TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 11/14 pp552–565
DOI: 10.26599/TST.2022.9010029
Volume 28, Number 3, June 2023

From Coarse to Fine: Two-Stage Indoor Localization with Multisensor Fusion

Li Zhang*, Jinhui Bao, Yi Xu, Qiuyu Wang, Jingao Xu, and Danyang Li

Abstract: Increasing attention has been paid to high-precision indoor localization in dense urban and indoor environments. Previous studies have shown single indoor localization methods based on WiFi fingerprints, surveillance cameras or Pedestrian Dead Reckoning (PDR) are restricted by low accuracy, limited tracking region, and accumulative error, etc., and some defects can be resolved with more labor costs or special scenes. However, requesting more additional information and extra user constraints is costly and rarely applicable. In this paper, a two-stage indoor localization system is presented, integrating WiFi fingerprints, the vision of surveillance cameras, and PDR (the system abbreviated as iWVP). A coarse location using WiFi fingerprints is done advanced, and then an accurate location by fusing data from surveillance cameras and the IMU sensors is obtained. iWVP uses a matching algorithm based on motion sequences to confirm the identity of pedestrians, enhancing output accuracy and avoiding corresponding drawbacks of each subsystem. The experimental results show that the iWVP achieves high accuracy with an average position error of 4.61 cm, which can effectively track pedestrians in multiple regions in complex and dynamic indoor environments.

Key words: indoor localization; WiFi fingerprints; computer vision; Pedestrian Dead Reckoning (PDR)

1 Introduction

Current navigation systems, such as Global Positioning System (GPS) or GLObal NAvigation Satellite System (GLONASS), can provide high accuracy in outdoor environments but extremely limited accuracy in indoor environments[1]. Therefore, accurate indoor localization and tracking have become a hot spot and spawned a series of applications, such as intelligent advertisement, customer navigation, and augmented reality. Increasing interests have been paid in indoor positioning and tracking technologies, including the use of wireless signals, cameras, Inertial Measurement Units (IMU), etc., as witnessed by recent publications of plentiful research[2–15].

Appealing WiFi-based indoor positioning[16–18] requires extensive, critical, and challenging pre-deployment efforts, but its positioning accuracy is limited. Pedestrian Dead Reckoning (PDR) realized by IMU is one of the most popular solutions due to the prevalence of smartphones. However, PDR suffers from accumulative errors seriously, leading to non-negligible deviation[19]. Computer vision is a promising solution for indoor localization[20–24]. Fusion indoor localization methods are presented to eliminate the above shortcomings. The widely installed indoor surveillance cameras are combined with smartphone IMU data to achieve indoor localization with high accuracy[25–27]. Furthermore, WiFi and PDR integrated with an extended Kalman filter are presented to achieve high-ranking accuracy[28]. WiFi, vision, and PDR are fused by a particle filter to locate pedestrians, known as iVR[29]. Nevertheless, most of the processes only locate a

• Li Zhang, Jinhui Bao, Yi Xu, and Qiuyu Wang are with the School of Mathematics, Hefei University of Technology, Hefei 230009, China. E-mail: lizhang@hfut.edu.cn; 2020111411@mail.hfut.edu.cn; yixu@mail.hfut.edu.cn; 2020111428@mail.hfut.edu.cn.
• Jingao Xu and Danyang Li are with the School of Software and BNRist, Tsinghua University, Beijing 100084, China. E-mail: xujingao13@gmail.com; lidanyang1919@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2022-03-19; revised: 2022-06-29; accepted: 2022-08-08

limited range and require special hardware, especially in complex indoor environments which are prone to pedestrian identity mismatch problems. Therefore, a three-in-one combination method (indoor surveillance cameras, smartphone, and WiFi) is desirable to obtain efficient and high-accuracy indoor localization, thus achieving higher accuracy.

However, it is not easy to transform the above idea into a practical system. It faces four important challenges:

(1) Incorrespondence of identification. The user ID provided by vision-based methods cannot be directly associated with IMU-based methods. This association is a prerequisite to integrate multimodal data.

(2) Frequency Line Of Sight (LOS) blockage. Frequently blocked pedestrian makes it impossible to locate the target. The LOS blockages decrease the efficiency of detection and tracking.

(3) Multiple types of sensor data fusion. Traces are directly generated by individual systems independently, and then aligned to distinguish the user's WiFi and obtain a fused trajectory. But all the data are from different types of devices. Fusing vision data and IMU data becomes a key point.

(4) The coordinate transformation model. The pixel coordinates of pedestrians are obtained through the fusion of multiple types of sensor data. Since most of the results are not useful, the transforming model from pixel coordinate to world coordinate needs to be obtained, and then the world coordinate of pedestrians is calculated according to the surveillance camera's internal and external parameters.

To tackle the above challenges, an indoor localization and tracking system is presented to achieve localization and tracking of pedestrians, integrating WiFi fingerprints, the vision of surveillance cameras, and PDR (the system abbreviated as iWVP). Firstly, the client sends Received Signal Strength Indicator (RSSI) data to the server, and then the system uses Bayes filter to determine a region and invokes the surveillance camera to get pedestrian positions and tracks. The video pedestrian sequence and the PDR sequence are associated to confirm the user's ID after the client sends IMU data to the server in the system.

iWVP is tested on the Windows server and many commercial smartphones. Extensive experiments are conducted in multi-story buildings. Pedestrians are localized and tracked in complex indoor environments. Evaluations demonstrate that iWVP reaches better precision and performs well in terms of detecting and tracking than former algorithms in complex indoor environments. Furthermore, iWVP is evaluated through the transforming model, and the average positioning error is about 4.61 cm.

The main contributions are summarized as follows:

- We use Bayes filter to lock pedestrians' range regions in multiple regions and use surveillance cameras to locate and track pedestrians, associating PDR and vision trace to confirm user ID.
- We design a real-time and reliable pedestrian tracking system by combining PDR with visual tracking, which makes up for the dilemma of pedestrian loss, thus different types of data have been deeply integrated.
- We show the design and implementation of the system on commercial servers and smartphones. Compared with other systems, iWVP is robust and accurate in both localization and tracking of pedestrians through the coordinate transformation model, as is shown in Fig. 1.

2 Related Work

2.1 Localization method based on WiFi

WiFi-based indoor localization method can be roughly classified into two categories: modeling-based algorithm and fingerprint-based algorithm. The former usually uses triangulation of arrival angle or time of flight[30] to determine current positions. To achieve high accuracy needs LOS with Access Points (APs), which is not always available due to many indoor obstacles. The latter is more robust because it has no requirement of LOS with APs. RADAR[31] and HORUS[32] pioneer the indoor positioning technology in terms of WiFi fingerprints. However, WiFi signal is not stable indoors because of environmental changes, which leads to the inaccurate collection of fingerprint and reduces the accuracy of

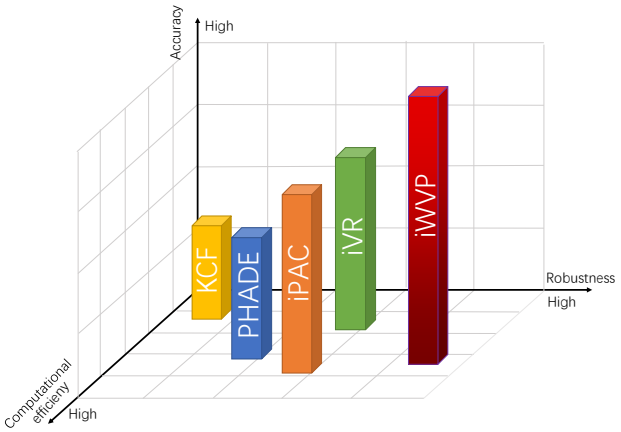

Fig. 1 Comparison of the state-of-the-art works.

localization. Yang et al.[33] proposed a WKNN indoor location algorithm based on spatial characteristics partition and localization restriction. Le et al.[34] proposed an Advanced Clustering (AC) strategy approximating location by clustering matching and achieving precise location by kernel ridge regression. Wu et al.[2] and Xu et al.[35] proposed ViVi and ViViplus, respectively, and their key idea is to exploit the spatial awareness of RSS values by formulating FSG profiles or RSG matrices as enhanced WiFi fingerprints. However, WiFi-based algorithm is difficult to obtain high location accuracy.

2.2 Localization method based on vision

Vision-based indoor localization methods achieve sub-meter level accuracy. The traditional method often resorts to SIFT or SURF image extraction algorithm[36, 37] to identify the target of each frame in the video, which provides an opportunity for accurate visual geometry calculation. Gu et al.[23] proposed "Spotlight" which performed passive localization using crowdsourced photos to achieve high accuracy. Yan et al.[38] presented and developed a novel 3D passive vision-aided PDR system using surveillance cameras and smartphone-based PDR, which could continuously track the user's movement on different floors by integrating results of inertial navigation and real-time pedestrian detection. It used large amounts of camera locations and embedded barometers to provide floor/height information to identify the user's positions in 3D space. Nevertheless, these relevant studies achieve high accuracy in indoor localization by high frame rates and suffer from frequent LOS blockage in indoor environments, which may lead to ineffective tracking algorithms.

2.3 Localization method based on data fusion

The inevitable accumulated error occurs when PDR data[39] are used for indoor positioning. However, researches show that better accuracy can be obtained by fusing PDR data with other sensor data. Poulose and Han[40] utilized the signals of magnetic field, Bluetooth, and WiFi as input data to train the model of the fingerprints. Chen et al.[41] proposed a multi-source data localization method that fuses WiFi, PDR, and indoor landmarks recognized by detecting a specific pattern of sensors. Xu et al.[29] presented iVR, an integrated vision and radio localization system with sub-meter accuracy, which utilizes particle filters to fuse raw estimates from multiple systems, including vision, radio, and inertial sensor systems. iPAC[42] is the indoor positioning system integrating vision and PDR, and employing a matching algorithm based on motion sequence to fuse raw estimates from both systems. And the identity of the pedestrian could be confirmed by a unique device ID in iPAC.

3 System Overview

The workflow of iWVP system consists of two main modules, the client module and the server module, as illustrated in Fig. 2. The positioning process can be outlined from the user's point of view. Firstly, the user starts the location application, RSS and sensor data will be collected continuously and automatically. Meanwhile, PDR data can be calculated locally and sent to the remote server along with the collected RSS data. Then, the current position will be obtained quickly. Once all the data are received, two stages will be applied to process the data. The first stage is WiFi region discrimination divided into the online phase and offline phase. In the offline phase, we need to collect WiFi signals in each region and calibrate the surveillance camera. Accurate region tracking will be executed in the second stage, where visual detection and tracking, and real-time PDR will be associated to confirm the user's identity by matching motion sequences. Finally, the server can locate pedestrians accurately and track them continuously through the new pixel coordinate.

4 Two-Stage Indoor Localization with Multisensor Fusion

4.1 Coarse positioning: Region discrimination based on WiFi

In the coarse positioning stage, region discrimination will be executed by combining WiFi signals with Bayes filter[43], which guarantees high accuracy in region positioning. Likewise, the method is divided into two phases: the offline phase and the online phase. In the offline phase, in each region, RSSI distribution table can be obtained once WiFi signals are captured. Gaussian curve is used to fit the RSSI distribution of each region. Meanwhile, RSSI data are divided into a training dataset and a test dataset for performance evaluation. The normalized histograms (for each access point (AP) and region) are calculated by using the training RSSI dataset. In the online phase, the client acquires the RSSI value of the environment as a query fingerprint and sends it to the server, the server runs Bayes filter to predict the correct

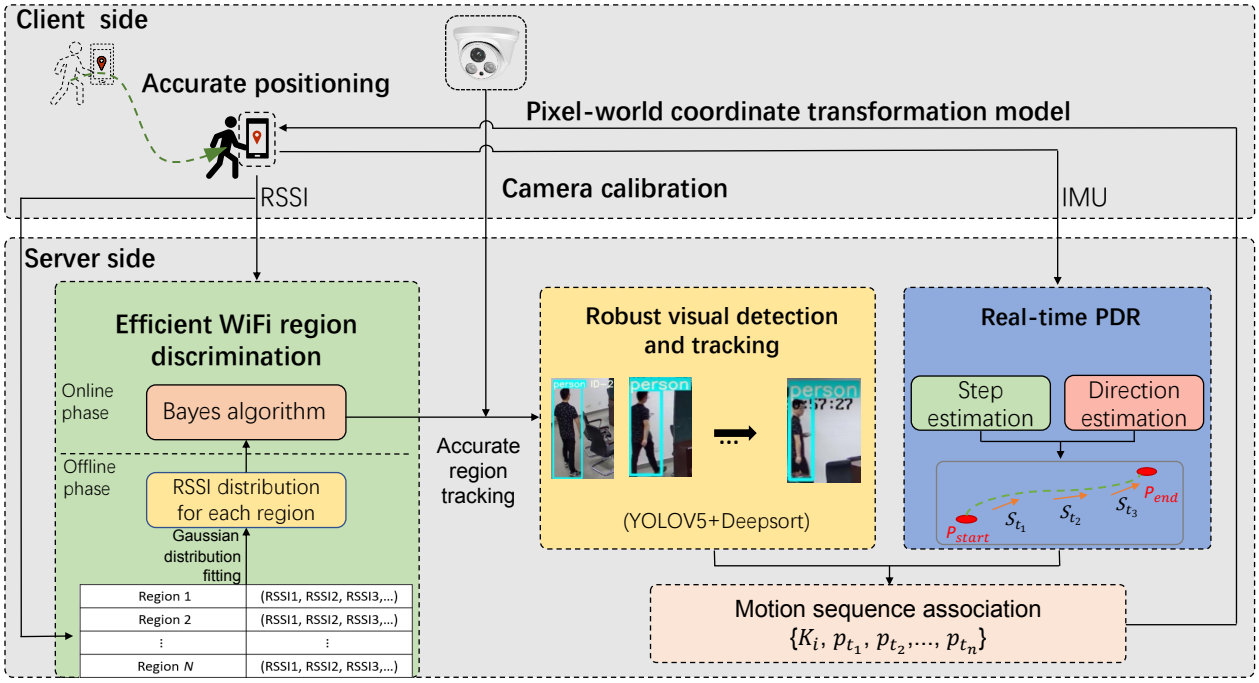

Fig. 2 System overview.

region and outputs the corresponding probability. Some symbols are explained in Table 1 for further study.

(1) Offline phase: RSSI-based studies have found that the distribution of RSSI follows the Gaussian normal distribution in many scenarios[44]. The following is the probability distribution function of Gaussian distribution:

$$P\left(s_j \mid z_i\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s_j-\mu)^2}{2\sigma^2}} \quad (1)$$

Four regions of the entire building have been selected to evaluate the region detection system as shown in Fig. 3. Regions 1–4 are corridor, hall, laboratory 1, and laboratory 2, respectively.

Figure 4 shows the original histograms and the cleaned histograms (Gaussian histograms) in Region 1. It can be seen that the original histograms can be well approximated by Gaussian Probability Density Functions (PDFs). The new normal distribution fills in the missing values of the original RSSI histogram

Table 1 Symbol meaning.

Symbol	Meaning
s	Received RSS signal from all APs
z_i	The i-th region
s_j	Received RSS signal from the j-th AP
N	Number of regions
μ	Mean value of RSS signal in a region
σ	Standard deviation of RSS signal in a region
$P(s_j \mid z_i)$	Distribution of RSS from the j-th AP in the i-th region

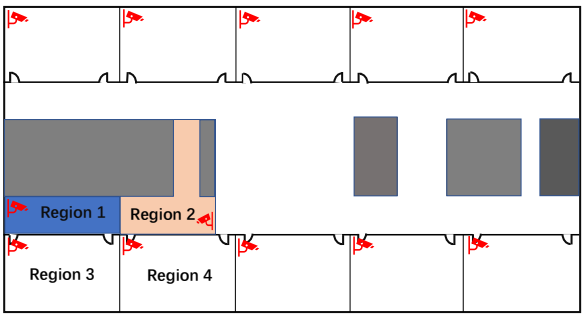

Fig. 3 Experimental areas.

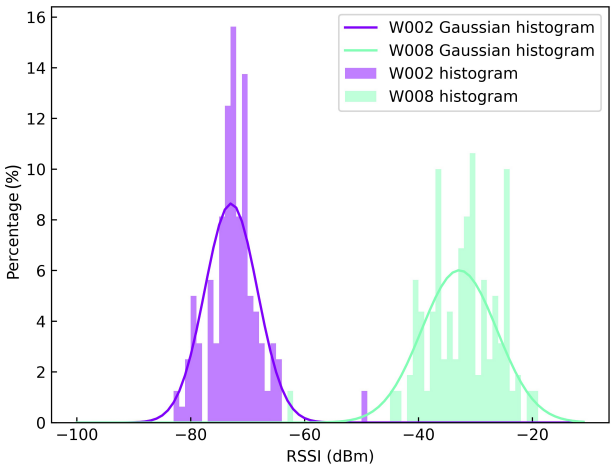

Fig. 4 APs histograms at Region 1.

and adjusts the noise values disturbed by reflection and scattering. A new dataset consisting of only Gaussian histograms is obtained after fitting the normalized

histograms. Each AP (access point) table will be formed with a combination of Gaussian histograms of different regions with the same AP as shown in Table 2, W_i represents the WiFi signal from Router 1.

(2) Online phase: Bayes filter can be used to infer the posterior probability, which uses obtained signals and prior knowledge,

$$P\left(z_i|s\right) = \frac{P\left(s|z_i\right)P(z_i)}{P(s)} \tag{2}$$

$$P(s) = \sum_{i=1}^{N} P\left(s|z_i\right)P\left(z_i\right) \tag{3}$$

In this way, the RSSI histogram tables (or AP tables) generated in the offline phase are used to calculate the $\frac{P(z_i|s)}{P(s)}$. By a series of iterations, the prior value $P(z_i)$ is updated with the current estimated posterior value $P(s \mid z_i)$ at the end of each iteration. The implementation of the Bayes-based estimator is shown in Algorithm 1. The AP tables are collected during the offline phase and current WiFi signals are uploaded by the mobile phone during the online phase. With the aforementioned as inputs, the estimated region and its probability will be returned by using Algorithm 1, which starts with a uniform prior probability distribution (Lines 5 and 6) and recursively calculates the probability of the posterior region based on AP tables (Line 10), and then the probability (Line 11) and predicted region (Line 12) are calculated. After evaluating all regions, the highest probability and its corresponding region are calculated (Lines 15 and 16). Finally, based on the previous results, the prior region is updated with the estimated posterior region (Line 20) to be used in the next iteration. In this way, the coarse positioning is completed. Surveillance cameras in this area will be further used in fine positioning.

4.2 Fine positioning: Fusion detection and tracking

(1) Pedestrian detection and tracking based on video images: Pedestrian detection in images is the basis for visual localization and tracking. iWVP uses YOLOv5 algorithm framework to detect pedestrians in video images. The algorithm uses a multi-scale pyramid structure to divide the original image into multiple equally spaced units which are detected on the feature map of three scales. Double upsampling is used to transfer the feature map on two adjacent scales. Each grid cell uses three anchor boxes to predict three bounding boxes. Each bounding box predicts the coordinates (x, y), the width and height of the target simultaneously.

Table 2 AP table for Router 1.

Region	Gaussian histogram
Region 1	W_1 Gaussian histogram at Region 1
Region 2	W_1 Gaussian histogram at Region 2
Region 3	W_1 Gaussian histogram at Region 3
⋮	⋮
Region N	W_1 Gaussian histogram at Region N

Algorithm 1 Bayes filter algorithm

1: R = number of APs
2: W_r= AP table of the r-th AP
3: **procedure** Bayes estimator $(\omega_1, \omega_2, \ldots, \omega_R)$
4: **start with uniform distribution**
5: $prior\ W_{1,2,\ldots,R} = \left[\frac{1}{N}; \frac{1}{N}; \ldots; \frac{1}{N}\right]_{N\times 1}$
6: probability $= (\frac{100}{N})\%$
7: **while** probability < probability threshold do
8: **perform Bayes**
9: **for** r from 1 to R **do**
10: $posteriorW_r$ = norm($priorW_r \times W_r$ [:, ω_r])
11: $prob_r$= max($posteriorW_r$)
12: $pred_r$= where($posteriorW_r = prob_r$)
13: **end for**
14: **find the highest probability**
15: probability = max($prob_{1,2,\ldots,R}$)
16: $r_$best = where($prob_{1,2,\ldots,R}$=probability)
17: prediction = $pred_{r\_best}$
18: **update the new prior**
19: **for** r from 1 to R **do**
20: $priorW_r = posteriorW_{r\_best}$
21: **end for**
22: **end while**
23: **return prediction and probability**
24: **end**

The YOLOv5 algorithm framework uses logistic regression to predict the probability that each bounding box contains objects. The probability of the anchor box is 1 if the overlap rate between the anchor box and the real target bounding box is greater than any other anchor box, but it is ignored if the overlap rate is greater than the threshold but less than the maximum overlap rate. Finally, the algorithm will select the best anchor box to assign to the target and use binary cross-entropy and logistic regression to predict its category.

When YOLOv5 provides detection boxes, Deep SORT is used for tracking. It has two basic modules. The first is the prediction module which uses the Kalman filter to predict the tracker. The second is the update module which includes matching, tracker updating, and feature set updating. In the update module, the fundamental method is using Intersection over Union (IoU) to match

the Hungarian algorithm. It uses a cascade matching algorithm to match different priorities. Finally, it adds Markov distance and cosine distance to compare the similarity between detector and tracker.

(2) PDR: Pedestrians have periodic acceleration changes when they walk normally. Therefore, the walking steps can be detected by the internal accelerometer in smartphones. The peak detection method uses the characteristics of peaks and valleys of acceleration data. It can eliminate the misjudgments when the user stops walking with a small number of calculations.

The main idea of PDR is calculating the number of detected steps, and then combining the estimation of step length and heading angle to obtain the pedestrian's relative position. The calculation formula is as follows:

$$x_{n+1} = x_n + L_n \sin\left(\sum_{i=1}^{n} \theta_i\right) \tag{4}$$

$$y_{n+1} = y_n + L_n \cos\left(\sum_{i=1}^{n} \theta_i\right) \tag{5}$$

where (x_{n+1}, y_{n+1}) represents the position coordinate after n steps; θ_i is the deflection angle of step i; L_n is the step size. The estimation of step size L_n is given in the following:

$$L_n = af' + b \tag{6}$$

where f' represents the step frequency; a and b are the coefficients.

In this paper, the inertial positioning method is used to obtain the relative displacements of pedestrians, and the results are used to assist visual passive positioning. The inertial sequence is matched with the visual sequence to determine the user's identity.

(3) Fusion tracking: After the region has been predicted, pedestrians' location and tracking will be carried out in this area.

During the initialization stage, the identifier of the mobile device can be used as the unique representation of the user's identity. However, the pedestrian detection and tracking in the video cannot correspond to the user directly. Therefore, in the initial stage, we need to associate the identifier of the mobile device with the pedestrian trajectory in the video to achieve one-to-one correspondence. In the subsequent fusion tracking stage, the pedestrian's trajectory calculated by PDR will contribute to accurate visual detection and tracking.

The key to matching is selecting some key features to match the data calculated by PDR and visual tracking. We select the user's motion sequence as the matching feature. According to the motion state tuple uploaded by the user, the motion state p_{t_j} of the user at a certain time t_j indicates whether the user is still or moving. We use the motion state over a period to construct the motion sequence $\{K_i, p_{t_1}, p_{t_2}, \ldots, p_{t_n}\}$, where K_i is the identifier of the mobile device. At the same time, the motion sequence of the pedestrian in the video is calculated through the visual tracking results. Finally, iWVP compares the motion sequence to complete the matching. The overview of fusion tracking is illustrated in Fig. 5.

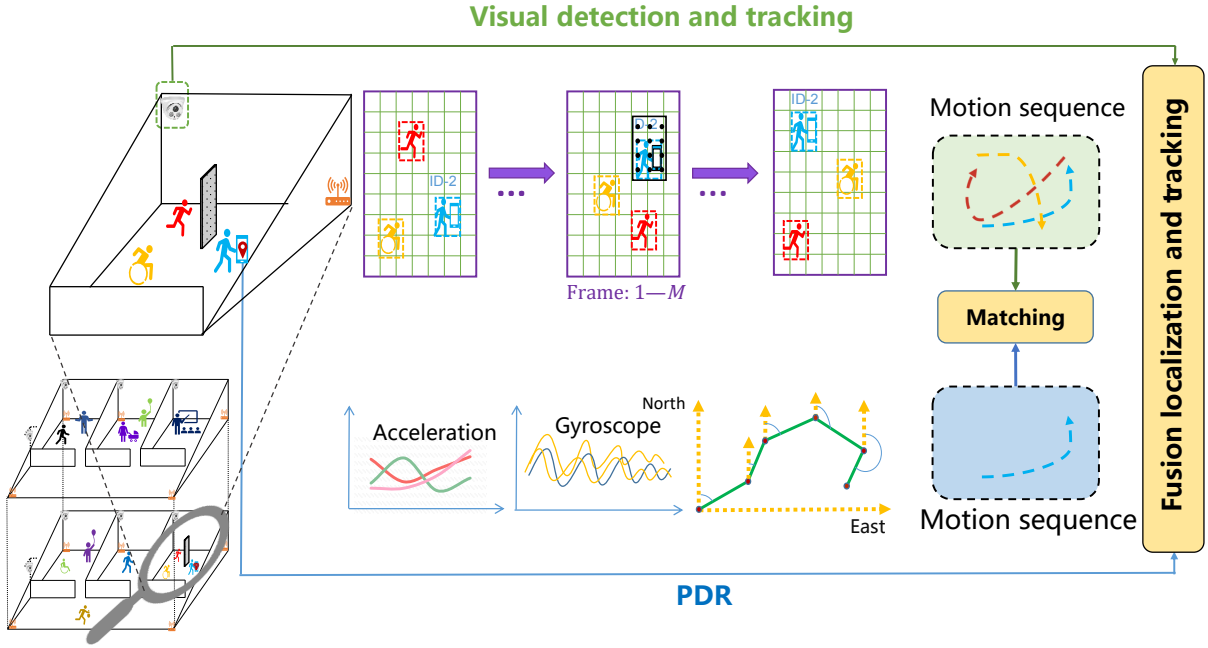

Fig. 5 Overview of fusion tracking.

In tracking stage, iWVP utilizes the trajectory calculated by PDR to assist visual tracking. The problem of inconsistent scale between the world coordinate and the pixel coordinate is solved by the coordinate transformation model in Section 5. iWVP uses the tuple $\langle m, p\rangle$ to record user information, where m is the user's identifier and p is the user's position. The server receives a motion state from a user every time, iWVP updates the tuple and checks the visual tracking result, where may appear three cases:

Case 1: The user is detected and positioned successfully in the frame, indicating that the visual tracking accuracy is higher than that of PDR. Thus, iWVP uses the result of visual tracking to update the user information tuple.

Case 2: If the detection result shows that the user's leaving the monitoring range, then visual positioning is ineffective and tracking accuracy relies more on PDR. iWVP will use PDR to continuously track the user and update the user information. When the user returns to the monitoring range, iWVP ensures that the user can be identified correctly and tracked continuously.

Case 3: When the user does not leave the monitoring area and the visual tracking fails to detect location due to blockage, iWVP keeps tracking until the obstacles disappear or the user leaves the blocked area. As shown in Fig. 6, the user information tuple is used to mark the corresponding location in the video to indicate the user's possible position through PDR data.

5 Coordinate Transformation Model

Most indoor cameras are monocular, but it is really hard for a monocular camera to obtain the depth information of the target. In general, the pixel coordinates of the target obtained by the monocular camera cannot correspond to its real spatial coordinate position. Therefore, the problem we need to solve is to achieve the conversion from the pixel coordinates under the monocular camera to the world coordinates (describing the real spatial position of the target). Based on the camera imaging principle, the camera calibration method based on a checkerboard is studied, and a scene-assisted monocular camera coordinate transformation model is proposed to realize the coordinate transformation of "pixel to world". The accuracy of the coordinate transformation model is evaluated by experiments.

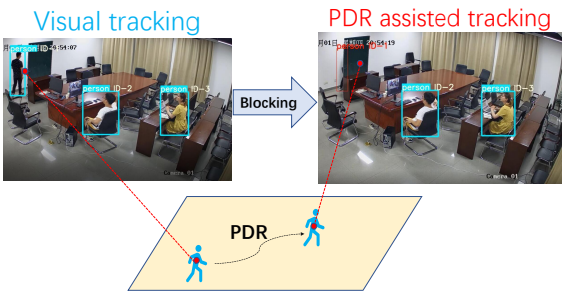

Fig. 6 PDR assisted tracking.

The pinhole model, an approximate model of the camera, is the simplest among various camera imaging models. The pinhole model does not take into account the distortion of the camera, so it actually only includes perspective projection transformation and rigid body transformation. In order to describe the process of determining the target's position in 3D scene space from the pixels in the image, it is necessary to understand the four coordinate systems in the camera model: pixel plane coordinate system, image plane coordinate system, camera coordinate system, and world coordinate system.

(1) Pixel plane coordinate system: Each digital image is composed of pixels, and the pixel plane coordinate represents the positions of pixels in the image. As shown in Fig. 7, the Cartesian coordinate system is defined on the image, whose origin O_0 is located at the upper left corner of the image. The image's width and height are w and h, respectively. The (u_0, v_0) is the pixel plane coordinate.

(2) Image plane coordinate system: The pixel plane coordinate only represents the number of rows and columns in the image array where the pixel is located, but has no actual physical meaning to represent the position of the pixel in the image. Therefore, it is essential to define the actual physical 2D image plane coordinate system $O_1 - xy$. As shown in Fig. 8, the origin O_1 represents the intersection of the camera optical axis and the image imaging plane. The x and y axes are parallel

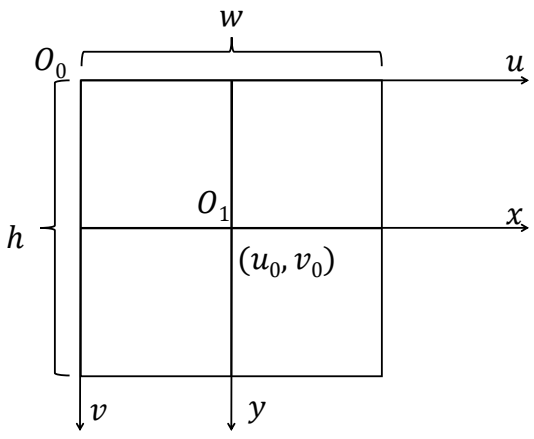

Fig. 7 Plane of pixel and imaging.

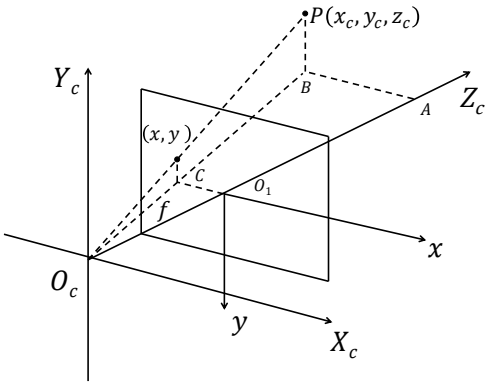

Fig. 8 Camera imaging model.

to the u and v axes of the pixel plane coordinate system, respectively, and their coordinate system unit is mm.

The origin of the plane coordinate system O_1 is generally located at the center of the image. Let O_1 be (u_0, v_0) in the coordinate of the pixel plane. Let each pixel in u and v axes on the physical size are d_x and d_y, respectively. Therefore, without considering distortion, the relationship between the pixel plane coordinate system and the image plane coordinate system can be expressed as $u = x/d_x + u_0$, $v = y/d_y + v_0$. The above relationship can be expressed in matrix in the following:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_x} & 0 & u_0 \\ 0 & \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{7}$$

(3) Camera coordinate system: As shown in Fig. 8, the origin of the camera coordinate system is the optical center of the camera, and its axes are parallel to the axes of the image plane coordinate system. The optical axes of the camera are perpendicular to the imaging plane. The distance of f is the focal length of the camera.

The transformation from camera coordinate system to image plane coordinate system belongs to perspective projection. As shown in Fig. 8, the projection of any point in the space on the image plane is the intersection of the camera's optical center line and the image plane. From the similarity transformation, we have

$$\frac{AB}{O_1C} = \frac{AO_c}{O_1O_c} = \frac{PB}{PC} \tag{8}$$

therefore

$$\frac{x_c}{x} = \frac{z_c}{f} = \frac{y_c}{y} \tag{9}$$

Equation (9) can be expressed in matrix form as follows:

$$z_c \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \tag{10}$$

(4) World coordinate system: World coordinate system is to describe the position of any object in a 3D scene. Any position in the environment can be chosen as the origin. The coordinate system is also known as the absolute coordinate system, used to represent the absolute coordinates of the scene.

The transformation from the world coordinate system to the camera coordinate system is a rigid body transformation as shown in Fig. 9, and their relationship can be described by a 3×3 orthogonal unit rotation matrix R and a 3D translation vector T. Therefore, let the coordinates of P in the camera coordinate system and the world coordinate system be (x_c, y_c, z_c) and (x_w, y_w, z_w), respectively. Then the relationship is shown in the following:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} R & T \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{11}$$

The conversion relation from the pixel plane coordinate system to the world coordinate system is expressed as following:

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f \times \frac{1}{d_x} & 0 & u_0 \\ 0 & f \times \frac{1}{d_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & T \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{12}$$

According to Eq. (12),the ground is taken as the xoy plane of the world coordinate system, that is to say $z_w = 0$, then world coordinates are calculated by pixel coordinates. The accurate location of the pedestrian is obtained. In our experiment, the center pixel coordinates of the detected pedestrian detection frame's bottom are regarded as the pedestrian pixel coordinates as shown in Fig. 10.

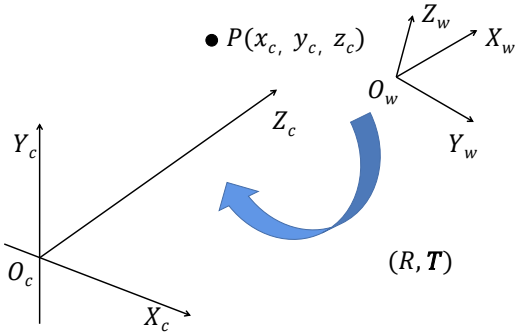

Fig. 9 Rigid body transformation.

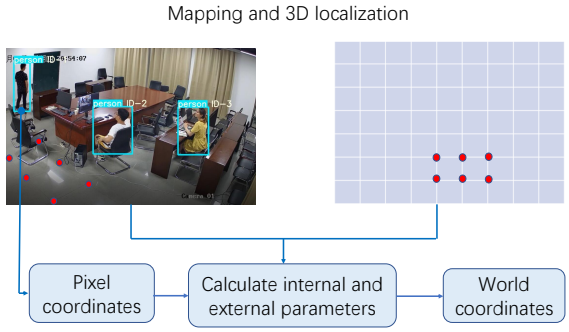

Fig. 10 Architecture of the coordinate transformation mode.

6 Implementation and Evaluation

6.1 Experimental scenarios

Extensive experiments have been conducted in laboratories, halls, and corridors. Surveillance cameras and wireless sensors are installed in the whole scenario. These experiments have different floor layouts, diverse wireless environments, and distinct user behavior patterns, which are quite complicated.

6.2 Experimental setup

The client of iWVP is implemented on the Android platform with all the devices mentioned above. The camera is HIKIVISION-H100 with a frame size of 1080 pixel × 960 pixel which is used as the IP camera to continuously monitor the region and send video streams to the server. One surveillance camera is deployed in each region of the experimental scenario. The server we use is DELL t3640 with i9-10900k CPU and 64 GB RAM, running Windows 10 operating system.

6.3 Performance

Our algorithm is inspired by iPAC, so we compare iWVP with iPAC in our experiments. In each region, RSSI signals are collected several times at different periods to obtain WiFi distribution, WiFi signals are collected 100 times, and Bayes algorithm is used to predict the region. Figure 11 shows the confusion matrix of the evaluation error obtained by the experiment, where each element represents the probability that the predicted region is the real region. The average regional prediction accuracy is 95.652%, virtually unaffected by time and environmental changes. The positioning area can be well expanded due to the wide use of surveillance cameras, combined with the high precision of WiFi area positioning.

The main performance of fusion detection and tracking has been tested. Figures 12–14 depict pedestrian detection's performances of iWVP and iPAC in four

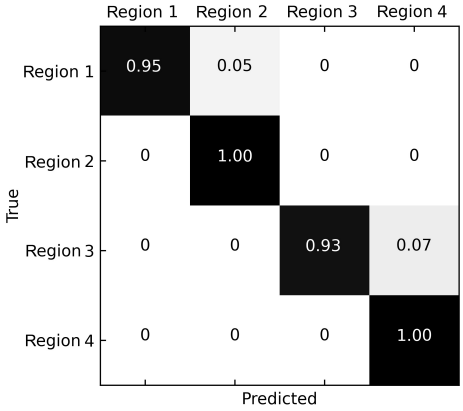

Fig. 11 Confusion matrix.

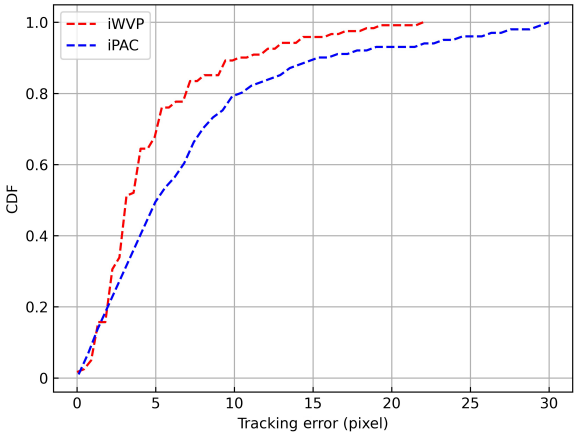

Fig. 12 Tracking error, where CDF denotes cumulative distribution function.

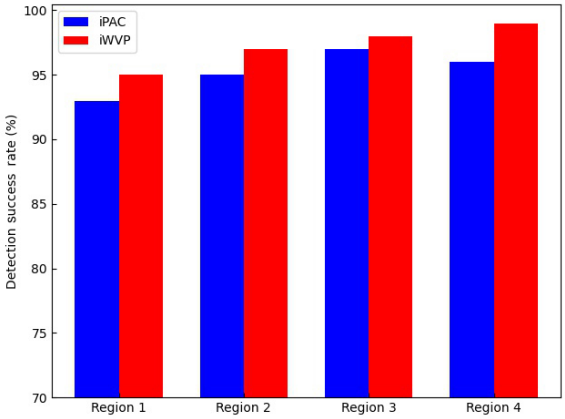

Fig. 13 Detection success rate in different areas.

different indoor environmental regions. In contrast with iPAC, our system achieves a better performance in detection success rate, exceeding 95% in all regions, the average tracking error of iWVP is less than 5 pixel as shown in Fig. 12, and the tracking success rates of iWVP in four regions are 97%, 98%, 98%, and 99%, respectively, as illustrated in Fig. 14. Also, iWVP performs very well in complex environments and indicates the given algorithm is more robust.

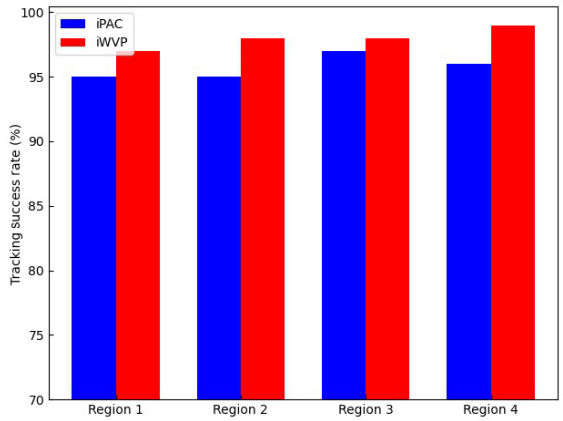

Fig. 14 Tracking success rate in different areas.

We further examine the positioning performance of iWVP and iPAC in three different experimental regions (Regions 1–3) as illustrated in Fig. 3, including one laboratory (Trajectory 1), one corridor (Trajectory 2), and one hall (Trajectory 3), as shown in Figs. 15–21 and Tables 3 and 4.

As shown in Figs. 15–17, the positioning accuracy of iWVP is studied in three different regions (laboratory, corridor, and hall) compared with iPAC. iWVP is positioned closer to the ground truth and performs effectively in different complex indoor environments. This is mostly because the given method can avoid pixel jitter effectively. As shown in Figs. 18–20, iWVP achieves better accuracy, yielding the 95th percentile error of 13.4 cm, 17.3 cm, and 12.7 cm, respectively.

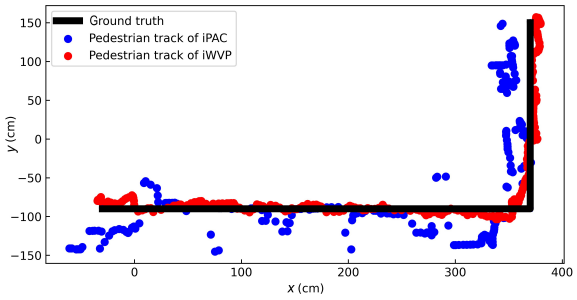

Fig. 15 Schematics of Trajectory 1.

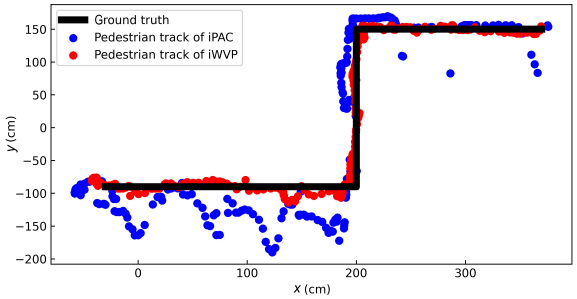

Fig. 16 Schematics of Trajectory 2.

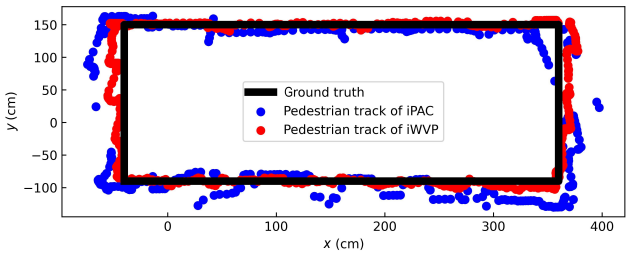

Fig. 17 Schematics of Trajectory 3.

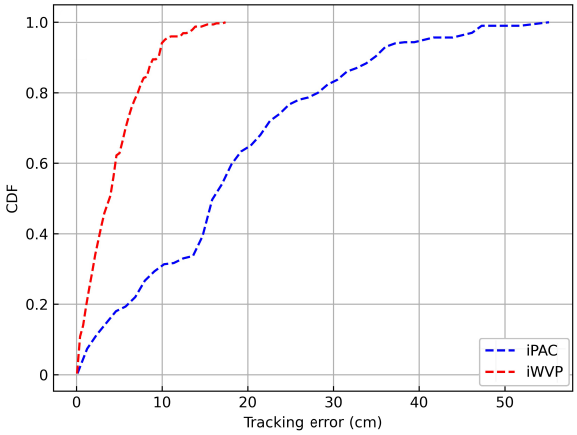

Fig. 18 Position error of Trajectory 1.

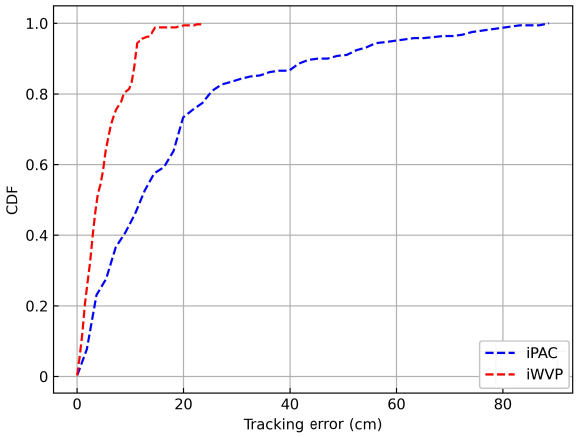

Fig. 19 Position error of Trajectory 2.

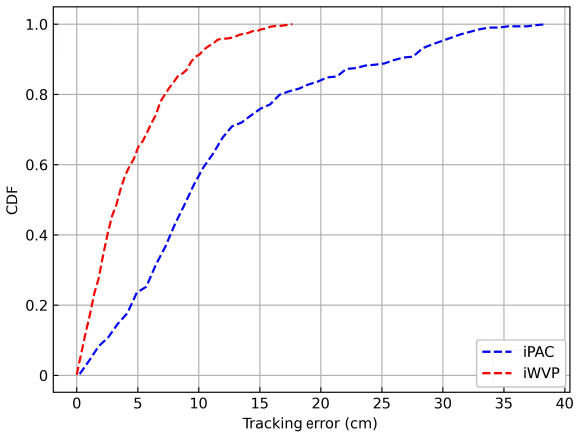

Fig. 20 Position error of Trajectory 3.

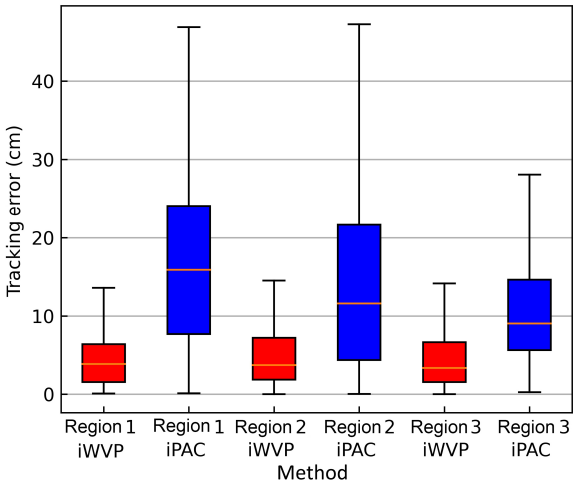

Fig. 21 Comparison of position errors between iWVP and iPAC in Regions 1–3.

Table 3 Positioning error of iWVP in different regions.
(cm)

Trajectory	Max error	Mean error
1	17.41	4.37
2	23.98	5.04
3	17.67	4.41

Table 4 Positioning error of iPAC in different regions.
(cm)

Trajectory	Max error	Mean error
1	55.16	17.52
2	88.68	17.67
3	38.48	11.27

Figure 21 shows the error assessment of iWVP and iPAC. It can be seen that the error of iWVP is lower than that of iPAC, indicating that iWVP plays a better role in positioning and tracking.

The maximum and average positioning errors of the three trajectories are shown in Tables 3 and 4. The maximum error of iWVP is about 20 cm, while that of iPAC fluctuates greatly, which also reflects the robustness of iWVP. As seen, iWVP yields an average error of 4.37 cm in the laboratory, 5.04 cm in the corridor, and 4.41 cm in the hall. The total average error is 4.61 cm. The results indicate that iWVP performs well regardless of the environmental difference in precise positioning.

We also evaluate the robustness of iWVP by introducing different visual noises. Three different levels of noise are introduced as shown in Fig. 22. iWVP with high noise (multiple pedestrian environments) performs almost as well as the case with low noise (individual pedestrian), and even with high noise, iWVP's average positioning error is still less than 7 cm. Besides, we also

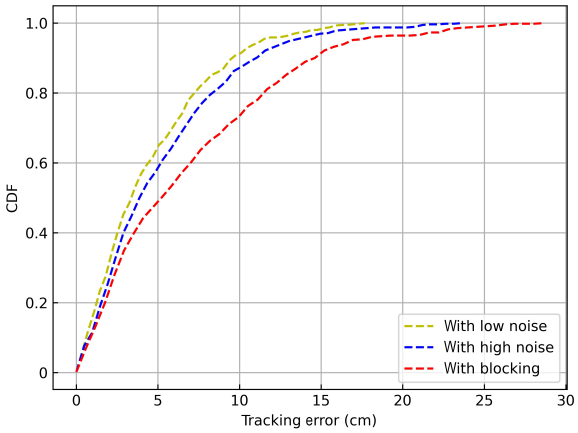

Fig. 22 Different noise strengths.

measure the performance when the user is completely blocked, in which case iWVP predicts the user position in the video by PDR. The average positioning error of iWVP under complete blockage is still less than 10 cm. The results indicate that even in complex environments with severe occlusion, iWVP can accurately track pedestrians with fusion tracking.

Although iWVP can achieve real-time performance (30 fps) in most commercial servers, we also measure the impact of the frame processing rate of iWVP to estimate the performance in large-scale deployment. As illustrated in Fig. 23, the average error decreases from 7.14 cm to 4.14 cm and the maximum error decreases from 28.64 cm to 16.66 cm when the frame processing rate increases from 15 fps to 30 fps. The results demonstrate that iWVP performs well enough even when the frame processing rate drops to 15 fps.

7 Conclusion

In this paper, an accurate two-stage indoor localization

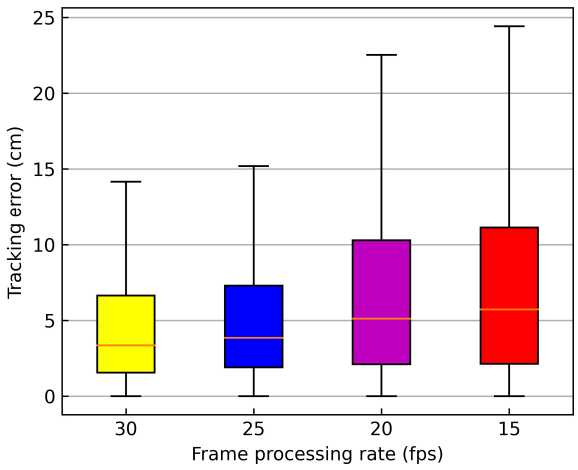

Fig. 23 Different frame processing rates.

system iWVP is studied. It integrates WiFi fingerprints, the vision of surveillance cameras, and PDR, avoiding the shortcomings of previous single indoor localization methods. For the monitoring region, the coarse location of the pedestrian is done via WiFi under a certain camera, and then the deep learning tracking and PDR fusion are added to achieve a fine location. The world coordinates of the pedestrian are obtained through the transforming model. Experimental results show that iWVP achieves an overall tracking success rate of 97% and an average positioning error of 4.61 cm in complex indoor environments. Meanwhile, iWVP can permit accurate and robust positioning and tracking of the specified pedestrian even under complete blockage. Also, the performance of iWVP is effectively validated by being implemented on commodity mobile devices.

Acknowledgment

This work was supported by the National Key Research and Development Program (No. 2018YFB2100301) and the National Natural Science Foundation of China (No. 61972131).

References

[1] H. Motte, J. Wyffels, L. De Strycker, and J. P. Goemaere, Evaluating GPS data in indoor environments, *Adv. Electr. Computer Eng.*, vol. 11, no. 3, pp. 25–28, 2011.

[2] C. S. Wu, J. G. Xu, Z. Yang, N. D. Lane, and Z. W. Yin, Gain without pain: Accurate WiFi-based localization using fingerprint spatial gradient, *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–19, 2017.

[3] C. S. Wu, Z. Yang, C. W. Xiao, C. F. Yang, Y. H. Liu, and M. Y. Liu, Static power of mobile devices: Self-updating radio maps for wireless indoor localization, in *Proc. 2015 IEEE Conf. Computer Communications* (*INFOCOM*), Hong Kong, China, 2015, pp. 2497–2505.

[4] Z. Yang, C. S. Wu, and Y. H. Liu, Locating in fingerprint space: Wireless indoor localization with little human intervention, in *Proc. 18th Ann. Int. Conf. Mobile Computing and Networking*, Istanbul, Turkey, 2012, pp. 269–280.

[5] P. Dollár, R. Appel, and W. Kienzle, *Crosstalk Cascades for Frame-rate Pedestrian Detection*. Berlin, Germany: Springer, 2012.

[6] B. Y. Liu, J. Z. Huang, Y. Lin, and C. Kulikowski, Robust tracking using local sparse appearance model and K-selection, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 1313–1320.

[7] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, Globallyoptimal greedy algorithms for tracking a variable number of objects, in *Proc. 2011 IEEE Conf. Computer Vision and Pattern Recognition* (*CVPR*), Washington, DC, USA, 2011, pp. 1210–1208.

[8] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, Zee: Zero-effort crowdsourcing for indoor localization, in *Proc. 18th Ann. Int. Conf. Mobile Computing and Networking*, Istanbul, Turkey, 2012, pp. 293–304.

[9] C. S. Wu, Z. Yang, and C. W. Xiao, Automatic radio map adaptation for indoor localization using smartphones, *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 517–528, 2017.

[10] Z. Yang, C. S. Wu, Z. M. Zhou, X. L. Zhang, X. Wang, and Y. H. Liu, Mobility increases localizability: A survey on wireless indoor localization using inertial sensors, *ACM Comput. Surv.*, vol. 47, no. 3, p. 54, 2015.

[11] Q. Shi, S. H. Zhao, X. W. Cui, M. Q. Lu, and M. D. Jia, Anchor self-localization algorithm based on UWB ranging and inertial measurements, *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 728–737, 2019.

[12] Q. Z. Lin and Y. Guo, Accurate indoor navigation system using human-item spatial relation, *Tsinghua Science and Technology*, vol. 21, no. 5, pp. 521–537, 2016.

[13] Q. X. Chen, D. D. Ding, and Y. Zheng, Indoor pedestrian tracking with sparse RSS fingerprints, *Tsinghua Science and Technology*, vol. 23, no. 1, pp. 95–103, 2018.

[14] L. M. Ni, Y. H. Liu, Y. C. Lau, and A. Patil, LANDMARC: Indoor location sensing using active RFID, in *Proc. 1st IEEE Int. Conf. Pervasive Computing and Communications*, Fort Worth, TX, USA, 2003, pp. 407–415.

[15] Y. H. Liu, J. L. Wang, Y. T. Zhang, L. S. Cheng, W. Y. Wang, Z. Wang, W. M. Xu, and Z. J. Li, Vernier: Accurate and fast acoustic motion tracking using mobile devices, *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 754–764, 2021.

[16] S. G. Wei, J. K. Wang, and Z. H. Zhao, Poster abstract: LocTag: Passive WiFi tag for robust indoor localization via smartphones, in *Proc. IEEE INFOCOM 2020-IEEE Conf. Computer Communications Workshops* (*INFOCOM WKSHPS*), Toronto, Canada, 2020, pp. 1342–1343.

[17] S. Saloni and A. Hegde, WiFi-aware as a connectivity solution for IoT pairing IoT with WiFi aware technology: Enabling new proximity based services, in *Proc. 2016 Int. Conf. Internet of Things and Applications* (*IOTA*), Pune, India, 2016, pp. 137–142.

[18] W. Gong and J. C. Liu, SiFi: Pushing the limit of time-based WiFi localization using a single commodity access point, *Proceedings of the ACM on Interactive*, *Mobile*, *Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 10, 2018.

[19] B. Shin, S. Lee, C. Kim, J. Kim, T. Lee, C. Kee, S. Heo, and H. Rhee, Implementation and performance analysis of smartphone-based 3D PDR system with hybrid motion and heading classifier, in *Proc. 2014 IEEE/ION Position, Location and Navigation Symp.-PLANS 2014*, Monterey, CA, USA, 2014, pp. 201–204.

[20] X. C. Liu, Y. R. Jiang, P. Jain, and K. H. Kim, TAR: Enabling fine-grained targeted advertising in retail stores, in *Proc. 16th Ann. Int. Conf. Mobile Systems, Applications, and Services*, Munich, Germany, 2018, pp. 323–336.

[21] J. Teng, B. Y. Zhang, J. D. Zhu, X. F. Li, D. Xuan, and Y. F. Zheng, EV-Loc: Integrating electronic and visual signals for accurate localization, *IEEE/ACM Trans. Network.*, vol. 22, no. 4, pp. 1285–1296, 2013.

[22] W. Ma, Q. Q. Li, B. D. Zhou, W. X. Xue, and Z. D. Huang, Location and 3-D visual awareness-based dynamic texture updating for indoor 3-D model, *IEEE Internet Things J*., vol. 7, no. 8, pp. 7612–7624, 2020.

[23] J. X. Gu, J. L. Wang, L. Zhang, Z. W. Yu, X. Z. Xin, and Y. H. Liu, Spotlight: Hot target discovery and localization with crowdsourced photos, *Tsinghua Science and Technology*, vol. 25, no. 1, pp. 68–80, 2020.

[24] X. Q. Teng, D. K. Guo, Y. L. Guo, X. L. Zhou, and Z. Liu, CloudNavi: Toward ubiquitous indoor navigation service with 3D point clouds, *ACM Trans. Sensor Networks*, vol. 15, no. 1, p. 1, 2019.

[25] A. R. Jimenez, F. Seco, C. Prieto, and J. Guevara, A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU, in *Proc. 2009 IEEE Int. Symp. Intelligent Signal Proc*., Budapest, Hungary, 2009, pp. 37–42.

[26] Z. D. Li, Z. B. Su, and T. T. Yang, Design of intelligent mobile robot positioning algorithm based on IMU/Odometer/Lidar, in *Proc. 2019 Int. Conf. Sensing, Diagnostics, Prognostics, and Control* (*SDPC*), Beijing, China, 2019, pp. 627–631.

[27] R. Harle, A survey of indoor inertial positioning systems for pedestrians, *IEEE Commun. Surv. Tut*., vol. 15, no. 3, pp. 1281–1293, 2013.

[28] M. Q. Zhan and Z. H. Xi, Indoor location method of WiFi/PDR fusion based on extended Kalman filter fusion, *J. Phys.: Conf. Ser*., vol. 1601, no. 4, p. 042004, 2020.

[29] J. G. Xu, H. J. Chen, K. Qian, E. Q. Dong, M. Sun, C. S. Wu, L. Zhang, and Z. Yang, iVR: Integrated vision and radio localization with zero human effort, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*., vol. 3, no. 3, p. 114, 2019.

[30] K. S. Wu, J. Xiao, Y. W. Yi, M. Gao, and L. M. Ni, FILA: Fine-grained indoor localization, in *Proc. 2012 IEEE INFOCOM*, Orlando, FL, USA, 2012, pp. 2210–2218.

[31] P. Bahl and V. N. Padmanabhan, RADAR: An in-building RF-based user location and tracking system, in *Proc. IEEE INFOCOM 2000. Conf. Computer Communications. Nineteenth Ann. Joint Conf. IEEE Computer and Communications Societies* (*Cat. No. 00CH37064*), Tel Aviv, Israel, 2000, pp. 775–784.

[32] M. Youssef and A. Agrawala, The Horus WLAN location determination system, in *Proc. 3rd Int. Conf. Mobile Systems, Applications, and Services*, Washington, DC, USA, 2005, pp. 205–218.

[33] H. F. Yang, Y. B. Zhang, Y. L. Huang, H. M. Fu, and Z. H. Wang, WKNN indoor location algorithm based on zone partition by spatial features and restriction of former location, *Pervasive Mob. Comput*., vol. 60, p. 101085, 2019.

[34] Y. F. Le, H. N. Zhang, W. B. Shi, and H. Yao, Received signal strength based indoor positioning algorithm using advanced clustering and kernel ridge regression, *Front. Inform. Technol. Electron. Eng*., vol. 22, no. 6, pp. 827–838, 2021.

[35] J. G. Xu, Y. Zheng, H. J. Chen, Y. H. Liu, X. C. Zhou, J. B. Li, and N. Lane, Embracing spatial awareness for reliable WiFi-based indoor location systems, in *Proc. 2018 IEEE 15th Int. Conf. Mobile Ad Hoc and Sensor Systems* (*MASS*), Chengdu, China, 2018, pp. 281–289.

[36] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[37] H. Bay, T. Tuytelaars, and L. Van Gool, SURF: Speeded up robust features, in *Proc. 9th European Conf. Computer Vision-Volume Part I*, Graz, Austria, 2006, pp. 404–417.

[38] J. J. Yan, G. G. He, A. Basiri, and C. Hancock, 3-D passive-vision-aided pedestrian dead reckoning for indoor positioning, *IEEE Trans. Instrum. Meas*., vol. 69, no.4, pp. 1370–1386, 2020.

[39] X. L. Gan, B. G. Yu, H. Zhang, L. Huang, and Y. N. Li, Indoor combination positioning technology of pseudolites and PDR, in *Proc. 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services*, Wuhan, China, 2018, pp. 1–7.

[40] A. Poulose and D. S. Han, Indoor localization using PDR with Wi-Fi weighted path loss algorithm, in *Proc. 2019 Int. Conf. Information and Communication Technology Convergence* (*ICTC*), Jeju, Republic of Korea, 2019, pp. 689–693.

[41] Z. H. Chen, H. Zou, H. Jiang, Q. C. Zhu, Y. C. Soh, and L. H. Xie, Fusion of WiFi, smartphone sensors and landmarks using the Kalman filter for indoor localization, *Sensors*, vol. 15, no. 1, pp. 715–732, 2015.

[42] D. Y. Li, Y. M. Lu, J. G. Xu, Q. Ma, and Z. Liu, iPAC: Integrate pedestrian dead reckoning and computer vision for indoor localization and tracking, *IEEE Access*, vol. 7, pp. 183514–183523, 2019.

[43] D. N. Fernández, Implementation of a WiFi-based indoor location system on a mobile device for a university area, in *Proc. 2019 IEEE XXVI Int. Conf. Electronics, Electrical Engineering and Computing* (*INTERCON*), Lima, Peru, 2019, pp. 1–4.

[44] L. Chruszczyk, Statistical analysis of indoor RSSI readouts for 433 MHz, 868 MHz, 2.4 GHz and 5 GHz ISM bands, *Int. J. Electron. Telec*., vol. 63, no. 1, pp. 33–38, 2017.

Jinhui Bao received the BS degree in applied mathematics from Jiaxing University, Jiaxing, China in 2019. He is currently a master student at the School of Mathematics, Hefei University of Technology. His main research interest is Internet of Things (IoTs).

Yi Xu received the BS degree in applied mathematics from Guangxi University, Nanning, China in 2020. He is currently a master student at the School of Mathematics, Hefei University of Technology. His main research interest is IoTs.

Li Zhang received the BS degree in applied mathematics from Anhui Normal University, Wuhu, China in 1999, and the MS and PhD degrees from Hefei University of Technology, Hefei, China in 2004 and 2009, respectively. From 2012 to 2013, she held a postdoctoral position in computer science and technology at Arizona State University, Phoenix, AZ, USA. She is currently a professor at the School of Mathematics, Hefei University of Technology. Her current research interests include computer-aided geometric design, computer graphics, and image processing.

Qiuyu Wang received the BS degree in applied mathematics from Anqing Normal University, Anqing, China, in 2019. She is currently a master student at the School of Mathematics, Hefei University of Technology. Her main research interest is IoTs.

Jingao Xu received the BEng degree from Tsinghua University, China in 2017, where he is currently a PhD candidate. His research interests include IoTs and mobile computing.

Danyang Li received the BEng degree from Yanshan University, China in 2019. He is currently a PhD candidate at the School of Software and BNRist, Tsinghua University. His research interests include IoTs and mobile computing.