PAPER

Fusion of binocular vision, 2D lidar and IMU for outdoor localization and indoor planar mapping

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **34** (2023) 025203 (15pp) https://doi.org/10.1088/1361-6501/ac9ed0

Fusion of binocular vision, 2D lidar and IMU for outdoor localization and indoor planar mapping

Zhenbin Liu[1,2], Zengke Li[1,2,*], Ao Liu[1,2], Yaowen Sun[1,2] and Shiyi Jing[1]

[1] School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, People's Republic of China
[2] MNR Key Laboratory of Land Environment and Disaster Monitoring, China University of Mining and Technology, Xuzhou 221116, People's Republic of China

E-mail: zengkeli@yeah.net

Received 28 June 2022, revised 26 September 2022
Accepted for publication 31 October 2022
Published 21 November 2022

Abstract

Emergent fields such as Internet of Things applications, driverless cars, and indoor mobile robots have brought about an increasing demand for simultaneous localization and mapping (SLAM) technology. In this study, we design a SLAM scheme called BVLI-SLAM based on binocular vision, 2D lidar, and an inertial measurement unit (IMU) sensor. The pose estimation provided by vision and the IMU can provide better initial values for the 2D lidar mapping algorithm and improve the mapping effect. Lidar can also assist vision to provide better plane and yaw angle constraints in weak texture areas and obtain higher precision 6-degree of freedom pose. BVLI-SLAM uses graph optimization to fuse the data of the IMU, binocular camera, and laser. The IMU pre-integration combines the visual reprojection error and the laser matching error to form an error equation, which is processed by a sliding window-based bundle adjustment optimization to calculate the pose in real time. Outdoor experiments based on KITTI datasets and indoor experiments based on the trolley mobile measurement platform show that BVLI-SLAM has different degrees of improvement in mapping effect, positioning accuracy, and robustness compared with VINS-Fusion and Cartographer, and can solve the problem of positioning and plane mapping in indoor complex scenes.

Keywords: 2D lidar, binocular vision, IMU, simultaneous localization and mapping, graph optimization

(Some figures may appear in colour only in the online journal)

1. Introduction

Simultaneous localization and mapping (SLAM) technology means that mobile carriers only carry their sensors to complete real-time self-positioning and map a perceived environment [1–3]. SLAM has been developed for nearly 30 years. Research has intensified in the last decade with the popularity of industries, such as autopilot, unmanned aerial vehicles, various service robots, and virtual reality/augmented reality, all of which have SLAM as their core basic function.

With the rapid development of SLAM in the past decade, it has evolved into two types of schemes, which are mainly based on visual cameras and lidar sensors [4]. Two types of SLAM solutions, namely, visual SLAM and lidar SLAM, have evolved. The constructed map is divided into a 2D grid map, a 3D sparse map, and a 3D dense map according to the form of sensors. In cases where the environment has adequate lighting, sufficient texture, and is composed of static rigid bodies, the existing visual SLAM can operate satisfactorily, meeting

* Author to whom any correspondence should be addressed.

the needs of localization and mapping with an error of 5 cm. However, visual SLAM technology cannot work well in environments with poor textures and insufficient lighting.

SLAM technology based on 2D lidar is mainly used for indoor floor mapping. Compared with 2D lidar, multi-threaded lidar can perform more robust localization and mapping in both indoor and outdoor environments, but itis more expensive [5]. Lidar-based SLAM technology is not affected by changes in ambient light texture, etc, but positioning errors are often encountered in a structured environment because the point cloud data at different locations have the same coordinate information. The researchers found that vision and lidar have very good complementary properties, and the SLAM fusion of vision and lidar is a current research hotspot [6].

However, researchers mostly focus on the fusion of multi-threaded lidar and camera, ignoring 2D lidar. Applications in indoor environments often only require 2D lidar, such as sweeping robots, while 3D lidar has a larger volume, which greatly increases its cost. However, mapping large-scene indoor environments is subject to factors such as the structural environment, glass curtain walls, and uneven ground. Only relying on 2D lidar to complete high-precision localization and mapping is very difficult. To cope with more indoor environments, this paper pursues the combination of low-cost 2D lidar, camera, and inertial measurement unit (IMU) to achieve high-precision 6-degree of freedom (DOF) pose estimation and 2D plane mapping. Given that 2D lidar is suitable for flat areas, obtaining sufficient motion excitation during IMU initialization is not easy. Therefore, this paper selects a binocular camera to replace the monocular camera to assist the initialization process, and the depth estimation based on the former has natural advantages over the latter. Based on 2D lidar, a more accurate heading angle can be obtained, compensating for the unobservable nature of the IMU heading angle. Therefore, a SLAM scheme based on binocular vision, 2D lidar, and IMU fusion is a low-cost optimal choice for high-precision real-time positioning and planar mapping. This paper presents a general framework, hereinafter referred to as BVLI-SLAM, based on factor graph optimization, which integrates a binocular camera, 2D lidar, and IMU sensors for high-precision real-time positioning and 2D mapping.

The main contributions of this paper are as follows.

(a) A theoretical framework is proposed based on graph optimization for the fusion of the binocular camera, 2D lidar, and IMU;
(b) The accuracy and robustness of the algorithm are verified through experiments on datasets and indoor complex scenes.

The remainder of the paper is organized as follows. Section 2 reviews the related work of SLAM technology of vision and laser fusion. Section 3 presents a detailed introduction to the proposed BVLI-SLAM scheme. Section 4 discusses the experimental tests performed on localization and mapping. Finally, a conclusion is given in section 5.

2. Related work

2.1. Visual SLAM

Visual SLAM systems that can run in real time, such as MonoSLAM [7] and PTAM [8], appeared in 2007. Since 2013, with the open-sourcing of many excellent visual SLAM schemes, such as SVO [9], LSD-SLAM [10], and ORB-SLAM2 [11], visual SLAM has received extensive attention and undergone rapid progress. If the environment is limited to scenarios with sufficient light, texture, and static rigid bodies, existing visual SLAM solutions can achieve centimeter-level positioning accuracy. Visual SLAM is too dependent on the environmental texture and other information, and positioning accuracy is heavily dependent on the environment. Therefore, some scholars proposed using IMU sensors to assist visual SLAM. Vision and IMU fusion methods include filtering and graph optimization. MSCKF [12] is a representative method based on filtering. Given that SLAM is a highly nonlinear system, the graph-based optimization method has been proven to achieve better accuracy than the algorithm based on filtering under the same computing power [13]. Therefore, the framework based on graph optimization has been widely studied by researchers since it was proposed, and many excellent open-source SLAM schemes have been obtained, represented by OKVIS [14], VINS-Fusion [15], and ORB-SLAM3 [16]. Table 1 summarizes some of the most representative open-source visual SLAM schemes, based on which most of the present research work is carried out.

2.2. Lidar SLAM

Compared with visual SLAM, the research on 2D lidar SLAM schemes was developed much earlier. The earliest lidar SLAM was mainly based on 2D lidar. The lidar SLAM schemes that rely on 2D lidar to establish maps can be divided into filter- and graph-based optimization according to the solution method. The filter-based method, derived from Bayesian estimation theory, is an early method to solve the SLAM problem. At present, the filter-based lidar SLAM scheme is mainly used in 2D indoor small-scale scenes. The SLAM scheme based on graph optimization considers more pose state and environmental observation information and uses a graph formed by nodes and edges to represent a series of mobile robot poses and constraints, which is a more efficient and popular optimization method. The filtering-based representative scheme EKF-SLAM [17] is computationally complex and has poor robustness to build maps. FastSLAM [18] was the first to realize the real-time output of grid maps, but it has disadvantages, such as memory consumption and serious particle dissipation. Gmapping [19] alleviates particle dissipation but relies heavily on odometer information. The optimal RBPF [20] further reduces the particle degradation problem. In the 1990s, SLAM

Table 1. Representative visual SLAM solutions.

Scheme name	Release time	Sensor form	Characteristics
MonoSLAM	2007	[a]	First real-time visual SLAM, EKF + sparse corners
PTAM	2007	[a]	Keyframe + BA, first optimized for back-end
SVO	2014	[a]	Sparse direct method
LSD-SLAM	2014	[a]	Direct method + semi-dense map
ORB-SLAM2	2015	[b]	ORB feature point + three-thread structure
MSCKF	2007	[c]	EKF-based VIO
OKVIS	2015	[c]	Optimized keyframe VIO
ROVIO [14]	2015	[c]	EKF-based VIO
VINS-Fusion	2019	[c]	Optical flow method + optimized back-end
ORB-SLAM3	2021	[c]	IMU initialization and fusion estimation, and submap function

[a] indicates support for monocular camera.
[b] indicates support monocular, binocular, and RGB-D cameras.
[c] indicates support for vision and IMU sensor fusion.

based on the pose graph was first proposed, but it was not popular because of its high computational complexity. In 2010, researchers realized the sparsity of the pose graph, which greatly reduced the computational complexity of SLAM based on the pose graph [21]. In 2014, Hector [22] based on scan to map was proposed, which is sensitive to the initial value and struggles to handle the closed loop. In 2016, Google's open-source solution Cartographer used a graph-based optimization framework and the branch and round approach method to accelerate the closed-loop solution process, which is the best solution today [23].

2.3. SLAM of vision and lidar fusion

Vision is rich in information and has good complementary properties to lidar, and the SLAM research that integrates vision and lidar has become a new research hotspot. V-LOAM [24] is a vision and laser fusion SLAM scheme based on the optimization method. The scheme assumes uniform velocity, has no loopback, and has good algorithm robustness. LVIO [25] uses vision, lidar, and IMU sensor fusion. It runs three modules in multiple layers in sequence to generate real-time self-motion estimation and processes coarse-to-fine data to generate high-frequency pose estimates and build low-drift maps over long distances. LVI-SAM [26] adopts the tight coupling scheme of the lidar, vision, and IMU fusion, and is a fusion of the lidar–inertial odometer (LIO)-SAM and VINS-Mono schemes. R3LIVE [27], the upgraded version of R2LIVE [28] from the MARS Laboratory of the University of Hong Kong, adopts a filtering method to integrate lidar, camera, and IMU. The above review summarizes the SLAM schemes based on vision and laser, as well as the excellent representative SLAM schemes of vision laser fusion.

This section provides a brief overview of SLAM, showing that SLAM research based on vision and lidar has made rapid progress. SLAM based on vision and lidar fusion has also been the subject of some research. However, compared with SLAM based on lidar, SLAM based on vision and laser fusion has no significant improvement in mapping effect, and the potential of vision and laser fusion has not been fully explored. In addition, researchers have focused on the fusion of 3D lidar and vision sensors, ignoring the development of multi-source fusion such as 2D lidar sensors and vision.

3. Principles and models

The BVLI-SLAM system designed in this paper comprises five parts: sensor data pre-processing, initial state estimation, local sliding window optimization, closed-loop detection, and global optimization. The overall framework of the BVLI-SLAM system is shown in figure 1. The constraint relationship between sensors is shown in figure 2. The functions and implementation ideas of each module are then introduced separately. In this paper, the external parameter calibration of 2D lidar, a binocular camera, and IMU is realized, and the calibration is considered reliable.

(a) Sensor data pre-processing. An image pyramid is constructed for each frame of the image acquired by the camera. Harris feature points are extracted for each layer of the image, quadtree is used to uniformize the feature points to obtain evenly distributed feature points, and the tracked feature points are pushed to the image queue. The IMU data are integrated to obtain the position, velocity, and rotation at the current moment. The pre-integration increment of adjacent image frames that will be used in the back-end optimization is calculated, as are the Jacobian matrix and covariance of the pre-integration error matrix item. The 2D lidar point cloud is de-distorted according to the IMU pre-integration positioning result, and the point cloud data of one frame is unified into the coordinate system of the first laser point.
(b) Initial state estimation. This part includes the LIO generated by the fusion of 2D laser and IMU, and the visual–inertial odometer (VIO) integrated with binocular vision and IMU. Using the pose estimation results of binocular vision, the acceleration bias and angular velocity bias of the IMU are calculated. At the same time, the binocular and IMU fusion results are aligned with the gravity vector. The 2D lidar data after de-distortion are projected onto the plane, and inter-frame matching based on correlation scan match (CSM) and gradient optimization are performed.

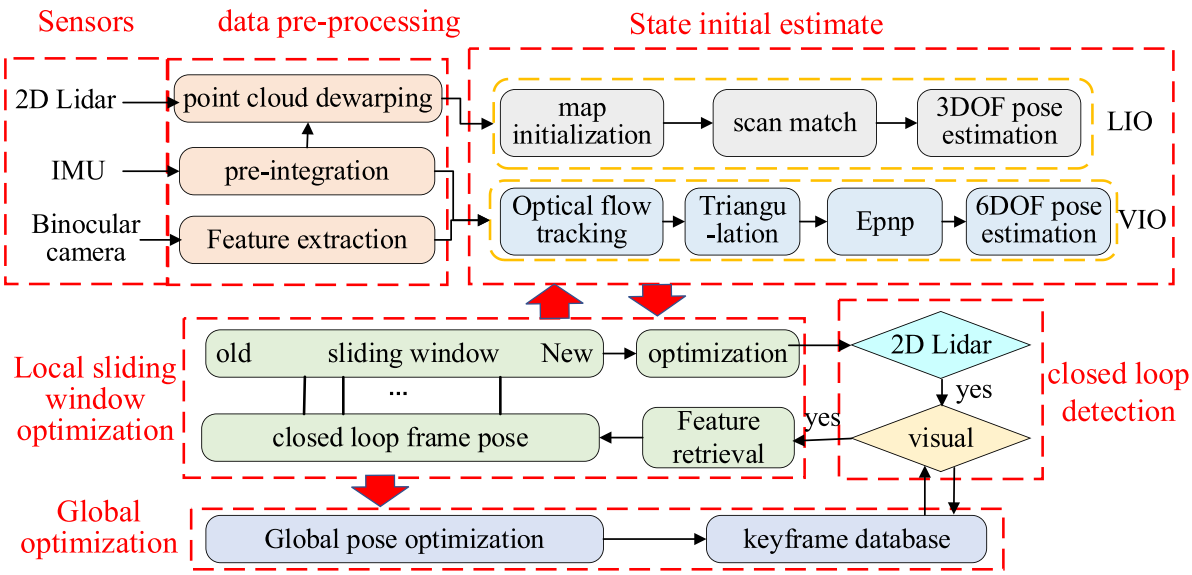


Figure 1. Overall framework of BVLI -SLAM scheme.

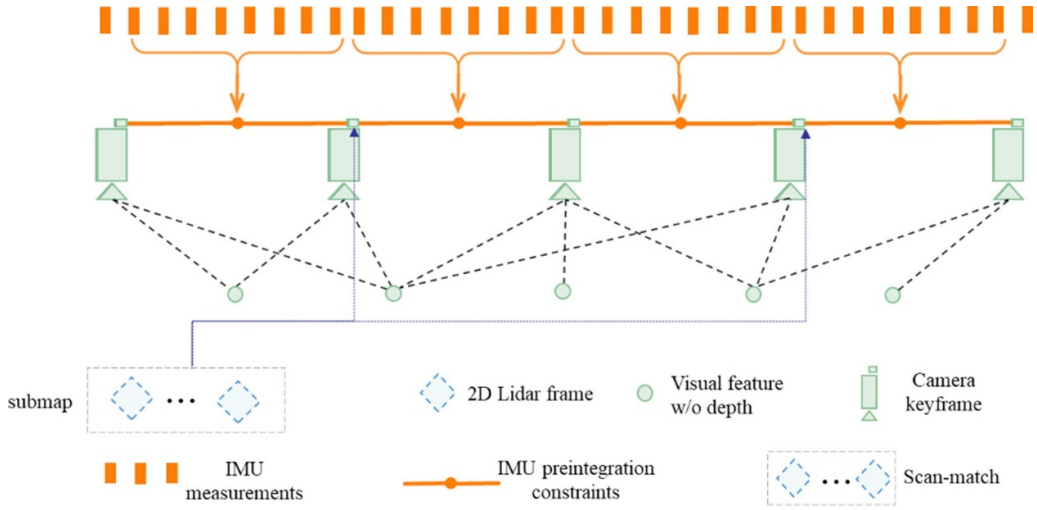


Figure 2. Sensor constraint structure diagram.

(c) Local sliding window optimization. The objective function is constructed for nonlinear optimization of IMU pre-integration constraints, 2D lidar constraints, and visual constraints. To maintain the real-time performance of the calculation, the optimization method of the sliding window is used for real-time pose calculation, and the optimized result is fed back to the initial state estimation.

(d) Closed-loop detection. The closed-loop detection algorithm based on 2D lidar (frame and database subgraph matching) and the vision-based bag of words model (Dbow3) algorithm is used for closed-loop detection. Only when the constraints of these two methods are satisfied at the same time is it considered a closed loop. Closed-loop constraints are then added to the global optimization.

(e) Global optimization. A separate thread is opened for the global optimization of keyframe-based pose graphs.

3.1. Symbol description

$(\cdot)^w$ is the world coordinate frame, and the gravity vector is aligned with the z-axis. $(\cdot)^b$ is the carrier system, which coincides with the IMU system. $(\cdot)^c$ is the camera coordinate system. Using R and the quaternion q to represent the rotation, the quaternion corresponds to the Hamiltonian notation. q_b^w and p_b^w represent the rotation and translation of the body system to the world coordinate system, respectively. b_k represents the body coordinate system when the kth image was taken, and c_k represents the camera coordinate system when the kth image was taken. $\otimes$ represents the multiplication of two quaternions, and $g^w = \begin{bmatrix} 0 & 0 & g \end{bmatrix}^T$ represents the representation of the gravity vector in the world coordinate system. $R \in SO(3)$ represents the rotation matrix, $P \in R^3$ represents the position vector, v represents the velocity vector, and b

represents the IMU bias. The transformation matrix $T \in SE(3)$ is expressed as $T = [R|\mathrm{p}]$.

The variables of the sliding window are expressed as follows. x_k represents the state of the kth frame in the sliding window, including the position $p^w_{b_k}$, velocity $v^w_{b_k}$, attitude $q^w_{b_k}$, acceleration bias b_a, and angular velocity bias b_g in the world coordinate system. $x^b_c = \left[p^b_c, q^b_c\right]$ represents the external parameter from the camera to the IMU, λ represents the inverse depth of the feature point

$$\begin{aligned} \chi &= \left[x_0, x_1, \ldots x_n, x^b_c, \lambda_0, \lambda_1, \ldots, \lambda_m\right] \\ x_k &= \left[p^w_{b_k}, v^w_{b_k}, q^w_{b_k}, b_a, b_g\right], k \in [0, n] \\ x^b_c &= \left[p^b_c, q^b_c\right] \end{aligned} .$$

3.2. IMU pre-integration factor

The angular velocity and acceleration observation model of IMU is defined as follows:

$$\hat{a}_t = a_t + b_{a_t} + R^t_w g^w + n_a \tag{1}$$

$$\hat{a}_t = a_t + b_{a_t} + R^t_w g^w + n_a \tag{2}$$

$\hat{a}_t$ and $\hat{w}_t$ represent the raw measurements of the IMU sensor. The accelerometer noises n_a and n_w are assumed to obey white Gaussian noise, $n_a \sim \eta(0, \sigma^2_\alpha)$, $n_w \sim \eta(0, \sigma^2_w)$. The accelerometer bias and gyroscope bias follow random walks, $n_{b_a} \sim \eta(0, \sigma^2_{b_a})$, and $n_{b_w} \sim \eta(0, \sigma^2_{b_w})$

$$\begin{aligned} \dot{b}_{a_t} &= n_{b_a} \\ \dot{b}_{w_t} &= n_{b_w} \end{aligned} . \tag{3}$$

Between the key frames b_k and b_{k+1} of the two frames of images, in the time range $[t_k, t_{k+1}]$, multiple IMU observation data are present, and the pre-integration formula under continuous time is as follows:

$$\begin{aligned} \alpha^{b_k}_{b_{k+1}} &= \int\int_{t\in[t_k,t_{k+1}]} R^{b_k}_t(\hat{a}_t - b_{a_t})dt^2 \\ \beta^{b_k}_{b_{k+1}} &= \int_{t\in[t_k,t_{k+1}]} R^{b_k}_t(\hat{a}_t - b_{a_t}) \\ \gamma^{b_k}_{b_{k+1}} &= \int_{t\in[t_k,t_{k+1}]} \frac{1}{2}\Omega(\hat{w}_t - b_{w_t})\gamma^{b_k}_t dt \end{aligned} \tag{4}$$

where

$$\Omega(w) = \begin{bmatrix} -\lfloor w \rfloor_\times & w \\ -w^T & 0 \end{bmatrix}, \lfloor w \rfloor_\times = \begin{bmatrix} 0 & -w_z & w_y \\ w_z & 0 & -w_x \\ -w_y & w_x & 0 \end{bmatrix}. \tag{5}$$

The discretized integral form of the above formula is as follows:

$$\begin{aligned} \hat{a}^{b_k}_{i+1} &= \hat{a}^{b_k}_i + \hat{\beta}^{b_k}_i \delta t + \frac{1}{2}R(\hat{\gamma}^{b_k}_i)(\hat{\alpha}_i - b_{a_i})\delta t^2 \\ \hat{\beta}^{b_k}_{i+1} &= \hat{\beta}^{b_k}_i + R(\hat{\gamma}^{b_k}_i)(\hat{\alpha}_i - b_{a_i})\delta t \\ \hat{\gamma}^{b_k}_{i+1} &\otimes \begin{bmatrix} 1 \\ \frac{1}{2}(\hat{w}_i - b_{w_i})\delta t \end{bmatrix} \end{aligned} \tag{6}$$

where i and $i+1$ correspond to two adjacent data observed by the IMU, and δt represents the interval between moments i and $i+1$. The error equation based on the IMU sensor is as follows:

$$\begin{aligned} {}_{\mathrm{B}}\left(\hat{z}^{b_k}_{b_{k+1}}, \chi\right) &= \begin{bmatrix} \delta\alpha^{b_k}_{b_{k+1}} \\ \delta\beta^{b_k}_{b_{k+1}} \\ \delta\theta^{b_k}_{b_{k+1}} \\ \delta b_a \\ \delta b_g \end{bmatrix} \\ &= \begin{bmatrix} R^{b_k}_w\left(p^w_{b_{b+1}} - p^w_{b_k} + \frac{1}{2}g^w\Delta t^2_k - v^w_{b_k}\Delta t_k\right) - \hat{\alpha}^{b_k}_{b_{k+1}} \\ R^{b_k}_w\left(v^w_{b_{k+1}} + g^w\Delta t_k - v^w_{b_k}\right) - \hat{\beta}^{b_k}_{b_{k+1}} \\ 2\left[q^{w^{-1}}_{b_{k+1}} \otimes q^w_{b_{k+1}} \otimes \left(\hat{\gamma}^{b_k}_{b_{k+1}}\right)^{-1}\right]_{xyz} \\ b_{a_{b_{k+1}}} - b_{a_{b_k}} \\ b_{w_{b_{k+1}}} - b_{w_{b_k}} \end{bmatrix} \end{aligned} . \tag{7}$$

According to the dynamic equation of the IMU, the covariance propagation equation in its discrete form can be further deduced, and its form can be found in [29]. The IMU magnetometer bias can be calculated from equation (8), and the initial value of the acceleration bias is set to zero, and it is solved in the back-end optimization process. After the initial value of the gyroscope bias is determined, the integration needs to be re-integrated, and this process is only performed once:

$$\begin{aligned} &\min_{\delta b_w} \sum_{k\in\mathrm{B}} \left\| {q^{c_0}_{b_{k+1}}}^{-1} \otimes q^{c_0}_{b_k} \otimes \gamma^{b_k}_{b_{k+1}} \right\|^2 \\ &\gamma^{b_k}_{b_{k+1}} \approx \hat{\gamma}^{b_k}_{b_{k+1}} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}J^\gamma_{b_w}\delta b_w \end{bmatrix}. \end{aligned} \tag{8}$$

Here, B represents all visual frames within the sliding window. $q^{c_0}_{b_{k+1}}$ and $q^{c_0}_{b_k}$ can be calculated by visual matching. $\gamma^{b_k}_{b_{k+1}}$ represents the IMU pre-integration value between b_k and b_{k+1}, and $J^\gamma_{b_w}$ represents the partial derivative of the pre-integration value with respect to the magnetometer bias

$$\begin{aligned} a^{b_k}_{b_{k+1}} &\approx \hat{a}^{b_k}_{b_{k+1}} + J^a_{b_a}\delta b_{a_k} + J^a_{b_w}\delta b_{w_k} \\ \beta^{b_k}_{b_{k+1}} &\approx \hat{\beta}^{b_k}_{b_{k+1}} + J^\beta_{b_a}\delta b_{a_k} + J^\beta_{b_w}\delta b_{w_k} \\ \gamma^{b_k}_{b_{k+1}} &\approx \hat{\gamma}^{b_k}_{b_{k+1}} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}J^\gamma_{b_w}\delta b_{w_k} \end{bmatrix}. \end{aligned} \tag{9}$$

Subsequent IMU pre-integration values are approximated using equation (9), and repeated pre-integration is not performed. $J^a_{b_a}$ and $J^a_{b_w}$ represent the partial derivatives of $a^{b_k}_{b_{k+1}}$ with respect to the accelerometer bias and the magnetometer bias. $J^\beta_{b_a}$ and $J^\beta_{b_w}$ is the partial derivative of $\beta^{b_k}_{b_{k+1}}$ with respect to the accelerometer bias and the magnetometer bias.

3.3. VIO

This part of VIO is based on sliding window optimization composed of binocular vision and IMU. The positional relationship between the camera and the IMU is shown in figure 3.

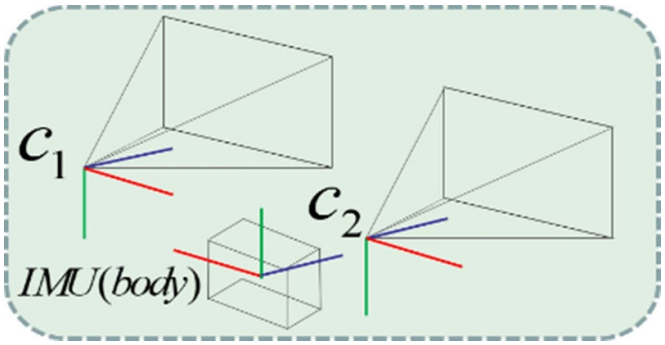


Figure 3. Schematic diagram of vision and IMU coordinate system.

The correlation with the feature points of the right-eye image is achieved by performing optical flow tracking based on the image pyramid on the left-eye image. Then, the construction of the initial map is completed using the triangulation of the binocular baseline length and feature points. The image-matching pose calculation of the 3D–2D (PnP) algorithm is performed according to the successfully initialized map points. With the movement of the carrier, the number of successfully tracked feature points will gradually decrease. When it is less than the threshold, new map points are triangulated according to the feature points associated with the left and right images to ensure that the number of successfully tracked feature points is not less than the threshold, and the random sample consensus algorithm [30] is used to eliminate the tracking error feature points. The extraction of image keyframes is mainly based on the number of successfully tracked feature points.

This paper assumes that the camera obeys the pinhole camera model, that the image observation value of the feature point l in the ith frame is $(\hat{u}_l^{c_i},\hat{v}_l^{c_i})$, and that the point is projected to the jth frame through the result of the front-end visual odometry is expressed as $p_l^{c_j}$ (equation (10)). Through optical flow tracking of the feature point l, the coordinates of this point on j are obtained as $\hat{p}_l^{c_j}$ $(\hat{u}_l^{c_j},\hat{v}_l^{c_j})$, so the reprojection form is constructed as in equation (11). ${K_c}^{-1}$ represents the inverse of the camera internal parameter matrix, projecting the phase plane coordinate system to the camera coordinate system:

$$p_l^{c_j} = R_b^c\left(R_w^{b_j}\left(R_{b_i}^w\left(R_c^b\frac{1}{\lambda_l}{K_c}^{-1}\left(\begin{bmatrix}\hat{u}_l^{c_i}\\ \hat{v}_l^{c_i}\end{bmatrix}\right)+p_c^b\right)+p_{b_i}^w-p_{b_j}^w\right)-p_c^b\right) \tag{10}$$

$$r_c(\hat{z}_l^{c_l},\chi) = \hat{p}_l^{c_j} - \frac{p_l^{c_j}}{\left\|p_l^{c_j}\right\|}. \tag{11}$$

The constraint relationship between the camera and the IMU is shown in figure 4. This constraint relationship is equivalent to building a nonlinear error equation by combining the residual factors of vision and IMU. By minimizing equation (12), the maximum *a posteriori* estimate can be obtained, and the error equation can be optimally solved by the L–M method:

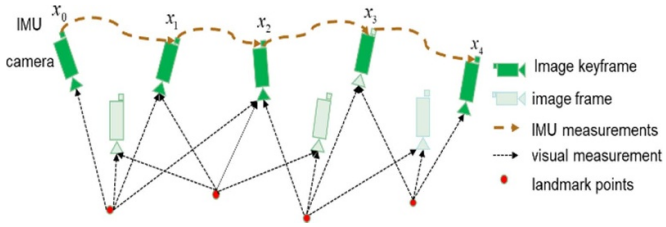


Figure 4. Structure diagram of visual inertial restraint.

$$\min_{\chi}\left\{\|r_p - H_p\chi\|^2 + \sum_{k\in \mathrm{B}}\left\|r_{\mathrm{B}}\left(\hat{z}_{b_{k+1}}^{b_k},\chi\right)\right\|_{p_{b_{k+1}}^{b_k}}^2 + \sum_{(l,i)\in}\rho\left\|r_c\left(\hat{z}_l^{c_j},\chi\right)\right\|_{P_l^{c_j}}^2\right\}. \tag{12}$$

The first term $\{r_p - H_p\chi\}$ in equation (12) represents the marginalized prior information. As the number of keyframes increases, the oldest frame is marginalized by the Schur compensation algorithm, keeping the number of keyframes in the sliding window constant.

3.4. Lidar–inertial odometer

The scanning matching process based on 2D lidar adopts the mainstream CSM method [23]. Before initialization, IMU preintegration provides a more accurate initial pose. After VIO initialization, the pose obtained by VIO provides a more accurate initial pose estimation for laser matching. According to the 6-DOF pose information provided, the lidar point cloud is projected onto the plane. The pose information can also provide more accurate initial values for CSM, narrow the search range, and improve the efficiency of scanning matching. To further obtain a more accurate pose, optimization is adopted; that is, the matching relationship between the current frame laser and the 2D grid map is obtained by minimizing equation (14). Given that it has a relatively accurate initial value, it can effectively prevent local optimization and can converge after a few iterations. To ensure computational efficiency, a submap composed of a certain frame is maintained to match the current frame:

$$S_i(T) = \begin{bmatrix}\cos T_\theta & -\sin T_\theta & T_x\\ \sin T_\theta & \cos T_\theta & T_y\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}p_{ix}\\ p_{iy}\\ 1\end{bmatrix} \tag{13}$$

$$r_L\left(\hat{z}_i^{l_j},\chi\right) = 1 - M(S_i(T)) \tag{14}$$

$$\min_{\chi}\sum_{(i,j)}\left\|r_L(\hat{z}_i^{L_j},\chi\right\|_{P_i^{l_j}}. \tag{15}$$

$T = (T_x, T_y, T_\theta)$ represents the pose and is also a variable that needs to be calculated. $p_i = (p_{ix}, p_{iy})$ represents the coordinates of the 2D laser point collected by the lidar, and $M(S_i(T))$ represents the pixel value of the map grid. An error

exists in the T calculated by the IMU pre-integration or by the VIO, so the laser point cloud and grid map cannot be accurately matched, and equation (14) is not equal to zero. By adjusting the pose T, equation (15) is minimized, the laser point cloud is matched with the grid map as much as possible, and the maximum posterior estimation of the pose is obtained.

3.5. Global optimization

Global optimization refers to combining visual reprojection constraints, IMU constraints, and 2D lidar constraints to construct a large nonlinear error equation. The joint solution process is shown in equation (16). In pursuit of efficient computing effects, 2D lidar matching constraints can be added at a lower frequency:

$$\min_{\chi}\left\{\begin{array}{l}\left\|r_p - H_p\chi\right\|^2 + \sum\limits_{k\in \mathrm{B}}\left\|r_{\mathrm{B}}\left(\hat{z}_{b_{k+1}}^{b_k},\chi\right)\right\|_{p_{b_k}^{b_{k+1}}}^2 + \\ \sum\limits_{(l,i)\in c}\left\|r_c\left(\hat{z}_l^{c_j},\chi\right)\right\|_{P_l^{c_j}}^2 + \sum\limits_{(i,j)}\left\|r_L\left(\hat{z}_i^{L_j},\chi\right)\right\|_{P_i^{l_j}}\end{array}\right\}. \quad (16)$$

This paper is based on the correlation scan matching algorithm of 2D lidar for loop closure detection. When a new scan is obtained, the optimal matching frame is searched in a certain range around it. If the optimal matching frame meets the requirements, it is considered a closed loop. On this basis, the closed-loop test of the visual word bag model is added. When the visual current frame also obtains the correct closed-loop with the historical frame, the closed-loop detection is considered successful, and the closed-loop constraint is added for back-end optimization.

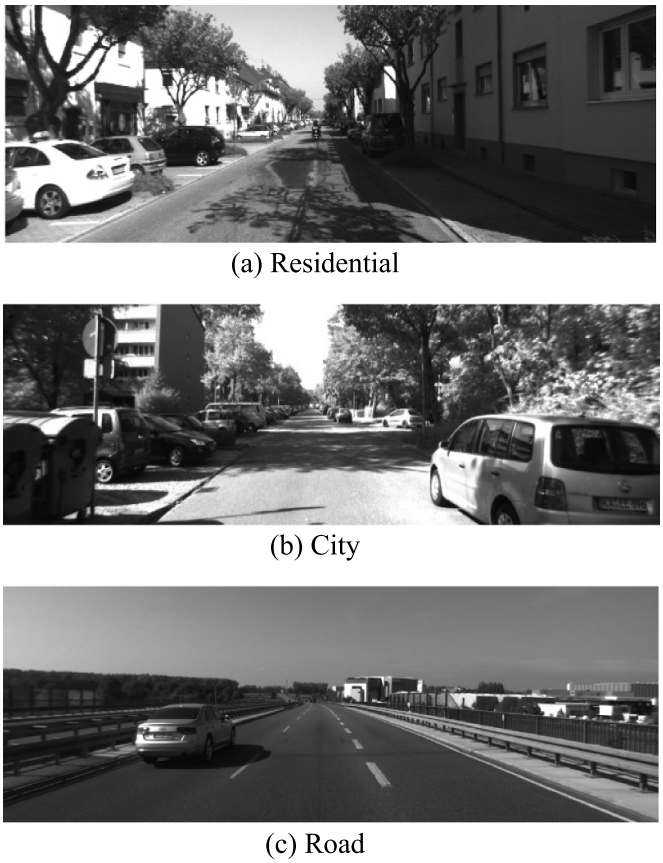
4. Results and analyses

In order to verify the algorithm proposed in this paper, outdoor and indoor experiments were designed. The outdoor scene is tested with the KITTI public dataset [31]. Because the outdoor environment cannot establish an effective 2D map like the indoor environment, and the quality cannot be judged, the outdoor environment based on the KITTI dataset is mainly analyzed for positioning accuracy. Thanks to the real trajectory information of the dataset, this paper uses the EVO tool [32] to complete the accuracy analysis of the proposed algorithm.

In the indoor scene, it is difficult to get the real track information of the mobile carrier because there is no external high-precision equipment. Therefore, in indoor environments, this paper verifies the effectiveness of this algorithm through the mapping quality. The specific experimental process is as follows.

4.1. Experiment on KITTI dataset

To validate our proposed algorithm, this paper conducts outdoor localization tests using the KITTI public dataset. Figure 5 shows some scenarios of the KITTI dataset. The source KITTI raw_data provides a binocular camera, 3D lidar, IMU, and Global Positioning System (GPS) data. For datasets with extract as a suffix, the frequency of the IMU is 100 Hz. For datasets with sync as a suffix, the frequency of IMU and GPS is the same as 10 Hz. In order to make full use of IMU data, the IMU data in the sync dataset need to be replaced with the IMU data in the extract dataset. Using the processed sync datasets as test data, this paper uses only one horizontal scan plane in the 3D lidar data to simulate the 2D lidar data. All experiments are carried out in an Intel i7-107500H CPU test environment with 16 GB RAM.



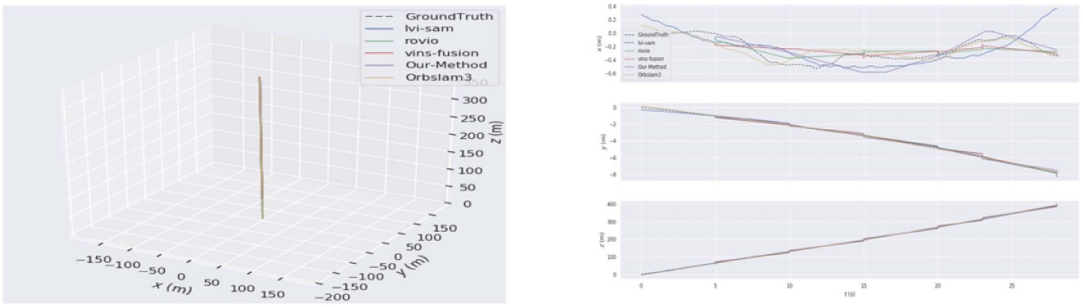
(a) Residential

(b) City

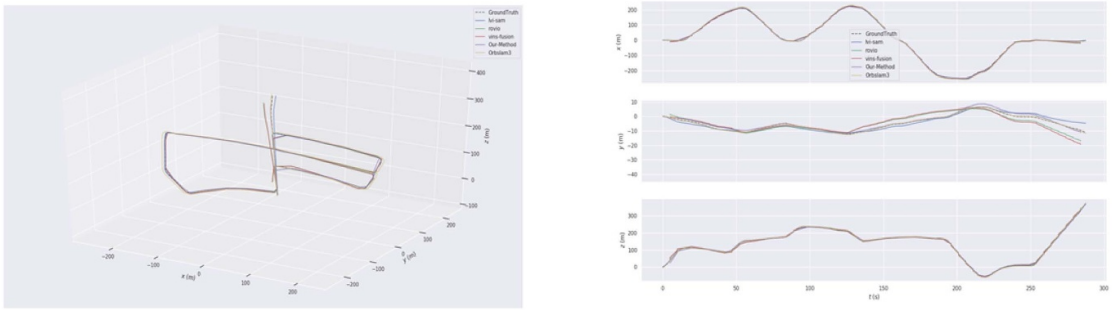
(c) Road

Figure 5. Some scenes of the KITTI dataset.

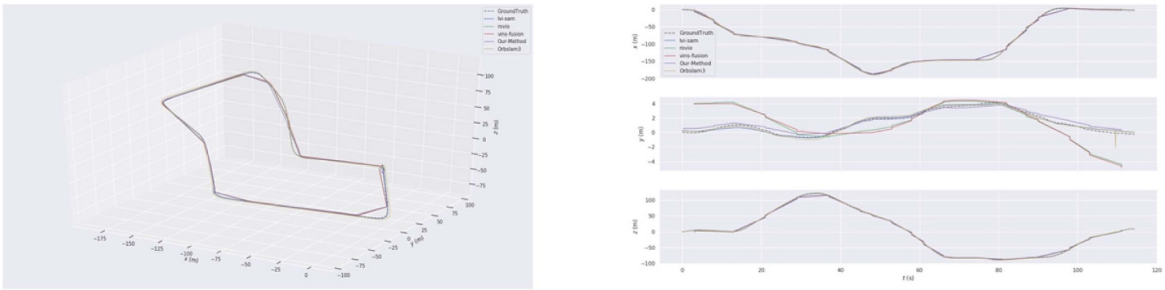
In this experiment, we compared the accuracy of our method with VINS-Fusion, ORB-SLAM3, Rovio, and LVI-SAM, which are all representative SLAM schemes, and verified the performance of our method. For the different sequences of the KITTI public dataset, the positioning results are shown in figure 6. From figure 6 and table 2, we can see that our method is comparable in accuracy to VINS-Fusion in the dataset of 2011_09_30_drive_0016 and 2011_10_03_drive_0042. This is because the above two sequences are sequences in a highway scene. In this case, because the scene is relatively monotonous and the highway lacks geometric changes, the shapes of the obtained local subgraphs are almost the same. Another reason is that the



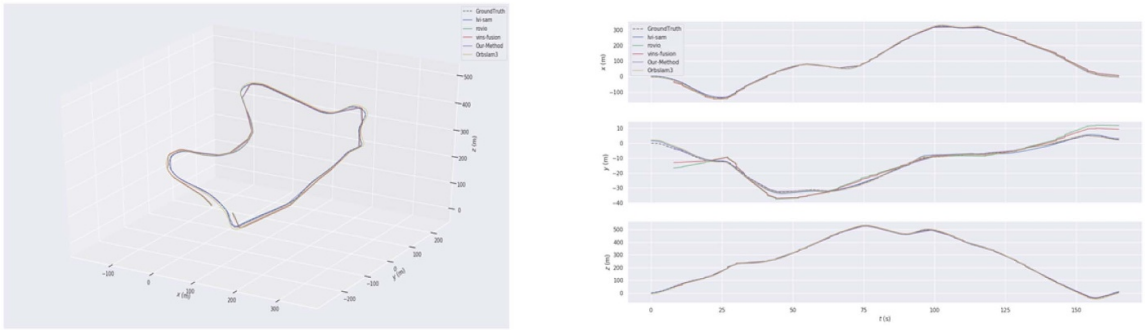
(a) 2011_09_30_drive_0016



(b) 2011_09_30_drive_0018

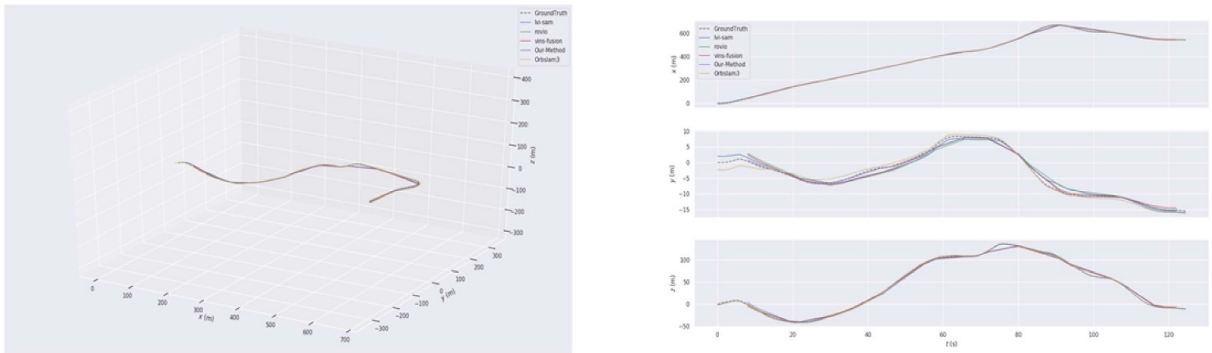


(c) 2011_09_30_drive_0027

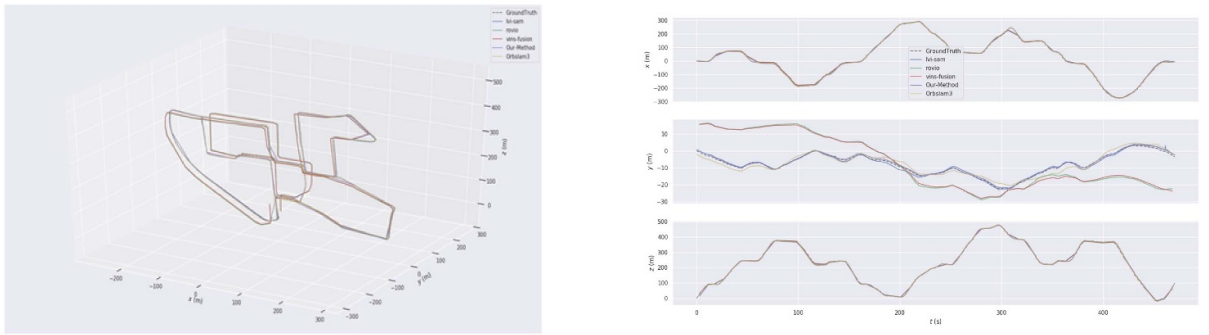


(d) 2011_09_30_drive_0033

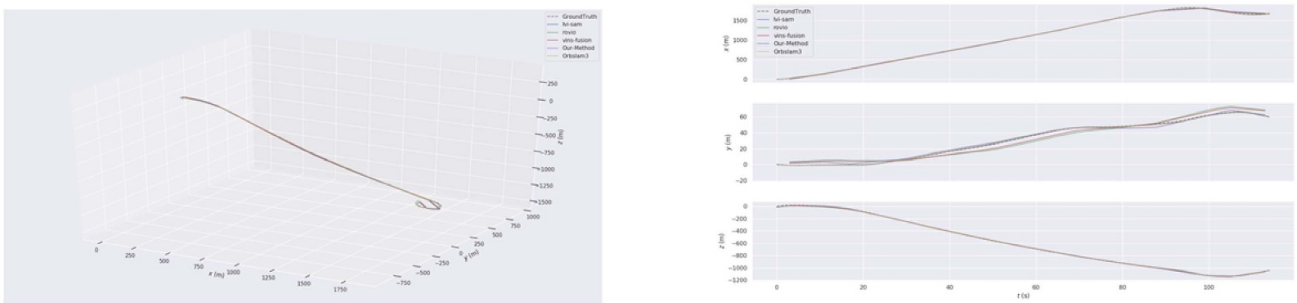
Figure 6. Comparing the positioning results of VINS-Fusion, ORB-SLAM3, Rovio, LVI-SAM, and our method based on KITTI datasets.



(e) 2011_09_30_drive_0034



(f) 2011_10_03_drive_0027



(g) 2011_10_03_drive_0042

Figure 6. (Continued.)

Table 2. Comparing the accuracy results of VINS-Fusion, ORB-SLAM3, Rovio, LVI-SAM, and our method based on KITTI datasets.

Kitti_raw dataset	Our method (rmse) m	LVI-SAM (rmse) m	ORB-SLAM3 (rmse) m	VINS-Fusion (rmse) m	Rovio (rmse) m
2011_09_30_drive_0016	0.81	0.85	0.79	0.80	0.84
2011_09_30_drive_0018	0.67	0.65	0.70	0.79	1.01
2011_09_30_drive_0027	0.65	0.63	0.72	0.83	1.05
2011_09_30_drive_0033	0.89	0.82	0.94	1.19	1.43
2011_09_30_drive_0034	0.67	0.66	0.73	0.84	1.11
2011_10_03_drive_0027	0.82	0.78	0.89	1.06	1.15
2011_10_03_drive_0042	1.31	1.22	1.27	1.29	1.46

(a) Flat map collection trolley

(b) OMD30M-R2000

(c) MYNT EYE S1040-IR-120/Mono

Figure 7. Planar mapping equipment integrating 2D lidar, binocular camera, and IMU.

horizontal scan cannot obtain valid data and cannot form effective plane constraints. In other datasets, the best accuracy of this paper is between LVI-SAM and ORB-SLAM3, which is obviously better than VINS-Fusion and Rovio. This situation shows that in an unstructured environment, 2D lidar can provide good pose constraints and effectively improve the positioning accuracy.

4.2. Indoor environment experiment

To verify the mapping effect of our method, an indoor plane mapping experiment is designed in this paper, and the hardware equipment used is Pepperl r2000 2D lidar and an MYNT camera (binocular + IMU) (figure 7). Cartographer is an excellent representative solution based on 2D lidar plane mapping. Therefore, this experiment uses Cartographer as a reference for comparison. To verify the accuracy of BVLI-SLAM

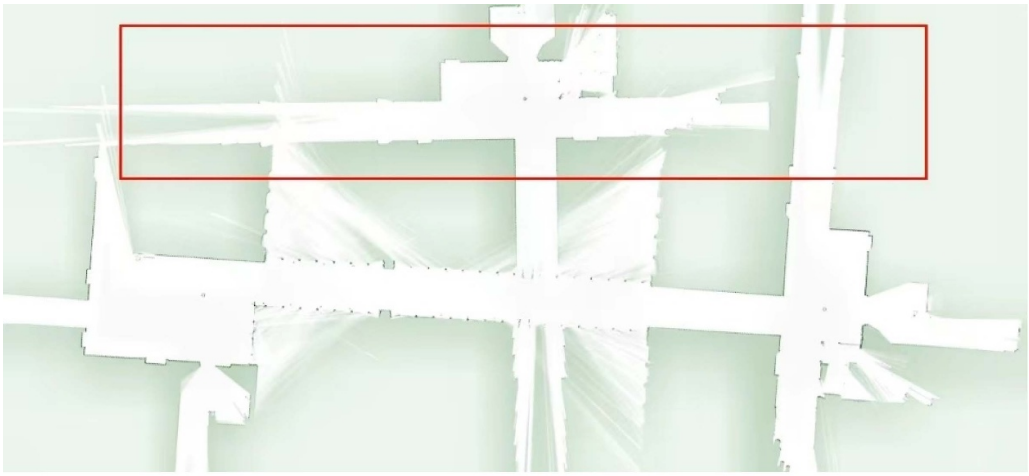
Figure 8. Cylindrical markers.

Figure 9. Image of corridor I.

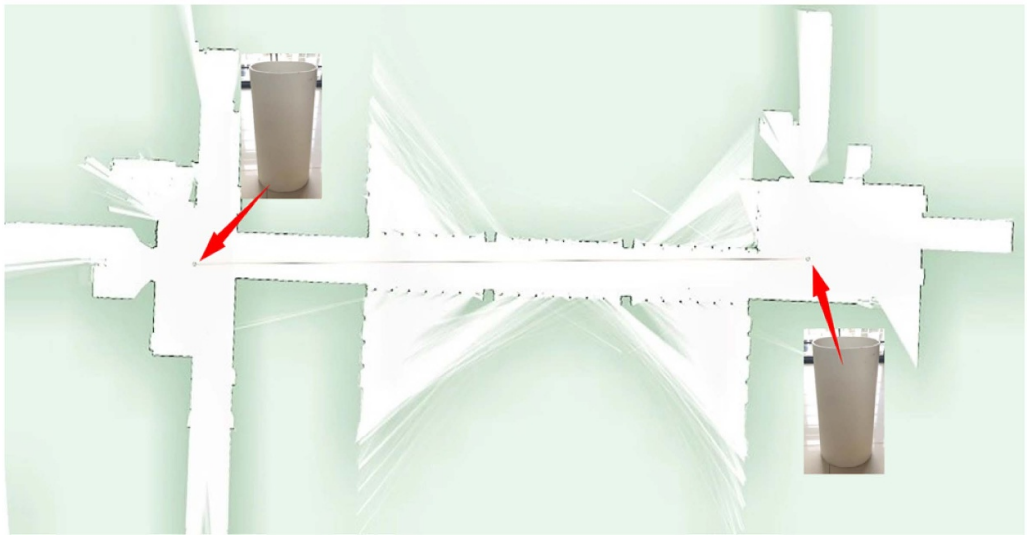
and Cartographer, cylindrical markers are arranged in the experimental scene in this paper (figure 8), and the horizontal distance between the markers is measured by a total station and used as the true distance. To show the robustness of this algorithm in indoor environment mapping, two challenging long corridor fields are selected as experimental scenes (figures 9 and 10). This paper conducts mapping experiments based on Cartographer and our algorithm for these two scenarios. The horizontally measured value distance of the two markers is obtained by the corresponding positions on the grid map (the red lines corresponding to the two points in figures 11 and 12). The measured value is compared with the real value to achieve the purpose of mapping accuracy analysis.

Figure 11 shows that the 2D grid map created by the Cartographer algorithm has more parts than the real scene, such as the part marked by the red frame. This is because the scene has a lot of glass, and 2D lidar cannot obtain enough laser points for matching, causing the algorithm to fail. Thus, the Cartographer-based algorithm cannot solve the mapping problem well in this scenario. With the assistance of vision and IMU, our proposed algorithm can still perform good pose estimation and 2D grid construction even if the number of 2D laser points is missing. The results in figure 12

Figure 10. Image of corridor II.



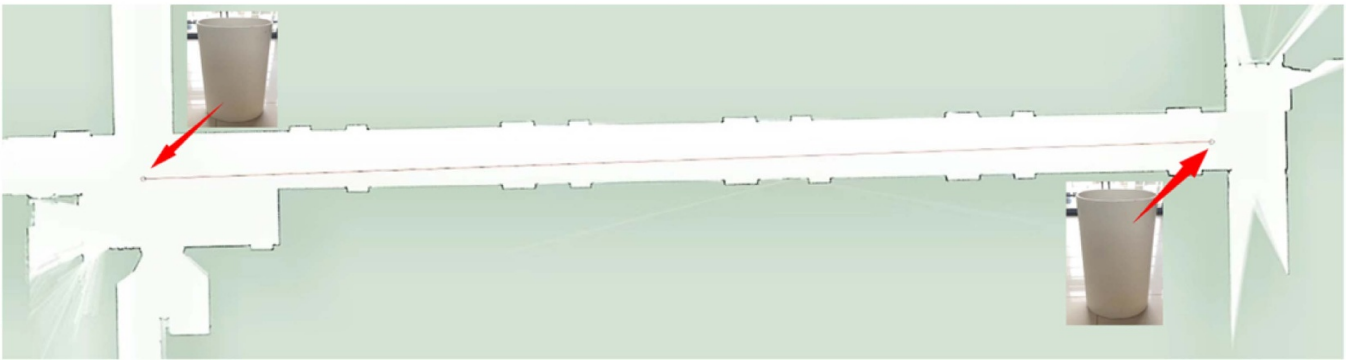
(a) 2D plane mapping result based on cartographer



（b）2D plane mapping result based on BVLI-SLAM

Figure 11. 2D plane mapping result of corridor I.

(a) 2D plane mapping result based on cartographer



(b) 2D plane mapping result based on BVLI-SLAM

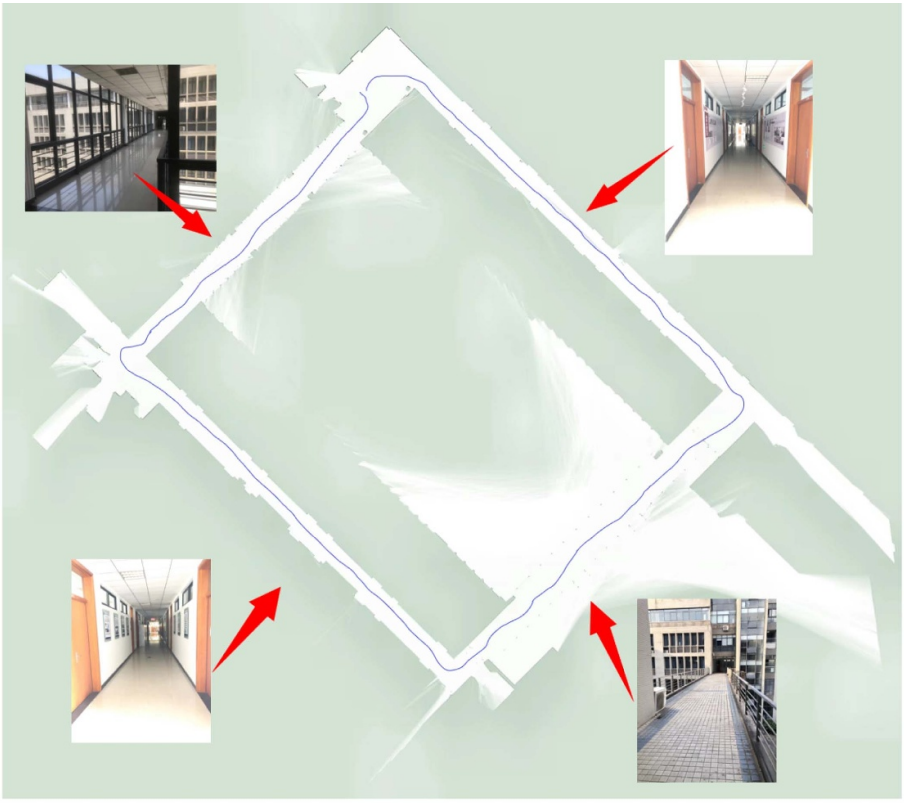
Figure 12. 2D drawing plane mapping result of corridor II.

Table 3. Accuracy comparison.

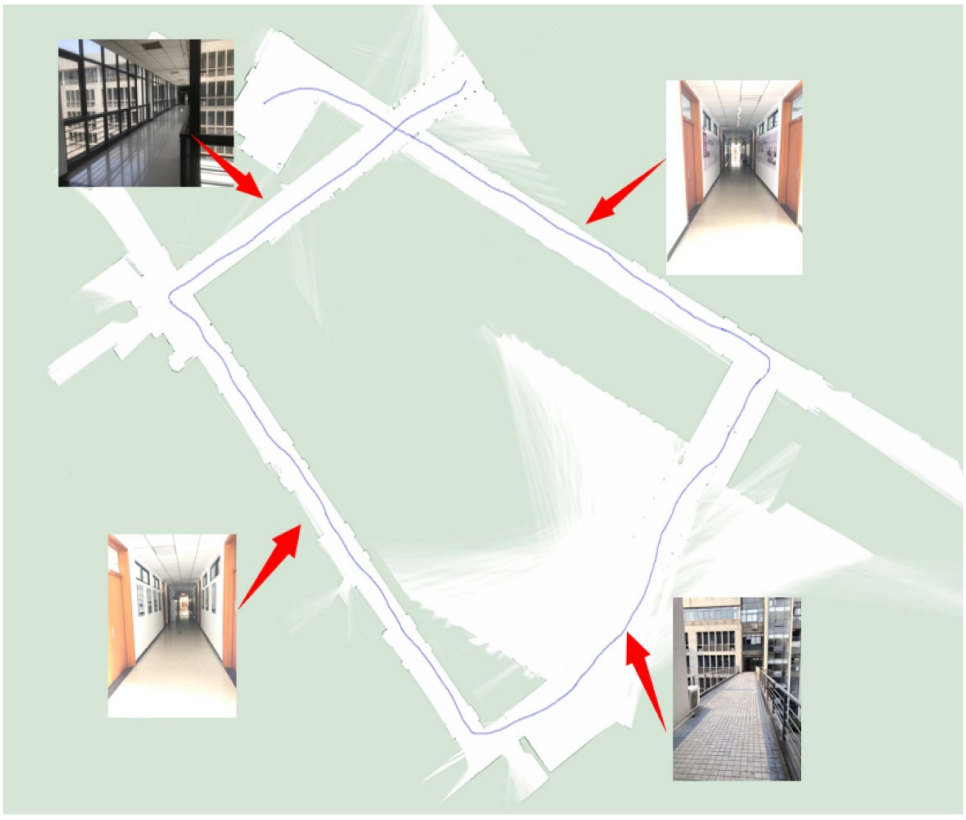
Scenes	Cartographer	BVLI-SLAM (m)	Ground-truth (m)
Corridor I	Fail	35.69	35.673
Corridor II	43.30	43.26	43.243

show that BVLI-SLAM proposed in this paper and Cartographer can build the map successfully, but the outline of the 2D grid map obtained based on the former is clearer. Table 3 further indicates that the algorithm proposed in this paper has advantages in accuracy and robustness compared with Cartographer.

To further test the robustness of the algorithm's mapping, this paper selects a complex scene with a glass corridor and an outdoor corridor for mapping testing. The length and width of the scene are about 40 m $\times$ 50 m. The experimental results show that the algorithm proposed in this paper can still run robustly and obtain a 2D grid map with clear outlines (as shown in figure 13(a), and it can be seen from figure 13(b) that Cartographer has obvious angle deviation in the outdoor corridor because it cannot scan enough effective point clouds.



(a) 2D plane mapping result based on cartographer



(b) 2D plane mapping result based on cartographer

Figure 13. 2D plane mapping result based on BVLI-SLAM and Cartographer in complex scenes. (The blue curve represents the motion track of the trolley in the x and y directions).

5. Conclusions

In this paper, we designed a real-time pose estimation and 2D mapping scheme based on the fusion of 2D lidar, a binocular camera, and IMU based on graph optimization, making full use of the performance of different sensors. An experiment on the KITTI dataset shows that the positioning accuracy of the BVLI-SLAM scheme designed in this paper is between that of LVI-SAM and ORB-SLAM3, and BVLI-SLAM can improve the positioning accuracy by more than 15% compared with VINS-Fusion in the outdoor unstructured environment. In outdoor structured environments, localization results with an accuracy comparable to that of VINS-Fusion can be achieved. In addition, this paper also designs the mapping experiment in an indoor environment based on BVLI-SLAM and Cartographer schemes, and the results show that the proposed BVLI-SLAM can obtain more robust and accurate 2D raster maps than Cartographer.

Through the above analysis, the algorithm proposed in this paper can provide high-precision real-time positioning and mapping, but it should also be noted that due to the limitation of 2D lidar, beautiful and useful 2D grid maps can only be created in an indoor environment.

Data availability statement

The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

The authors would like to thank Siqi Liu for her academic support enabling successful completion of this work.

Funding statement

The study was partially sponsored by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX22_2592) and sponsored by the Graduate Innovation Program of China University of Mining and Technology (Grant No. 2022WLKXJ109).

Conflict of interest

The authors declare that they have no conflicts of interest.

ORCID iD

Zhenbin Liu https://orcid.org/0000-0003-0237-3663

References

[1] Smith R C and Cheeseman P 1986 On the representation and estimation of spatial uncertainty *Int. J. Robot. Res.* **5** 56–68
[2] Cadena C, Carlone L, Carrillo H, Latif Y and Scaramuzza D N J 2016 Past, present, and future of simultaneous localization and mapping: toward the robust-perception age *IEEE Trans. Robot.* **32** 1309–32
[3] Koller D and Friedman N 2009 *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA: MIT Press)
[4] Xu X, Zhang L, Yang J, Cao C, Wang W, Ran Y, Tan Z and Luo M 2022 A review of multi-sensor fusion slam systems based on 3D LIDAR *Remote Sens.* **14** 2835
[5] Tee Y K and Han Y C 2021 Lidar-based 2D SLAM for mobile robot in an indoor environment: a review *Int. Conf. on Green Energy, Computing and SustainableTechnology (GECOST)* pp 1–7
[6] Debeunne C and Vivet D 2020 A review of visual-lidar fusion based simultaneous localization and mapping *Sensors* **20** 2068
[7] Davison A J, Reid I D, Molton N D and Stasse O 2007 MonoSLAM: real-time single camera SLAM *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 1052–67
[8] Klein G and Murray D 2008 Parallel tracking and mapping for small AR workspaces *Int. Symp. on Mixed and Augmented Reality* pp 225–34
[9] Forster C, Pizzoli M and Scaramuzza D 2014 SVO: fast semi-direct monocular visual odometry *IEEE Int. Conf. on Robotics and Automation* pp 15–22
[10] Engel J, Schps T and Cremers D 2014 LSD-SLAM: large-scale direct monocular SLAM *European Conf. on Computer Vision* pp 834–49
[11] Mur-Artal R and Tardós J D 2017 Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras *IEEE Trans. Robot.* **33** 1255–62
[12] Mourikis A I and Roumeliotis S I 2007 A multi-state constraint kalman filter for vision-aided inertial Navigation *IEEE Int. Conf. on Robotics and Automation* p 6
[13] Strasdat H, Montiel J and Davison A J 2012 visual slam: why filter? *Image Vis. Comput.* **30** 65–77
[14] Leutenegger S, Lynen S, Bosse M, Siegwart R and Furgale P 2015 Keyframe-based visual-inertial odometry using nonlinear optimization *Int. J. Robot. Res.* **34** 314–34
[15] Qin T, Cao S, Pan J and Shen S 2019 A general optimization-based framework for global pose estimation with multiple sensors (arXiv:1901.03642)
[16] Campos C, Elvira R, Rodríguez J J G, Montiel J and Tardós J D 2020 Orb-slam3: an accurate open-source library for visual, visual-inertial and multi-map slam *IEEE Trans. Robot.* **37** 1874–90
[17] Thrun S 2002 Probabilistic robotics *Commun. ACM* **45** 52–57
[18] Montemerlo M and Thrun S 2003 Simultaneous localization and mapping with unknown data association using fast SLAM *IEEE Int. Conf. on Robotics and Automation* pp 1985–91
[19] Grisetti G, Stachniss C and Burgard W 2007 Improved techniques for grid mapping with rao-blackwellized particle filters *IEEE Trans. Robot.* **23** 34–46
[20] Blanco J L, Gonzalez J and Fernandez-Madrigal J A 2010 Optimal filtering for non-parametric observation models: applications to localization and slam *Int. J. Robot. Res.* **29** 1726–42
[21] Konolige K, Grisetti G, Kümmerle R, Burgard W and Vincent R 2010 Efficient sparse pose adjustment for 2D mapping *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems* pp 18–22

[22] Kohlbrecher S, Stryk O V, Meyer J and Klingauf U 2011 A flexible and scalable SLAM system with full 3D motion estimation *IEEE Int. Symp. on Safety* pp 155–60
[23] Hess W, Kohler D, Rapp H and Andor D 2016 Real-time loop closure in 2D LIDAR SLAM *2016 IEEE Int. Conf. on Robotics and Automation (ICRA)* pp 1271–8
[24] Ji Z and Singh S 2015 Visual-lidar odometry and mapping: low-drift, robust, and fast *IEEE Int. Conf. on Robotics & Automation (ICRA)* pp 2174–81
[25] Yupeng J 2021 Lvio-fusion: a self-adaptive multi-sensor fusion SLAM framework using actor-critic method *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)* pp 286–93
[26] Shan T, Englot B, Ratti C and Rus D 2021 LVI-SAM: tightly-coupled lidar-visual-inertial odometry via smoothing and mapping *IEEE Int. Conf. on Robotics and Automation (ICRA)* pp 5692–8
[27] Lin J and Zhang F 2021 R^3LIVE: a robust, real-time, RGB-colored, lidar-inertial-visual tightly-coupled state Estimation and mapping package (arXiv:2109.07982)
[28] Lin J, Zheng C, Xu W and Zhang F 2021 R^2LIVE: a robust, real-time, LiDAR-inertial-visual tightly-coupled state estimator and mapping *IEEE Robot. Autom. Lett.* **6** 7469–76
[29] Solà J 2017 Quaternion kinematics for the error-state Kalman filter (arXiv:1711.02508)
[30] Fischler M A and Bolles R C 1981 Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography *Commun. ACM* **24** 381–95
[31] Andreas G 2012 Are we ready for autonomous driving? The KITTI vision benchmark suite *Proc. 2012 IEEE Conf. on Computer Vision and Pattern Recognition* pp 3354–61
[32] Grupp M 2017 EVO: python package for the evaluation of odometry and SLAM (available at: https://github.com/MichaelGrupp/evo)