A Lightweight Indoor Localization Method Integrating Keyframe Recognition and Inertial Navigation

Chenzhe Wang [1], Kai Bi [1*] , Shu Peng [1] , Jie Zhu [1] , Zhaolong Li [2], Han Liu [2], Shuaiyi Shi [3], Yujia Chen [4], Shiliang Tao [5]

[1] National Geomatics Center of China, Beijing, China - wangchenzhe@ngcc.cn, 24677958@qq.com, pengshu@ngcc.cn, zhujie@ngcc.cn
[2] Beijing University of Civil Engineering and Architecture, China - lzlfofficial@163.com, 331179177@qq.com
[3] Moganshan Geospatial Information Laboratory, Huzhou, China - shisy1@mgslab.com
[4] North China University of Science and Technology, Tangshan, China - chenyujia@ncst.edu.cn
[5] Tencent, Beijing, China - 505165517@qq.com

Keywords: Indoor Positioning System, Inertial Navigation, Key Frame Classification, Feature Point Recognition, XFeat, MobileNet V3-Small.

Abstract

Most traditional indoor localization schemes necessitate the pre-installation of hardware devices, leading to uncontrollable costs and the requirement for ongoing maintenance. While pure vision-based localization solutions offer the advantages of low cost and deployment-free implementation, they still encounter two major technical bottlenecks. Firstly, vision-based systems relying solely on point cloud data impose substantial computational burdens, which creates difficulties in meeting the real-time performance requirements of mobile terminals. This challenge stems partly from the intensive operations required for point cloud processing—including feature extraction and spatial alignment—whose complexity often exceeds the hardware capabilities of portable devices designed for energy efficiency. Secondly, image matching schemes based on key frames are prone to position jumps, particularly in dynamic scenes or areas with insufficient features. To address the aforementioned constraints, this paper proposes a lightweight indoor positioning framework that establishes tight coupling between visual data and inertial measurements. This framework is structured into three sequential phases: data preprocessing, real-time visual localization computation, and fused positioning result output. During the data preprocessing phase, image data covering the entire indoor scene is acquired, and representative key frames are selected to train a key frame recognizer—with the aim of reducing redundant information and improving subsequent matching efficiency. Concurrently, feature point descriptors of these selected key frames are extracted and organized to construct a structured environmental feature information database. In the real-time visual localization computation phase, based on externally input real-time video streams, relatively precise position estimation is achieved through key frame matching and feature point correspondence, by leveraging the pre-established environmental feature database to accelerate the matching process. Finally, in the fused localization result output stage, based on visual localization, the system integrates data from an Inertial Measurement Unit (IMU) to construct a position estimation framework using the Extended Kalman Filter, outputting smooth and continuous precise positions. Compared with conventional vision-based solutions, this system optimizes the motion trajectory through the recursive propagation of inertial data under the constraints of visual features, thereby significantly enhancing the spatiotemporal continuity of the localization results while maintaining the accuracy of visual localization.

1. Introduction

As the foundational infrastructure of the internet of things (IoT) era, indoor localization technologies hold substantial application potential in smart venues, transportation hubs, and similar environments. At present, the dominant indoor positioning technologies mainly include Wi-Fi, Bluetooth beacons, pedestrian dead reckoning (PDR), geomagnetic positioning, and ultra-wideband (UWB). Alongside the continuous evolution of image recognition algorithms and the advancement of deep learning frameworks, vision-based indoor localization technology has undergone gradual maturation, emerging as an increasingly competitive option in the field of indoor positioning. Although contemporary vision-based localization paradigms offer advantages in cost-effectiveness and ease of deployment, they encounter two critical technical limitations: (1) the high computational demands of pure point-cloud-based vision systems render real-time performance unsustainable for mobile terminals, and (2) keyframe-dependent image matching schemes are prone to positional discontinuities, particularly in dynamic environments or feature-deficient regions.

To mitigate the aforementioned challenges, this study puts forward a lightweight visual-inertial integrated framework characterized by tight coupling between visual data and inertial measurements. This methodology synergizes feature-based keyframe matching with inertial measurement unit (IMU) data through an extended Kalman filter (EKF)-based estimation architecture. By fusing visual feature constraints with inertial data propagation, the system achieves trajectory optimization that preserves the precision of vision-aided localization while substantially enhancing the spatiotemporal continuity of the positioning outputs. This hybrid approach provides a robust solution for resource-constrained platforms that require continuous, high-accuracy indoor positioning.

2. System Process Overview

The proposed lightweight indoor localization system that integrates keyframe recognition and inertial navigation, is illustrated in Fig. 1. As shown in the workflow diagram, the

framework is structurally partitioned into three modular components: (1) a vision-based localization module, (2) an inertial navigation module, and (3) a fusion engine for optimized pose estimation. The specific implementation procedures of the proposed framework are outlined below.

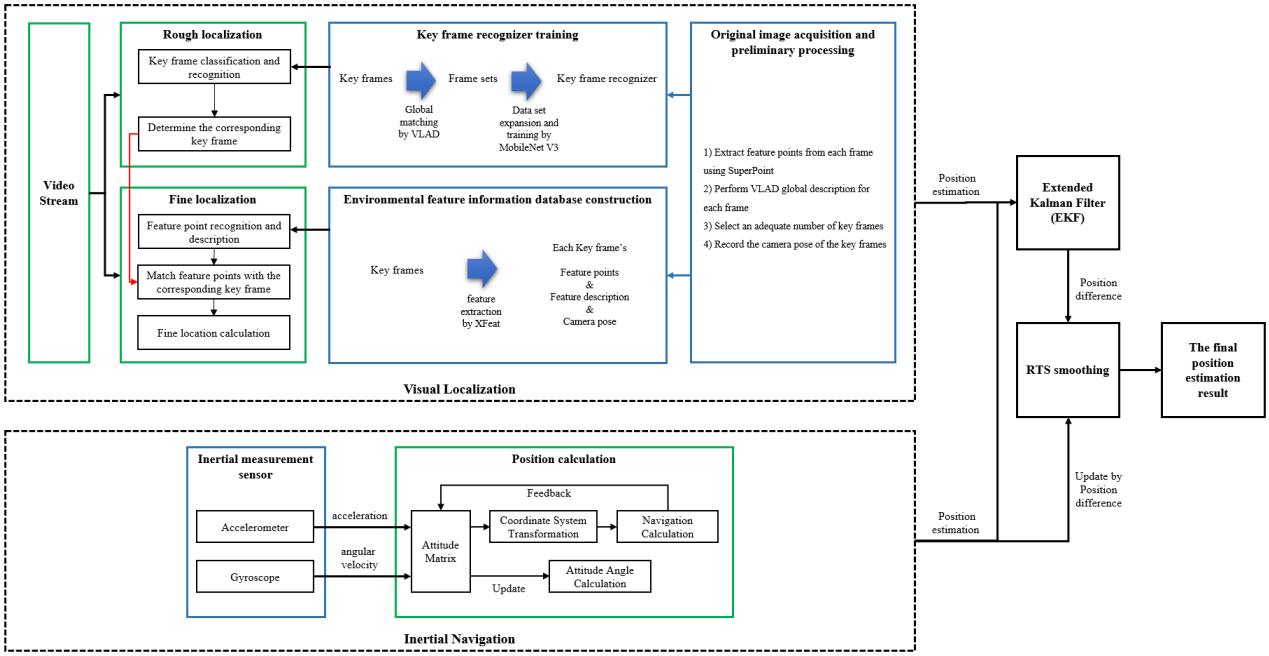

Figure 1. Workflow Diagram of the Indoor Positioning System Integrating Keyframe Recognition and Inertial Navigation Technology.

2.1 Visual Localization

2.1.1 Data Preprocessing: The environmental feature database is constructed through the following procedure. First, video sequences are repeatedly captured (3-4 times per position) using video acquisition equipment. Each video is subsequently frame-decomposed to extract the XFeat features and VLAD global descriptors. From these, N keyframes with salient feature representations are uniformly selected, and the corresponding camera pose parameters are recorded. Global feature matching is performed for each keyframe to establish similar image sets. The training dataset is augmented using geometric (rotation and affine transformation) and photometric (grayscale conversion and filtering) transformations. Leveraging transfer learning, a pre-trained MobileNet V3-Small architecture is fine-tuned to develop a keyframe recognition model. Finally, all XFeat descriptors are connected to their spatial pose information to establish an integrated environmental feature-pose-mapping database.

2.1.2 Real-time Positioning: Real-time video streams are ingested into a pre-trained keyframe recognizer to obtain the coarse positional region and corresponding reference keyframe. Subsequently, the XFeat descriptors are extracted from the query frame and matched against the environmental feature database. Pose refinement is achieved through affine transformation parameter estimation and homography matrix decomposition, enabling centimeter-level localization accuracy.

2.2 Inertial Navigation

Localization was achieved by integrating tri-axial accelerometers and gyroscopes within an inertial measurement unit (IMU). The linear acceleration components along each axis are double-integrated to compute the real-time position and velocity vectors of the moving platform. Gyroscopes measure either vibrational phase differentials (as in Coriolis-effect gyroscopes) or angular velocities, which are integrated using rotation vector algorithms to resolve the three-dimensional attitude angles (pitch, roll, and yaw) within the navigation coordinate frame.

2.3 Fusion-based Localization Output

The proposed tightly coupled vision-inertial architecture based on an extended Kalman filter (EKF) achieves enhanced localization performance through mutual constraint mechanisms. The inertial navigation system (INS) serves as the primary kinematic constraint, leveraging its high-frequency output capability to bridge perceptual gaps in vision-based localization and ensure continuous pose estimation during visual feature degradation. Concurrently, the vision-based localization module mitigates INS drift accumulation through environmental feature re-localization, establishing a bidirectional constraint mechanism. This synergistic framework ultimately yields accurate and continuous positioning outputs by combining the complementary strengths of both sensory modalities.

3. Detailed Elaboration of Key Technologies for this study

3.1 Development of the Key Frame Recognizer

3.1.1 Lightweight Feature Point Detection: Traditional feature point extraction approaches that do not rely on convolutional neural networks—including SURF and ORB—are predominantly dependent on static algorithmic frameworks for feature point extraction, thereby providing minimal scope for adaptive adjustments to varying environmental conditions or application-specific requirements. Consequently, each algorithm can only excel in its specific domain of expertise and fails to universally accommodate all scenarios. Additionally, these algorithms exhibit high computational complexity, resulting in suboptimal performance on mobile devices with average computational capabilities. In order to achieve superior performance, this paper employs Xfeat as the feature point recognition algorithm.

XFeat adopts a novel convolutional neural network architecture, utilizing meticulously designed strategies for keypoint detection and local feature extraction, aiming to minimize computational overhead while maintaining robustness and accuracy. Furthermore, XFeat is applicable to both sparse feature matching based on keypoints and dense matching of coarse feature maps. In comparison with other image matching approaches, XFeat achieves a superior balance between matching precision and computational efficiency. It outperforms the vast majority of lightweight deep learning-based local feature methods in terms of speed, while simultaneously attaining accuracy comparable to that of larger-scale models such as SuperPoint

The network architecture of XFeat comprises three major modules: a lightweight backbone network, a dual-branch feature extractor (for keypoint detection and descriptor generation), and a semi-dense matching refinement module, as illustrated in Figure 2.

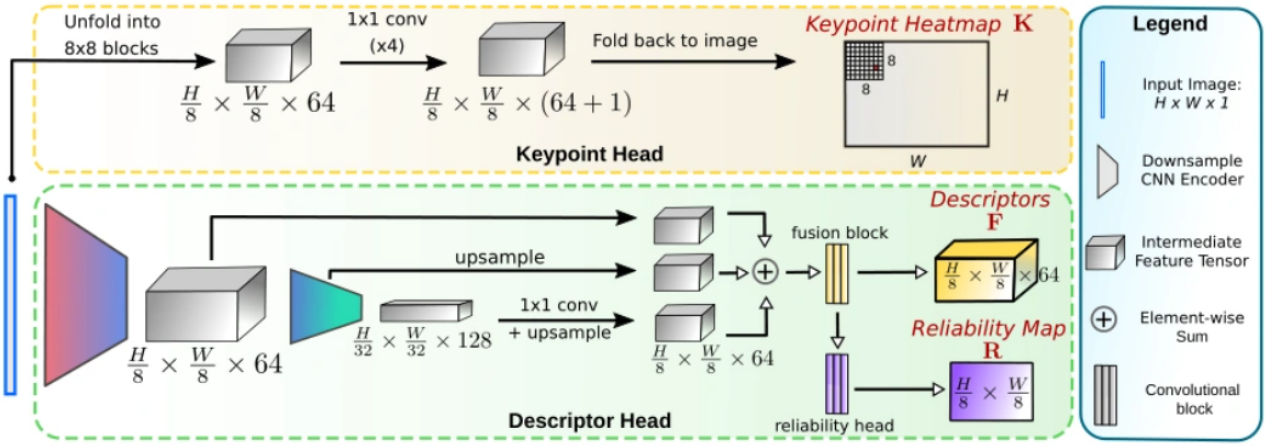

Figure 2. Schematic illustration of the architectural structure of the XFeat network.

1) Featherweight Backbone: A channel optimization strategy is employed, wherein the initial layer utilizes only 4 channels. As the spatial resolution decreases (with a stride of 2 for each subsequent layer), the number of channels is progressively increased up to 128. The formalized computational cost is expressed as:

$$F_{ops} = H_i \cdot W_i \cdot C_i \cdot C_{i+1} \cdot k^2 \quad (1)$$

Where H_i and W_i represent the spatial resolution, C_i denotes the number of channels, and k is the size of the convolutional kernel.

By reducing the number of channels at an early stage, the computational load is significantly decreased. Feature maps at three scales, specifically 1/8, 1/16, and 1/32, are fused by upsampling them to a 1/8 resolution through bilinear interpolation, followed by summation. This process enhances the robustness against variations in viewpoint.

2) Dual-branch feature extractor: The keypoint detection branch of XFeat operates independently from the descriptor branch, thereby avoiding mutual interference during joint training. It partitions the input image into an 8x8 grid and regresses the coordinates of keypoints within each grid through 1x1 convolution, enabling sub-pixel level localization. Additionally, a "dustbin" classification mechanism is introduced to filter out invalid regions.

The descriptor generation branch of XFeat outputs a 64-dimensional dense descriptor map (Dense Descriptor Map), which is combined with a reliability map (Reliability Map) to select high-confidence features. The reliability map, regressed through convolutional blocks, represents the unconditional probability of successful feature point matching.

3) Semi-dense matching refinement module: The matching strategy is divided into two modes. The sparse mode involves extracting 4,096 high-confidence keypoints and rapidly matching them through mutual nearest neighbor search (MNN). The semi-dense mode entails extracting 10,000 feature regions and utilizing a lightweight Multi-Layer Perceptron (MLP) to predict pixel-level offsets, thereby achieving sub-pixel matching.

4) Training strategy and loss function: The model is trained using a mixed dataset of Megadepth and COCO, with a 6:4 ratio to balance real-world scenes and synthetic deformation data. Pixel-level correspondences are employed to supervise the learning of feature descriptors and keypoint locations. By leveraging the Dual-Softmax Loss, the similarity of matched feature pairs is maximized:

$$L_{ds} = \sum_i \log(softmax_r(S)_{ii}) - \log(softmax_r(S^T)_{ii}) \quad (2)$$

3.1.2 Global Feature Representation and Matching Mechanisms: For the training of the key-frame recognizer, it is necessary to extract frame subsets with high similarity that match each key frame from the raw image dataset, which serves as the training data. Simple similarity retrieval based on elementary feature points is insufficient; instead, the features of each frame need to be aggregated into a consolidated global descriptor vector. In this study, the VLAD algorithm was utilized for both the generation of global descriptors and the retrieval of similar frames, ensuring the effectiveness of the training data construction process.

The VLAD algorithm is predicated on the assumption that each image frame encompasses local feature points with a dimensionality of $N \times D$(where N may be substantial and varies per image). The objective is to derive a compact $K \times D$ -dimensional global descriptor (with K being a predefined value, e.g., 128) from these $N \times D$ -dimensional features. The core procedural steps are outlined as follows:

1) K-means clustering is applied to all $N \times D$-dimensional local features to derive K cluster centers, designated as C_k.

2) The $N \times D$-dimensional local features are transformed into a global feature V, which exhibits a feature vector dimensionality of $K \times D$ —where k ranges over K and j over D . The corresponding formula is given as follows:

$$V(j,k) = \sum_{i=1}^{N} a_k\,(x_i)(x_i(j) - c_k(j)) \qquad (3)$$

In this formula, x_i denotes the i-th local feature extracted from the image, and c_k represents the k-th cluster center—with both x_i and c_k being D -dimensional vectors. Meanwhile, $a_k(x_i)$ denotes a binary indicator function: it takes a value of 1 if and only if x_i is assigned to the cluster center c_k, and 0 in all other cases.

Ultimately, this process yields a dimensionally reduced global descriptor for the image. Through the application of a suitable Euclidean distance threshold, all images bearing similarity to the key frame can be retrieved with high efficiency.

3.1.3 Key Frame Identification Model: The central system module of this research was implemented on mobile devices, with computational efficiency emerging as a pivotal factor given the algorithm's intensive processing requirements. Consequently, MobileNetV3-Small was adopted as the backbone network architecture. Through architectural optimizations, MobileNetV3 achieves superior accuracy compared to most large-scale neural networks while maintaining significantly fewer parameters and lower computational overhead. The latest iteration, MobileNetV3-Small, processes images in just 22 ms-substantially faster than conventional deep networks. MobileNet V3-Small comprises 12 distinct Bneck layers, one standard convolutional layer, and two pointwise convolutional layers. Notably, it exhibits the following attributes.

1) Depthwise Separable Convolution (MobileNetV1):
Employing depthwise separable convolution instead of standard convolution reduces both parameter count and computational requirements by approximately 90% while preserving comparable model accuracy.

2) Inverted Residual with Linear Bottleneck (MobileNetV2):
This enhanced architecture further decreases parameter size and computational costs by 30-50% compared to standard convolution through its innovative spatial-channel optimization approach.

3) Squeeze-and-Excitation Attention Modules:
The integrated lightweight attention mechanism dynamically recalibrates channel-wise feature responses, amplifying relevant features while suppressing redundant ones through learned channel interdependencies.

4) h-swish Activation Function:
Implementation of the h-swish activation function, as validated by Google AI research, demonstrates approximately 15% computational efficiency gains while maintaining numerical stability in mobile-optimized networks.

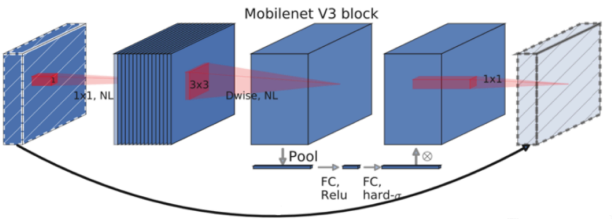

Figure 3. Distinctive Bneck Structure in MobileNet V3.

3.1.4 Transfer Learning for Key Frame Identification: In the current research, the parameters of the pre-trained model were derived via the transfer of well-calibrated weights from MobileNet V3, which had undergone prior training on the ImageNet dataset—this transfer strategy leverages the generic feature extraction capabilities already encoded in the model through large-scale image learning. To further adapt the model to the specific task of keyframe recognition, a fine-tuning process was implemented: specific layers of the pre-trained model were selected for retraining to capture task-specific patterns, while the remaining layers were kept frozen to preserve the pre-learned generic features that are critical for robust visual representation. Through this targeted adjustment, the intended keyframe recognizer, optimized for the indoor positioning scenario, was ultimately obtained.

3.2 Development of the Environmental Feature Data Database

By using the XFeat algorithm, feature points in all keyframe images are identified, and their corresponding feature descriptors are computed. Subsequently, by integrating the positional information of the keyframe images along with the camera's pose information at the time of capturing each keyframe, an environmental feature information database can be constructed.

3.3 Vision-based Precise Position Determination

3.3.1 By utilizing the key-frame recognizer, the key-frame corresponding to the current location is identified. Following this, XFeat is utilized to extract and characterize feature points from the video stream. Next, the feature points within the current video stream are matched against those of the respective key frame. In the present study, stable corresponding point pairs were first screened from the initially matched corresponding point pairs to eliminate spurious matches, and the homography matrix $\boldsymbol{H}$ (formulated as Equation (4)) was subsequently computed based on these reliable point correspondences. To derive the camera's pose parameters, singular value decomposition (SVD)—a classic numerical method for resolving pose parameters—was employed to decompose the homography matrix, through which the rotation matrix and translation matrix were acquired, followed by the calculation of the camera attitude angle $\theta_{attitude}$.

Further, the set of stable feature points was subjected to random perturbation to simulate potential variations in real-world imaging scenarios, and the positional differences between the perturbed and original feature points were quantified. Inverse-weighted coefficients were assigned to each feature point according to the magnitude of these differences—with smaller differences corresponding to higher weights—to suppress the impact of abnormal values. On this basis, the weighted mean value of the positional distance differences $\delta_{distance}$ between feature points in the template image and those in the real-time video stream image was computed.

By integrating key imaging parameters and pre-calibrated data, including the known camera focal length f, the previously derived camera attitude angle $\theta_{attitude}$, the position coordinates P_0 of the template image, and the shooting distance z_0 of the template image, the spatial distance between the mobile device and the current template image was calculated. Through this multi-step optimization and integration of geometric constraints and feature matching results, the high-precision position coordinates P (detailed in Equation (5)) of the mobile device in the indoor environment were ultimately determined.

$$\boldsymbol{H} = \boldsymbol{K}(\boldsymbol{R} + \boldsymbol{T}\frac{1}{d}\boldsymbol{N}^T)\boldsymbol{K}^{-1} \tag{4}$$

Herein, $\boldsymbol{K}$ denotes the intrinsic parameter matrix of the camera—a core component that encapsulates the optical and geometric properties (e.g., focal length, principal point coordinates) of the imaging system in the indoor positioning scenario. d stands for the inter-center distance, which is relevant to the spatial configuration of the imaging setup and critical for accurate coordinate mapping. $\boldsymbol{N}$ represents the normal vector of the camera's image plane, a parameter that characterizes the orientation of the imaging plane relative to the 3D spatial coordinate system of the indoor environment. $\boldsymbol{R}$ refers to the extrinsic rotation matrix of the camera, responsible for describing the rotational posture of the camera relative to the global coordinate system established for indoor localization. $\boldsymbol{T}$ denotes the extrinsic translation vector of the camera, quantifying the linear displacement of the camera relative to the origin of the global coordinate system and enabling the conversion between 2D image coordinates and 3D spatial coordinates.

$$\boldsymbol{P} = \boldsymbol{P_0} + [(\delta_{distance} f + z_0)sin\theta_{attitude} \quad (\delta_{distance} f + z_0)cos\theta_{attitude}]^T \tag{5}$$

3.4 Inertial Navigation Algorithm

Inertial Navigation System (INS) is grounded in Newtonian mechanics, utilizing inertial sensors to measure the angular velocity and linear acceleration of a vehicle's motion. Through real-time computation by a computer, it derives navigation information such as the vehicle's three-dimensional attitude, velocity, and position. INS primarily comprises two components: the inertial measurement sensor unit and the computational processing unit.

3.4.1 Inertial Measurement Unit (IMU): Inertial navigation and positioning technology employs accelerometers and gyroscopes to quantify the linear acceleration and angular velocity of a moving object, respectively, and derives the object's position and attitude through integral computations. The accelerometer determines linear acceleration by measuring the inertial force generated by the object, often employing micro-electro-mechanical system (MEMS) technology. The gyroscope, on the other hand, measures the angular velocity by detecting the rotational speed of the object as it rotates about a certain axis, capable of providing information regarding the rotational rates of the object around three orthogonal directions.

3.4.2 Quaternion-based Attitude Update: The Inertial Navigation System (INS) necessitates the transformation between the body-fixed coordinate system and the navigation coordinate system, which can be achieved by following a specific sequence of rotations based on the attitude angles. Accurate attitude determination is crucial for the positioning model. Generally, there are three commonly used methods to represent attitude, namely the Euler angle method, the direction cosine matrix (DCM) method, and the quaternion method. This paper adopts the quaternion method for attitude updating. The formula is as follows:

$$q_{b,k}^n = q_{b,k-1}^n \circ \left[\cos\|0.5\Delta\theta_k\| \quad \frac{(\Delta\theta_k)^T}{\|\Delta\theta_k\|}\sin\|0.5\Delta\theta_k\|\right]^T \tag{6}$$

Where $\circ$ denotes quaternion multiplication; $\Delta\theta_k$ represents the angular increment output by the gyroscope from time instant $k-1$ to time instant k; $\Delta\theta_k \approx \omega_k^b \Delta t$, ω_k^b is the angular velocity at a certain time instant; and Δt is the sampling interval. The corresponding rotation matrix can be obtained by transforming the quaternion $q_b^n = [q_1 \ q_2 \ q_3 \ q_4]^T$.

$$C_b^n = \begin{bmatrix} q_1^2+q_2^2-q_3^2-q_4^2 & 2(q_2q_3-q_1q_4) & 2(q_2q_4+q_1q_3) \\ 2(q_2q_3+q_1q_4) & q_1^2-q_2^2+q_3^2-q_4^2 & 2(q_3q_4-q_1q_2) \\ 2(q_2q_4-q_1q_3) & 2(q_3q_4+q_1q_2) & q_1^2-q_2^2-q_3^2+q_4^2 \end{bmatrix} \tag{7}$$

By leveraging such attitude information, the velocity update equation and the position update equation may be formulated as follows:

$$v_k^n = v_{k-1}^n + C_{b,k}^n\left(\frac{\Delta v_k^b + (\Delta\theta_k^b \times \Delta v_k^b)}{2}\right) - g^n\Delta t \tag{8}$$

$$p_k^n = v_{k-1}^n\Delta t + C_{b,k}^n\left(\Delta v_k^b + \frac{\Delta\theta_k^b \times \Delta v_k^b}{2}\right)\Delta t - 0.5g^n\Delta t^2 \tag{9}$$

$$\Delta\theta_k^b = \left(\omega_k^b - b_g\right)\Delta t \tag{10}$$

$$\Delta v_k^b = \left(f_k^b - b_f\right)\Delta t \tag{11}$$

3.5 Fusion Localization Algorithm Based on Extended Kalman Filter

The Extended Kalman Filter (EKF) is primarily employed to address the state estimation problem in nonlinear systems. The localization process of the Pedestrian Dead Reckoning (PDR) technology is nonlinear and less susceptible to indoor environmental influences; however, it suffers from accumulated errors, leading to poor localization stability and making it unsuitable for standalone use over extended periods. In contrast, visual localization technology is easily affected by indoor environmental factors but does not accumulate errors, thus maintaining stable localization accuracy during prolonged use. This paper adopts the Extended Kalman Filter to fuse these two localization technologies, utilizing the localization results from visual localization as observed values and those from PDR localization as state estimates to construct a localization system that achieves the integration of both technologies.

The state vector and observation vector of the positioning system are herein defined as follows:

$$\begin{cases} X_k = [x_k, y_k, \varphi_k]^T \\ Z_k = [\dot{x}_k, \dot{y}_k]^T \end{cases} \tag{12}$$

Where x_k and y_k denote the predicted position coordinates at the k-th step, φ_k represents the predicted pedestrian heading angle at the k-th step, $\dot{x}_k$ and $\dot{y}_k$ indicate the localization coordinates obtained from visual localization at time instant k.

The state equation for the positioning system is formulated as follows:

$$X_k = \begin{bmatrix} x_k \\ y_k \\ \varphi_k \end{bmatrix} = \begin{bmatrix} x_{k-1} + s_{k-1} \cdot \sin\varphi_{k-1} \\ y_{k-1} + s_{k-1} \cdot \cos\varphi_{k-1} \\ \varphi_{k-1} + \Delta\varphi \end{bmatrix} + W \tag{13}$$

The observation equation is given by:

$$Z_k = \begin{bmatrix} \dot{x}_k \\ \dot{y}_k \end{bmatrix} + V \tag{14}$$

Where W represents the additive Gaussian white noise vector associated with the system's state equation, and V denotes the Gaussian white noise vector of the system observation equation, with both being mutually independent. x_{k-1} and y_{k-1} signify the fused localization coordinates obtained at time instant k−1, s_{k-1} indicates the step length determined at time instant k−1, φ_{k-1} represents the pedestrian heading angle acquired at time instant k−1, and $\Delta\varphi$ denotes the anticipated increment of the pedestrian heading angle. By performing a Taylor expansion on the nonlinear portion, the state transition matrix A_k can be derived as follows:

$$A_k = \begin{bmatrix} 1 & 0 & s_{k-1} \cdot \sin\varphi_{k-1} \\ 0 & 1 & s_{k-1} \cdot \cos\varphi_{k-1} \\ 0 & 0 & 1 \end{bmatrix} \tag{15}$$

The observation matrix H_k is given by:

$$H_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{16}$$

The initial covariance matrix P_1 is given by:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{17}$$

The system process covariance noise matrix Q and the observation noise covariance matrix R are defined as follows:

$$Q = \begin{bmatrix} \delta_x^2 & 0 & 0 \\ 0 & \delta_y^2 & 0 \\ 0 & 0 & \delta_\varphi^2 \end{bmatrix} \tag{18}$$

$$R = \begin{bmatrix} \delta_x^2 & 0 \\ 0 & \delta_y^2 \end{bmatrix} \tag{19}$$

Where, δ_x^2 and δ_y^2 respectively denote the localization variances along the X-axis and Y-axis for both Pedestrian Dead Reckoning (PDR) localization and visual localization, while δ_φ^2 represents the variance of the heading angle.

Following the complete setup of all the aforementioned initial conditions, iterative computations are conducted to derive the localization outcomes of the fusion localization algorithm that is based on the Extended Kalman Filter.

4. Experiment and Results

To rigorously validate the proposed indoor localization technology, a systematic experimental protocol was designed to evaluate its functional feasibility and positioning accuracy. The experimental workflow comprises software deployment, localization testing, and quantitative analysis. Furthermore, the experiment not only assessed the positioning performance of the system proposed in this study but also compared its accuracy against that of traditional WiFi positioning and standalone PDR positioning. All tests were conducted under identical environmental conditions, with strict control over factors such as indoor signal interference and pedestrian movement speed to ensure fair comparison.

4.1 Experimental Hardware Environment

To validate the feasibility and accuracy of the indoor positioning system that integrates visual and inertial navigation proposed in this paper, corresponding experiments were designed. The positioning terminal employed in the experiments was a mobile phone (model: Samsung Galaxy S23), which is equipped with both an IMU (Inertial Measurement Unit) module and a camera, thus meeting the hardware requirements of the proposed system.

4.2 Image Data Acquisition and Preprocessing

Environmental images of the experimental site were uniformly collected, and XFeat was employed to identify and describe feature points from all the acquired environmental image data. Frames with prominent features were selected from all the environmental images as key frames, ensuring a relatively uniform distribution of these key frames across the experimental site. The camera pose data was documented at the moment each key frame was captured. Ultimately, the development of a key-frame identifier and a database containing environmental feature information was undertaken.

4.3 Analysis of Experimental Results

Dynamic localization experiments were performed at the designated test site. For three different localization schemes (the

system proposed in this paper, traditional WiFi localization, and single PDR localization), identical routes were tested at least 30 times. The average deviation of each scheme from the true trajectory was ultimately calculated to obtain the average error for each approach. Traditional WiFi localization exhibited an average error of approximately 1.5 meters, while the single PDR scheme yielded an average error of around 2.3 meters. In contrast, the scheme proposed in this paper managed to control the average error within approximately 0.1 meters, and was capable of outputting localization results for at least 5 frames per second. The experimental data fully demonstrated the feasibility, high efficiency, and high precision of the scheme proposed in this study.

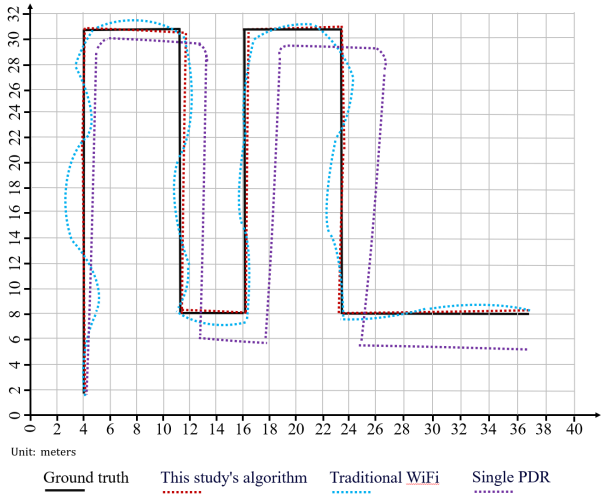

Figure 3. Comparison between real track and test track..

5. Conclusion

This paper presents a lightweight localization scheme that tightly couples visual and inertial information. Based on feature-based key-frame matching, this scheme integrates data from an Inertial Measurement Unit (IMU) to construct position estimation using an Extended Kalman Filter. While maintaining the accuracy of visual localization, it significantly enhances the spatiotemporal continuity of the localization results.

Experimental findings demonstrate that the localization error associated with the proposed approach maintains a consistent range of approximately 0.1 meters. Compared to traditional indoor localization schemes, this approach offers multiple advantages, including the elimination of the need for pre-installed hardware, higher accuracy, lower computational overhead, and improved localization continuity. Further in-depth research will be conducted in this paper with the aim of further enhancing the accuracy and stability of localization.

Acknowledgements

This work is supported by the research and development project on interactive decision-making and management technology for urban sustainable development (No. 2022YFC3802904).

References

Guilherme, P., Felipe, C., Andre, A., Renato, M., Erickson, R. N., 2024. XFeat: Accelerated Features for Lightweight Image Matching. *CVPR 2024.*, arXiv:2404.19174.

Yao, H., Wang, X., Qi, H., and Liang, X., 2022: Tightly coupled indoor positioning using uwb/mmwave radar/imu, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVI-3/W1-2022, 323–329, https://doi.org/10.5194/isprs-archives-XLVI-3-W1-2022-323-2022.

Wang, C., Bi, K., Zhao, B., Li, M., Chen, Y., Tao, S., and Yang, J., 2024: Lightweight Indoor Positioning System Based on Multiple Self-Learning Features and Key Frame Classification, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-4-2024, 373–379, https://doi.org/10.5194/isprs-annals-X-4-2024-373-2024, 2024.

Zhou, B., Wu, Z., Chen, Z., Liu, X., and Li, Q., 2023: Wi-Fi RTT/Encoder/INS-Based Robot Indoor Localization Using Smartphones, *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6683–6694, May 2023, doi: 10.1109/TVT.2023.3234283.

Mansour, A., Chen, W., Weng, D., Yang, Y., and Wang, J., 2023: Leveraging human mobility and pervasive smartphone measurements-based crowdsourcing for developing self-deployable and ubiquitous indoor positioning systems, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W2-2023, 1119–1125, https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1119-2023.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010: Aggregating local descriptors into a compact image representation, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, 10.1109/CVPR.2010.5540039.

Howard, A., Sandler, M., Chu, G., 2019: Searching for mobilenetv3, *Proce edgings of the IEEE/CVF International Conference on Computer Vision.*, 2019: 1314-1324.