Wi-Fi, CCTV and PDR Integrated Pedestrian Positioning System

Max J. L. Lee [1], Meiling Su [1] and Li-Ta Hsu [1]

[1] *Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*

Abstract

Navigation inside a closed area with no GNSS-signal accessibility is a highly challenging task. In order to tackle this problem, the Wi-Fi beacon-based methods have grabbed the attention of many researchers. However, it suffers from lack of accuracy and robustness against environmental changes and dynamic object movements. With the advancement of inference speed in real-time deep learning applications, in this paper, a multi-modal end-to-end system for large-scale indoor positioning has been proposed, namely Wi-Fi, CCTV and PDR integrated pedestrian positioning system, which increases positioning accuracy by overcoming the difficulties of environment changes. Firstly, a user request positioning and sends the detect access points (APs) and associated signal strength to a server. A Wi-Fi one-shot model is used to estimate an initial position for the user, then the nearby CCTV cameras are activated to detect and position pedestrians. A CCTV and PDR integrated particle filter is used to determine a valid correspondence to the user, simultaneously collecting Wi-Fi fingerprints from the user to passively update the Wi-Fi one-shot support set in real-time. By implementing the proposed system, we could achieve a highly accurate integrated indoor positioning with precision level of 0.3m.

Keywords

Indoor Positioning, Visual Positioning, Particle Filter, Wi-Fi Positioning

1. Introduction

Indoor positioning systems have attracted a great interest from the researchers over the past decade. These systems can provide positioning, navigation, tracking services where global navigation satellite systems (GNSSs) could not reach [1]. To overcome the limitation of GNSS in indoor environments, various indoor positioning systems have been developed using Wi-Fi, Bluetooth (BLE), ultra-wideband (UWB), and radio-frequency identification (RFID). Typically, beacon nodes are a prerequisite to localize in the indoor environment [2]. Out of the above mentioned radio-frequency based systems, Wi-Fi is the most popular one due to its scalability [3]. The Wi-Fi localization is usually performed by four main techniques: triangulation, received signal strength (RSS), scene analysis (fingerprinting) and proximity based. However, the radio ranging measurements contain noise on the order of several meters [4]. This noise occurs because radio propagation tends to be highly non-uniform. Physical obstacles such as walls, furniture, etc. reflect and absorb radio waves [5]. As a result, the distance estimation is deteriorated by the diffractions and reflections. The fingerprinting method attempts to overcome this problem by making use of diffractions and reflections as additional features for positioning, however, it suffers when there are changes in the environment, as the features also change. Hence, an alternative positioning method is suggested in this paper to overcome the difficulties of environment changes.

IPIN 2022 WiP Proceedings, September 5 - 7, 2022, Beijing, China
EMAIL: maxjl.lee@connect.polyu.hk (M. J. L. Lee); meiling.su@connect.polyu.hk (M. Su); lt.hsu@polyu.edu.hk (L.T. Hsu)
ORCID: 0000-0002-5524-6724 (M. J. L. Lee); 0000-0002-9095-5681 (M. Su); 0000-0002-0352-741X (L.T. Hsu)


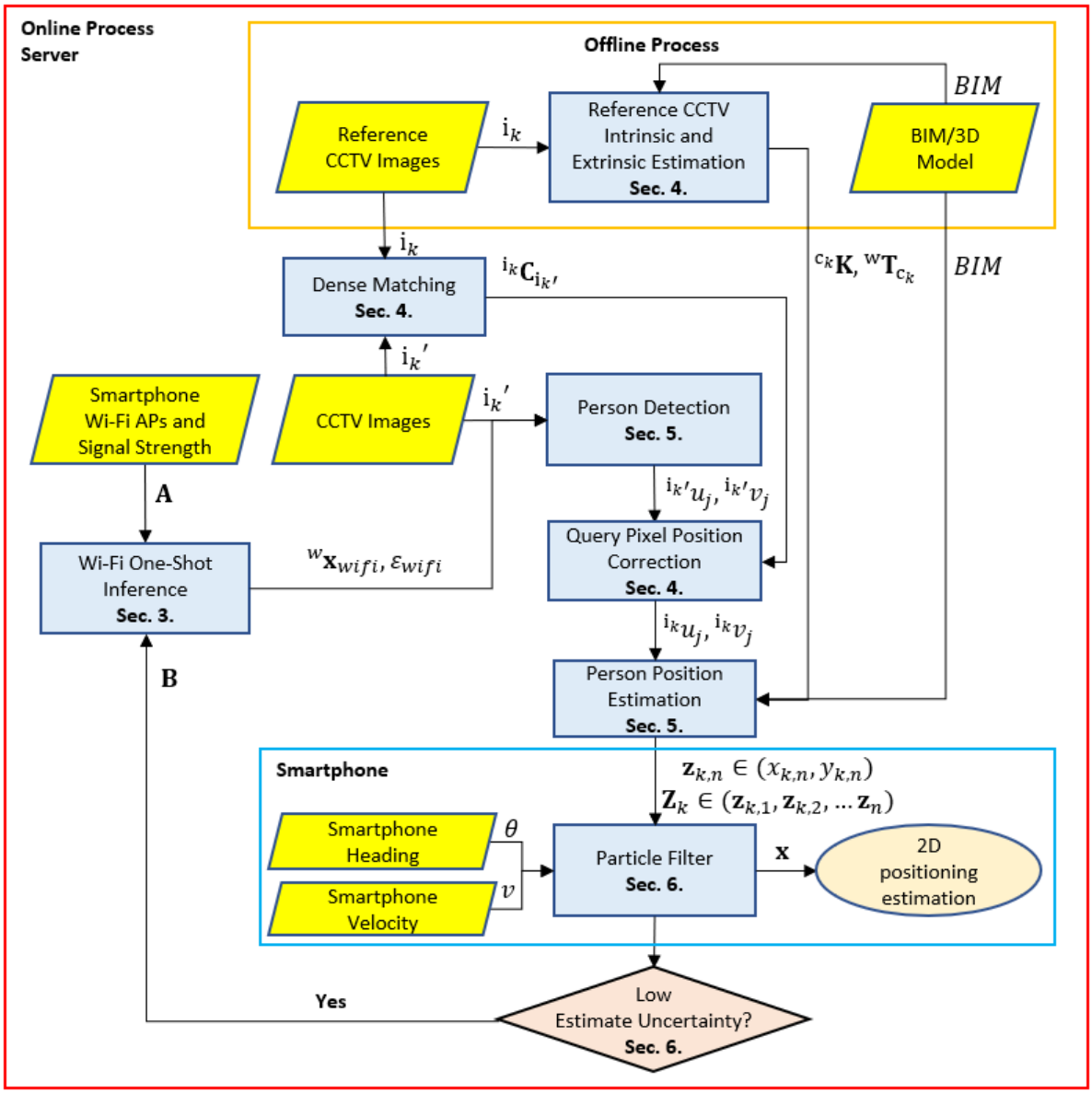

Figure 1: Overview of The Proposed Wi-Fi and CCTV Integrated Pedestrian Positioning System.

In the recent years, CCTV cameras have matured to a stage where it is playing an important role in monitoring and security [6]. These CCTV cameras can capture real-time information of the environment, which provides a reliable basis for setting up a digital twin (DT) because they can integrate information ranging from geometric changes in the building layout to the occupancy and use of rooms and spaces. DT for buildings can be seen as an extension to capture real-world data and feed it back into a 3D model, thus neatly closing the information loop [7-9]. In addition, CCTV cameras also capture the position of people and assets, hence, a position of the objects relative to the camera (and therefore the global position) can be obtained. If the correspondence to the user is found, it can directly transmit the person's position to the person's smartphone. This position can also be used to collect real-time data for Wi-Fi fingerprinting. Therefore, the main objective of the study presented in this paper is to develop a Wi-Fi, CCTV and PDR integrated pedestrian positioning system that enables navigation of the indoor environment using in-built Wi-Fi access points (APs) and CCTV cameras.

The proposed Wi-Fi, CCTV and PDR integrated pedestrian positioning system attempts to make full use of the existing infrastructure of the indoor environment for positioning. The proposed method offers several major advantages over the existing Wi-Fi standalone positioning methods.

- Firstly, the integration of CCTV and PDR can provide 0.3m positioning accuracy, enough for pedestrian positioning purposes.
- Secondly, the positioning can be used to update the Wi-Fi one-shot support set, eliminating the need for continuous manual Wi-Fi fingerprinting collection.
- Thirdly, as the CCTV is integrated for positioning, it is less susceptible to environment changes and dynamic object movements, overcoming the poor positioning due to the change in Wi-Fi features.

2. The Proposed Indoor Positioning System

The flowchart of the proposed Wi-Fi and CCTV integrated pedestrian positioning system is shown in Fig. 1. The algorithm can be divided into the offline and online processes. The offline stage includes estimating the CCTV cameras' intrinsic parameter by a checkboard, and extrinsic parameter using perspective and point from the known pose of visible landmarks in the indoor environment and the BIM (Building Information Modelling model) in Sec. 4. In the online stage, the smartphone requests a position by performing a Wi-Fi scan, then sending the detected APs and signal strength to the server. The server calculates an initial position based on a trained one-shot learning model (Sec. 3). The CCTV cameras near the initial position is activated to detect people (Sec. 5). The query pixel position of each detected person classified as "holding a phone" are corrected (Sec. 4), then their 3D positions are calculated based on the inverse perspective projection (Sec. 5). Assuming there are multiple detected persons (candidate positions), the problem is reformulated to estimate the correct correspondence from a list of candidates. We used a particle filter (PF) to estimate the correct correspondence based on the smartphone heading and velocity (Sec. 6). The PF positioning solution will then be used to estimate the state, and if given a small state covariance, the position will be used to update the Wi-Fi few-shot support set (Sec. 6).

Several assumptions were made in the research:
- Firstly, the Wi-Fi one-shot model was pre-trained in another environment as describe in [10]. The research only updates the support set.
- Secondly, the research requires initial Wi-Fi fingerprints for the Wi-Fi on-shot model support set. However, it can eliminate the need for continuous Wi-Fi fingerprinting collection.
- Thirdly, the world coordinate system was assumed to be the BIM 3D cartesian coordinate system for ease of calculation. The calculated cartesian positions can be converted to WGS84.
- Fourthly, the BIM/3D model used to estimate the extrinsic parameter of the CCTV was assumed to be accurate and up to date.
- Fifthly, the query pixel position correction assumes the query image shares the approximate same scene with the reference image at different viewpoints.
- Sixthly, the pedestrian that request positioning are holding a smartphone for the deep learning model to classify pedestrians "holding a smartphone".
- Seventhly, the calculated position of pedestrian is assumed to be the bottom centre pixel of the detected bounding box.
- Eighthly, the smartphone heading estimation is assumed to provide within ±90° uncertainty.

The paper is organized as follow: Section 3 introduces the Wi-Fi one-shot positioning training and inference. Section 4 introduces the BIM and CCTV cameras' intrinsic and extrinsic parameters estimation. Section 5 presents the person detection and position estimation. Section 6 presents the PF. Section 7 tests the proposed method with data obtained in a laboratory. Finally, conclusions and future perspectives are presented in Sections 8 and 9, respectively.

3. Wi-Fi One-shot Positioning Training and Inference

A key challenge to machine learning-based fingerprinting approaches for wireless indoor localization is data labeling [10]. This includes the need to collect timely labeled data in one environment as the

environment dynamics may change over time (e.g., for real-time localization), and the need to collect a new set of labeled data for localization in each new environment (e.g., for multi-environment localization). This time-sensitive and environment-dependent nature of wireless data incurs significant data collection and maintenance cost for supervised learning which requires a large amount of labeled data. The recently introduced few-shot transfer learning overcomes this problem by reformulating classification as a similarity [11]. More specifically, a graph neural network (GNN) model can be trained to measure the similarity of a query data to each of the n classes in the support set. Inspired by the recent advance in few-shot learning, the proposed Wi-Fi, CCTV and PDR integrated pedestrian positioning system attempts to eliminate the need for data collection by passively collecting Wi-Fi fingerprints from users, while obtaining the approximate ground truth location estimated from the PF (detailed in Section 6). The estimated position and uncertainty using Wi-Fi signals are detailed in [10] and described as Eq. (1) in this paper.

$$ {}^{w}\mathbf{x}_{wifi}, \varepsilon_{wifi} = OneShot(\mathbf{A}, \mathbf{B}) \tag{1} $$

Where ${}^{w}\mathbf{x}_{wifi}$ is the estimated Wi-Fi 2D state (x, y) in the BIM 3D cartesian coordinate system, and ε_{wifi} is the estimated uncertainty. **A** refers to the query data, a list of detected access points and signal strength from an unknown location. **B** refers to the support set, where each class corresponds to one known location, and each class has a list of detected access points and signal strength at that known location. $OneShot$ is the function that accepts a query data and calculates the similarity to each class in the support set, the class with the highest similarity is ${}^{w}\mathbf{x}_{wifi}$.

Once an accurate position is estimated from the PF (Sec. 6), the collected Wi-Fi fingerprints will be used to update the support set (**B**) of the Wi-Fi one-shot model.

4. CCTV Intrinsic and Extrinsic Estimation

To calculate the position of pedestrians based on the cameras, the intrinsic and extrinsic parameters of the cameras must be known. To estimate the intrinsic parameter (**K**), a checkerboard can be used to calibrate each CCTV camera. To estimate the extrinsic parameter (**T**), we made use of visible landmarks in the indoor environment to infer the pose of the camera. A BIM model stores information of objects in the indoor environment, including the objects' class, shape, appearance, size and pose in the world coordinate frame [12]. Therefore, we can use perspective and point from the BIM to estimate the pose of the CCTV camera. Once this is complete, we can obtain an intrinsic and extrinsic associated reference image for each CCTV camera as described in Eq. 2.

$$ \begin{aligned} {}^{c_k}\mathbf{x}_j &= \left({}^{w}\mathbf{T}_{c_k}\right)^{-1} \cdot {}^{w}\mathbf{x}_j \\ {}^{i_k}\mathbf{p}_j &= {}^{c_k}\mathbf{K} \cdot {}^{c_k}\mathbf{x}_j \end{aligned} \tag{2} $$

Where w and c is the world and camera coordinate frame where each coordinate is expressed as a 3D cartesian vector ${}^{c_k}\mathbf{x}, {}^{w}\mathbf{x} \in (x, y, z)$. i is the image coordinate frame, where each coordinate is expressed a pixel vector $\mathbf{p} \in (u, v)$. ${}^{c_k}\mathbf{K}$ is the 3x3 intrinsic matrix of camera k. ${}^{w}\mathbf{T}_{c_k}$ is the 4x4 extrinsic matrix of camera k in the world coordinate frame. We can therefore take a point j in the world frame ${}^{w}\mathbf{x}_j$ and express it as a pixel vector ${}^{i_k}\mathbf{p}_j$ in the image frame (i_k) for camera k. The expanded equation is expressed in Eq. (3).

$$ \begin{aligned} \begin{bmatrix} {}^{c_k}x_j \\ {}^{c_k}y_j \\ {}^{c_k}z_j \\ 1 \end{bmatrix} &= \left({}^{w}\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{c_k} \right)^{-1} \begin{bmatrix} {}^{w}x_j \\ {}^{w}y_j \\ {}^{w}z_j \\ 1 \end{bmatrix} \\ {}^{c_k}z_j \begin{bmatrix} {}^{i_k}u_j \\ {}^{i_k}v_j \\ 1 \end{bmatrix} &= {}^{c_k}\begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} {}^{c_k}x_j \\ {}^{c_k}y_j \\ {}^{c_k}z_j \end{bmatrix} \end{aligned} \tag{3} $$

It is assumed that each camera's intrinsic and extrinsic parameter may change overtime. The former can be caused by optical zoom, whereas the latter from camera tilt and panning. We denote c_k as the

reference camera, and the query (changed) camera as $c_k{}'$. The query camera's intrinsic and extrinsic parameters are difficult to estimate simultaneously from the reference camera due to many unknowns, leading to local minima [13]. However, in this paper, the position of the pedestrian is assumed and calculated based on the 3D position of a pixel, as described in Sec. 5.

Therefore, rather than estimating the query camera parameters to calculate the 3D position of a pixel, we estimate the pixel-to-pixel correspondence between the new and reference image. Since the 3D position is known for each pixel in the reference image, once a pixel is detected in the query image, it can correspond to a pixel in the reference image to estimate its 3D position. State-of-the-art dense matching was used to estimate the pixel-to-pixel correspondence as described in [14] and shown in Fig. 2. We assume the Percentage of Correct Key-points (PCK) is above 70% as evaluated in [14].

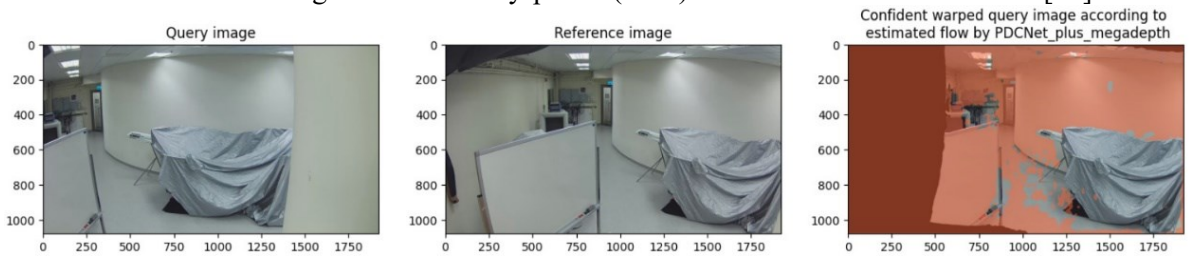

Figure 2: Example of Using a Dense Matching to Match Pixels of Query Image with Reference Image.

The detailed equation can be found in [14] and described as Eq. (4) in this paper.

$$ {}^{i_k}\mathbf{C}_{i_k'} = DenseMatching(i_k, i_k') \tag{4}$$

$$\begin{bmatrix} {}^{i_k}u_j \\ {}^{i_k}v_j \\ 1 \end{bmatrix} = {}^{i_k}\mathbf{C}_{i_k'} \cdot \begin{bmatrix} {}^{i_k'}u_j \\ {}^{i_k'}v_j \\ 1 \end{bmatrix}$$

Where ${}^{i_k}\mathbf{C}_{i_k'}$ is the correspondence matrix to transform pixel coordinate in new image coordinate frame $i_k{}'$ to reference image coordinate frame i_k.

5. Person Detection and Position Estimation

In the online process, pedestrians are first detected using a primary PeopleNet network [15, 16]. To reduce the number of candidate positions (hypothesizes), we also assumed pedestrians that request positioning are holding a smartphone. Therefore, we employed a secondary classification model, that classifies a person either "holding a phone" or "not holding a phone" after being detected. Given a person is detected and is "holding a phone", we calculate the person's position using the center bottom pixel of the bounding box. It is assumed that the detection accuracy is 80% or above as stated in [15].

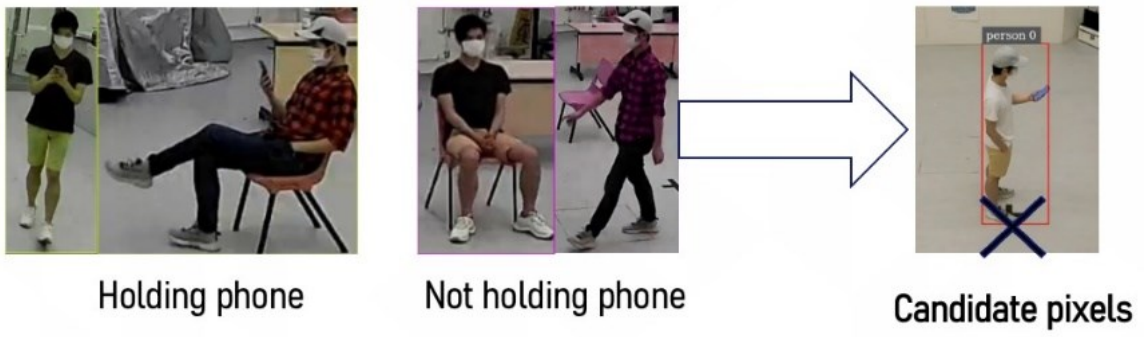

Figure 3: Classification and Pedestrian Detection.

The 3D position of the pixel is the intersection between the ray emanating from the normalized pixel and the 3D ground plane. We define a plane in Eq. (5).

$$ax + by + cz + d = 0 \tag{5}$$

Since the world coordinate ground plane is known from the BIM model, we can transform it into the camera coordinate frame in Eq. 6.

$$\begin{bmatrix} {}^{c_k}a_g \\ {}^{c_k}b_g \\ {}^{c_k}c_g \\ {}^{c_k}d_g \end{bmatrix} = \left(\left(\left({}^{w}\mathbf{T}_{c_k}\right)^{-1}\right)^{T}\right)^{-1} \cdot \begin{bmatrix} {}^{w}a_g \\ {}^{w}b_g \\ {}^{w}c_g \\ {}^{w}d_g \end{bmatrix} \tag{6}$$

Where g denotes the ground plane. The candidate pixel is normalized by multiplying the inverse intrinsic parameters. We can express the normalized pixel in the camera coordinate frame in Eq. (7).

$$\begin{bmatrix} {}^{c_k}\hat{x}_j \\ {}^{c_k}\hat{y}_j \\ {}^{c_k}\hat{z}_j \end{bmatrix} = \left({}^{c_k}\mathbf{K}\right)^{-1} \cdot \begin{bmatrix} {}^{i_k}u_j \\ {}^{i_k}v_j \\ 1 \end{bmatrix} \tag{7}$$

The camera coordinate of the normalized pixel is ${}^{c_k}\hat{\mathbf{x}}_j \in ({}^{c_k}\hat{x}_j, {}^{c_k}\hat{y}_j, {}^{c_k}\hat{z}_j)$. The camera coordinate of the intersection between the normalized pixel and ground plane is calculated in Eq. (8).

$${}^{c_k}\mathbf{x}_j = t \cdot {}^{c_k}\hat{\mathbf{x}}_j; \quad t = \frac{-{}^{c_k}d_g}{{}^{c_k}a_g\,{}^{c_k}\hat{x}_j + {}^{c_k}b_g\,{}^{c_k}\hat{y}_j + {}^{c_k}c_g} \tag{8}$$

Then the camera coordinate of the intersection can be expressed in the world coordinate as Eq. (9).

$${}^{w}\mathbf{x}_{k,j} = {}^{w}\mathbf{T}_{c_k} \cdot {}^{c_k}\mathbf{x}_j \tag{9}$$

The calculated positions are then filtered based on the Wi-Fi initial guess search area ${}^{w}\mathbf{x}_{wifi}, \varepsilon_{wifi}$. In the remainder, as the positions are located on the ground plane, they will be expressed as a sequence of measurements in 2D Cartesian world coordinate frame as $\mathbf{z}_{k,n}$ for ease of notation in Eq. (10).

$$\begin{gathered} \mathbf{z}_{k,n} \in (x_{k,n}, y_{k,n}) \\ \mathbf{Z}_k \in (\mathbf{z}_{k,1}, \mathbf{z}_{k,2}, \ldots \mathbf{z}_n) \end{gathered} \tag{10}$$

Where k is the index of the camera. n is the index of the filtered positions from j that are within the Wi-Fi search area, and $\mathbf{Z}_k$ are the measurement inputs of camera k. $\mathbf{Z}_{1\ldots k}$ is the measurement inputs of all cameras.

6. Particle Filter

We formulated a recursive state estimation problem by using a PF. The goal is to estimate a 2D state vector $\mathbf{x}$. More specifically, the goal is to track the hidden state sequence $\{\mathbf{x}_t\}$ of a dynamical system, where t is a discrete time step. The process model that encodes prior knowledge on how the state $\mathbf{x}$ is expected to evolve over time can be written as Eq. (11).

$$\begin{gathered} x_t = x_{t-1} + v_{t-1} \cdot \cos(\theta_{t-1}) \cdot dt + \epsilon_x \\ y_t = y_{t-1} + v_{t-1} \cdot \sin(\theta_{t-1}) \cdot dt + \epsilon_y \\ \mathbf{u} \in (v, \theta) \end{gathered} \tag{11}$$

Where v is the pedometer velocity and θ is the magnetometer heading from the client smartphone [17]. The cumulative uncertainty is assumed to be within ±0.2m every second. The measurements $\mathbf{Z}_{1\ldots k,t}$ is defined in Sec. 5. The particle filter is described in Alg. 1.

Algorithm 1 Particle Filter

Input: Prior particles set $\{S_{t-1} =< \mathbf{x}_{i,t-1}, w_{i,t-1} >, \mathbf{u}_t, \mathbf{Z}_{1\ldots k,t}\}$
Output: S_t
1. $S_t = \emptyset, \eta = 0$
2. For $j = 1 \ldots n$
3. Sample particles $i(j)$ from the discrete distribution given by w_{t-1}
4. Sample $\mathbf{x}_{j,t}$ from $p(\mathbf{x}_t|\mathbf{x}_{i(j),t-1}, \mathbf{u}_t)$
5. $w_{j,t} = p(\mathbf{Z}_{1\ldots k,t}|\mathbf{x}_{j,t})$
6. $\eta = \eta + w_{j,t}$
7. $S_t = S_t \cup \{< \mathbf{x}_{j,t}, w_{j,t} >\}$
8. For $j = 1 \ldots n$
9. $w_{j,t} = \frac{w_{j,t}}{n}$

Where S is a set of prior particles ($\mathbf{x}_{i,t-1}$) with a corresponding weight ($w_{i,t-1}$), input ($\mathbf{u}_t$) and measurements ($\mathbf{Z}_{1\ldots k,t}$). η is the normalization constant. Then, hidden state $\mathbf{x}_t$ is calculated based on the particle with the highest weight ($w_{j,t}$). To calculate the weight ($w_{j,t}$) of each particle, we need to understand the observation uncertainty ($\sigma_{k,n,t}$) using the known height of the camera (${}^{c_k\prime}z_t$) and measured 3D distance from the camera to the detected person $d({}^{c_k\prime}\mathbf{x}_t, \mathbf{z}_{k,n,t})$. A Monte-Carlo simulation was used to sample the correlation.

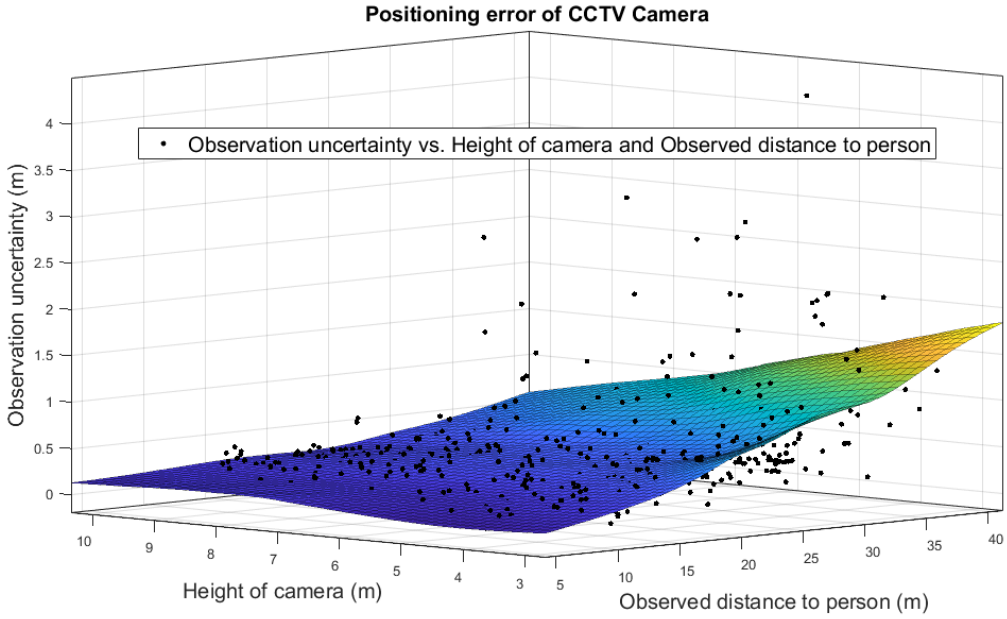

Figure 4: Observation Uncertainty of CCTV Positioning System and Fit for Eq. 12.

A lowess (linear) smoothing regression was used to fit the data samples. It can be seen that when the height of camera is low and measured distance increases, the positioning error also increases. This is likely due to the pixels representing more distance when the height of camera is low. The fit has the following properties:

Table 1
Lowess Fit Properties of Observation Uncertainty of CCTV Positioning System

Fit Type	SSE (m^2)	R-square	DFE	Adj R-sq	RMSE (m)	Coeff
Lowess	78.77	0.3481	694	0.3397	0.3369	10

The fit was then used to estimate the positioning error in real-time for each detected person in Eq. 12.

$$\sigma_{k,n,t} = f({}^{c_k}z_t, d({}^{c_k}\mathbf{x}_t, \mathbf{z}_{k,n,t})) \tag{12}$$

We assumed a multivariable gaussian distribution of the positioning errors, where the covariance matrix variance is equal to $\sigma_{k,n,t}$ to generate a likelihood for a given position in Eq. 13.

$$\mu_{k,n,t} = [x_{k,n,t}, y_{k,n,t}]; \boldsymbol{\Sigma}_{k,n,t} = \begin{bmatrix} \sigma_{k,n,t} & 0 \\ 0 & \sigma_{k,n,t} \end{bmatrix} \tag{13}$$

$$\ell_{\mathbf{x},k,n,t} = \frac{1}{\sqrt{|\boldsymbol{\Sigma}_{k,n,t}|(2\pi)^2}} exp\left(-\frac{1}{2}(\mathbf{x}-\mu_{k,n,t})\boldsymbol{\Sigma}_{k,n,t}{}^{-1}(\mathbf{x}-\mu_{k,n,t})^{\mathrm{T}}\right)$$

Where $\mu_{k,n,t}$ is the mean and $\boldsymbol{\Sigma}_{k,n,t}$ is the covariance matrix of the multivariable gaussian distribution. $\ell_{\mathbf{x},k,n,t}$ is the calculated likelihood of a given position $\mathbf{x}$ from camera (k) and person (n). All distribution were compared such that the highest likelihood at each position was used to combine into a global distribution, then normalized such that the total area under the distribution is equal to 1. It is important to note that there is no minimum distance required between two pedestrians as each pedestrian is

represented as a distribution. The final weighting ($w_{j,t}$) in Alg. 1 of a given position can be calculated based on Alg. 2.

Algorithm 2 Calculation of $w_{j,t}$

Input: $\mathbf{Z}_{1\ldots k,t}, \mathbf{x}_{j,t}$

Output: $w_{j,t}$

1. Calculate $\ell_{\mathbf{x},1\ldots k,1\ldots n,t}$ for every position
2. Find the maximum $\ell_{\mathbf{x},1\ldots k,1\ldots n,t}$ for every position and create new distribution such that $\ell_{\mathbf{x},t}^{max}$ for every position
3. Normalize distribution such that $\hat{\ell}_{\mathbf{x},t}^{max}$ for every position
4. $w_{j,t} = \hat{\ell}_{\mathbf{x}_j,t}^{max}$ for specific position index (j)

The weighting calculation is visualized in Fig. 5. Where the likelihood distribution of each person is combined into a normalized global distribution.

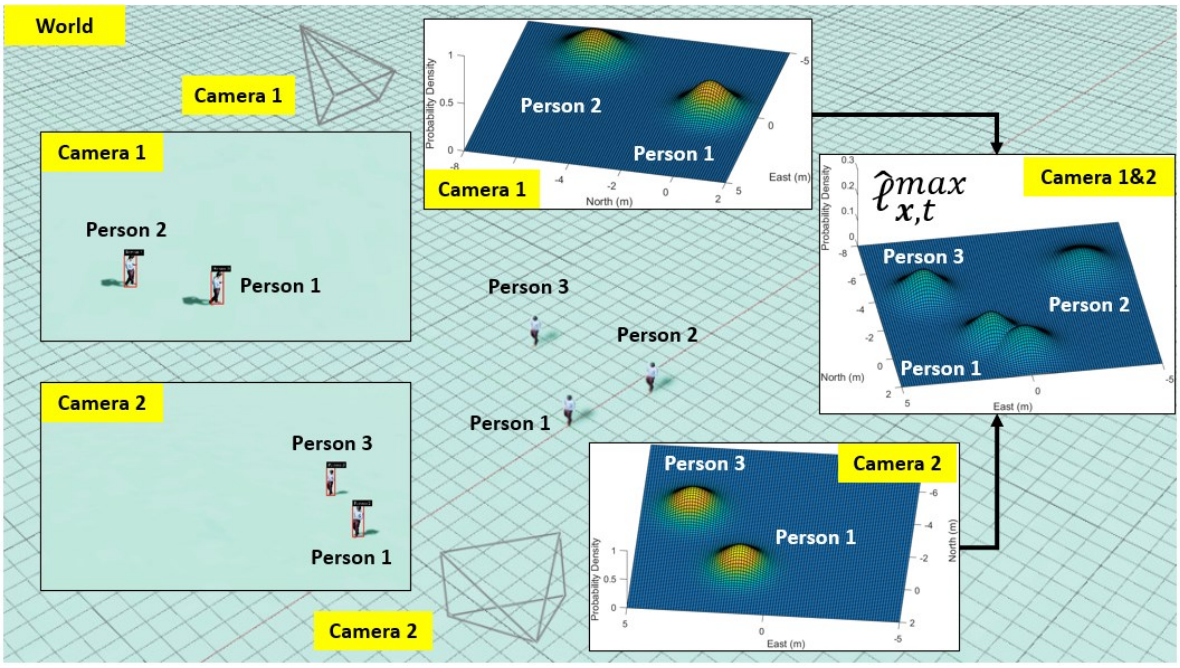

Figure 5: Example of A Likelihood Function Heatmap of 3 Pedestrians Captured by Two CCTV Cameras.

The particle with the highest weighting as described in algorithm 1 will be the PF estimated state $\boldsymbol{x}_t$. Given that the particle estimated state covariance is less than 1m, it will be used to update the support set described in Sec. 3.

7. Experimental Results

In this study, the experimental trajectory was conducted in a laboratory as shown in Fig. 8. The laboratory is a small indoor environment which contains numerous dynamic objects. Three commercial CCTV cameras are placed across the indoor environment to test the feasibility of the proposed positioning system. The ground truth positions of the trajectory were recorded, and the positioning quality of the proposed method was analyzed with other positioning methods, including:

1. Ground Truth, provided by markers labelled on the floor and timestamp from the CCTV camera.
2. Pedestrian Dead Reckoning (PDR), provided by Samsung Galaxy Note 20 smartphone and PDR app in [17].
3. CCTV, provided by commercial Tapo smart cameras [18].
4. Proposed Particle Filter (CCTV & PDR).

5. Wi-Fi Extended Naïve Bayes Positioning (ML) [19].
6. Wi-Fi One-Shot Positioning (One-Shot) [10].
7. Particle Filter (Wi-Fi ML & PDR).
8. Particle Filter (Wi-Fi One-Shot & PDR).

Methods 5, 6 are obtained through a succession of fixed points, whereas the other methods integrated with PDR are obtained during a movement.

To compare with traditional Wi-Fi ML positioning, Wi-Fi data were collected one month in advance at each location with 2-meter separation as shown in Fig. 6. 200 data were collected at each location to train an Extended Naive Bayes positioning model to classify new Wi-Fi fingerprints to a location. To compare with the Wi-Fi One-Shot positioning, two Samsung Galaxy smartphones were used to collect Wi-Fi data simultaneously during the experiment, where one smartphone collects data for the support set input, and the other phone will act as query input for the one-shot model.

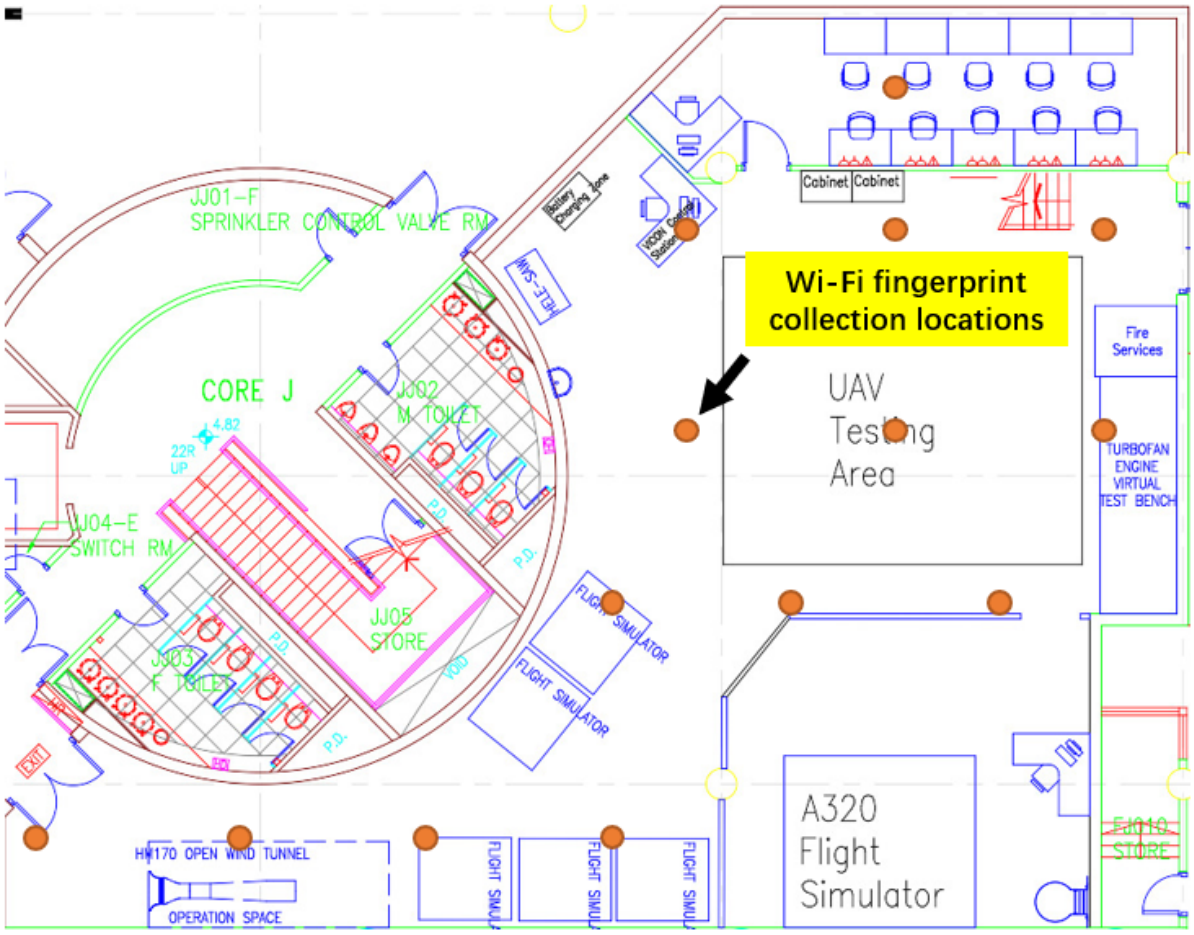

Figure 6: Positions That Collected Wi-Fi Signals to Train A Wi-Fi Extended Naïve Bayes Model.

In addition, to mimic a real indoor environment with multiple pedestrians, in this study we have a pedestrian walk around the client to create a false-positive dynamic likelihood to the CCTV measurements as shown in Fig. 7. A false-positive static likelihood at a single location was also added post-experiment to the CCTV measurement. The former represents a pedestrian walking, and the latter represents a pedestrian standing, commonly seen in all indoor environments. This was performed to test whether the correct correspondence to the client can be found using the proposed method.

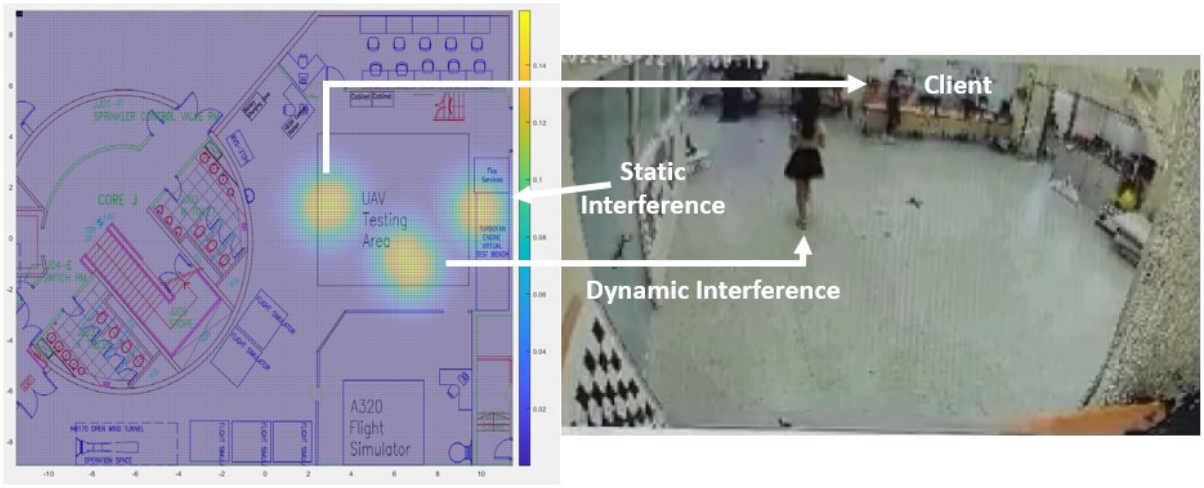

Figure 7: Measurement Likelihood with Dynamic and A Mimic Static Interference.

The positioning results are plotted onto a bird's-eye view of the laboratory in Fig. 8. There are two performance metrics used: mean and standard deviation (SD) of the 2D positioning error.

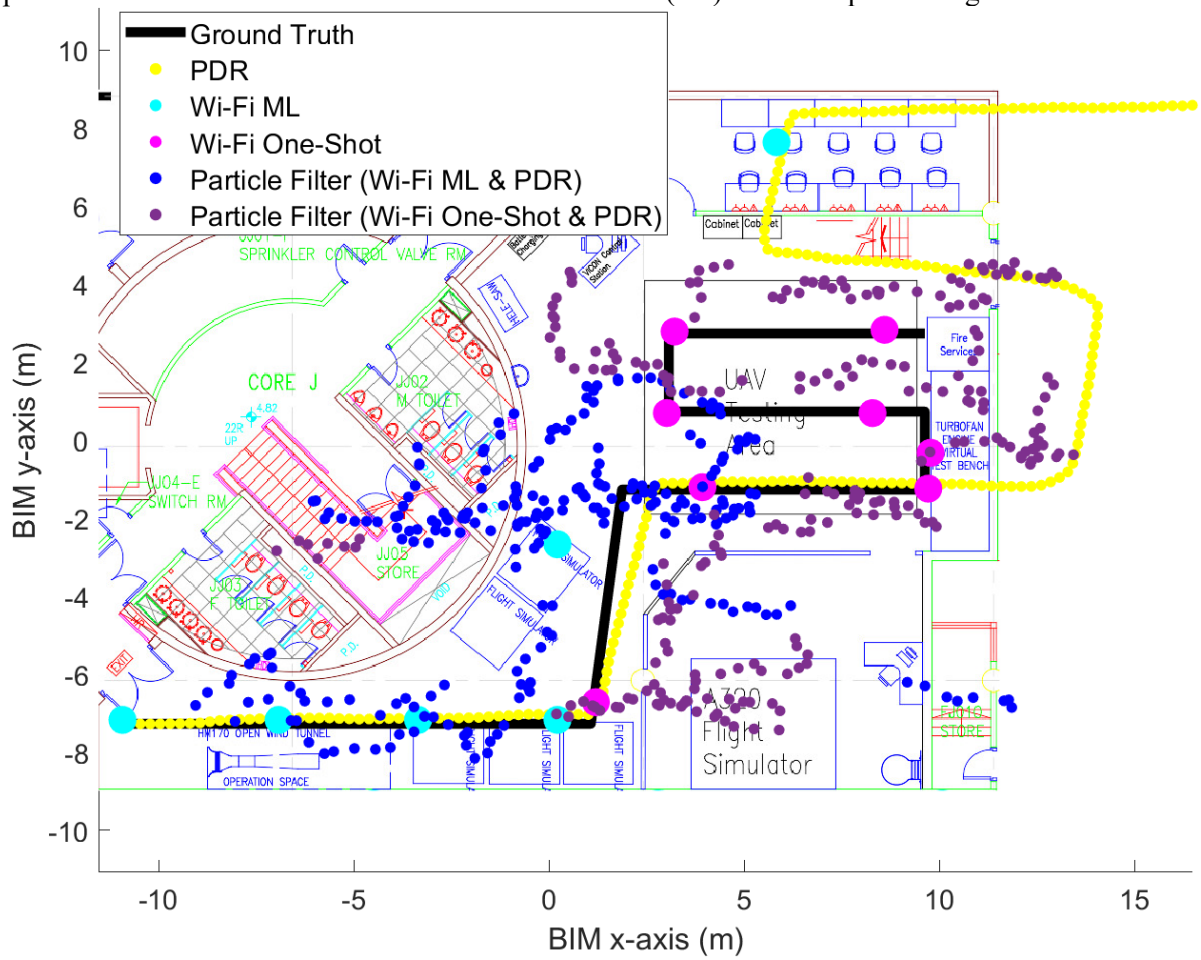

Figure 8: Positioning Results of PDR and Wi-Fi Based Positioning Methods.

The PDR requires an initial position and was set to the ground truth initial position. The results show that the PDR (yellow marker) begins with great positioning accuracy but starts to accumulate drift error as it progresses as shown after time epoch 20. This has led to a significant cumulative positioning mean error of 2.4m. Therefore, it needs to be integrated with an absolute positioning measurement to correct its cumulative errors. The Wi-Fi Extended Naïve Bayes point positioning (cyan marker) is on average 3.81m to the ground truth (black marker) as analyzed in Figs. 8, 9 and Table II. This is mainly due to

the change in the physical environment from dynamic objects, which are common in many indoor environments, thus creating NLOS and multipath signals different from the signals captured previously at the location. The integration of the Wi-Fi Extended Naïve Bayes and PDR (blue marker) is relatively accurate from time 0 to 14 seconds was perhaps due to the more unique Wi-Fi signatures then compared to the open space counterpart at time 20 seconds and after, leading to 3.64m positioning inaccuracy. The collected Wi-Fi signals from the CCTV & PDR PF, were then used as the support set of the Wi-Fi One-Shot positioning, improving the accuracy to 1.89m mean error and 1.56 standard deviation. It was then integrated with PDR to provide 1.98m positioning accuracy with a smaller 1.11m standard deviation.

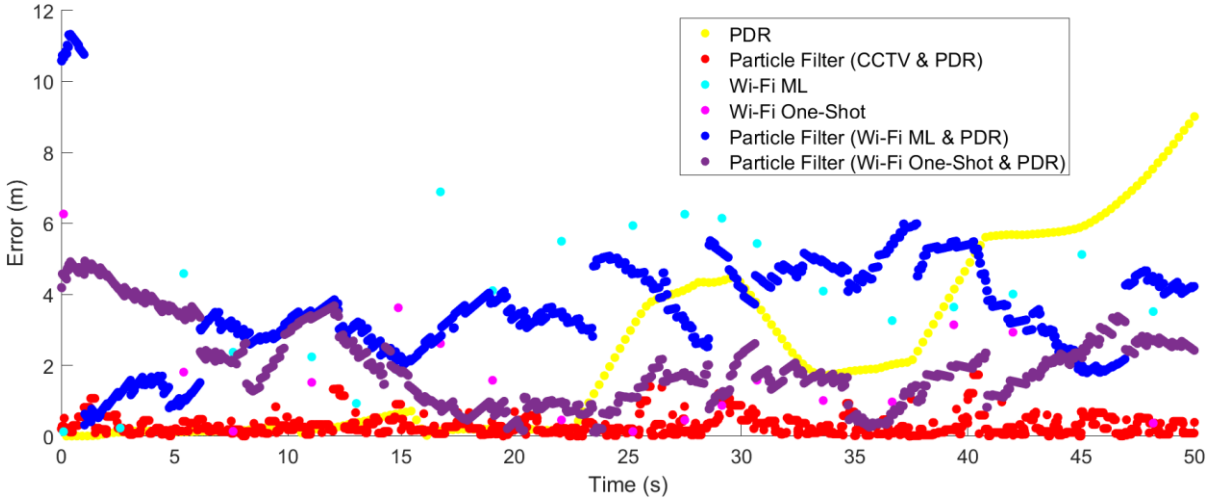

Figure 9: Positioning Error of The Proposed Positioning System and Other Positioning Systems.

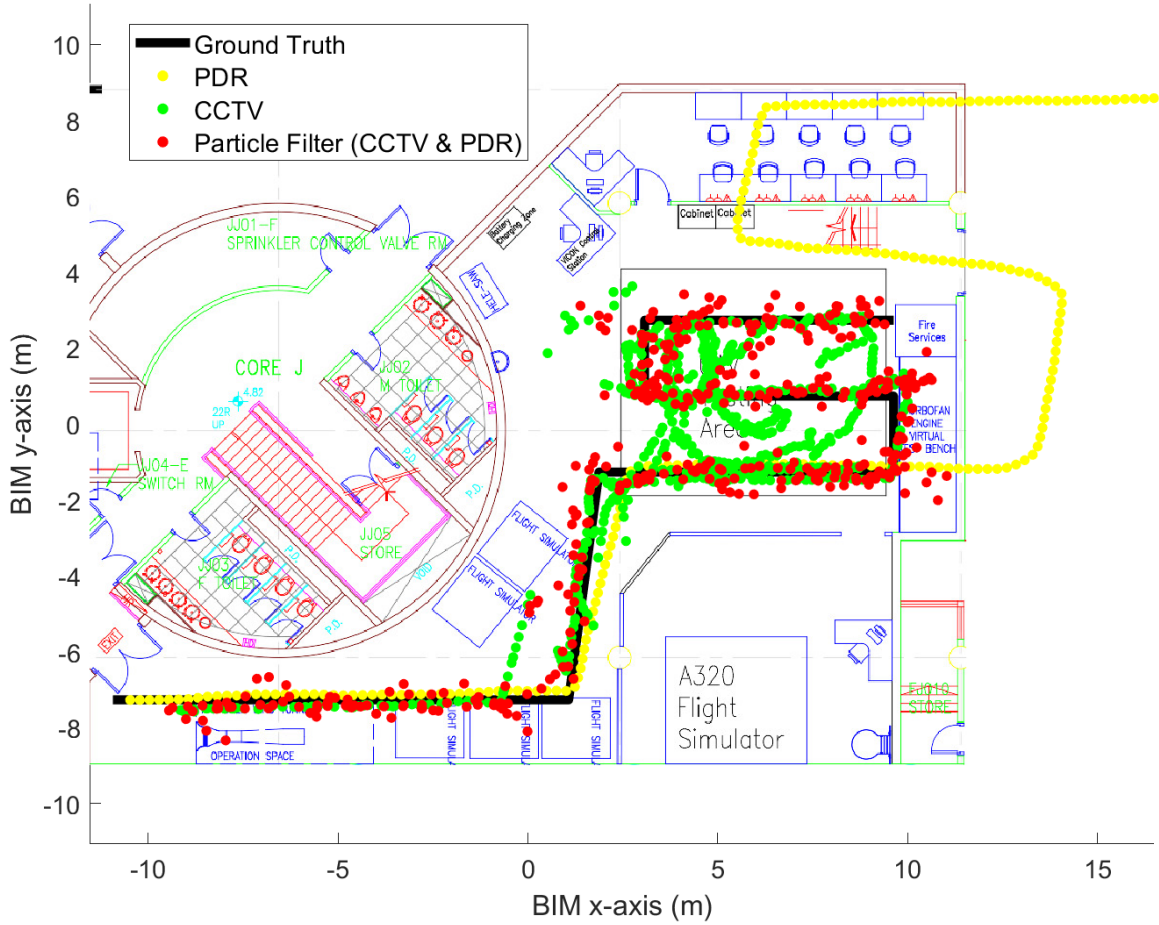

Figure 10: Positioning Results of The Proposed Positioning System and Other Positioning Systems.

The CCTV positioning solution is shown in the green marker, where a correct correspondence is required for positioning. The proposed method (red marker) integrates the PDR with CCTV positioning solution via PF. Consequently, the PDR can estimate the correspondence for the CCTV, whereas the CCTV can correct the positioning error of the PDR. The results of the proposed method provide an accurate and continuous positioning of 0.29m mean error. The collected Wi-Fi fingerprints were then used as the support set of the Wi-Fi One-Shot positioning, improving the accuracy to 1.89m.

Table 2
Accuracy of The Compared Positioning Systems

Method	2D position error	
	Mean (m)	SD (m)
PDR	2.40	2.50
PF (CCTV & PDR)	**0.29**	**0.27**
PF (Wi-Fi ML & PDR)	3.64	1.62
PF (Wi-Fi One-Shot & PDR)	1.98	1.11
Wi-Fi ML	3.81	2.01
Wi-Fi One-Shot	1.89	1.56

8. Conclusions

This paper proposes a Wi-Fi and CCTV integrated indoor positioning solution for a 2D position estimation. In short, pedestrians were detected, and a particle filter was used to estimate the correct correspondence to the user. The estimated state can then be used to update a Wi-Fi one-shot model. The potential advantages of the proposed method are:

- The integration of Wi-Fi, CCTV and PDR improves positioning accuracy compared to Wi-Fi standalone positioning.
- The Wi-Fi one-shot positioning can improve upon the traditional Wi-Fi positioning due to the use of real-time Wi-Fi data.

In this paper, a laboratory indoor environment was used to test the proposed method, however, in the real world there would be additional challenges that needs to be addressed including:

- Up-to-date BIM model that reflects the building to calculate the accurate extrinsic parameters of the CCTV cameras.
- Areas with no Wi-Fi or CCTV coverage.
- True negatives detection of pedestrians, in other words failure to detect pedestrians.

Considering the preliminary results presented in this paper, we believe the proposed method can provide accurate positioning and to support various indoor applications, that can be extended to digital twin.

9. Future Works

We will conduct a real-time experiment in a large-scale mall in Hong Kong to validate the proposed method. Several potential future developments on the proposed method are suggested.

- To make use of state-of-the-art visual trackers that can provide higher weighting to correspondences.
- To make use of 3D object detection algorithms that are more robust to positioning estimation compared to 2D object detection [20].

References

[1] A. Hameed and H. A. Ahmed, "Survey on indoor positioning applications based on different technologies," in *2018 12th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, 24-25 Nov. 2018 2018, pp. 1-5, doi: 10.1109/MACS.2018.8628462.

[2] V. Renaudin *et al.*, "Evaluating Indoor Positioning Systems in a Shopping Mall: The Lessons Learned From the IPIN 2018 Competition," *IEEE Access,* vol. 7, pp. 148594-148628, 2019, doi: 10.1109/ACCESS.2019.2944389.

[3] M. Xue *et al.*, "Locate the Mobile Device by Enhancing the WiFi-Based Indoor Localization Model," *IEEE Internet of Things Journal,* vol. 6, no. 5, pp. 8792-8803, 2019, doi: 10.1109/JIOT.2019.2923433.

[4] Q. Yang, S. Zheng, M. Liu, and Y. Zhang, "Research on Wi-Fi indoor positioning in a smart exhibition hall based on received signal strength indication," *EURASIP Journal on Wireless Communications and Networking,* vol. 2019, no. 1, p. 275, 2019/12/17 2019, doi: 10.1186/s13638-019-1601-3.

[5] E. R. Magsino, I. W. H. Ho, and Z. Situ, "The effects of dynamic environment on channel frequency response-based indoor positioning," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 8-13 Oct. 2017 2017, pp. 1-6, doi: 10.1109/PIMRC.2017.8292442.

[6] D. Acharya, K. Khoshelham, and S. Winter, *Real-time detection and tracking of pedestrians in CCTV images using a deep convolutional neural network*. 2017.

[7] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the Digital Twin: A systematic literature review," *CIRP Journal of Manufacturing Science and Technology,* vol. 29, pp. 36-52, 2020/05/01/ 2020, doi: https://doi.org/10.1016/j.cirpj.2020.02.002.

[8] M. J. L. H. H. Y. H. L.-T. A. S. L. M. Lee, "BIPS: Building Information Positioning System," presented at the In Proceedings of the 2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Online, 2021.

[9] M. J. L. Lee, L.-T. Hsu, H.-F. Ng, and S. Lee, "Semantic-Based VPS for Smartphone Localization in Challenging Urban Environments," 2020. [Online]. Available: http://arXiv.org/abs/
https://arxiv.org/ct?url=https%3A%2F%2Fdx.doi.org%2F10.1109%2FTRO.2021.3075644&v=a8cc9408.

[10] B.-J. Chen and R. Y. Chang, "Few-Shot Transfer Learning for Device-Free Fingerprinting Indoor Localization," 2022. [Online]. Available: http://arXiv.org/abs/.

[11] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning," 2016. [Online]. Available: http://arXiv.org/abs/.

[12] S. Alizadehsalehi, A. Hadavi, and J. C. Huang, "From BIM to extended reality in AEC industry," *Automation in Construction,* vol. 116, p. 103254, 2020/08/01/ 2020, doi: https://doi.org/10.1016/j.autcon.2020.103254.

[13] S. Tang, C. Tang, R. Huang, S. Zhu, and P. Tan, "Learning Camera Localization via Dense Scene Matching," 2021. [Online]. Available: http://arXiv.org/abs/.

[14] P. Truong, M. Danelljan, L. V. Gool, and R. Timofte, "Learning Accurate Dense Correspondences and When to Trust Them," 2021. [Online]. Available: http://arXiv.org/abs/.

[15] NVIDIA. "PeopleNet." https://catalog.ngc.nvidia.com/orgs/nvidia/models/tlt_peoplenet (accessed 2022).

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020. [Online]. Available: http://arXiv.org/abs/.

[17] N. Patel. "Dead Reckoning, a location tracking app for Android smartphones." GitHub. https://github.com/nisargnp/DeadReckoning (accessed 2022).

[18] Tapo, "Pan/Tilt Home Security Wi-Fi Camera," 2022. [Online]. Available: https://www.tapo.com/en/product/smart-camera/tapo-c210/.

[19] Find3. "Framework for Internal Navigation and Discovery." https://github.com/schollz/find3 (accessed 2022).

[20] J. Ku, A. Pon, and S. Waslander, *Monocular 3D Object Detection Leveraging Accurate Proposals and Shape Reconstruction*. 2019.