Graph-Based Indoor 3D Pedestrian Location Tracking With Inertial-Only Perception

Shiyu Bai, *Member, IEEE*, Weisong Wen, *Member, IEEE*, Dongzhe Su, and Li-Ta Hsu, *Senior Member, IEEE*

***Abstract*—Pedestrian location tracking in emergency responses and environmental surveys of indoor scenarios tend to rely only on their own mobile devices, reducing the usage of external services. Low-cost and small-sized inertial measurement units (IMU) have been widely distributed in mobile devices. However, they suffer from high-level noises, leading to drift in position estimation over time. In this work, we present a graph-based indoor 3D pedestrian location tracking with inertial-only perception. The proposed method uses onboard inertial sensors in mobile devices alone for pedestrian state estimation in a simultaneous localization and mapping (SLAM) mode. It starts with a deep vertical odometry-aided 3D pedestrian dead reckoning (PDR) to predict the position in 3D space. Environment-induced behaviors, such as corner-turning and stair-taking, are regarded as landmarks. Multi-hypothesis loop closures are formed using statistical methods to handle ambiguous data association. A factor graph optimization fuses 3D PDR and behavior loop closures for state estimation. Experiments in different scenarios are performed using a smartphone to evaluate the performance of the proposed method, which can achieve better location tracking than current learning-based and filtering-based methods. Moreover, the proposed method is also discussed in different aspects, including the accuracy of offline optimization and proposed height regression, and the reliability of the multi-hypothesis behavior loop closures. The video (YouTube) or (BiliBili) is also shared to display our research.**

***Index Terms*—Indoor localization, pedestrians, inertial perception, SLAM, factor graph optimization.**

I. Introduction

ACCESSIBLE indoor pedestrian location tracking [1], [2] is vital to a wide range of applications in unknown scenarios, such as emergency response and environmental survey [3], [4]. Location tracking approaches that utilize camera or light detection and ranging (LiDAR) have become popular in recent years as they provide accurate positioning solutions in unknown environments. However, it causes high power consumption and hardware requirements, which is unsuitable for portable mobile devices. The inertial measurement unit (IMU) based on micro-electromechanical systems (MEMS) processing technology is a low-cost and widely deployed sensor in current mobile devices like smartphones. More importantly, it is self-contained and can provide a continuous location in any environment [5], attracting broad attention for indoor pedestrian location tracking.

Location Tracking using Inertial Measurements: Although MEMS-IMU presents a clear advantage for indoor localization, strap-down inertial navigation using MEMS-IMU suffers from severe drift due to the high noise level. Pedestrian dead reckoning (PDR) is an effective alternative for improving positioning accuracy with MEMS-IMU. Pedestrian step can be detected by smartphone built-in accelerometers, then the step length can be obtained to calculate the location based on dead reckoning [6]. However, the drift issue remains in PDR, leading to unusable positioning solutions in long-distance applications. Like simultaneous localization and mapping (SLAM) based on visual or LiDAR information, IMU data can also be fully used in a SLAM mode to suppress error accumulation. For example, human behaviors, such as sitting or standing still, detected by smartphone IMU can be used as landmarks to correct the PDR error [7]. Behavior SLAM using smartphone IMU has proven to be an effective manner to track pedestrian locations without external aiding sources. In addition, the generated behavior map can also be employed for the following navigation application, avoiding the need for the floorplan.

Challenges of Behavior SLAM: However, there are still some issues limiting the performance of behavior SLAM. The first one is the limited availability in 3D scenarios. In the behavior SLAM, odometry is achieved by PDR. However, most PDR methods based on smartphones are mainly applied to 2D scenes. Although the vertical position can be derived by recognizing the behavior of taking stairs, it is hard to achieve accurate detection with only a smartphone IMU. In addition, the barometer is usually required to provide the height. However, the barometer is easily affected by environments, and only some smartphones are equipped with a barometer. Another issue is that the existing behavior SLAM is mainly based on the particle filter. However, filtering-based methods marginalize past states and only estimate current states, which decreases the estimation accuracy as past and current states are connected in SLAM.

Our Contribution: To address the above-mentioned problems, this paper proposes a novel behavior SLAM for 3D pedestrian location tracking. The main contributions of this paper are as follows:

Received 4 January 2024; revised 10 August 2024; accepted 2 January 2025. Date of publication 6 January 2025; date of current version 4 April 2025. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110771 and in part by the University Grants Committee of Hong Kong under the scheme Research Impact Fund under Grant R5009-21. This research was also supported by the Faculty of Engineering, The Hong Kong Polytechnic University under the project "Perception-based GNSS PPP-RTK/LVINS integrated navigation system for unmanned autonomous systems operating in urban canyons". Recommended for acceptance by R. Zhang. *(Corresponding author: Li-Ta Hsu.)*

Shiyu Bai, Weisong Wen, and Li-Ta Hsu are with the Hong Kong Polytechnic University, Hong Kong (e-mail: shiyu.bai@polyu.edu.hk; welson.wen@polyu.edu.hk; lt.hsu@polyu.edu.hk).

Dongzhe Su is with the Hong Kong Applied Science and Technology Research Institute (ASTRI), Hong Kong (e-mail: dzsu@astri.org).

Digital Object Identifier 10.1109/TMC.2025.3526196

1) *Inertial perception-only indoor 3D pedestrian SLAM via smartphones:* A deep vertical odometry-aided 3D PDR system is developed, and environment-induced behaviors are utilized to enhance tracking accuracy using only IMU data.
2) *Graph optimization with multi-hypothesis behavior loop closures:* This paper employs optimization to improve SLAM accuracy. Multi-hypothesis behavior loop closures are formulated to deal with ambiguous data associations, leading to improved reliability.
3) *Comprehensive experimental validation:* Experiments in different scenarios are conducted for the comparison. Moreover, the proposed method is evaluated from multiple perspectives.

The rest of the paper is structured as follows: In Section II, the relevant work is discussed. In Section III, the methodology overview of the proposed method is introduced. The proposed method is described in detail in Sections IV and V. In Section VI, experiments and discussions are given. Finally, Section VII concludes the paper.

II. Related Work

In indoor location-based applications, wireless signal, such as Wi-Fi [8] and Bluetooth Low Energy (BLE) [9] are still mainstream as it can provide absolute positioning solution. However, this approach requires deployment and calibration of infrastructure. It cannot be used for special applications like responses and surveys because the availability of wireless signals cannot be ensured.

A. Pedestrian Dead Reckoning

Nowadays, IMU has been widely integrated into smartphones to achieve screen rotation, step count, and virtual/augmented reality (VR/AR) experiences [10], [11], and it can also be utilized for indoor pedestrian tracking. The pedestrian's position can be updated by accumulating displacements derived from the step, a process known as PDR. PDR comprises the step detection, step length, and heading estimation [12]. In most cases, PDR is utilized for the horizontal position estimation, and a barometer is required for the vertical estimation [13]. To reduce the dependence on a barometer, Itzik proposed a vertical positioning based on smartphone IMU [14]. The specific force magnitude detects the stair period to estimate vertical step length. Boim presented two methods for estimating the vertical position using smartphone accelerometers [15]. The integration-based and peak-based height difference via vertical acceleration can be obtained to form a 3D PDR. However, these methods depend on determining the stair period, which easily fails due to high noise and unobvious data representation with only a smartphone.

B. Learning-Based Dead Reckoning

Besides traditional model-based PDR, machine learning (ML) has recently been exploited to achieve dead reckoning based on inertial measurements. Chen developed IONet [16], in which a bidirectional LSTM is formulated to regress the polar vector using a sequence of IMU raw data. The dataset for training is collected by an IMU on a smartphone, and the motion capture system (MCS) is used to provide the pose ground truth. Yan proposed RIDI [17], in which the support vector machine (SVM) is utilized to classify the phone placement. Corrections in linear accelerations can be estimated using regressed velocity vectors. Then, the corrected linear acceleration is employed to obtain the pedestrian position. Herath introduced RoNIN for estimating horizontal position change using a sequence of IMU data [18]. IMU data is first transformed to the reference frame, and the displacements in the reference frame can be regressed. Liu proposed TLIO [19], where position changes and corresponding uncertainties in the reference frame are regressed using ResNet. Then, a stochastic cloning extended Kalman filter (EKF) is applied to solve the pose and sensor bias. Among learning-based methods, the availability of datasets for training is a general problem. MCS and visual-inertial odometry (VIO) are typically used to provide the ground truth. The former can output accurate pose, but it can only be used in a confined space and cannot cover the multilayers. The latter typically uses a Google Tango phone [18], [20] to run VIO. This manner can be utilized in multiple scenarios with stairs. However, it brings forward high requests for training and testing data. On the one hand, it is cumbersome to collect ideal training data since the pose accuracy for the reference cannot be fully guaranteed with a VIO, and the misalignment between different devices needs to be compensated, which is not easy to estimate in most cases. On the other hand, the accuracy of heading estimation in the testing data can be worse; it can cause error-prone transformation, leading to seriously compromised horizontal position regressions.

C. Behavior-Based Map Matching

More significantly, drift is always an essential issue in dead reckoning. When there is no external data, the environment-induced behavior is fully exploited as an observation. Some structured features exist in indoor scenes, such as corners and stairs. When a user turns at a corner or takes the stairs, these behaviors occur at the corner or stairs. The locations of these features are fixed, which can be associated with corresponding behaviors to suppress the error accumulation [21]. A sequence of behaviors is used to achieve the map matching [22]. Hidden Markov model (HMM) [23] is employed to match the behavior sequence to nodes in an indoor map network. Gu proposed the integration of PDR and context-based behaviors based on smartphones for urban city areas [24], in which stores, escalators, elevators, corners, and corridors are considered for map matching. Although these methods reduce the dependence on external information, a prior floorplan is needed, leading to limited application if the indoor map is unavailable or scenes are changed.

D. Behavior-Based SLAM

SLAM can alleviate the positioning issue in an environment without maps [25]. As for indoor pedestrian tracking, there have been corresponding SLAM algorithms using Wi-Fi, BLE, and magnetic signals [26], [27], [28]. Similarly, human behaviors can be regarded as landmarks to achieve the positioning

when the prior map is unavailable. Hardegger proposed ActionSLAM [29], in which foot-mounted and wrist-mounted IMUs are used to recognize pedestrian behaviors as landmarks. A particle filter is then employed for state estimation. ActionSLAM on a mobile phone [7] was proposed to use pedestrian behaviors recognized by a smartphone, such as sitting and standing, as observations, which can be utilized for tracking in home and office scenarios. Abdelnasser presented SemanticSLAM [30], an indoor tracking involving seed and organic landmarks. Seed landmarks refer to specific objects like elevators and stairs. Shokry introduced DynamicSLAM [31]. Besides frequently used landmarks, some human anchors based on encounters are introduced to improve the density of landmarks, boosting the accuracy. Existing SLAMs based on human behaviors are mainly extended from FastSLAM [32], a particle filter-based state estimator. Compared with visual or LiDAR-based SLAM, achieving reliable data association with behavior landmarks is more challenging due to insufficient features. A particle filter maintains multiple particles in which different hypotheses with different data associations are included, which can improve the reliability of data associations. However, like other filtering methods, the particle filter also marginalizes past states, reducing estimation accuracy.

E. Graph Optimization

Recently, graph optimization has been a popular method for state estimation in different communities [33], [34]. Historical states are reserved in the graph optimization. They can be iteratively optimized with the following information, which is more suitable for SLAM as the loop closure can connect past and current states. This strong constraint can fully use all the information to improve the estimation accuracy. As for behavior-based SLAM, Liu proposed a collaborative SLAM based on Wi-Fi fingerprint similarity and turning motion [35]. Iterative closest point (ICP) [36] is employed to match turning behaviors. However, this manner easily causes the wrong match as only one hypothesis exists. If one wrong loop closure is built, all the following position estimations are influenced, and the proper trajectory cannot be recovered. Graph optimization with multiple hypothesis tracking can be used to achieve robust data association. Hsiao proposed a multi-hypothesis iSAM to handle the ambiguity in SLAM [37], which can avoid the wrong estimates in existing approaches based on a single hypothesis. Bernreiter presented a semantic SLAM system based on factor graphs and a multiple hypothesis tracking mapping for dealing with ambiguous data associations [38]. Moreover, the proposed resampling method enables a more robust solution and requires fewer hypotheses. Although some graph-based SLAM methods use multi-hypothesis tracking, they aim for SLAM using vision or LiDAR. There is still no corresponding algorithm for inertial perception-only SLAM, whose models and data associations differ from current ones.

F. Our Difference

To our knowledge, this is the first paper to investigate graph optimization-based indoor 3D SLAM using behavior landmarks with only a smartphone IMU. This work is different from our previous work on 2D scenarios [39]. In this research, we develop a deep vertical odometry-aided 3D PDR. We consider stair-taking behaviors and formulate multi-hypothesis behavior loop closures in 3D environments to suppress drift. Real-world tests in different scenarios are performed to evaluate the performance of the proposed method. Furthermore, this paper discusses the proposed method from different perspectives, including the batch accuracy, regression accuracy, and estimation reliability, which cannot be found in existing literature.

III. Overview of the Proposed Method

The problem description, overall structure, and contribution are indicated in Fig. 1. Indoor pedestrian location tracking with mobile devices and low power consumption is always in demand, and it also tries to reduce the dependence on external infrastructure by utilizing internal sensors. IMU can be utilized in any scene, but it is typically integrated with other sensors to achieve indoor localization. Although many researchers have exploited the potential of inertial-only indoor localization, the challenges of limited availability in 3D scenarios and reduced accuracy still exist.

The proposed method comprises 3D PDR, multi-hypothesis behavior loop closures, and factor graph optimization, which are marked with red dotted lines to highlight the differences. 3D PDR is like an odometry to predict the pedestrian's location. Model-based PDR on a flat surface is achieved using gyroscope and accelerometer data. A neural network is leveraged to regress the vertical displacements with acceleration. Then, an inertial-only 3D PDR can be achieved to provide the location prediction and relative constraint. As for behavior loop closures, stair-takings and corner-turnings are built as landmarks. Multiple hypotheses with different probabilities are maintained, forming different data associations to ensure reliability. In addition, N-scan back pruning [40] is employed to avoid the unlimited growth of the number of hypotheses. Finally, multiple factor graphs are built to achieve the state estimation. The optimization is based on the Georgia Tech Smoothing and Mapping library (GTSAM) [41], a library of C++ classes that implements sensor fusion based on factor graphs. Then, the hypothesis with the highest probability is the output.

The typical symbols utilized throughout the paper are defined in Table I. The coordinates in this paper include the body and world frame. The body frame, denoted as b, is aligned with three axes of IMU. The world frame, denoted as w, originates in the body frame at the initial instant. Its three axes point to east, north, and up, respectively.

IV. Inertial Perception-Only 3D Pedestrian SLAM

This part introduces the proposed inertial perception-only 3D SLAM from two perspectives: a deep vertical odometry-aided 3D PDR and multi-hypothesis behavior loop closures. The former is used to predict the pedestrian's location, while the latter is built and adopted as an observation to suppress the error accumulation in PDR.

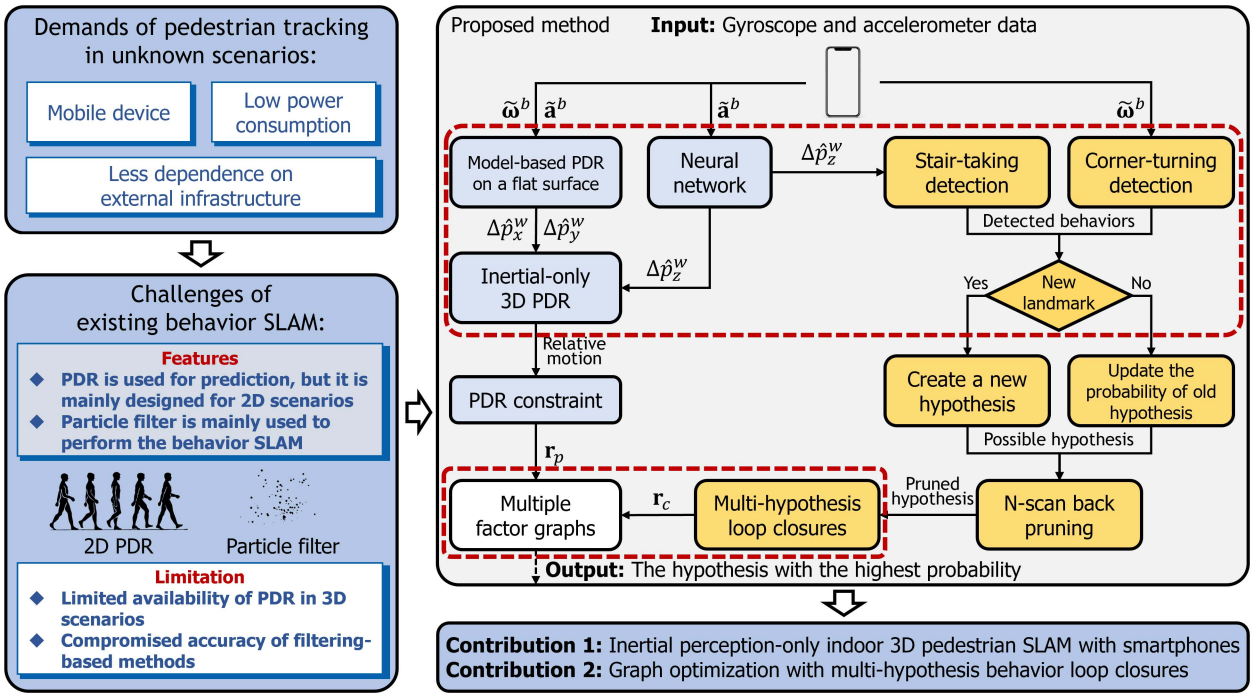


Fig. 1. Problem description, overall structure, and contribution. In the proposed method, the input is the gyroscope and accelerometer data from a smartphone built-in IMU, and the output is the hypothesis with the highest probability.

TABLE I
THE DEFINITION OF TYPICAL SYMBOLS

Symbol	Definition
$\boldsymbol{\omega}_k^b$	Angular velocity at the instant t_k in the body frame
$\mathbf{a}_k^b$	Specific force at the instant t_k in the body frame
$\mathbf{l}_k^w$	Motion acceleration at the instant t_k in the world frame
$\Delta p_{d,k}^w$	Position change at the instant t_k along the d-axis in the world frame
$\mathbf{v}_k^w$	Velocity at the instant t_k in the world frame
$\mathbf{R}_k$	Rotation matrix from the body frame to world frame at the instant t_k
g^w	Gravity value in the world frame
$\mathbf{r}_p$	3D PDR residual
$\mathbf{r}_c$	Behavior loop closure residual
$\tilde{d}_k$	Measured value at the instant t_k
$\hat{d}_k$	Estimated value at the instant t_k

A. Deep Vertical Odometry-Aided 3D PDR

Traditional 3D PDR using smartphones can be mainly divided into two types. In the first one, the horizontal position is derived using a model-based PDR algorithm, and a barometer provides the height. This way requires an additional barometric pressure sensor and suffers from environmental factors as the pressure is easily affected. In the second one, the position on a flat surface is still derived by a model-based PDR method, while the height is calculated by recognizing the motion pattern. However, the differentiation of smartphone IMU data is not as distinct when comparing walking on a flat surface to going up stairs. Consequently, it becomes challenging to accurately detect the motion and predict the location.

The potential of machine learning on smartphone-based PDR has been fully exploited. However, they are mainly designed for 2D scenarios. Moreover, existing end-to-end PDR methods put high demands on the datasets, and a rapid drift in the horizontal position often occurs due to uncontrollable data. Therefore, this paper proposes a deep vertical odometry-aided 3D PDR to solve the above problems, which can fully make the most of the model and data to perform self-contained position prediction in indoor 3D environments. Meanwhile, the proposal can achieve a light data collection. A specific analysis process of the advantage in the deep vertical odometry and proposed 3D PDR model are given next.

Let the velocity in the world frame at the instant t_0 is $\mathbf{v}_0^w$, the velocity at the instant t_1 in the world frame can be expressed as follows:

$$\mathbf{v}_1^w = \mathbf{v}_0^w + \mathbf{l}_1^w \Delta t, \tag{1}$$

where $\mathbf{l}_1^w$ is the motion acceleration at the instant t_1 in the world frame, and Δt is the sampling interval between the instant t_0 and t_1. Let intervals between all successive instants be Δt, then the velocity at the instant t_k in the world frame can be given as

$$\mathbf{v}_k^w = \mathbf{v}_0^w + \mathbf{l}_1^w \Delta t + \mathbf{l}_2^w \Delta t + \cdots + \mathbf{l}_k^w \Delta t. \tag{2}$$

Let us suppose a period T consists of n instants; then its displacement in the world frame can be expressed as

$$\begin{aligned} \Delta \mathbf{L} = n\mathbf{v}_0^w \Delta t + (n-1)\,\mathbf{l}_1^w \Delta t^2 \\ + (n-2)\,\mathbf{l}_2^w \Delta t^2 + \cdots + \mathbf{l}_{n-1}^w \Delta t^2, \end{aligned} \tag{3}$$

where $\Delta \mathbf{L}$ is the displacement over a period T that consists of n instants. Considering that $\mathbf{a}_k^b$ is defined as the non-gravitational

force sensed with respect to an inertial frame, it includes both the motion acceleration and an opposing acceleration relative to gravitational acceleration, known as the restoring acceleration on land. Therefore, the restoring acceleration should be removed from the accelerometer data to obtain the motion acceleration $\mathbf{l}_k^w$

$$\mathbf{l}_k^w = \mathbf{R}_k \mathbf{a}_k^b - \mathbf{G}^w, \tag{4}$$

where $\mathbf{G}^w$ is $[0, 0, g^w]^{\mathrm{T}}$, and $\mathbf{R}_k$ is the rotation matrix. $\mathbf{R}_k$ comprises the roll, pitch, and heading at the instant t_k, denoted as r_k, p_k, and h_k. Three angles affect horizontal motion accelerations, and the vertical motion acceleration is related to the roll and pitch. Unlike the heading, roll and pitch can be corrected using accelerometers. On the one hand, they are accessible in multiple environments with any smartphone. On the other hand, they are more reliable compared to heading estimation. Thus, the vertical component of the (3) is extracted for further analysis, which can be extracted as

$$\begin{aligned}\Delta L_z = {} & n v_{z,0}^w \Delta t + (n-1)\, l_{z,1}^w \Delta t^2 \\ & + (n-2)\, l_{z,2}^w \Delta t^2 + \cdots + l_{z,n-1}^w \Delta t^2,\end{aligned} \tag{5}$$

where

$$\begin{aligned}l_{z,k}^w = {} & -\sin(r_k)\cos(p_k)\, a_{x,k}^b + \sin(p_k)\, a_{y,k}^b \\ & + \cos(r_k)\cos(p_k)\, a_{z,k}^b - g^w,\end{aligned} \tag{6}$$

where $v_{z,0}^w$ is the initial vertical speed in the world frame of the period T. $l_{z,k}^w$ is the vertical motion acceleration at the instant t_k in the world frame. $a_{d,k}^b$ denotes the d-axis specific force at the instant t_k in the body frame.

When a pedestrian is walking on a flat surface, $v_{z,0}^w$ can be regarded as zero. Furthermore, $v_{z,0}^w$ can also be considered zero if the period is long enough, even if the pedestrian ascends or descends the stairs. Thus, the (5) can be simplified as

$$\begin{aligned}\Delta L_z = {} & (n-1)\, l_{z,1}^w \Delta t^2 + (n-2)\, l_{z,2}^w \Delta t^2 \\ & + \cdots + l_{z,n-1}^w \Delta t^2.\end{aligned} \tag{7}$$

It can be noticed that the vertical displacement is only related to vertical motion accelerations obtained using accelerometer data, roll, and pitch, in which the roll and pitch can be accessible by fusing the gyroscope and accelerometer data. However, the accuracy of vertical displacement ΔL_z cannot be guaranteed in actual systems if the accelerometer data is directly employed in (7) due to the sensor error and its accumulation. Thus, a neural network is adopted in this paper to regress the vertical displacement based on the mapping relation in (7).

1) Architecture and Training Data Collection: This paper leverages a 1D version of ResNet-18 architecture, adding a fully connected layer at the end to regress the vertical displacement. The input dimension of the network is $(n-1) \times 1$, which includes $n-1$ vertical motion acceleration samples, and the output is the vertical displacement, which is expressed as

$$\Delta \hat{L}_{z,n} = f_{ResNet}\left(\hat{\mathbf{l}}_{z,1:n-1}^w\right). \tag{8}$$

The loss function is defined as the Mean Square Error (MSE) form, which is expressed as

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta \hat{L}_{z_i,n} - \Delta L_{z_i,n}\right)^2, \tag{9}$$

where $\Delta \hat{L}_{z_{i,n}}$ and $\Delta L_{z_{i,n}}$ are regressed and referenced vertical displacements. N is the amount of data in the training dataset.

An iPhone 12 was used to collect the training dataset, in which only a built-in IMU and a barometric pressure sensor are needed. Although the barometer is easily affected by environments, it is effortless to calibrate the observed value in the training process by measuring the stairs, reducing the effort for collecting the dataset. The dataset consists of over 30 sequences with 10 hours total, covering stairs of different height changes. The training and validation datasets include 22 and 9 sequences, and the rest is used for testing. The Adam optimizer is used for training with an initial learning rate of 0.0001. The model is trained for a total of 5,000 epochs using an NVIDIA GeForce RTX 4090. Since only an iPhone 12 was used for data collection, the regression performance of vertical position change degrades when other smartphones are used with the trained model. However, this paper aims to explore the feasibility of tracking with inertial-only perception. Therefore, only a limited dataset was formed for training.

Unlike the horizontal movement, the vertical position change in short period is less obvious. Therefore, this paper extends the period T to capture the relation, leading to better regression. In this paper, the period T is set as 10 s. When the sampling rate of IMU is 100 Hz, the number of vertical acceleration samples for the network is 999. It is important to note that the neural network is used for regression after 10 s. During the initial 10 s, we assume that the pedestrian moves on a flat surface without any vertical motion, thus setting the vertical displacement to zero.

2) Proposed 3D PDR Model: In the proposed deep vertical odometry-aided 3D PDR model, the horizontal position is implemented based on the traditional PDR algorithm, in which step detection is used to estimate the step length. The step length $\Delta \hat{D}_k$ is obtained using the Weinberg model [42], which can be expressed as

$$\Delta \hat{D}_k = K \sqrt[4]{\hat{a}_{z,\max}^w - \hat{a}_{z,\min}^w}, \tag{10}$$

where K is a pre-set coefficient. $\hat{a}_{z,\max}^w$ and $\hat{a}_{z,\min}^w$ denote the vertical acceleration peak and valley in the world frame. 3D PDR model can be expressed as

$$\begin{aligned}\hat{p}_{x,k}^w &= \hat{p}_{x,k}^w + \Delta \hat{D}_k \cos\hat{\theta}_k \\ \hat{p}_{y,k}^w &= \hat{p}_{y,k}^w + \Delta \hat{D}_k \sin\hat{\theta}_k \\ \hat{p}_{z,k}^w &= \hat{p}_{z,k}^w + \Delta \hat{H}_k,\end{aligned} \tag{11}$$

where $\hat{\mathbf{p}}_i^w = [\hat{p}_{x,i}^w, \hat{p}_{y,i}^w, \hat{p}_{z,i}^w]^{\mathrm{T}}$ represents the position estimate of the pedestrian at the instant t_i in the world frame. $\hat{\theta}_k$ stands for the heading at the instant t_k. $\hat{\theta}_k$ is calculated by integrating the angular rate. To restrain heading drift, the gyroscope bias must first be removed. $\Delta \hat{H}_k$ is the estimated vertical displacement

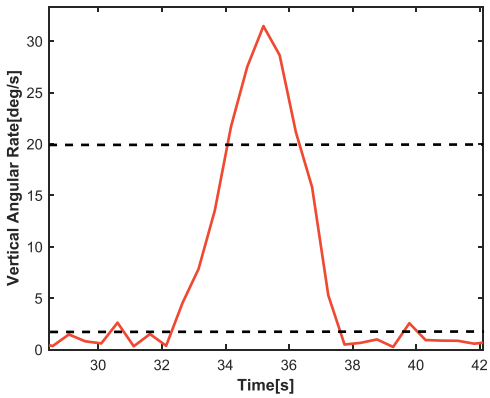


Fig. 2. The corner-turning behavior sequence detection, in which the sequence can be determined by two valleys of vertical angular rates expressed in the world frame.

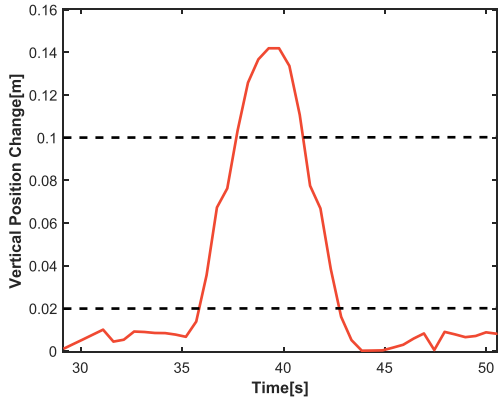


Fig. 3. The stair-taking behavior sequence detection, in which the sequence can be determined by two valleys of vertical position changes.

between two successive instants, which can be calculated by

$$\Delta\hat{H}_k = \Delta\hat{L}_{z,k} - \Delta\hat{L}_{z,k-1} + \hat{p}^w_{z,k-T} - \hat{p}^w_{z,k-1-T}. \quad (12)$$

The covariance $\hat{\mathbf{\Sigma}}_k$ can express the uncertainty of the current position and its propagation model can be expressed as

$$\hat{\mathbf{\Sigma}}_k = \hat{\mathbf{F}}_k \hat{\mathbf{\Sigma}}_{k-1} \hat{\mathbf{F}}_k^{\mathrm{T}} + \hat{\mathbf{G}}_k \mathbf{Q} \hat{\mathbf{G}}_k^{\mathrm{T}}, \quad (13)$$

where $\hat{\mathbf{F}}_k$ is the state transition matrix that can be described by a 3D identity matrix. $\mathbf{Q}$ denotes the pre-set covariance matrix of the noise. $\hat{\mathbf{G}}_k$ is the noise matrix that can be expressed as

$$\hat{\mathbf{G}}_k = \begin{bmatrix} \cos\hat{\theta}_k & -\Delta\hat{D}_k \sin\hat{\theta}_k & 0 \\ \sin\hat{\theta}_k & \Delta\hat{D}_k \cos\hat{\theta}_k & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (14)$$

Equation (13) shows that the uncertainty of pedestrian position increases with the movement. Therefore, $\hat{\mathbf{\Sigma}}_k$ is updated based on the posterior density $\mathbf{P}(\mathbf{X}|\mathbf{Y})$ after each optimization [41].

B. Multi-Hypothesis Behavior Loop Closures

3D PDR can provide continuous location tracking but suffers from drift in long-distance movement as it is a dead reckoning algorithm. Human behaviors detected by inertial data are used as landmarks in SLAM mode in this paper, and the loop closure can be triggered to calibrate PDR errors when the user revisits the same landmark. The multi-hypothesis behavior loop closure is formed to improve the reliability of the correction.

1) Behavior Landmark and Multi-Hypothesis Model Formulation: This paper uses corner-turning and stair-taking, widespread in structured indoor 3D scenarios, to build landmarks. These two types of behaviors are detected based on empirical data from the IMU. By having the pedestrian move several times while holding the smartphone, we can identify the pattern and determine if a behavior occurs.

The corner-turning behaviors are detected with the angular rate and heading. The gyroscope data is first transformed to the world frame using the rotation matrix $\mathbf{R}_k$. Then, the vertical angular rate can be extracted, as indicated in Fig. 2. The threshold for the turning detection is set to 20 $\mathrm{deg/s}$ based on empirical data when a person takes a turn. The peak is the maximum vertical angular rate above the threshold, whose instant is marked as t_o. The valley is the first angular rate of less than 2 $\mathrm{deg/s}$ on both sides of the peak. There are two valleys, and their instants are marked as t_{o-a} and t_{o+b}. Thus, the turning sequence can be determined by $[t_{o-a}, t_{o+b}]$.

The turning behavior can be divided into normal turn and U-turn. The former is the corner-turning, while the latter refers to the turning like a U shape. The U-turn cannot be formulated as a landmark because it can occur anywhere. Thus, the heading is introduced to differentiate the type of the turning, which can be expressed as

$$\Delta\hat{\theta} = abs\left(mean\left(\hat{\boldsymbol{\theta}}_{o-a:o}\right) - mean\left(\hat{\boldsymbol{\theta}}_{o:o+b}\right)\right), \quad (15)$$

where $\hat{\boldsymbol{\theta}}_{o-a:o}$ and $\hat{\boldsymbol{\theta}}_{o:o+b}$ are the estimated heading sequence from the instant t_{o-a} to t_o, and instant t_o to t_{o+b}, respectively. The normal turn and U-turn can be differentiated by

$$\begin{cases} Normal\ turn & \Delta\hat{\theta} < \theta_T \\ U-turn & \Delta\hat{\theta} > \theta_T \end{cases}, \quad (16)$$

where θ_T is the threshold set as 90 deg in this paper, which is set according to our test results.

The stair-taking behaviors are detected based on the estimated vertical position change using the proposed 3D PDR shown in Fig. 3. The threshold for the stair-taking detection is set to 0.1 m. The peak is the maximum vertical position change above the threshold, whose instant is marked as t_o. The valley is the first position change of less than 0.02 m on both sides of the peak. There are two valleys, and their instants are marked as t_{o-a} and t_{o+b}. Therefore, the stair-taking sequence can also be determined by $[t_{o-a}, t_{o+b}]$.

It is necessary to point out that the proposed method assumes the smartphone is held in a texting mode, which is a relatively stable holding mode to ensure accurate behavior detection. Once the corner-turning and stair-taking are detected, they are regarded as landmarks, and their positions are recorded with the pedestrians' positions based on the assumption that the behavior occurs on the moving track. Unlike vision or LiDAR, an IMU cannot sense the surrounding environment to detect

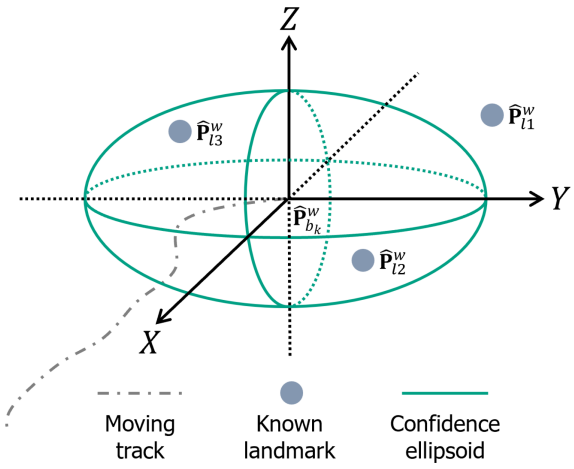


Fig. 4. The relationship among the moving track, confidence ellipsoid, and known landmarks. Known landmarks inside the ellipsoid are regarded as candidates for loop closures.

if the same landmark is revisited. In traditional methods, the nearest point or similarity is often used to detect the loop closure. However, behavior features at different locations are similar, and unexpected behavior landmarks may exist due to incorrect detection, which can easily cause false loop closures. Therefore, a multi-hypothesis behavior loop closure is proposed in this paper.

First, a statistic-based approach [43] is designed to determine possible revisited landmarks. A confidence ellipsoid is obtained according to the position covariance. If one landmark is in the ellipsoid, it can be built as a potential loop closure, a hypothesis, as shown in Fig. 4.

Let the positions of the ith landmark and the pedestrian at the current instant be $\hat{\mathbf{P}}^w_{li}$ and $\hat{\mathbf{P}}^w_k$, then transform the landmark to a coordinate in which $\hat{\mathbf{P}}^w_k$ is the origin by

$$\hat{\mathbf{P}}^w_c = \hat{\mathbf{P}}^w_{li} - \hat{\mathbf{P}}^w_k. \tag{17}$$

Let the covariance matrices of $\hat{\mathbf{P}}^w_{li}$ and $\hat{\mathbf{P}}^w_k$ be $\hat{\mathbf{\Sigma}}_{li}$ and $\hat{\mathbf{\Sigma}}_k$, the covariance matrix of $\hat{\mathbf{P}}^w_c$ can be expressed as

$$\hat{\mathbf{\Sigma}}_c = \hat{\mathbf{\Sigma}}_{li} + \hat{\mathbf{\Sigma}}_k. \tag{18}$$

Then, multiply the transformed coordinate $\hat{\mathbf{P}}^w_c$ by the inverse of its covariance matrix to obtain mutually independent coordinate, which can be expressed as

$$\hat{\mathbf{P}}^{w\mathrm{T}}_{nc} = \hat{\mathbf{P}}^{w\mathrm{T}}_c \hat{\mathbf{\Sigma}}^{-1}_c. \tag{19}$$

The square of the Euclidean distance based on the new coordinate can be calculated, shown as

$$r_d = \hat{\mathbf{P}}^{w\mathrm{T}}_{nc} \hat{\mathbf{P}}^w_{nc}. \tag{20}$$

The confidence probability is set as 0.95in this paper, making the significance level 0.05. We choose this value according to the 68-95-99.7 rule in statistics. A confidence probability of 0.95 can avoid the loss of loop closures and increased computational time. The freedom of the position is three. Therefore, the chi-square critical value is 7.815 by querying the chi-square distribution table. Whether the landmark is within the scope of the confidence ellipsoid can be expressed as

$$\begin{cases} Inside & r_d < 7.815 \\ Outside & r_d > 7.815 \end{cases}, \tag{21}$$

r_d of all the known landmarks will be calculated to determine the number of the potential loop closures.

There may be multiple known landmarks within the ellipsoid when a behavior is detected, such as $\hat{\mathbf{P}}^w_{l2}$ and $\hat{\mathbf{P}}^w_{l3}$ shown in Fig. 4. In addition, it is also possible that the current behavior occurs at a new place given the inaccuracy of the covariance. Therefore, different hypotheses with corresponding probability need to be formed. According to the multi-hypothesis tracking theory [44], the probability of the jth hypothesis can be calculated by

$$p\left(\Omega^j_k\right) = \eta \left(f\left(\mathbf{z}_k\right) p^{det}_k\right)^{\delta} \circ \lambda^{1-\delta}_{new} \circ p\left(\Omega^j_{k-1}\right), \tag{22}$$

where Ω^j_k is the jth hypothesis at the instant t_k. η denotes the normalizer. $f(\mathbf{z}_k)$ and p^{det}_k signify the probability density of measurement and the inverse of the number of behaviors in the ellipsoid. λ_{new} is the average rate of behavior. When a known landmark is found, δ is 1. Otherwise, it is 0. In the (22), the probability density can be assumed as a normal distribution by

$$f\left(\mathbf{z}_k\right) = \mathcal{N}\left(\mathbf{z}_k; \hat{\mathbf{z}}_k, \hat{\mathbf{U}}_k\right), \tag{23}$$

where $\mathbf{z}_k$ and $\hat{\mathbf{z}}_k$ are the actual and predicted measurements. $\hat{\mathbf{U}}_k$ is the covariance matrix of $\hat{\mathbf{z}}_k$. In the proposed approach, the measurement is the position difference between the landmark and the pedestrian. Since the landmark is on the moving track, the actual measurement can be expressed as

$$\mathbf{z}_k = [0, 0]^{\mathrm{T}}, \tag{24}$$

$\hat{\mathbf{z}}_k$ can be expressed as

$$\hat{\mathbf{z}}_k = \hat{\mathbf{P}}^w_{li} - \hat{\mathbf{P}}^w_k, \tag{25}$$

$\hat{\mathbf{U}}_k$ can be calculated by

$$\hat{\mathbf{U}}_k = \hat{\mathbf{H}}_k \hat{\mathbf{\Sigma}}_k \hat{\mathbf{H}}^{\mathrm{T}}_k + \hat{\mathbf{H}}_{li}\, \hat{\mathbf{\Sigma}}_{li} \hat{\mathbf{H}}^{\mathrm{T}}_{li}, \tag{26}$$

where $\hat{\mathbf{H}}_k$ and $\hat{\mathbf{H}}_{li}$ denote the Jacobian matrices of the (25) relative to $\hat{\mathbf{P}}^w_{b_k}$ and $\hat{\mathbf{P}}^w_{li}$.

According to (22), the probability of each hypothesis can be obtained. It is noted that the probability is updated only at the instant t_o in the behavior sequence. In addition, it shows that the number of hypotheses grows each time a behavior is detected, leading to an increased computational burden. Therefore, possible hypotheses will be cut using the N-scan back pruning algorithm to maintain a certain number of hypotheses.

2) Key Point Data Association: In the proposed method, the data association for loop closures is formulated using the corner-turning and stair-taking behavior sequence mentioned in Figs. 2 and 3. There are three types of key points in the sequence: entry point, middle point, and exit point, which correspond to t_{o-a}, t_o, and t_{o+b}. First, two corner-turning trajectories are considered, shown in Fig. 5. Let the red and blue lines denote the trajectories where a pedestrian passes and revisits the same corner.

In (a), two trajectories are similar. Therefore, three key points are all employed to formulate the data association, like the black

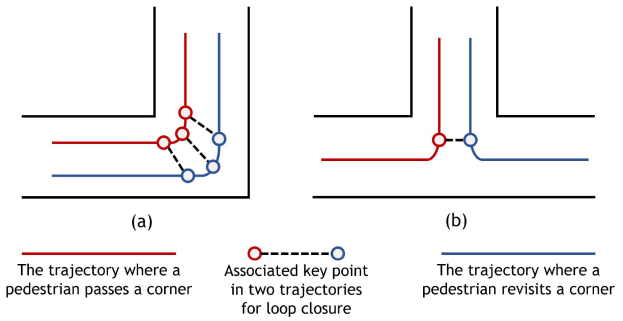


Fig. 5. Two corner-turning trajectories when a person passes and revisits the same corner. In (a), the red and blue trajectories are similar. The entry, middle, and exit point are associated. In (b), the red and blue trajectories are different, and only one key point is chosen for association.

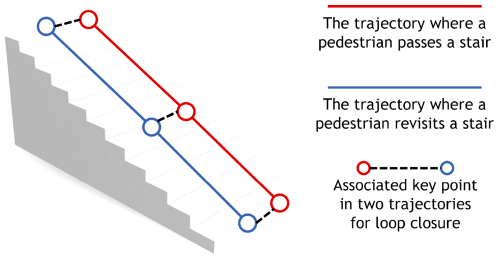


Fig. 6. The stair-taking trajectories when a person passes and revisits the same stair. Red and blue lines are two trajectories.

dotted lines. Two middle points are associated, while different trajectories' entry and exit points are matched according to the moving direction. For example, if the moving directions of the pedestrian on these two trajectories are similar, the entry points of these two lines are matched, and the exit points are associated. Otherwise, the entry point will be associated with the exit point. In (b), two trajectories are different, so only one key point will be chosen for the data association, which is also decided by the moving direction of the pedestrian.

As for the stair-taking behavior, the red and blue lines denote the trajectories where a pedestrian passes and revisits the same stair, as shown in Fig. 6. In this situation, all key points can be utilized to achieve the data association given the motion feature of ascending and descending stairs. Two middle points in red and blue lines can always be associated to form loop closures. The associations of entry and exit points depend on the moving direction of the pedestrian.

V. Factor Graph Optimization Based on 3D PDR and Behavior Loop Closures

Unlike traditional methods, this paper first adopts the factor graph to achieve state estimation in indoor 3D SLAM with only smartphone built-in IMU. In SLAM, past and current states are connected by loop closures, which means that state estimation accuracy for historical states directly affects current estimation. Filtering-based estimators deal with past states only once and cannot optimize past states with the following data. It signifies that historical errors will be introduced when loop closures are triggered, leading to a compromised solution.

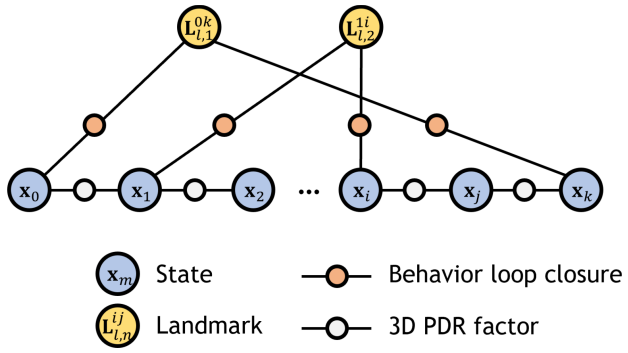


Fig. 7. The structure of the factor graph in the proposed method. 3D PDR factor constrains successive instants, and behavior loop closure connects two states when the pedestrian revisits the same place.

A factor graph $\mathbf{F} = (\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ consists of two types of nodes: the factor node $\mathbf{f}_i \in \boldsymbol{\mathcal{U}}$ and variable node $\mathbf{x}_i \in \boldsymbol{\mathcal{V}}$. When a factor contains a variable, an edge $\mathbf{e}_{ij} \in \boldsymbol{\mathcal{E}}$ exists between the factor node and the variable node [45]. Let the noise follow a Gaussian distribution, then the factor node can be given as:

$$\mathbf{f}_i\left(\mathbf{X}\right) = \left\|\mathbf{z}_i - \mathbf{h}_i\left(\mathbf{X}\right)\right\|^2_{\boldsymbol{\Sigma}_i}, \tag{27}$$

where $\mathbf{X}$ is the state set that needs to be estimated. $\mathbf{z}_i$ and $\mathbf{h}_i(\mathbf{X})$ are the ith actual and predicted measurements. $\boldsymbol{\Sigma}_i$ is the covariance matrix. $\mathbf{d}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{d}$ is the Mahalanobis distance, in which $\mathbf{d}$ is the residual.

Nonlinear optimization is used to solve the state by minimizing the global error formed by available factors, which is expressed by

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \left(\sum_i \mathbf{f}_i\left(\mathbf{X}\right)\right), \tag{28}$$

where $\hat{\mathbf{X}}$ is the estimated global optimal solution.

The structure of the proposed method is shown in Fig. 7. The blue circle with $\mathbf{x}_m$ represents the variable node, and the yellow circle with $\mathbf{L}^{ij}_{l,n}$ is the landmark, in which the subscript l, n is the nth landmark and the superscript ij indicates that states at the instant t_i and t_j are connected. The state vector in the proposed method is given as

$$\begin{aligned} \mathbf{X}_m &= \left[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_m\right] \\ \mathbf{x}_k &= \left[p^w_{x,k}, p^w_{y,k}, p^w_{z,k}\right]^{\mathrm{T}}, \end{aligned} \tag{29}$$

where $\mathbf{X}_m$ is the state set, and $\mathbf{x}_k$ is the state at the instant t_k that includes the position. The cost function of the proposed method is expressed as

$$\mathbf{X}^* = \operatorname{argmin}\left(\sum_{i=1}^{a}\left\|\mathbf{r}_p\left(\hat{\mathbf{z}}_p, \mathbf{X}\right)\right\|^2_{\boldsymbol{\Sigma}_p} + \sum_{i=1}^{b}\left\|\mathbf{r}_c\left(\hat{\mathbf{z}}_c, \mathbf{X}\right)\right\|^2_{\boldsymbol{\Sigma}_c}\right), \tag{30}$$

where $\mathbf{r}_p(\hat{\mathbf{z}}_p, \mathbf{X})$ and $\mathbf{r}_c(\hat{\mathbf{z}}_c, \mathbf{X})$ denote the 3D PDR and behavior loop closure residuals. $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_c$ are the covariance matrices of 3D PDR and behavior loop closure models. a and b signify the numbers of corresponding residuals. According to the (11), the residual of 3D PDR can be expressed

as

$$\mathbf{r}_p\left(\hat{\mathbf{z}}_p, \mathbf{X}\right) = \mathbf{p}_k^w - \mathbf{p}_{k-1}^w - \Delta\hat{\mathbf{p}}_k^w = \begin{bmatrix} p_{x,k}^w - p_{x,k-1}^w - \Delta\hat{D}_k \cos\hat{\theta}_k \\ p_{y,k}^w - p_{y,k-1}^w - \Delta\hat{D}_k \sin\hat{\theta}_k \\ p_{z,k}^w - p_{z,k-1}^w - \Delta\hat{H}_k \end{bmatrix}. \quad (31)$$

The behavior loop closure residual can be given as

$$\mathbf{r}_c\left(\hat{\mathbf{z}}_c, \mathbf{X}\right) = \mathbf{p}_i^w - \mathbf{p}_j^w = \begin{bmatrix} p_{x,i}^w - p_{x,j}^w \\ p_{y,i}^w - p_{y,j}^w \\ p_{z,i}^w - p_{z,j}^w \end{bmatrix}, \quad (32)$$

where $\mathbf{p}_i^w$ and $\mathbf{p}_j^w$ represent the two location points connected by loop closures.

To solve (30) and obtain the estimation $\mathbf{X}^*$, we first choose $\mathbf{X}_0$ as a priori estimate based on previous estimations. The Jacobian matrix of each residual in (30) is then calculated based on $\mathbf{X}_0$. The Levenberg-Marquardt algorithm is used to solve the optimization, and the iterative process is performed until the termination condition is met. The estimated $\mathbf{X}^*$ is a set of states that includes the states at different instants. In $\mathbf{X}^*$, the state at the latest instant represents the real-time location, while the other states denote the optimized historical locations.

VI. Experimental Evaluation

The experiments were carried out to evaluate the performance of the proposed method. In the experiments, we used an iPhone 12 and a LiDAR-inertial integrated system called Mid-360 from Livox Technology Company. iPhone's IMU data was utilized to achieve the inertial-only 3D pedestrian SLAM. Mid-360 was used to run the FAST-LIO [46] to provide accurate ground truth. Considering that our proposed method only uses the smartphone IMU, the Mid-360 is sufficient as the reference system. In the Mid-360, the gyroscope and accelerometer noise levels are approximately 4.5 $\mathrm{mdps}/\sqrt{\mathrm{Hz}}$ and 100 $\mu\mathrm{g}/\sqrt{\mathrm{Hz}}$, respectively. Experimental scenarios include an office building and a broader site composed of stairs with different vertical position changes on campus. The experimental scheme is shown in Fig. 8. The smartphone is held in a texting mode to collect the data.

A. Evaluation Methods and Metrics

As for the comparison, the following methods are included, which can be summarized as:

1) *3D-PDR* [14]: which tracks the pedestrian's location with the model-based dead reckoning.
2) *RoNIN*: [18] which calculates the pedestrian's location based on the regressed displacements from the neural network. Specifically, we used the RoNIN ResNet model, hereafter referred to as RoNIN. We trained a model using a self-collected dataset for RoNIN, adjusting the window size to adapt to 100 Hz IMU data. ARKit was employed to provide the reference for training.
3) *PF-SLAM* [7], [30]: which tracks the pedestrian's location based on the 3D-PDR and behavior loop closures with a particle filter framework.

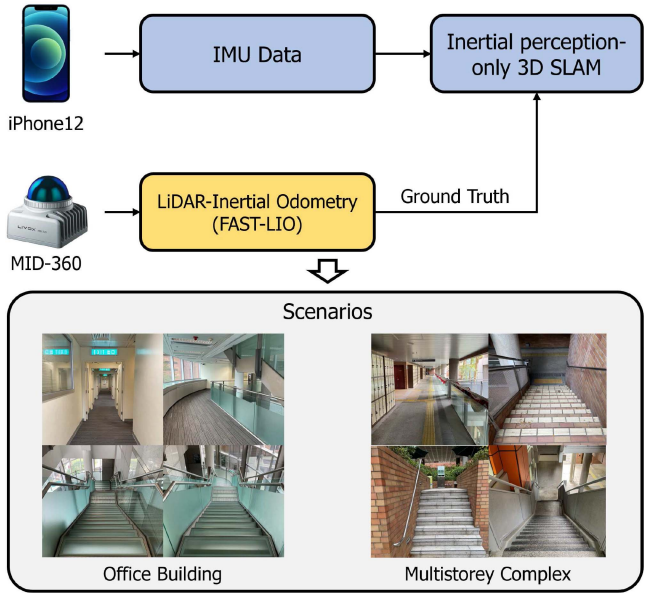


Fig. 8. The experimental scheme. A LiDAR-inertial integrated system was used to provide the ground truth. Test scenarios include the office building and multistorey complex.

4) *Pro-RO*: the proposed method in a real-time estimation mode, in which only past and current information are used.
5) *Pro-PO*: the proposed method in an off-line mode, where all information is used to recover the pedestrian's trajectory.

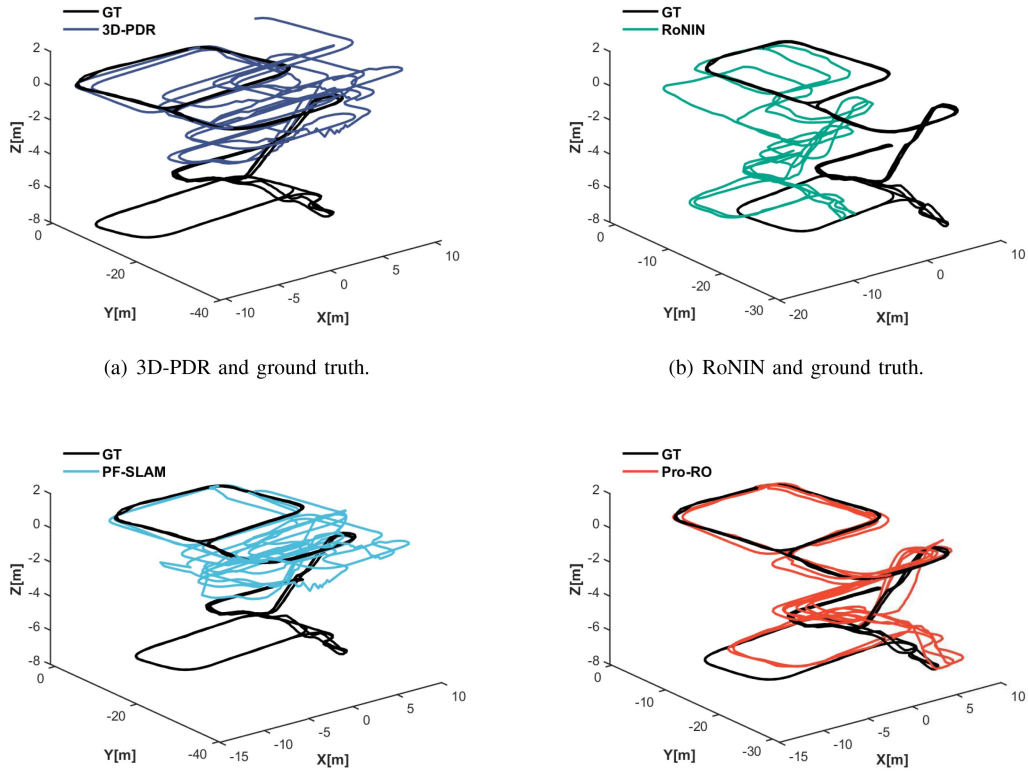
To qualify the positioning error, the absolute trajectory error (ATE) and relative trajectory error (RTE) are used.

B. Experiment in the Office Building

An experimental test was first carried out in an office building, covering a three-floor moving track, whose length is about 593 m.

The 3D trajectory comparison among 3D-PDR, RoNIN, PF-SLAM, and Pro-RO in the office building is shown in Fig. 9. Horizontal and vertical positioning error comparison is given in Fig. 10. The horizontal positioning error is calculated as the square root of the sum of the squares of the errors in the X and Y directions. It can be noticed that the trajectory estimations in 3D-PDR and PF-SLAM are worse than RoNIN and Pro-RO due to the significant errors in the vertical position. Furthermore, the erroneous vertical position estimation easily gives rise to false loop closures, influencing the estimation for horizontal position, like the increased horizontal positioning error of PF-SLAM in Fig. 10. Although RoNIN can achieve better vertical position estimation than 3D-PDR and PF-SLAM, its horizontal position has more significant errors since RoNIN directly regresses the displacements in the world frame, which is easily affected by the heading uncertainty.

The proposed method shows better trajectory estimation than traditional approaches, with more minor horizontal and vertical positioning errors. First, the proposed method can provide a better vertical position solution to guarantee the prior position for loop closure formulation is in a reasonable range, avoiding



(a) 3D-PDR and ground truth.

(b) RoNIN and ground truth.

(c) PF-SLAM and ground truth.

(d) Pro-RO and ground truth.

Fig. 9. The trajectory comparison among 3D-PDR, RoNIN, PF-SLAM, and Pro-RO in the office building.

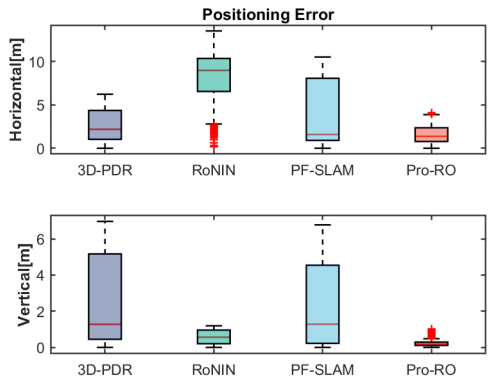


Fig. 10. The horizontal and vertical positioning error comparison among 3D-PDR, RoNIN, PF-SLAM, and Pro-RO in the office building.

the case of PF-SLAM in which a wrong trajectory is calculated, even the multi-hypothesis tracking is used in the particle filter. Second, the proposed method can make the most of behavior loop closures to suppress the error accumulation. The 3D positioning ATE and RTE comparison is presented in Table II.

C. Experiment in the Multistorey Complex

Another experiment was conducted in a multistorey complex to evaluate the proposed method's performance further. In this scenario, the length of the moving tracking is about 583 m.

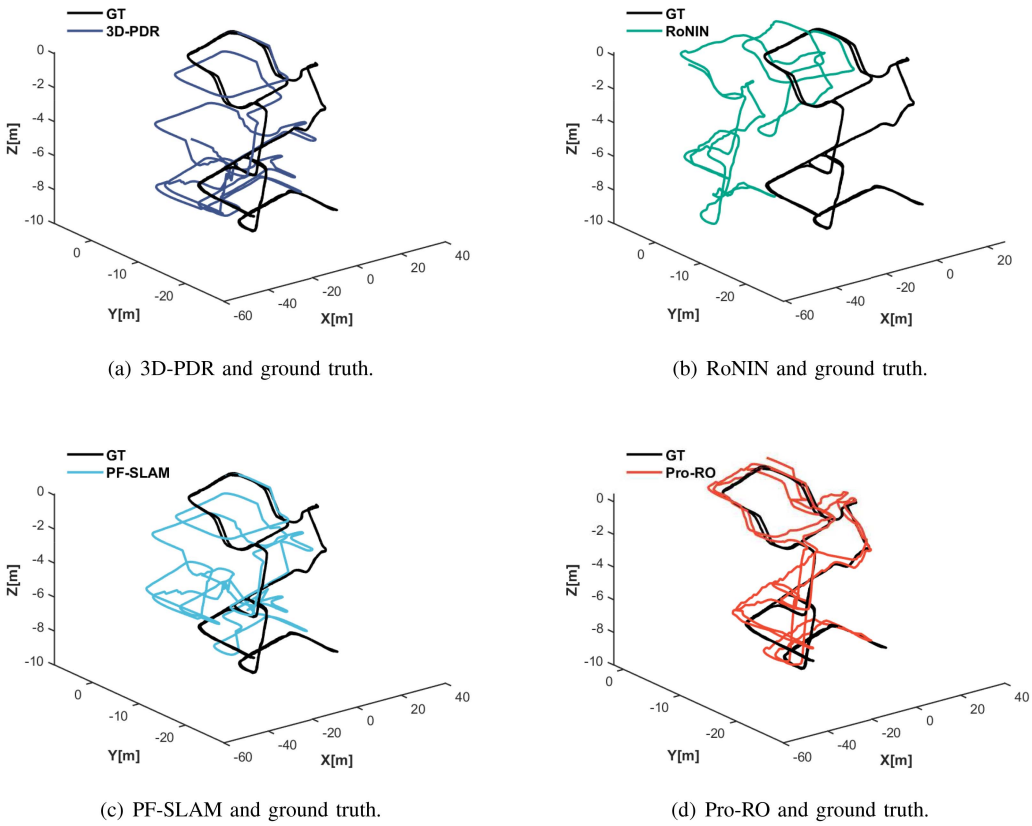
The trajectory comparison among different methods is shown in Fig. 11. The error comparison for the horizontal and vertical position estimation is given in Fig. 12. It can be noticed that the vertical positioning error in 3D-PDR and PF-SLAM still shows worse estimation results than RoNIN and Pro-RO. Moreover, it can be observed that the trajectory of RoNIN deviates from the ground truth due to the compromised horizontal position. As for Pro-RO, it shows that its trajectory is closer to the ground truth than other methods. Although there are drifts in some segments, the PDR errors can be calibrated when the behavior loop closure is triggered. The 3D positioning ATE and RTE comparison is shown in Table III.

TABLE II
THE 3D POSITIONING ATE AND RTE COMPARISON AMONG 3D-PDR, RONIN, PF-SLAM, AND PRO-RO IN THE OFFICE BUILDING

Method	ATE (m)	RTE (m)
3D-PDR	3.40	3.51
RoNIN	6.27	3.46
PF-SLAM	4.33	4.78
Pro-RO	1.30	1.70

TABLE III
THE 3D POSITIONING ATE AND RTE COMPARISON AMONG 3D-PDR, RONIN, PF-SLAM, AND PRO-RO IN THE MULTISTOREY COMPLEX

Method	ATE (m)	RTE (m)
3D-PDR	4.51	3.61
RoNIN	18.62	11.60
PF-SLAM	6.07	5.95
Pro-RO	1.84	1.89



(a) 3D-PDR and ground truth.

(b) RoNIN and ground truth.

(c) PF-SLAM and ground truth.

(d) Pro-RO and ground truth.

Fig. 11. The trajectory comparison among 3D-PDR, RoNIN, PF-SLAM, and Pro-RO in the multistorey complex.

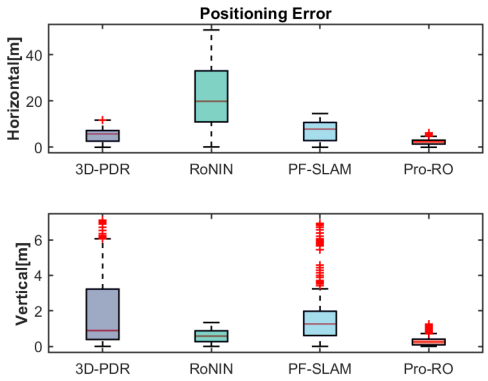


Fig. 12. The horizontal and vertical positioning error comparison among 3D-PDR, RoNIN, PF-SLAM, and Pro-RO in the multistorey complex.

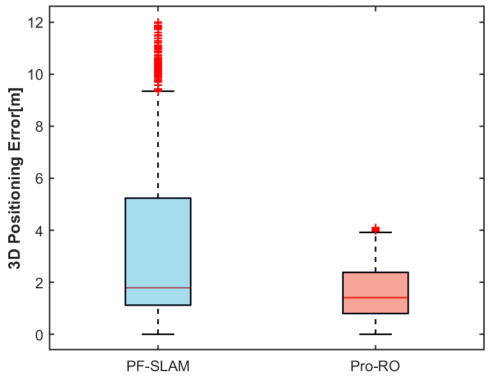


Fig. 13. The 3D positioning error comparison between PF-SLAM and Pro-RO in the office building.

D. Discussion: Estimation Performance Between the Particle Filter and Factor Graph Optimization

In Sections B and C, the pedestrian's vertical position in PF-SLAM is estimated by detecting stair-taking behaviors and accumulating the height of each stair step, and it can cause false loop closures, leading to severely reduced positioning accuracy. In this part, the proposed deep vertical odometry-aided 3D PDR is also leveraged in PF-SLAM to achieve the position prediction. Based on the variable-controlling scheme, we can evaluate the advantage of the factor graph optimization compared with the particle filter.

Fig. 13 present 3D positioning error comparisons between the PF-SLAM and Pro-RO in the office building. It is observed that the positioning error of PF-SLAM is decreased compared to the counterpart in Section B as more accurate position prediction can be employed to determine loop closures. However, Pro-RO still outperforms PF-SLAM. The main reason is that the particle filter can only estimate current states, which means that the past states cannot be further optimized with the following information. When the pedestrian revisits

TABLE IV
THE 3D POSITIONING ATE AND RTE COMPARISON BETWEEN PF-SLAM AND PRO-RO IN THE OFFICE BUILDING AND MULTISTOREY COMPLEX

Method	Office building		Multistorey complex	
	ATE (m)	RTE (m)	ATE (m)	RTE (m)
PF-SLAM	3.29	4.84	4.85	4.21
Pro-RO	1.30	1.70	1.84	1.89

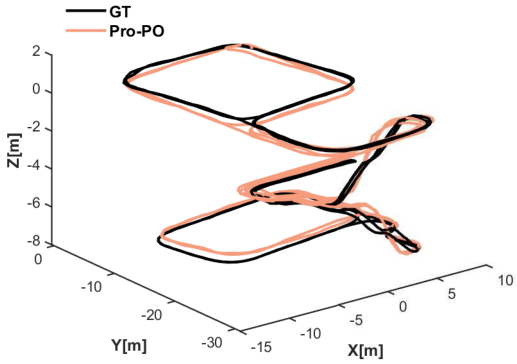


Fig. 14. The trajectory comparison between Pro-PO and ground truth in the office building.

the same place, the known landmark is used to update the current position and the landmark's error will also be introduced into the current estimation. On the contrary, the factor graph can iteratively optimize the past state, reducing the error in visited landmarks. Thus, a better state estimation can be achieved.

The 3D positioning ATE and RTE comparison in the office building and multistorey complex is given in Table IV. It shows that the ATE of Pro-RO is reduced by 60% and 62% compared with PF-SLAM.

E. Discussion: Estimation Performance Between the Offline and Real-Time Optimization

In above parts, Pro-RO is used to compare in real-time situations. Compared to filtering, factor graph optimization can improve the estimation performance for the whole trajectory via offline batch optimization, which can be used for post-analysis and track reuse for other applications.

In this part, the estimation performance of Pro-PO is given, which is shown in Fig. 14. Unlike Pro-RO in Fig. 9, the trajectory estimated by Pro-PO is smoother than real-time results. The main reason is that Pro-PO can use all information to obtain a trajectory, while only past data relative to the current instant can be used by Pro-RO. Therefore, it can be noticed that some segments in Pro-RO present drift as the position cannot be calibrated until the subsequent loop closure is triggered.

The 3D positioning ATE and RTE comparison between Pro-RO and Pro-PO in the office building and multistorey complex is shown in Table V. The ATE of Pro-PO is reduced by 48% and 34% compared with Pro-RO.

TABLE V
THE 3D POSITIONING ATE AND RTE COMPARISON BETWEEN PRO-RO AND PRO-PO IN THE OFFICE BUILDING AND MULTISTOREY COMPLEX

Method	Office building		Multistorey complex	
	ATE (m)	RTE (m)	ATE (m)	RTE (m)
Pro-RO	1.30	1.70	1.84	1.89
Pro-PO	0.68	1.01	1.21	1.53

F. Discussion: Height Estimation Performance Between RoNIN and the Proposed Method

In this part, the estimation effect of the proposed deep vertical odometry is evaluated. The height change regressed by RoNIN and the proposed method is compared in Fig. 15. In RoNIN, the accelerometer data in the world frame is directly utilized as the input for the neural network, and the period of IMU data for regression is relatively short. It is difficult to accurately differentiate the IMU data in a short interval, especially in the vertical direction. It is noticed that the RoNIN regresses some height change even though the pedestrian is walking on a flat surface. In the proposed method, we use vertical linear acceleration as the input to guarantee the learnable terms in deep learning [47]. Moreover, we extend the period to let the network capture the prolonged features of IMU data. It indicates that the proposed approach can regress height change more accurately, as shown in Fig. 15(b).

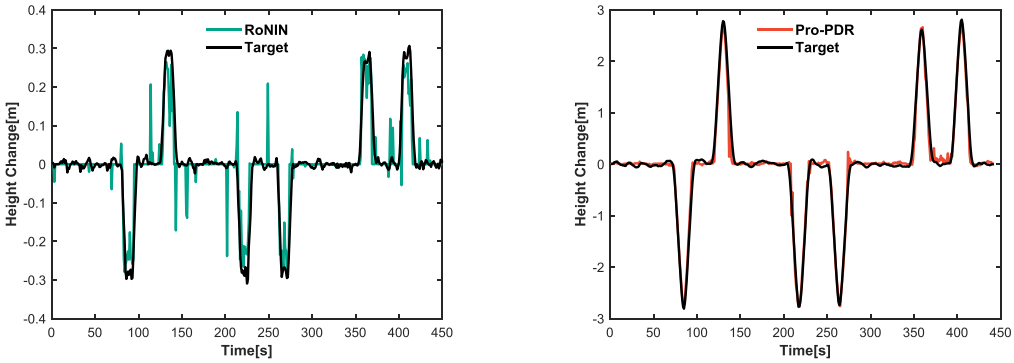
The vertical position estimation comparison between RoNIN and the proposed method is presented in Fig. 16. It shows that the vertical position estimated by the proposed method is closer to the ground truth. Given that the less accurate height changes are regressed by RoNIN, thus leading to a compromised vertical position.

G. Discussion: Reliability of the Proposed Method

This paper formulates a multi-hypothesis factor graph for the ambiguous behavior-based data association. This part first gives the advantage of multi-hypothesis tracking compared to general factor graph optimization, which is based on a single hypothesis, denoted as S-FGO.

The trajectory comparison between S-FGO and ground truth in the multistorey complex is given in Fig. 17. The loop closure in the S-FGO is determined by the matching method in [35]. It can be noticed that S-FGO cannot correctly estimate the pedestrian's position. There are some sudden changes in the track due to the false loop closure. In this method, the false loop closure often occurs as only one hypothesis is considered, easily leading to the wrong data association.

In addition, this paper discusses the performance of the proposed method when confronted with false loop closures. Pro-PO(lr) is introduced in this section. In fact, Pro-PO(lr) and Pro-PO share the same algorithm. The difference lies in the average rate of behavior, which is smaller in Pro-PO(lr). Fig. 18 shows the partial trajectory from the beginning to the point where the first false loop closure occurs. On the right side of Fig. 18, the ground truth indicates a stair-taking behavior. However, there is a false loop closure in Pro-PO(lr), leading to an incorrect trajectory estimation. Fig. 19 shows the entire trajectory, and we can see



(a) RoNIN. (b) Proposed method.

Fig. 15. The height change regression comparison between RoNIN and the proposed method.

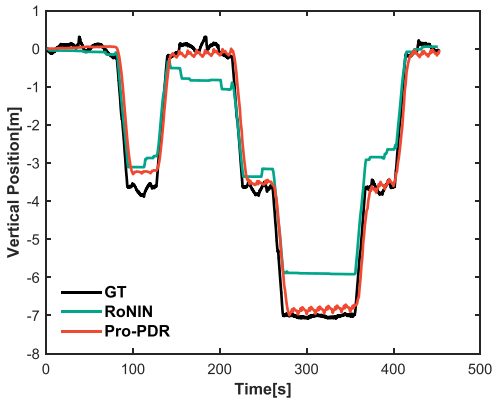


Fig. 16. The vertical position estimation comparison between RoNIN and the proposed method.

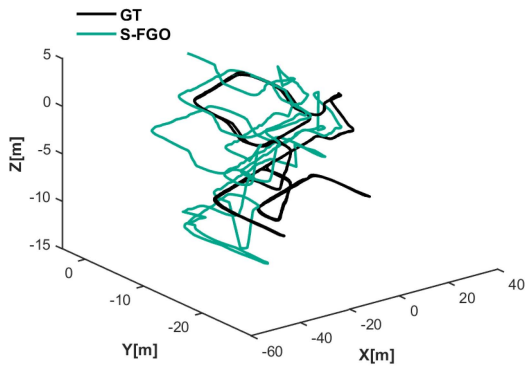


Fig. 17. The trajectory comparison between S-FGO and ground truth in the multistorey complex.

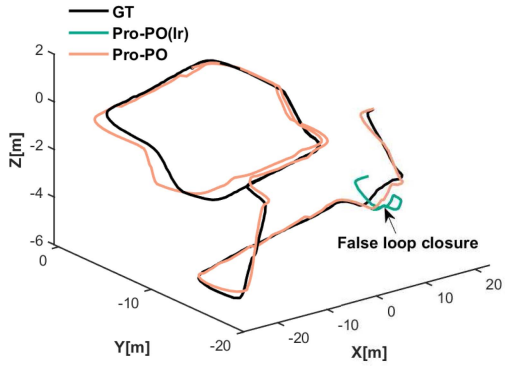


Fig. 18. The false loop closure in Pro-PO(lr) in the multistorey complex.

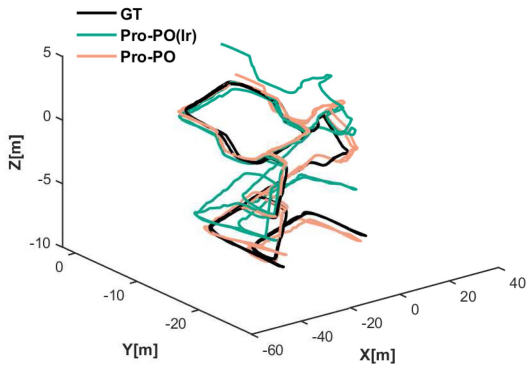


Fig. 19. The trajectory comparison between Pro-PO(lr) and Pro-PO in the multistorey complex.

that Pro-PO(lr) fails to track the location accurately, whereas Pro-PO succeeds.

Even though the chance of incorrect loop closures is small, it can still be considered the output if other probabilities are smaller. If a false hypothesis dominates, the method cannot recover from it. However, increasing the average rate of behavior can address this issue. A higher average rate makes it more likely to represent a new landmark. While this may result in the loss of loop closures, it prevents incorrect trajectory due to unexpected data associations. When more behaviors are detected, the probability of the hypothesis with correct data association will increase. Then, it will be used as the output.

VII. Conclusion

In this paper, we propose a graph-based indoor 3D pedestrian SLAM system with inertial-only perception. We first propose a deep vertical odometry-aided 3D PDR method to address the prediction failure in current model-based and learning-based

methods. Then, we form a graph optimization with multi-hypothesis behavior loop closures to handle the compromised estimation accuracy and ambiguous data association in existing behavior SLAM. Experimental results indicate that the proposed method can suppress the drift and outperform existing methods. Moreover, this paper discusses the advantage of the factor graph in pedestrian SLAM over the filtering-based algorithm. The offline trajectory estimation is also discussed. It shows that the proposed method based on batch optimization can achieve a better solution. Finally, the height regression and reliability of the proposed method are analyzed.

In the future, we will introduce more behaviors to increase the density of landmarks for SLAM, such as door opening, elevator taking, and escalator taking. Therefore, deep learning needs to be introduced to recognize more complex behaviors and enhance detection accuracy when the smartphone is not held stably. Considering that the proposed method may fail when the pedestrian starts taking stairs at the beginning, we will also design a transfer. This transfer can use fewer seconds of data to regress the vertical displacement at the start. In addition, we will collect IMU data from different phone models to develop a more comprehensive model, allowing this method to be used on multiple smartphones.

REFERENCES

[1] Z. Yang, X. Feng, and Q. Zhang, "Adometer: Push the limit of pedestrian indoor localization through cooperation," *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2473–2483, Nov. 2014.

[2] J. Choi, G. Lee, S. Choi, and S. Bahk, "Smartphone based indoor path estimation and localization without human intervention," *IEEE Trans. Mobile Comput.*, vol. 21, no. 2, pp. 681–695, Feb. 2022.

[3] F. Gu, S. Valaee, K. Khoshelham, J. Shang, and R. Zhang, "Landmark graph-based indoor localization," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8343–8355, Sep. 2020.

[4] M. G. Puyol, D. Bobkov, P. Robertson, and T. Jost, "Pedestrian simultaneous localization and mapping in multistory buildings using inertial sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1714–1727, Aug. 2014.

[5] C. Chen, C. X. Lu, J. Wahlström, A. Markham, and N. Trigoni, "Deep neural network based inertial odometry using low-cost inertial measurement units," *IEEE Trans. Mobile Comput.*, vol. 20, no. 4, pp. 1351–1364, Apr. 2021.

[6] R. W. Levi and T. Judd, "Dead reckoning navigational system using accelerometer to measure foot impacts," US Patent 5,583,776, Dec. 10, 1996.

[7] M. Hardegger, S. Mazilu, D. Caraci, F. Hess, D. Roggen, and G. Tröster, "ActionSLAM on a smartphone: At-home tracking with a fully wearable system," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, 2013, pp. 1–8.

[8] C. Wu, Z. Yang, and Y. Liu, "Smartphones based crowdsourcing for indoor localization," *IEEE Trans. Mobile Comput.*, vol. 14, no. 2, pp. 444–457, Feb. 2015.

[9] T.-M. T. Dinh, N.-S. Duong, and Q.-T. Nguyen, "Developing a novel real-time indoor positioning system based on BLE beacons and smartphone sensors," *IEEE Sensors J.*, vol. 21, no. 20, pp. 23055–23068, Oct. 2021.

[10] A. Steed and S. Julier, "Design and implementation of an immersive virtual reality system based on a smartphone platform," in *Proc. 2013 IEEE Symp. 3D User Interfaces*, 2013, pp. 43–46.

[11] Z. Yuan, D. Zhu, C. Chi, J. Tang, C. Liao, and X. Yang, "Visual-inertial state estimation with pre-integration correction for robust mobile augmented reality," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1410–1418.

[12] D. Yan, C. Shi, and T. Li, "An improved PDR system with accurate heading and step length estimation using handheld smartphone," *J. Navigation*, vol. 75, no. 1, pp. 141–159, 2022.

[13] Y. Yu et al., "A novel 3-D indoor localization algorithm based on BLE and multiple sensors," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9359–9372, Jun. 2021.

[14] K. Itzik and L. Yaakov, "Step-length estimation during movement on stairs," in *Proc. 27th Mediterranean Conf. Control Automat.*, 2019, pp. 518–523.

[15] S. Boim, G. Even-Tzur, and I. Klein, "Height difference determination using smartphones based accelerometers," *IEEE Sensors J.*, vol. 22, no. 6, pp. 4908–4915, Mar. 2022.

[16] C. Chen, X. Lu, A. Markham, and N. Trigoni, "IONet: Learning to cure the curse of drift in inertial odometry," in *Proc. AAAI Conf. Artif. Intell.*, 2018, Art. no. 792.

[17] H. Yan, Q. Shan, and Y. Furukawa, "RIDI: Robust IMU double integration," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 621–636.

[18] S. Herath, H. Yan, and Y. Furukawa, "RoNIN: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods," in *Proc. 2020 IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3146–3152.

[19] W. Liu et al., "TLIO: Tight learned inertial odometry," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 5653–5660, Oct. 2020.

[20] Y. Wang, H. Cheng, C. Wang, and M. Q.-H. Meng, "Pose-invariant inertial odometry for pedestrian localization," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 8503512.

[21] S. Khalifa and M. Hassan, "Evaluating mismatch probability of activity-based map matching in indoor positioning," in *Proc. 2012 Int. Conf. Indoor Positioning Indoor Navigation*, 2012, pp. 1–9.

[22] B. Zhou, Q. Li, Q. Mao, W. Tu, and X. Zhang, "Activity sequence-based indoor pedestrian localization using smartphones," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 5, pp. 562–574, Oct. 2015.

[23] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[24] Y. Gu, D. Li, Y. Kamiya, and S. Kamijo, "Integration of positioning and activity context information for lifelog in urban city area," *NAVIGATION: J. Inst. Navigation*, vol. 67, no. 1, pp. 163–179, 2020.

[25] K. Ebadi et al., "Present and future of SLAM in extreme environments: The DARPA SubT challenge," *IEEE Trans. Robot.*, vol. 40, pp. 936–959, 2023.

[26] J. Huang, D. Millman, M. Quigley, D. Stavens, S. Thrun, and A. Aggarwal, "Efficient, generalized indoor WiFi GraphSLAM," in *Proc. 2011 IEEE Int. Conf. Robot. Autom.*, 2011, pp. 1038–1043.

[27] P. Mirowski, T. K. Ho, S. Yi, and M. MacDonald, "SignalSLAM: Simultaneous localization and mapping with mixed WiFi, Bluetooth, LTE and magnetic signals," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, 2013, pp. 1–10.

[28] C. Gao and R. Harle, "MSGD: Scalable back-end for indoor magnetic field-based GraphSLAM," in *Proc. 2017 IEEE Int. Conf. Robot. Automat.*, 2017, pp. 3855–3862.

[29] M. Hardegger, D. Roggen, S. Mazilu, and G. Tröster, "ActionSLAM: Using location-related actions as landmarks in pedestrian SLAM," in *Proc. 2012 Int. Conf. Indoor Positioning Indoor Navigation*, 2012, pp. 1–10.

[30] H. Abdelnasser et al., "SemanticSLAM: Using environment landmarks for unsupervised indoor localization," *IEEE Trans. Mobile Comput.*, vol. 15, no. 7, pp. 1770–1782, Jul. 2016.

[31] A. Shokry, M. Elhamshary, and M. Youssef, "DynamicSLAM: Leveraging human anchors for ubiquitous low-overhead indoor localization," *IEEE Trans. Mobile Comput.*, vol. 20, no. 8, pp. 2563–2575, Aug. 2021.

[32] M. Montemerlo et al., "FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges," in *Proc. 18th Int. Joint Conf. Artif. Intell.*, 2003, pp. 1151–1156.

[33] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[34] S. Bai, J. Lai, P. Lyu, B. Ji, B. Wang, and X. Sun, "A novel plug-and-play factor graph method for asynchronous absolute/relative measurements fusion in multisensor positioning," *IEEE Trans. Ind. Electron.*, vol. 70, no. 1, pp. 940–950, Jan. 2023.

[35] R. Liu et al., "Collaborative SLAM based on WiFi fingerprint similarity and motion information," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1826–1840, Mar. 2020.

[36] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611, Bellingham, WA, USA: SPIE, 1992, pp. 586–606.

[37] M. Hsiao and M. Kaess, "MH-iSAM2: Multi-hypothesis iSAM using Bayes tree and hypo-tree," in *Proc. 2019 Int. Conf. Robot. Automat.*, 2019, pp. 1274–1280.

[38] L. Bernreiter, A. Gawel, H. Sommer, J. Nieto, R. Siegwart, and C. C. Lerma, "Multiple hypothesis semantic mapping for robust data association," *IEEE Robot. Automat. Lett.*, vol. 4, no. 4, pp. 3255–3262, Oct. 2019.
[39] S. Bai, W. Wen, L.-T. Hsu, and P. Yang, "Factor graph optimization-based smartphone IMU-only indoor slam with multi-hypothesis turning behavior loop closures," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 6, pp. 8380–8400, Dec. 2024.
[40] I. J. Cox and S. L. Hingorani, "An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 138–150, Feb. 1996.
[41] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Inst. Technol., Atlanta, GA, Tech. Rep. GT-RIM-CP&R-2012-002, 2012.
[42] H. Weinberg, "Using the ADXL202 in pedometer and personal navigation applications," *Analog Devices AN-602 Appl. Note*, vol. 2, no. 2, pp. 1–6, 2002.
[43] J. Hair, "Multivariate data analysis," *Exploratory Factor Anal.*, 2009.
[44] D. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, Dec. 1979.
[45] F. Dellaert et al., "Factor graphs for robot perception," *Found. Trends Robot.*, vol. 6, no. 1/2, pp. 1–139, 2017.
[46] W. Xu and F. Zhang, "FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 3317–3324, Apr. 2021.
[47] M. Zhang, M. Zhang, Y. Chen, and M. Li, "IMU data processing for inertial aided navigation: A recurrent neural network based approach," in *Proc. 2021 IEEE Int. Conf. Robot. Automat.*, 2021, pp. 3992–3998.

Shiyu Bai (Member, IEEE) received the PhD degree in navigation, guidance and control from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2022. He is currently a postdoctoral fellow with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. His research interests include inertial navigation, multi-sensor fusion, indoor positioning, and vehicular positioning.

Weisong Wen (Member, IEEE) received the BEng degree in mechanical engineering from Beijing Information Science and Technology University (BISTU), Beijing, China, in 2015, the MEng degree in mechanical engineering from China Agricultural University, in 2017, and the PhD degree in mechanical engineering from The Hong Kong Polytechnic University (PolyU), in 2020. He was also a visiting PhD student with the Faculty of Engineering, University of California, Berkeley (UC Berkeley), in 2018. Before joining PolyU as an assistant professor, in 2023, he was a research assistant professor with AAE of PolyU since 2021. His research interests include the trustworthy multi-sensory integration, LiDAR SLAM, GNSS positioning and autonomous systems.

Dongzhe Su received the bachelor's degree from the Huazhong University of Science and Technology, and the MPhil degree in computer science from The Hong Kong University of Science and Technology. He is now working toward the PhD degree with the Department of Aeronautical and Aviation Engineering, Hong Kong Polytechnic University. He is now director of Smart Mobility Technologies with the Communication Technologies Group, Hong Kong Applied Science and Technology Research Institute (ASTRI). He has been leading the system architecture in research and development of vehicle-to-everything (V2X) communication and application systems, connected autonomous vehicles (CAV) systems. His role has been to define the technical scope and overall system design for ASTRI's V2X&CAV system targeting various application scenarios.

Li-Ta Hsu (Senior Member, IEEE) received the BS and PhD degrees in aeronautics and astronautics from National Cheng Kung University, Taiwan, in 2007 and 2013, respectively. He is currently an associate professor with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, before he served as a postdoctoral researcher with the Institute of Industrial Science, University of Tokyo, Japan. In 2012, he was a visiting scholar with University College London, U.K. His research interests include GNSS positioning in challenging environments and localization for pedestrian, autonomous driving vehicle, and unmanned aerial vehicle.