, then the state variables to be optimized are defined as the motion states $\boldsymbol{x}_{B_{t_i}}^W (i = 0, 1, \cdots, N)$ within the sliding window range $N + 1$, camera extrinsic $\boldsymbol{x}_C^B$, 3D lidar extrinsic $\boldsymbol{x}_L^C$, and inverse depth d_j $(j = 1, 2, \cdots, m)$ for the observed visual features. The complete definition of state variables is

$$\boldsymbol{x} = [\boldsymbol{x}_{B_{t_0}}^W \quad \cdots \quad \boldsymbol{x}_{B_{t_N}}^W \quad \boldsymbol{x}_C^B \quad \boldsymbol{x}_L^C \quad d_1 \quad \cdots \quad d_m]. \tag{1}$$

Specifically, $\boldsymbol{x}_{B_{t_i}}^W$, $\boldsymbol{x}_C^B$, and $\boldsymbol{x}_L^C$ are in the forms:

$$\left.\begin{aligned} \boldsymbol{x}_{B_{t_i}}^W &= [\boldsymbol{R}_{B_{t_i}}^W \quad \boldsymbol{p}_{B_{t_i}}^W \quad \boldsymbol{v}_{B_{t_i}}^W \quad \boldsymbol{b}_{a,t_i} \quad \boldsymbol{b}_{\omega,t_i}] \\ \boldsymbol{x}_C^B &= [\boldsymbol{R}_C^B \quad \boldsymbol{p}_C^B] \\ \boldsymbol{x}_L^C &= [\boldsymbol{R}_L^C \quad \boldsymbol{p}_L^C] \end{aligned}\right\}, \tag{2}$$

where, $\boldsymbol{R}_{B_{t_i}}^W$ and $\boldsymbol{p}_{B_{t_i}}^W$ represent the orientation (described by a rotation matrix) and the position (described by 3D vectors) of the body frame at the time t_i, respectively; $\boldsymbol{v}_{B_{t_i}}^W$ is the velocity vector; $\boldsymbol{b}_{a,t_i}$ and $\boldsymbol{b}_{\omega,t_i}$ are the IMU accelerometer and gyroscope biases at the time t_i, respectively; $\boldsymbol{R}_C^B$ and $\boldsymbol{p}_C^B$ consist of the extrinsic between the camera and IMU; $\boldsymbol{R}_L^C$ and $\boldsymbol{p}_L^C$ are the extrinsic between the 3D lidar and camera.

Finally, it should be noted that the visual loop closure module of VINS-Mono is directly utilized to eliminate accumulated localization errors[6].

2.2 Adapted Sliding Window

The sliding window mechanism of our system is adapted from VINS-Mono[6]. The main difference lies in feature management, as shown in Fig. 2. Considering the measurement at the lowest frequency (cloud features) as reference, it always performs spatial-temporal alignment with the matching visual feature. However,

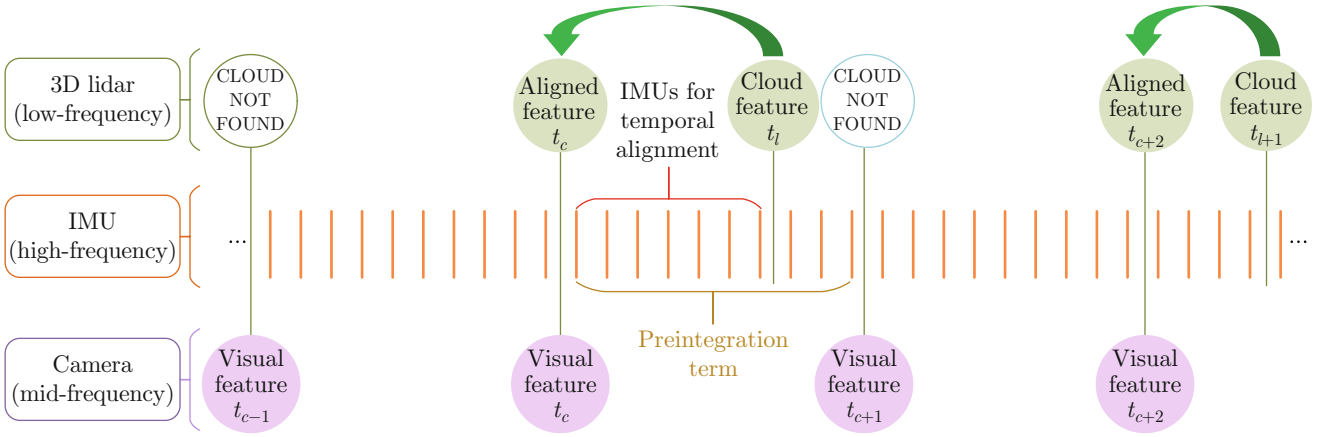


Fig. 2 Adapted sliding window: IMU measurements represented by orange bars, cloud features by green circles, and visual features by purple circles

because of its higher frequency, a visual feature may not always find a matching cloud feature (as indicated by the dashed circle stating "CLOUD NOT FOUND" in Fig. 2). We next describe the spatial-temporal alignment in detail.

When a new frame of cloud feature $\boldsymbol{P}_{t_l}^L$ arrives at the time t_i, it is first spatially aligned to the body frame, i.e.,

$$\boldsymbol{P}_{t_l}^B = \boldsymbol{R}_C^B(\boldsymbol{R}_L^C \boldsymbol{P}_{t_l}^L + \boldsymbol{p}_L^C) + \boldsymbol{p}_C^B. \tag{3}$$

After spatial alignment, $\boldsymbol{P}_{t_l}^B$ is temporally aligned to the most recent visual feature time t_c:

$$\left.\begin{aligned} & t_c = \max\{t_0, t_l, \cdots, t_N\} \\ & \text{s.t.} \quad t_c \leqslant t_l \end{aligned}\right\}. \tag{4}$$

By denoting $\boldsymbol{T}_X^Y$ as the transformation matrix between the frame Y and the frame X, the propagated transformation $\boldsymbol{T}_{B_{t_l}}^{B_{t_c}} = \begin{bmatrix} \boldsymbol{R}_{B_{t_l}}^{B_{t_c}} & \boldsymbol{p}_{B_{t_l}}^{B_{t_c}} \\ \boldsymbol{0} & \boldsymbol{1} \end{bmatrix}$ is computed using IMU measurements in the time interval $[t_c, t_l]$. IMU measurements are shown within the red bracket in Fig. 2.

The temporal alignment is then formulated as

$$\boldsymbol{P}_{t_c}^{B}=\boldsymbol{R}_{B_{t_l}}^{B_{t_c}}\boldsymbol{P}_{t_l}^{B}+\boldsymbol{p}_{B_{t_l}}^{B_{t_c}}. \tag{5}$$

After performing these two steps, initial spatial-temporal alignment is accomplished. However, $\boldsymbol{b}_a$ and $\boldsymbol{b}_\omega$ may change after nonlinear optimization; therefore, if the change in either bias exceeds a certain threshold, IMU re-propagation is performed to update $\boldsymbol{T}_{B_{t_l}}^{B_{t_c}}$ and realign cloud features.

2.3 Nonlinear Optimization with Online Calibration

Using the state variables defined in Eq. (1) and sensor measurements, we formulate a nonlinear least squares problem to jointly perform motion estimation and online calibration:

$$\begin{aligned}\bar{\boldsymbol{x}}=\arg\min_{\boldsymbol{x}}\{&\|r_p-H_p\boldsymbol{x}\|^2+\\&\sum_{i\in B}\|r_B(\hat{z}_{B_{i+1}}^{B_i},\boldsymbol{x})\|_{\boldsymbol{P}_{B_{i+1}}^{B_i}}^2+\\&\sum_{(l,j)\in C}\rho(\|r_C(\hat{z}_l^{C_j},\boldsymbol{x})\|_{\boldsymbol{P}_l^{C_j}}^2)+\\&\sum_{m\in E}\rho(\|r_E(\boldsymbol{M}_l^E,\hat{z}_m^E,\boldsymbol{x})\|_{\boldsymbol{P}_m^E}^2)+\\&\sum_{n\in P}\rho(\|r_P(\boldsymbol{M}_l^P,\hat{z}_n^P,\boldsymbol{x})\|_{\boldsymbol{P}_n^P}^2)+\\&\|r_{\mathrm{ex}}(\boldsymbol{x},\boldsymbol{x}_{\mathrm{ini}})\|_{\boldsymbol{P}^{\mathrm{ex}}}^2\},\end{aligned} \tag{6}$$

where, $\rho(\cdot)$ is the Huber function[22]; $\|r(\cdot)\|_{\boldsymbol{P}}$ is the covariance matrix of residual r; $r_p-H_p\boldsymbol{x}$ denotes the marginalized prior; $r_B(\hat{z}_{B_{i+1}}^{B_i},\boldsymbol{x})$ and $r_C(\hat{z}_l^{C_j},\boldsymbol{x})$ are the IMU pre-integration residual and visual reprojection residual, respectively. The derivations of $r_B(\hat{z}_{B_{i+1}}^{B_i},\boldsymbol{x})$ and $r_C(\hat{z}_l^{C_j},\boldsymbol{x})$ are presented in the discussions involving VINS-Mono[6]. In addition, $r_E(\boldsymbol{M}_l^E,\hat{z}_m^E,\boldsymbol{x})$ and $r_P(\boldsymbol{M}_l^P,\hat{z}_n^P,\boldsymbol{x})$ are residuals for lidar edge and planar features, respectively, which are discussed below; $r_{\mathrm{ex}}(\boldsymbol{x},\boldsymbol{x}_{\mathrm{ini}})$ is the residual term of the extrinsic prior. Equation (6) forms a maximum a posteriori (MAP) estimation minimizing the sum of the marginalized prior and the Mahalanobis norm of all measurement residuals, and the C++ library Ceres[23] is used for the nonlinear optimization implementation.

The lidar edge residual $r_E(\boldsymbol{M}_l^E,\hat{z}_m^E,\boldsymbol{x})$ is defined as the error between an edge feature point $\hat{z}_m^E$ (described as 3D vector $\boldsymbol{P}_{m,E}^L$) and a matched edge in the local edge feature map $\boldsymbol{M}_l^E$ under the state $\boldsymbol{x}$ of the frame l, where $\boldsymbol{M}_l^E$ is defined as the union of N edge cloud frames $\boldsymbol{P}_{l-i}^{W,E}(i=1,2,\cdots,N)$ in the world coordinate system, such that

$$\boldsymbol{M}_l^E=\boldsymbol{P}_{l-1}^{W,E}\cup\boldsymbol{P}_{l-2}^{W,E}\cup\cdots\cup\boldsymbol{P}_{l-N}^{W,E}. \tag{7}$$

The edge residual term is defined as the point-to-line error:

$$r_E(\boldsymbol{M}_l^E,\hat{z}_m^E,\boldsymbol{x})=\frac{(\boldsymbol{P}_{m,E}^W-\boldsymbol{P}_A)\times(\boldsymbol{P}_{m,E}^W-\boldsymbol{P}_B)}{\|\boldsymbol{P}_A-\boldsymbol{P}_B\|}, \tag{8}$$

where $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$ are the matched edge points in local map $\boldsymbol{M}_l^E$, and $\boldsymbol{P}_{m,E}^W\in\boldsymbol{P}_l^{W,E}$ represents the world coordinates of the edge feature extracted from the current lidar scan.

Similarly, we define $r_P(\boldsymbol{M}_l^P,\hat{z}_n^P,\boldsymbol{x})$ as the residual between a planar feature $\hat{z}_n^P$ (described as 3D vector $\boldsymbol{P}_{n,P}^L$) and the matched planar surfel (surface element) in the local planar feature map $\boldsymbol{M}_l^P$ consisting of N planar cloud frames $\boldsymbol{P}_{l-i}^{W,P}(i=1,2,\cdots,N)$, such that

$$\boldsymbol{M}_l^P=\boldsymbol{P}_{l-1}^{W,P}\cup\boldsymbol{P}_{l-2}^{W,P}\cup\cdots\cup\boldsymbol{P}_{l-N}^{W,P}. \tag{9}$$

The lidar planar residual $r_P(\boldsymbol{M}_l^P,\hat{z}_n^P,\boldsymbol{x})$ is defined as the point-to-plane error:

$$r_P(\boldsymbol{M}_l^P,\hat{z}_n^P,\boldsymbol{x})=-\boldsymbol{n}^{\mathrm{T}}\boldsymbol{P}_{n,P}^W+d, \tag{10}$$

where, $\boldsymbol{n}$ and d are the norm vector and constant term of the matched surfel, respectively; $\boldsymbol{P}_{n,P}^W\in\boldsymbol{P}_l^{W,P}$ represents the world coordinates of the planar feature in the current frame.

The extrinsic prior residual $r_{\mathrm{ex}}(\boldsymbol{x},\boldsymbol{x}_{\mathrm{ini}})$ is defined as the error between initial state $\boldsymbol{x}_{\mathrm{ini}}$ and current state $\boldsymbol{x}$:

$$r_{\mathrm{ex}}(\boldsymbol{x},\boldsymbol{x}_{\mathrm{ini}})=\boldsymbol{x}\boxminus\boldsymbol{x}_{\mathrm{ini}}, \tag{11}$$

where $\boxminus$ is the SE(3) 'translation and rotation in 3D' operator. In addition,

$$\boldsymbol{x}-\boldsymbol{x}_{\mathrm{ini}}=\begin{bmatrix}\ln(\boldsymbol{R}\boldsymbol{R}_{\mathrm{ini}}^{\mathrm{T}})^{\vee}\\ \boldsymbol{p}-\boldsymbol{p}_{\mathrm{ini}}\end{bmatrix}, \tag{12}$$

where $\ln(\cdot)$ is the logarithm map defined on the SO(3) '3D rotation group', and $(\cdot)^{\vee}$ is the operator mapping a skew-symmetric matrix to its corresponding vector.

3 Experiments and Analysis of Results

To verify the performance of the proposed method, several experiments were carried out on both public datasets and real scenarios. The proposed lidar-visual-inertial odometry is compared with other SOTA fusion methods including VINS-Mono[6], LIO-SAM[8], and LVI-SAM[20].

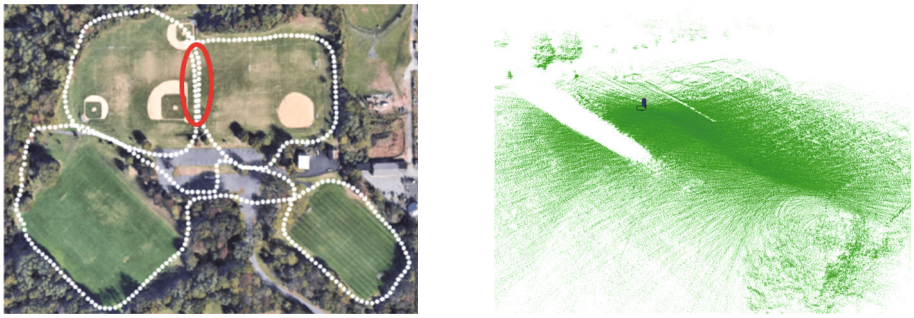
3.1 Experiments on Public Dataset

First, we evaluate the performance of the proposed system on a public dataset. The handheld dataset was collected on a baseball field and surrounding area with a handheld sensor suite consisting of a Velodyne VLP-16 3D lidar, FLIR BFS-U3-04S2M-CS monocular camera, MicroStrain 3DM-GX5-25 IMU, and Reach

RS+GPS[20]. The environment is shown in Fig. 3, where a lidar-degenerating scene on flat ground is denoted by the red ellipse in Fig. 3(a).

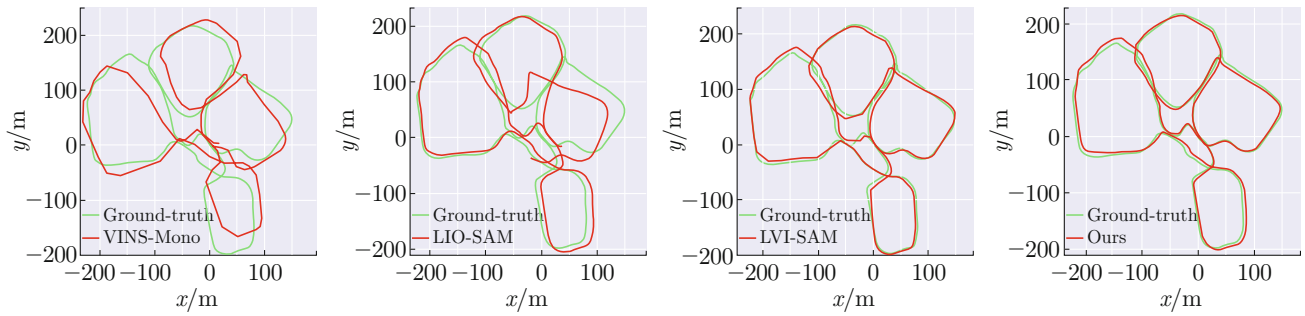
To verify the performance of the proposed method, we compare VINS-Mono[6], LIO-SAM[8], LVI-SAM[20], and our system using the dataset. Because lidar-based loop closure is not implemented in our system, we turn off the lidar-based loop closure of LVI-SAM to ensure an equitable comparison. All methods are implemented with C++ and executed on a personal computer with an Intel i7-8700 3.20 GHz CPU and 16 GB RAM.

The localization results for the entire area in Fig. 3(a) are shown in Table 1. All trajectories are evaluated using the absolute trajectory error (ATE). VINS-Mono exhibits the worst performance because there are few distinct visual features that can be used to find a loop; thus the localization error accumulates. When considering the flat ground, LIO-SAM cannot extract sufficient structural information, and thus the localization error is also large, whereas our system leverages visual features of the flat ground to perform accurate localization. Compared with our system, LVI-SAM is a relatively loosely-coupled system, and although it performs well, our system outperforms all three methods. The result shows that our method achieves the lowest root mean squared error (RMSE) ATE of 5.105 m, indicating that our system is robust, even in the degenerated flat area. Trajectories for all methods are shown in Fig. 4, where the ground-truth from GPS positioning measurements is shown in green.



(a) Handheld dataset environment (b) Flat ground

Fig. 3 Handheld dataset environment



(a) VINS-Mono (b) LIO-SAM (c) LVI-SAM (d) Ours

Fig. 4 Trajectories for different methods using the handheld dataset

Table 1 Comparison of localization results for the handheld dataset

Method	VINS-Mono[6]	LIO-SAM[8]	LVI-SAM[20]	Ours
RMSE ATE/m	28.470	14.372	7.570	5.105

3.2 Experiments in Real-World Scenario

We further evaluate the performance of the proposed system using real-world experiments. The experimental platform is a JiaoLong intelligent wheelchair equipped with a sensor box, as shown in Fig. 5. The sensor box includes a Robosense RS-163D lidar (10 Hz), Realsense D435i camera (30 Hz) with embedded IMU (200 Hz), and CHCNAV CGI-210 GNSS/INS device (to provide the ground-truth, 100 Hz), and all of the sensors are not synchronized beforehand.

Two individual experiments were carried out on the campus of Shanghai Jiao Tong University at night and midday (Fig. 6) to evaluate the localization robustness and online calibration effectiveness of the proposed method, respectively. Trajectory of the ground-truth (GNSS/INS) is the blue line in Fig. 6(a); Fig. 6(b) shows images collected at the same location under different illumination conditions (top-right: midday; bottom-right: night).

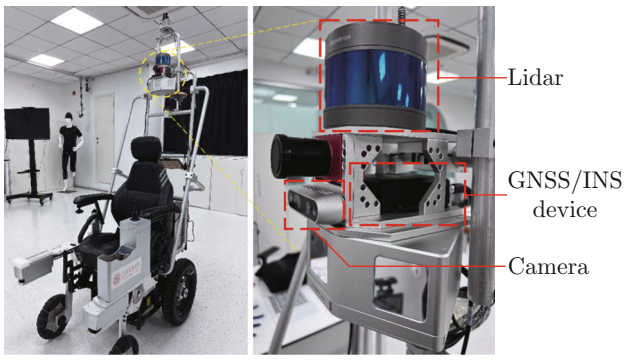
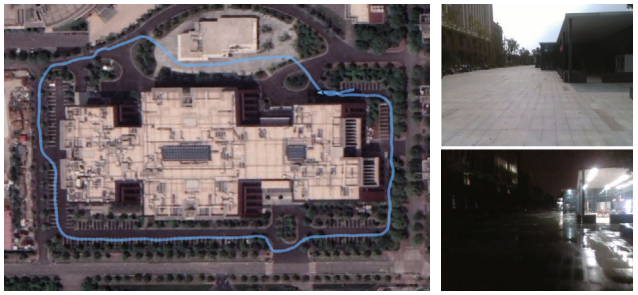


Fig. 5 JiaoLong wheelchair and hardware set-up



(a) Top view (b) Scene images

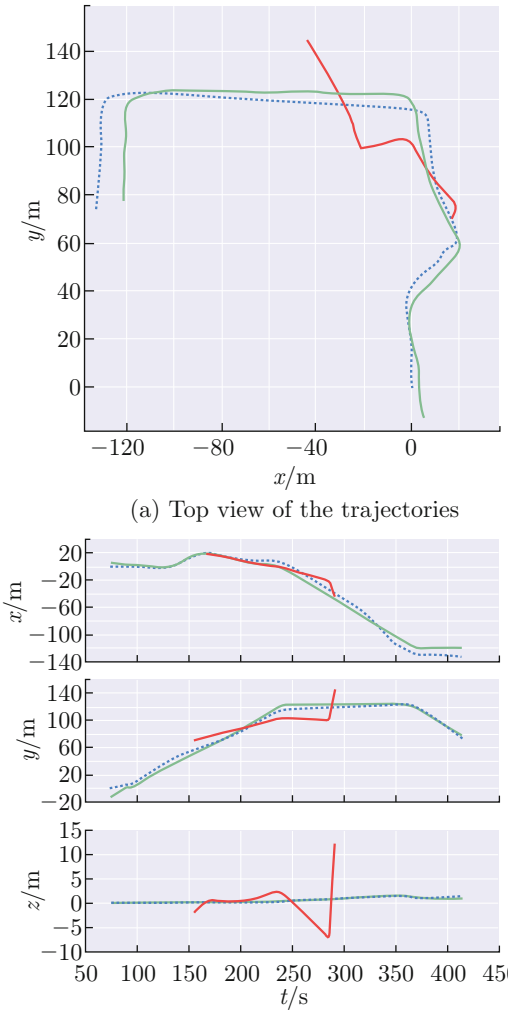
Fig. 6 Outdoor experiment area

To evaluate the robustness of our system under low illumination conditions, we compare our system with LVI-SAM during the night scene, as LVI-SAM configures the same sensors as ours. It should be noted that because visual-based methods will fail quickly in dark scenes, whereas lidar-based methods will not be affected under low illumination, we do not compare VINS-Mono and LIO-SAM. The localization results are shown in Table 2. Results demonstrate that the accuracy of LVI-SAM is much worse than that of our system. In fact, LVI-SAM quickly fails, whereas our system achieves robust pose estimation. On the one hand, LVI-SAM is a combination of two subsystems, i.e., VIO and LIO, which means that it cannot make full use of the correlations among the sensors. Although it possesses a failure detection mechanism, the indicators reflecting failure status are relatively simple. Under the low illumination conditions in our experiment, even when VIO failed to produce an accurate initial guess of pose for LIO, the mechanism did not report an error, which finally caused the entire system to fail. On the other hand, our system couples visual features, cloud features, and IMU measurements into one odometry, which is more robust when operating in sensor-degenerating environments. Trajectories for both methods are shown in Fig. 7.

To evaluate the performance of the online calibration used in our method, another experiment was executed at midday. Localization results are listed in Table 3, where we denote Ours-Fix as our system with fixed extrinsic, and Ours-OC indicates our system with online calibration. Evaluation metrics include maximum (Max), Mean, minimum (Min), and RMSE. Results show that with the help of online calibration, our system achieves better localization accuracy, especially

Table 2 Comparison of localization results for the real-world experiment

Method	LVI-SAM[20]	Ours
RMSE ATE/m	17.715	8.782



(a) Top view of the trajectories

(b) Trajectories along x, y, and z axes

Fig. 7 Localization results at night

Table 3 ATE results for online calibration using the proposed method

Method	Max/m	Mean/m	Min/m	RMSE/m
Ours-Fix	4.873	1.739	0.076	1.904
Ours-OC	3.767	1.645	0.052	1.821

in terms of maximum error.

4 Conclusion

We propose a lidar-visual-inertial odometry technique that performs robust 6-DoF pose estimations and can be calibrated online. An adapted sliding window mechanism is proposed to manage the measurements arriving at different frequencies from heterogeneous sensors. Sparse visual features, cloud features, and IMU measurements are combined to formulate residual terms, which are then used in a MAP problem to solve the nonlinear optimization. The experimental results indicate that our system achieves better localization accuracy when facing sensor-degenerating conditions than existing SOTA methods. In the future, loop closure based on lidar will be implemented to further improve localization precision.

References

[1] ZHUANG H Y, ZHOU X J, WANG C X, et al. Wavelet transform-based high-definition map construction from a panoramic camera [J]. *Journal of Shanghai Jiao Tong University* (*Science*), 2021, **26**(5): 569-576.

[2] CHEN J C, LI L, YANG X B. Efficient online vehicle tracking for real–virtual mapping systems [J]. *Journal of Shanghai Jiao Tong University* (*Science*), 2021, **26**(5): 598-606.

[3] AN L F, ZHANG X Y, GAO H B, et al. Semantic segmentation–aided visual odometry for urban autonomous driving [J]. *International Journal of Advanced Robotic Systems*, 2017, **14**(5): 1-11.

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

[5] LI M Y, MOURIKIS A I. High-precision, consistent EKF-based visual-inertial odometry [J]. *The International Journal of Robotics Research*, 2013, **32**(6): 690-711.

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

[7] ZHANG J, SINGH S. Low-drift and real-time lidar odometry and mapping [J]. *Autonomous Robots*, 2017, **41**(2): 401-416.

[8] SHAN T X, ENGLOT B, MEYERS D, et al. LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping [C]//*2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Las Vegas, NV: IEEE, 2020: 5135-5142.

[9] XU W, ZHANG F. FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter [J]. *IEEE Robotics and Automation Letters*, 2021, **6**(2): 3317-3324.

[10] PEREIRA M, et al. Self calibration of multiple LIDARs and cameras on autonomous vehicles [J]. *Robotics and Autonomous Systems*, 2016, **83**: 326-337.

[11] KIM D H, KIM G W. Efficient calibration method of multiple LiDARs on autonomous vehicle platform [C]//*2020 IEEE International Conference on Big Data and Smart Computing*. Busan: IEEE, 2020: 446-447.

[12] GOGINENI S. Multi-sensor fusion and sensor calibration for autonomous vehicles [J]. *International Research Journal of Engineering and Technology*, 2020, **7**(7): 1073-1078.

[13] BELTRÁN J, GUINDEL C, DE LA ESCALERA A, et al. Automatic extrinsic calibration method for LiDAR and camera sensor setups [J]. *Transactions on Intelligent Transportation Systems*, 2022, **23**(10): 1-13.

[14] KATO S, TAKEUCHI E, ISHIGURO Y, et al. An open approach to autonomous vehicles [J]. *IEEE Micro*, 2015, **35**(6): 60-68.

[15] FURGALE P, BARFOOT T D, SIBLEY G. Continuous-time batch estimation using temporal basis functions [C]//*2012 IEEE International Conference on Robotics and Automation*. Saint Paul, MN: IEEE, 2012: 2088-2095.

[16] LEVINSON J, THRUN S. Automatic online calibration of cameras and lasers [C]//*Robotics*: *Science and Systems IX*. Berlin: IEEE, 2013: 24-28.

[17] JIAO J H, YE H Y, ZHU Y L, et al. Robust odometry and mapping for multi-LiDAR systems with online extrinsic calibration [J]. *IEEE Transactions on Robotics*, 2022, **38**(1): 351-371.

[18] ZUO X X, GENEVA P, LEE W, et al. LIC-fusion: LiDAR-inertial-camera odometry [C]//*2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Macao: IEEE, 2019: 5848-5854.

[19] LIN J R, ZHENG C R, XU W, et al. R^2LIVE: A robust, real-time, LiDAR-inertial-visual tightly-coupled state estimator and mapping [J]. *IEEE Robotics and Automation Letters*, 2021, **6**(4): 7469-7476.

[20] SHAN T X, ENGLOT B, RATTI C, et al. LVI-SAM: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping [C]//*2021 IEEE International Conference on Robotics and Automation*. Xi'an: IEEE, 2021: 5692-5698.

[21] LUCAS B D, KANADE T. An iterative image registration technique with an application to stereo vision [C]//*7th International Joint Conference on Artificial Intelligence*. Vancouver: Morgan Kaufmann Publishers Inc, 1981: 674-679.

[22] HUBER P J. Robust estimation of a location parameter [M]//Breakthroughs in statistics. New York: Springer, 1992: 492-518.

[23] AGARWAL S, MIERLE K. Ceres solver: Tutorial & reference [EB/OL]. [2021-12-10]. http://www.ceres-solver.org.