Improvement of Person Tracking Accuracy in Camera Network by Fusing WiFi and Visual Information

Thi Thanh Thuy Pham
Academy of People Security, Hanoi, Vietnam
E-mail: thanh-thuy.pham@mica.edu.vn

Thi-Lan Le and Trung-Kien Dao
MICA International Research Institute, Hanoi University of Science and Technology
(HUST - CNRS/UMI-2954 - Grenoble INP), Hanoi, Vietnam
E-mail: {thi-lan.le, trung-kien.dao}@mica.edu.vn

Keywords: camera, WiFi, fusion method, person tracking by identification

Received: March 29, 2017

Person tracking in camera network is still an open subject nowadays. The main challenge for this problem is how to link exactly individual trajectories when people move in a camera FOV (Field of View) or switch to other ones. This refers to solve the problem of person re-identification (Re-ID) in tracking process. A popular method for this is assigning the current position with the previous one based on the minimum distance between them. This is called as person identification by tracking. In this work, we approach tracking by identification, which means the trajectory assignment is done by the person identity (ID) determined at each video frame. In order to improve the accuracy of vision-based person tracking, we focus on accuracy enhancement for person identification by adding ID of the WiFi-enabled device held by each person. A fusion scheme of WiFi and visual signals is proposed in this work for person tracking. An optimal assignment and Kalman filter are used in this combination to assign the position observations and predicted states from camera and WiFi systems. The correction step of Kalman filter is then applied for each tracker to give out state estimations of locations. The fusion method allows tracking by identification in non-overlapping cameras, with clear identity information taken from WiFi adapter. The evaluation on a multi-model dataset show outperforming tracking results of the proposed fusion method in comparison with vision-based only method.

Povzetek: Opisana je metoda sledenja osebam preko kamer s pomočjo zlivanja podatkov.

1 Introduction

There have been several attempts to combine camera and WiFi systems for indoor person tracking. A multi-modal system is reported in [1] using WiFi-based localization and tracking by stationary cameras. The combined system focuses on improving the positioning accuracy and confidence at room level. According to the authors' assessments, camera-based localization achieves higher positioning accuracy than WiFi-based system. However, blind points, occlusions and person identification are much more challenging for camera systems. WiFi systems give clearer identity information because each mobile device has a unique MAC address, but considered targets are required to hold mobile devices during tracking. In this work, RSSI property and fingerprinting method are used in WiFi system to locate mobile targets. In camera-based system, foreground segmentation is done by GMM (Gaussian Mixture Model) method. The region which contains person feet is then extracted from foreground and projected on the floor plane. Gaussian kernels are used to model the foot region. Each single module is executed depending on the availability of each sensor information. When both of them appear, a combined Bayes model with the corresponding confidence weights is done.

The authors in [2] reported another approach for object localization fusing images and WiFi signals. The system can be deployed in both indoor and outdoor environments. The algorithm of PlaceEngine [3] and the modified version of the Centroid algorithm [4] are used in this work for WiFi-based localization. The mixture of observation model based on Particle filter allows continuously track targets even in case they are occluded by other objects or temporarily disappear when moving in blind areas among disjoint cameras.

In [5], the authors proposed to combine RGB data with wireless signals emitted from a person's cell phone to locate and track individuals. The authors considered a unique MAC address of mobile device as a reliable cue of person's ID. Wireless data is efficiently embedded in RGB data as a ring image, which captures radius estimation, error bounds, and confidence level (noise detection) for each antenna. In

order to improve tracking algorithm, each MAC address is assigned to an observed tracklet and bipartite graph is proposed for data association problem. The testing results proved that performance of person localization and tracking can be improved by fusion RGB and wireless data.

In this paper, we propose a fusion method of WiFi and camera for person localization and Re-ID in a camera network. It allows to improve the vision-based person tracking in not only one camera FOV, but also among different camera FOVs by using the unique ID information from WiFi hardware.

The rest of paper is organized as follows. In Section II, a framework for multi-modal person tracking by fusion of WiFi and camera is presented. Section III and Section IV indicate each single person localization system based on visual and WiFi signals, respectively. A combined method of WiFi and camera is discussed in Section V. The comparative evaluations are shown in Section VI. Conclusion and future directions will be finally denoted in the last section.

2 Framework

Figure 1 shows the fusion framework for person localization and Re-ID in non-overlapping camera networks. The combined model is processed in the real scenario of a fully-automated person surveillance system, which is reported in our previous work [6].

In this system, the camera FOVs are covered by WiFi range. This means WiFi signals are always available for person localization, but disjointed camera shot areas cause intermittent positioning for vision-based system. In each camera FOV, person localization is done by three phases, i.e., human detection, tracking and localization to output person ID j by camera C (ID_j^C) and the corresponding position (P_j^C). Because WiFi range covers the camera FOVs, so in each camera FOV, the vision-based positioning result of person j will be combined with WiFi-based localization result of person i (P_i^W, ID_i^W) by a fusion algorithm in order to make effective decisions about position and identity of person in environments. When people switch from one camera FOV to another, they will be re-identified to update the ID for each individual trajectory. The trajectories through the cameras will be also linked to show the entire route in the environment. Addtionally, in the fusion model, WiFi-based localization results are used to activate the cameras which are in the positioning range returned by a WiFi-based system. The proposed mixture model allows to continuously localize and identify person moving in non-overlapping camera networks.

In the proposed system, the positioning processes are executed independently from each single model. The locations calculated from both models of WiFi and camera are shown on the uniform coordinate system of a 2D floor map. A fusion algorithm for person localization and Re-ID is proposed. It is based on Kalman filter model, together with an optimal assignment of estimated and observed locations from both models. The details for each single person localization system and the proposed fusion algorithm will be shown in the next sections.

3 Vision-based person localization and Re-ID

Camera-based person localization and Re-ID is a process of finding the positions and the corresponding ID of a person when he/she moves in one camera FOV or switches from one camera FOV to others in camera networks. It refers to linking person trajectories in the frame sequences captured from multiple cameras. These trajectories are then transformed to real-world coordinate system by a process called 3D localization.

3.1 Person localization

A camera-based person localization system includes three main steps of human detection, tracking and 3D localization. For each camera FOV, human detection is executed at each frame to output the human ROI (Region of Interest), which is presented by a rectangular bounding box containing the person. The person position on image is defined in this work as a middle point of the rectangle's bottom edge which has contact with the floor plane (see Figure 2). It is called a FootPoint position. Human tracking in a frame sequence captured from a camera FOV is considered as FootPoint tracking. In case of multi-person tracking, each detected FootPoint has to be assigned with the corresponding ID. 3D person localization is done by transforming FootPoint positions to real world locations on a predefined 2D coordinate system of the floor plane where the person moves.

First, a combination of HOG-SVM and GMM background subtraction techniques [6] is applied for human detection. In order to improve the performance of human detection, shadow removal method in [6] is used as a post-processing step for human detection.

Second, in each camera FOV, based on the detection results, FootPoint tracking is done by utilizing Kalman Filter and Hungarian data association algorithm [7] to improve the performance of track association. For each camera, a grid of the floor plane where people move in the camera FOV, namely detection grid (see Figure 3), is defined as a function $G(x, y)$:

$$G(x,y) = \begin{cases} 1 & \text{if } (x,y) \in C_T; \\ 0 & otherwise. \end{cases}$$

where C_T is a threshold region bounded by a contour line which is the border of camera FOV on the floor plane. As each detected person is represented by a FootPoint position, so a FootPoint position can belong to one of the positions of the detection grid where $G(x, y)$=1. Let (px_t, py_t) denote the pixel coordinates of a FootPoint position at time t in the grid, (mx_t, my_t) the pixel coordinates of a

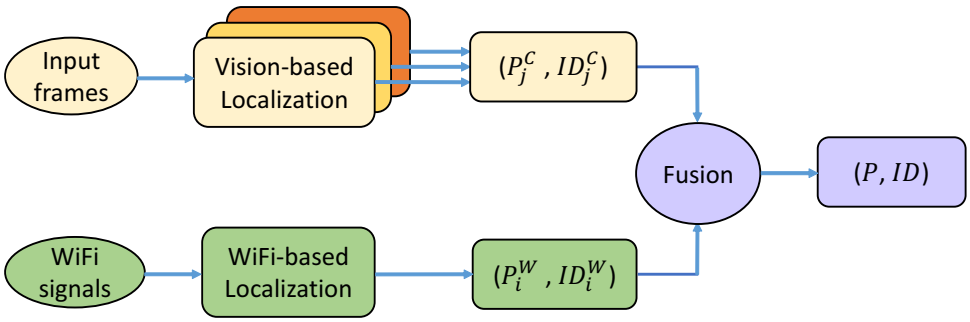


Figure 1: Framework of person localization and Re-ID using the combined system of WiFi and camera.

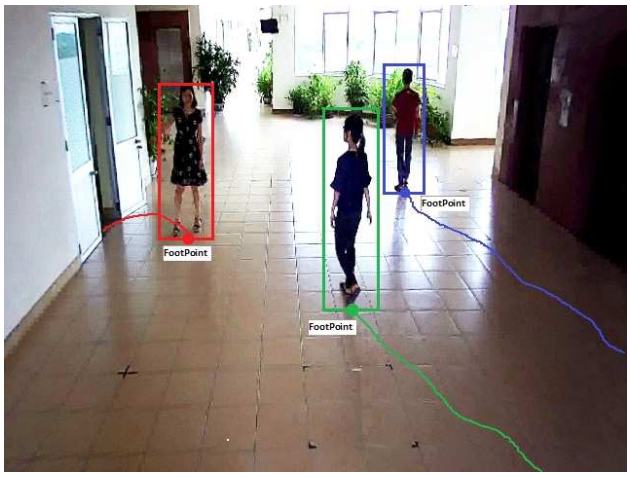


Figure 2: Examples of tracking lines which are formed by linking trajectories of corresponding FootPoint positions.

Figure 3: Example of a grid map and threshold region bounded by a contour line.

measurement in the grid, so that $G(mx_t, my_t) = 1$, and (vx_t, vy_t) velocity values at time t in x and y direction.

The state vector $\mathbf{x}_t$ of an user at time frame t can be characterized by the corresponding FootPoint location, and measurement vector $\mathbf{z}_t$ are defined as:

$$\mathbf{x}_t = (px_t, py_t, vx_t, vy_t) \quad (1)$$

$$\mathbf{z}_t = (mx_t, my_t) \quad (2)$$

Using the state and measurement update equations of Kalman filter, in conjunction with the initial conditions, at each time frame, the state vector and its covariance matrix are estimated. The 2D spatial coordinates of an estimated state $(\hat{p}_x, \hat{p}_y)$ (an estimated FootPoint position) refer to the position p of the user u.

In multi-person tracking, a separate Kalman filter is initialized and models each person's trajectory. A set $\mathcal{U}_t$ of individuals and a set $\mathcal{M}_t$ of measurements at time frame t are defined as:

$$\mathcal{U}_t = \{u_1, u_2, .., u_N\} \quad (3)$$

$$\mathcal{M}_t = \{m_1, m_2, .., m_L\} \quad (4)$$

with N is the number of people need to be tracked or trackers, and L is the number of available measurements at time t. In order to assign a person i to a measurement j, the Hungarian method is used.

Third, in order to locate people in real world coordinate system, we define a 2D map of the floor plane on which people move. This map contains all considered camera FOVs on the floor plane. We then calculate the coordinates of each FootPoint position on the 2D map on the basis of camera calibration and hormography transform [8]. The trajectories for each person through cameras are then linked by a method of wrapping multiple camera FOVs using a stereo calibration technique [9].

3.2 Person re-ID

In this paper, the person Re-ID problem is solved in the scenario of tracking by identification. This means that at each detected FootPoint position, we extract the human ROI, and a feature descriptor is built on this region. In this work, a robust KDES descriptor (Kernel Descriptor) which is proposed in our previous work [6], and an SVM classifier are used for person Re-ID in camera networks. The basic idea of KDES descriptor is to compute the approximate explicit feature map for kernel match function (see Figure 4). In other words, the kernel match functions are approximated

by explicit feature maps. This enables efficient learning methods for linear kernels to be applied to the non-linear kernels. Given a match kernel function $k(x, y)$, the feature map $\varphi(.)$ for the kernel $k(x, y)$ is a function mapping a vector $\mathbf{x}$ into a feature space so that $k(x, y) = \varphi(x)^\top \varphi(y)$. Given a set of basis vectors $B = \{\varphi(v_i)\}_{i=1}^{D}$, the approximation of feature map $\varphi(x)$ can be:

$$\phi(x) = Gk_B(x) \tag{5}$$

where $G^\top G = K_{BB}^{-1}$, K_{BB} is a $D \times D$ matrix with $\{K_{BB}\}_{ij} = k(v_i, v_j)$, and k_B is a $D \times 1$ vector with $\{k_B\}_i = k(x, v_i)$.

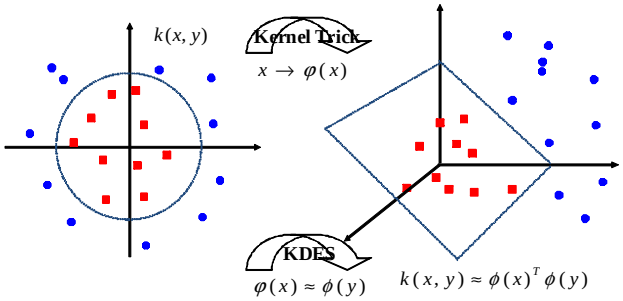


Figure 4: The basic idea of representation based on kernel methods.

Similar to [10], three match kernel functions for gradient, color and shape are built from different pixel attributes of gradient, color and local binary pattern (LBP). For each match kernel, feature extraction is done at three levels: pixel, patch and whole detected human region.

4 WiFi-based person localization

For WiFi, RSSI is the most popular attribute used in localization. However, the localization performance depends much on how well we can model the relationship between RSSI and the distance. Two main approaches have been proposed to solve this: pass-loss/radio propagation model [12, 13] and fingerprinting method [14]. The first one is still an open subject, because it is not easy to have an optimal model for relationship between RSSI and distance. The second one is time and workforce consuming but it is effective for localization, especially when the probabilistic methods are applied.

In this work, both of radio propagation model and fingerprinting method for WiFi-based localization are approached. A probabilistic propagation model (PPM) in [11], together with a new-defined radio map in fingerprinting database are used. The radio propagation model reflects the complex nature of indoor environments by taking into account the obstacles, such as walls and floors to model the relationship between RSSI value and the distance to a reference point (RP). The model is based on the empirical equation of radio-frequency signal strength in indoor environments and its uncertainty is considered by probabilistic characteristics. An optimization process based on genetic algorithm is also applied to tune system parameters for best fitting with the devices in use. Based on the probabilistic propagation model, the distance between a mobile user and APs is calculated. In fingerprinting database, a new radio map of distance features instead of RSSI values is defined in order to make the radio map more reliable and stable, with lower cost for setting and updating. Additionally, KNN matching method is applied with an additional coefficient reflecting temporal changes of fingerprinting data in environments. The flowchart of the proposed WiFi-based person localization system is illustrated in Figure 5, with two main phases of training and testing. The first phase is

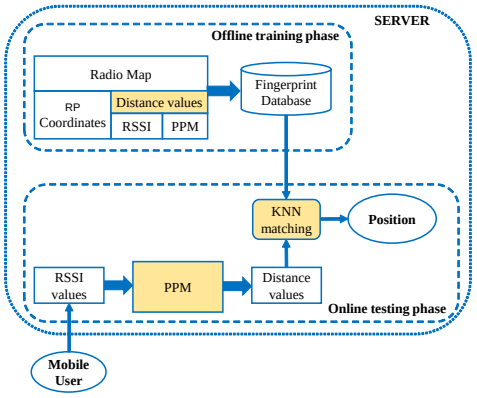


Figure 5: Diagram of the proposed WiFi-based object localization system.

processed off-line with radio maps are constructed to make fingerprint database. Normally, a radio map contains RP coordinates and corresponding RSSI values from available APs. However, in our proposed system, RSSI values are replaced by distance values. A distance value is defined as the distance $d_i(L)$ from the i^{th} RP to the L^{th} AP in range (see Figure 6) which is calculated from RSSI observations by using the PPM model. In the testing phase, a mobile de-

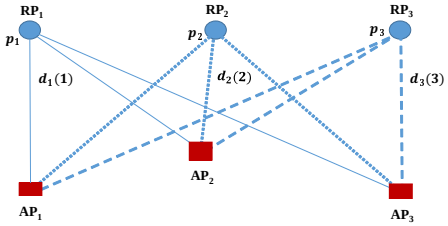


Figure 6: An example of radio map with a set of p_i RPs and the distance values $d_i(L)$ from each RP to L APs.

vice continuously scan signals from nearby APs and sends corresponding RSSI values to a server. These values are then transformed to distance values by a proposed probabilistic propagation model. Distance matchings are done with fingerprint database by methods of KNN to find the best candidates for mobile user location.

4.1 Probabilistic propagation model

The probabilistic propagation model which is formed by a deterministic model in Eq. 6 and a probabilistic model.

$$P = P_0 - 10nlog(\frac{r}{r_0}) - k_d \sum_{i=1}^{n_w} \frac{d_i}{cos\beta_i} \quad (6)$$

where n_w is the number of walls and floors in the middle of the AP and the receiver, d_i is the thickness of the i^{th} wall/floor, β_i is the angle of arrival corresponding to the i^{th} wall/floor, and k_d is an attenuation factor per wall/floor thickness unit, as illustrated in Figure 7.

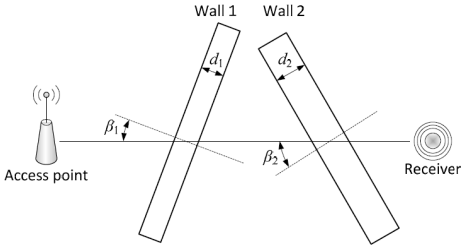


Figure 7: WiFi signal attenuation through walls/floors.

The deterministic model in Eq. 6 does not consider the uncertainty of RSSI values at a distance, so a probabilistic model (Eq. 7) is proposed. In reality, given RSSI P, the distance r might not be exactly the value calculated from Equation 6, but it is within a range around this value, which is denoted by $\bar{r}$. To be more precise, $\bar{r}$ will be the nominate value of the distance r with the highest probability. Given a RSSI P, the distribution of the distance is assumed to follow a normal (or Gaussian) distribution with median $\bar{r}$:

$$\rho(r, P) = P_r(r|P) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(r-\bar{r})^2}{2\sigma^2}} \quad (7)$$

where σ is a standard deviation, which is also a function of P. For simplicity, σ is assumed to be related to $\bar{r}$ by a linear relation:

$$\sigma = k_\sigma \bar{r} \quad (8)$$

In the proposed probabilistic propagation model, there are totally five parameters to be determined: P_0, r_0, n, k_d and k_σ. Excepting k_0, other parameters can be estimated separately from individual measurements in a straightforward manner. However, the values of these parameters can be slightly affected by the assumptions taken in the RF (Radio Frequency) propagation model. For this reason, a genetic algorithm (GA) [15] is used to find the optimal parameter set, all together. Genetic algorithms are global search techniques modeled after the natural genetic mechanism to find approximate or exact solutions for optimization and search problems. In a GA, each parameter to be optimized is represented by a gene. Moreover, each individual is characterized by a chromosome, which is actually the above set of parameters awaiting optimization. To assess the quality of an individual, a fitness function (objective function, or cost function) must be defined. For the localization module, the fitness function Ψ is defined as the root mean square of the localization error.

$$\Psi = (\frac{1}{N} \sum_{i=1}^{N} (\hat{x_i} - x_i)^2 + (\hat{y_i} - y_i)^2 + (\hat{z_i} - z_i)^2)^{1/2} \quad (9)$$

where N is the number of measurements, (x_i, y_i, z_i) and ($\hat{x_i}$, $\hat{y_i}$, $\hat{z_i}$) are the real and the estimated positions, respectively.

4.2 Fingerprinting database and KNN matching

Normally, a radio map in fingerprinting method is defined as follows:

$$\mathcal{R} \triangleq \{(\mathbf{p}_i, \mathbf{F}(\mathbf{p}_i)) \mid i = 1, .., N\} \quad (10)$$

where $\mathbf{p}_i \triangleq [p^x\ p^y\ p^z]^T$ is real world coordinates of the i^{th} RP and $\mathbf{F}(\mathbf{p}_i) \triangleq [\mathbf{r}_i(1)\ ,..,\mathbf{r}_i(n)]$ is the fingerprinting matrix, with n being the number of training samples at each RP. The vector $\mathbf{r}_i(t) \triangleq [r_i^1(t), .., r_i^L(t)]^T$ contains RSSI values that are scanned from L APs at time t and the location p_i. By using distance feature instead of RSSI, the radio map in Equation 10 then has a fingerprinting matrix $\mathbf{F}(\mathbf{p}_i) \triangleq [\mathbf{d}_i(1)\ ,..,\mathbf{d}_i(n)]$, with a vector $\mathbf{d}_i(t) \triangleq [d_i^1(t), .., d_i^L(t)]$ contains distance samples d_i from the i^{th} RP to L APs. This results in a reliable and stable radio map even in case some APs may be inactive at a certain point of time. Furthermore, the cost for setting and updating the radio map is much lower than using RSSI as usual. It is only rebuilt when we deploy new APs and RPs or discard them from the WiFi-based localization system.

In testing phase, the RSSI values scanned from nearby APs by a mobile device will be converted to the corresponding distance values by PPM model. They will be compared with the training data to find the best matches. The matching method used in this work is KNN. In KNN, prediction for a new instance is based on its nearest neighbors in the training data. There are three main ingredients associated with this method, those are (1) the similarity measure (the distance measurement) between the query patterns and training data; (2) the number of neighbors to be taken in the prediction; (3) the weight of the neighbors; Euclidean and Manhattan distances are two common geometric measures, in which Euclidean is the most used in WiFi-based localization system [16, 17]. In this work, KNN method is evaluated by Euclidean measure.

In the proposed radio map, each RP is represented by vector $\mathbf{d}_i(t) \triangleq [d_i^1(t), .., d_i^L(t)]^T$ in L dimensional space. In learning phase, all these training data $\mathbf{D}$ with their dependent variables are stored. In this case, the dependent variables are equivalent to the positions p_i of RPs in the environment. In prediction, for a new query pattern $\mathbf{z}$ and for each instance $\mathbf{d}$ in $\mathbf{D}$, the similarity between $\mathbf{d}$ and $\mathbf{z}$ is

computed by Euclidean distance measure:

$$l(d,z) = \sqrt{\sum_{i=1}^{n}(d_i - z_i)^2} \tag{11}$$

A set $NB(z)$ of the nearest neighbors of $\mathbf{z}$ with $|NB(z)| = k$ is also determined and then the estimated location for $\mathbf{z}$ is calculated. To find out an optimal k, we test on the empirical data with k in the range from 1 to 200 by an error function (12) for each k.

$$E_k = \sqrt{\sum_{i=1}^{n}(\frac{\hat{y} - y}{y})^2} \tag{12}$$

where $\hat{y}$ is the estimated position and y is true position. Finally, the predicted location of $\mathbf{z}$ is calculated by the weighted sum of the k neighbors (13).

$$y_z = \frac{\sum\limits_{d \in NB(z)} w(d,z) \times y_d}{\sum\limits_{d \in NB(z)} w(d,z)} \tag{13}$$

where w shows the weights that are chosen by (14).

$$w(d,z) = e^{-\theta \times l(d,z)} \times e^{-\lambda \times |t_i - t_0|} \tag{14}$$

where θ and λ are constants used to define the curve of exponential functions; t_0 belongs to the time a query instance is captured and t_i is the time of WiFi signal scanning at each corresponding RP in training phase; $l(d,z)$ is the dissimilarity between a query instance and the its neighbor. In Equation 14, beside the weight based on dissimilarity θ a new coefficient of λ is proposed to reflect the chronological changes of fingerprinting data in the environment. This means the recently-updated fingerprinting data with query instance will have higher weight than the older one.

5 Proposed fusion method

In order to improve the performance of person tracking in camera networks, for each camera FOV, person's locations determined by WiFi system are optimally assigned with positioning results from camera system. This allows to not only maintain the high accuracy of vision-based person localization, but also improve the performance of person tracking in camera networks by assigning clearer ID of WiFi adapter to each position determined by camera system.

Algorithm 1 shows the combined method of WiFi and camera system for people localization and identification. At time *t*, on the 2D floor map, a set of position observations from WiFi system ($\mathbf{z}^w_{i,t}$) or camera system ($\mathbf{z}^c_{j,t}$) for multiple targets are shown. Index *i* designates one among *N* targets located by WiFi system, and index *j* refers to one of *M* positions observed by camera system. We consider recursively two consecutive observations of the localization results from any available sensors. At time *t*, assuming that we have a set of location observations coming from WiFi system for *N* targets, with $\mathbf{z}^w_{i,t} = (X^w_{i,t}, Y^w_{i,t}, ID^w_{i,t})$. If at previous time step (*t-1*) we get the observations $\mathbf{z}^c_{j,t-1} = (X^c_{j,t-1}, Y^c_{j,t-1})$ for M positions from camera system. Without loss of generality, we can consider these observations as the state estimations at time *t-1*. The prediction step of the Kalman filter ($KalmanPrediction$) will be applied to estimate the next state $\mathbf{x}^c_{j,t}$ based on $\mathbf{z}^c_{j,t-1}$. An assignment algorithm is then utilized to find out optimal matchings between the estimated states $\mathbf{x}^c_{j,t}$ from camera system with observations ($\mathbf{z}^w_{i,t}$) from the WiFi system. Considering the result $K_{i,t}$ of the assignment is the observations at the current time t, then the predicted state x_t will be corrected by $KalmanCorrection$ step, by which WiFi-based positions will be augmented with the vision-based positions.

5.1 Kalman filter

In the proposed fusion algorithm, the step of state prediction in Kalman filter is used to estimate the process state at a certain time based on the position observation or measurement obtained from the previous time. The correction step of Kalman filter is done after doing optimal assignment between the estimated states and the observations at a certain time. In this case, a process state need to be estimated at a certain time is defined as a position p_t of a person in the real world coordinate system of 2D floor map. It is presented by a state vector $\mathbf{x}_t$ of location coordinates pX_t and pY_t on 2D floor map, together with their corresponding velocity values vX_t and vY_t:

$$\mathbf{x}_t = (pX_t, pY_t, vX_t, vY_t) \tag{15}$$

A position observation z_t is then defined as follows:

$$\mathbf{z}_t = (mX_t, mY_t) \tag{16}$$

By assumption of constant velocity and acceleration in movement of people, and the position is measured n times per second, the state equations are then defined as follows:

$$pX_t = pX_{t-1} + vX_{t-1}\Delta T \tag{17}$$

$$pY_t = pY_{t-1} + vY_{t-1}\Delta T \tag{18}$$

$$vX_t = vX_{t-1} \tag{19}$$

$$vY_t = vY_{t-1} \tag{20}$$

where $\Delta T = \frac{1}{n}$. The state transition matrix $\mathbf{A}$ and the state-measurement matrix $\mathbf{H}$ are then defined as:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Kalman-based tracking will be started after the first successful calculated position from WiFi or camera system,

Algorithm 1: Person tracking by fusion of position observations from WiFi and camera systems.

Input: position observations **z** from WiFi and camera localization systems
Output: position estimations **x**

Parameters initiation: **A**, **H**, $\mathbf{P}_1$, **Q**, **R**;
for *each set of position observations* $\boldsymbol{z}$ **do**
  if $\boldsymbol{z}_{i,t}$ *is from WiFi location system* $[\boldsymbol{z}^w_{i,t} = (X^w_{i,t}, Y^w_{i,t}, ID^w_{i,t})]$ **then**
    if $\boldsymbol{z}_{i,t-1}$ *is from camera location system* $[\boldsymbol{z}^c_{j,t-1} = (X^c_{i,t-1}, Y^c_{i,t-1})]$ **then**
      $[\mathbf{x}^c_{j,t}, \mathbf{P}_t]$ = ***KalmanPrediction***$(\mathbf{A}, \mathbf{Q}, \mathbf{z}^c_{j,t-1}, \mathbf{P}_{t-1})$;
      $K_{i,t}$ = ***Assignment***$(\mathbf{x}^c_{j,t}, \mathbf{z}^w_{i,t})$;
      $[\mathbf{x}^w_{i,t}, \mathbf{P}_t]$ = ***KalmanCorrection***$(\mathbf{H}, \mathbf{R}, K_{i,t}, \mathbf{x}_t, \mathbf{P}_t)$;
      Save $\mathbf{x}^w_{i,t}$ as a state estimation at time t;
    end
  else
    $[\mathbf{z}^c_{j,t} = (X^c_{j,t}, Y^c_{j,t})]$
    if $\boldsymbol{z}_{i,t-1}$ *is from WiFi localization system* $[\boldsymbol{z}^w_{i,t-1} = (X^w_{i,t-1}, Y^w_{i,t-1}, ID^w_{i,t-1})]$ **then**
      $[\mathbf{x}^w_{i,t}, \mathbf{P}_t]$ = ***KalmanPrediction***$(\mathbf{A}, \mathbf{Q}, \mathbf{z}^w_{i,t-1}, \mathbf{P}_{t-1})$;
      $K_{i,t}$ = ***Assignment***$(\mathbf{x}^w_{i,t}, \mathbf{z}^c_{j,t})$;
      $[\mathbf{x}^w_{i,t}, \mathbf{P}_t]$ = ***KalmanCorrection***$(\mathbf{H}, \mathbf{R}, K_{i,t}, \mathbf{x}_t, \mathbf{P}_t)$;
      Save $\mathbf{x}^w_{i,t}$ as a state estimation at time *t*;
    end
  end
end
return $\boldsymbol{x}^w_{i,t}$;

with the initial state vector x_1. The initial covariance matrix $\mathbf{P}_1$ for the initial state is:

$$\mathbf{P}_1 = \begin{bmatrix} \sigma^2_{x_1} & 0 & 0 & 0 \\ 0 & \sigma^2_{y_1} & 0 & 0 \\ 0 & 0 & \sigma^2_{vx_1} & 0 \\ 0 & 0 & 0 & \sigma^2_{vy_1} \end{bmatrix}$$

The state noise covariance matrix **Q** and the measurement noise covariance matrix **R** are defined as:

$$\mathbf{Q} = \begin{bmatrix} \sigma^2_{pX} & 0 & 0 & 0 \\ 0 & \sigma^2_{pY} & 0 & 0 \\ 0 & 0 & \sigma^2_{vX} & 0 \\ 0 & 0 & 0 & \sigma^2_{vY} \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \sigma^2_{mX} & 0 \\ 0 & \sigma^2_{mY} \end{bmatrix}$$

where σ^2 denotes deviation in centimeter from real values of each quantity. The measurement noise refers to the noise of calculated positions from WiFi or camera system, and the state noise is defined according to the motion of people. The initial covariance matrix $\mathbf{P}_1$ for the initial state $\mathbf{x}_1$, with assumption that the calculated position has the deviation of $\pm$5cm from real position in both X and Y directions, and the velocity has the deviation of $\pm$3cm. Similarly, the state noise covariance matrix **Q** is set with standard deviations of $\pm$5cm and $\pm$3cm for the determined position and its velocity, respectively. The measurement noise covariance matrix **R** is described with the standard deviation of 3cm for Foot-Point measurement in X and Y directions, and ΔT is set to 1, meaning that the position is measured every second.

5.2 Optimal assignment

After the Kalman prediction step, we have a position estimation of $\mathbf{x}^c_{j,t}$ or $\mathbf{x}^w_{i,t}$ for camera or WiFi system, respectively. Considering the first case of position estimation $\mathbf{x}^c_{j,t}$ at time t for camera system, it is estimated from the previous observation of vision-based location $\mathbf{z}_{j,t-1}$. Then, optimal assignment at time t between $\mathbf{x}^c_{j,t}$ and $\mathbf{z}^w_{i,t}$ is applied. Assuming that the assignment of an estimated position $\mathbf{x}_j$ and an observation $\mathbf{z}_i$ incurs a cost d_{ij} which is the Euclidean distance between them, then the matrix $D_{N\times L}$ of the costs or distances between every $\mathbf{x} \in \mathcal{M}$ and $\mathbf{z} \in \mathcal{N}$ is then defined as:

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & ... & d_{1N} \\ d_{21} & d_{22} & ... & d_{2N} \\ ... & ... & ... & ... \\ d_{M1} & d_{M2} & ... & d_{MN} \end{bmatrix}$$

where $d_{ij} = \sqrt{(X^c_j - X^w_i)^2 + (Y^c_j - Y^w_i)^2}$. The assignment is now formulated as a linear assignment problem:

$$min \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{M}} d_{ij} x_{ij} \tag{21}$$

subject to

$$\sum_{i\in\mathcal{N}} x_{ij} = 1 \quad \forall j \in \mathcal{M}$$

$$\sum_{j\in\mathcal{M}} x_{ij} = 1 \quad \forall i \in \mathcal{N}$$

$$x_{ij} \geq 0 \qquad \forall i \in \mathcal{N}, j \in \mathcal{M}$$

This optimal assignment is done with the following constraints:

- If $N = M$, for each pair of ($\mathbf{x}^c_{j,t}$, $\mathbf{z}^w_{i,t}$), we augment the position $\mathbf{x}^c_{j,t}$ with the identity $ID^w_{i,t}$ from $\mathbf{z}^w_{i,t}$;
- If $N > M$, all unassigned $\mathbf{z}^w_{i,t}$ will be kept up with their original coordinates which are computed from WiFi-based localization system;
- If $N < M$, all unassigned $\mathbf{x}^c_{j,t}$ are considered as false positives and will be discarded, because we assume in the surveillance system that all people coming in the monitoring areas hold WiFi-enabled devices and they have checked in at the entrance.

The overall formula for these constraints is given as follows:

$$K_{i,t} = \begin{cases} (X^c_{j,t}, Y^c_{j,t}, ID^w_{i,t}) & \text{if } \mathbf{z}^w_{i,t} \text{ is assigned;} \\ (X^w_{i,t}, Y^w_{i,t}, ID^w_{i,t}) & otherwise. \end{cases}$$

where $K_{i,t}$ denotes the association between position estimations $\mathbf{x}^c_{j,t}$ and observations $\mathbf{z}^w_{i,t}$. Each component $K_{i,t}$ is a random variable that takes its value among $\{0, .., N\}$. Based on this association, the location information from WiFi-based observations will be corrected according to the positions given by the camera system, and the corresponding ID from the WiFi system will be assigned. The correction step of the Kalman filter is applied to update the predicted state by the current position observation $K_{i,t}$.

The same procedure is done for the case in which WiFi-based location observations come before camera-based ones, and we have optimal assignment of an estimated position $\mathbf{x}_i$ from the WiFi system and an observation $\mathbf{z}_j$ from the camera system.

6 Dataset and evaluation

6.1 Testing dataset

In order to evaluate the combined algorithm for person tracking using both WiFi and camera systems, a multimodal dataset with two scripts are constructed in this work. Script 1 is set with simpler scenarios than Script 2. Two people are involved in Script 1, with their random routes of moving through two non-overlapping cameras. Some inter-person occlusions appeared but not as frequently as in Script 2. The visual data in Script 1 is used for person localization and Re-ID based on camera. Script 2 contains five scenarios referring to different number of people taking part in each scenario: one person, two, three, and five moving people. The data in Script 2 is very challenging for both WiFi-based and vision-based systems. People move through four different cameras. Severe occlusions happened because all people are required to move in close proximity with a fixed route (see Figure 8). Moreover, the similar human appearance is a challenge for visual processing problems.

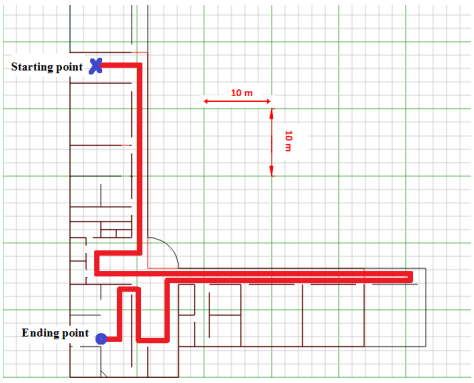


Figure 8: A 2D floor map of the testing environment in Figure 9, with the routing path of moving people in testing scenarios.

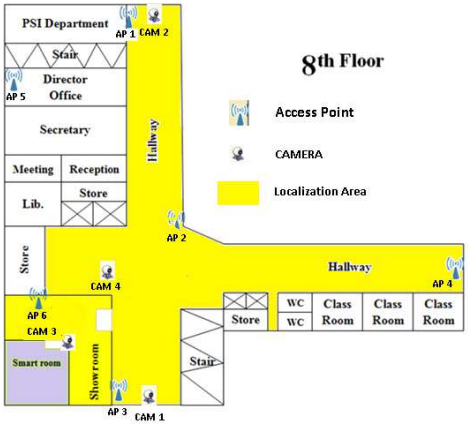


Figure 9: Testing environment.

The testing environment for building the dataset is shown in Figure 9, with 6 access points (APs) and 4 cameras are deployed in the environment. The APs are set to a same SSID, which assures continuous connectivity for mobile devices when people move from the range of one AP to another. The WiFi range for each AP is about 30–50 meters in radius, depending on walls and obstacles in the environment. The AP specifications are MAC address, AP position in X, Y and Z. All APs used in the testing are Linksys E1200 devices. A person holds a WiFi-enable device and moves in the testing environment, with a normal velocity of 1–1.3m/s.

The time duration for each scenario is from 3 to 5 minutes, with about 400 RSSI values are acquired from 6 APs and average time deviation between two consecutive samples is 2 seconds. The mobile devices and cameras are time synchronized to Internet time. This makes a synchronization of data captured from both camera and WiFi. Basing on this, we can compute real-world positions of a mobile user on the 2D floor map at each time. The time stamp for each person location calculated from camera or WiFi system will provide the basis for processing multi-model object localization. The WiFi data is scanned from the mobile devices and stored in XML files. These devices con-

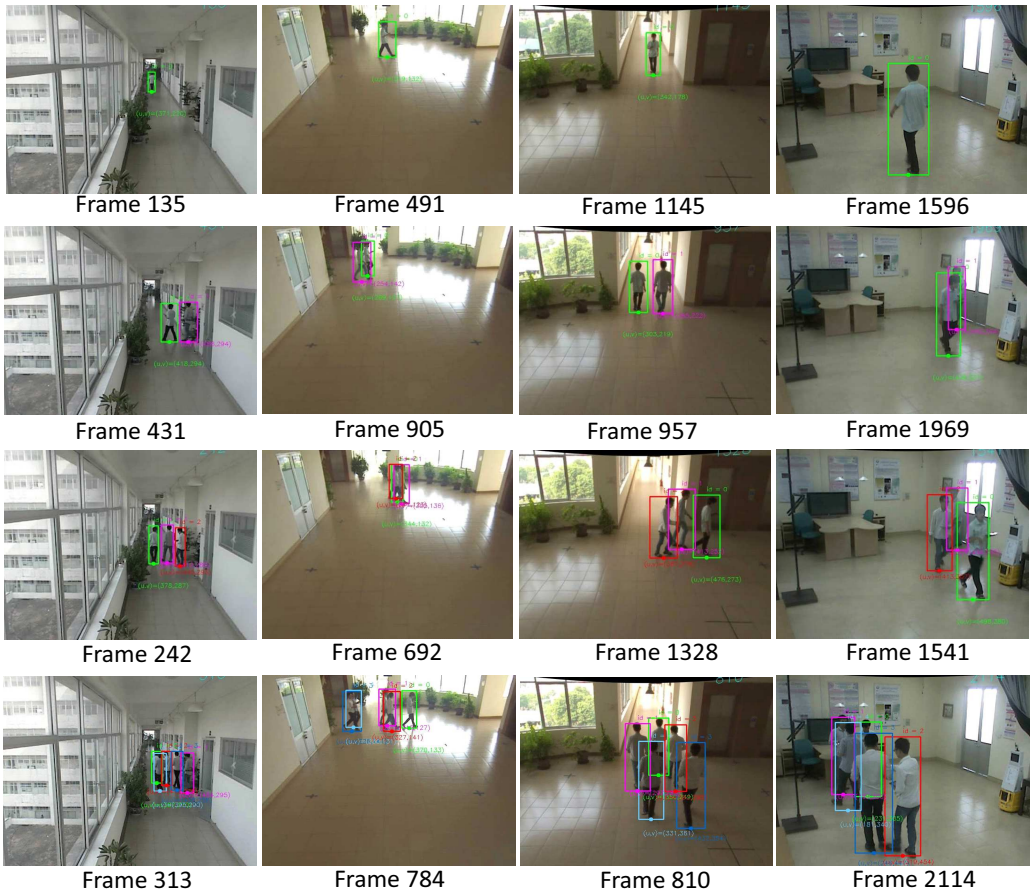


Figure 10: The visual examples in Script 2. The first row contains frames for the scenario of one moving person. The scenarios for 2, 3 and 5 moving people are shown in the second, third and fourth rows.

tinuously capture the signals from available APs in the environment. The AP specifications are saved as a record of scanning time, MAC address, AP name, and RSSI. The APs are distinguished by their own MAC addresses.

For visual data, we manually assign FootPoint positions on the captured frames with the corresponding time stamps and IDs. These positions are then automatically transformed into 2D locations on the floor map by using camera calibration and homography matrix. The person ID which is assigned in visual data is equivalent to the ID of WiFi adapter by predefined convention. In short, for each scenario, the ground truth data is achieved and saved as XML files which contain the following records:

- Frame number.
- Person ID.
- Coordinates of top left and bottom right positions of the bounding box containing the person.
- The image coordinates of FootPoint position.
- The corresponding coordinates of FootPoint positions on 2D floor map.

In case of no person detected, except frame number, all other records are set to *-1*.

Figure 10 illustrates examples in Script 2. The frames in the first row show the scenario of one moving person, while those in the second, third and fourth rows are frames for the scenarios of two, three and five moving people.

For WiFi data which is determined outside camera FOV, the ground truth of person locations in these regions are calculated by a pedestrian foot counting program. It takes input information from the acceleration and direction sensors that are available on smart phones or tablets [20]. Basically, the positions of mobile user in this region are computed by the route length that user passes through marking points or reference points. This distance is calculated by foot counter with the average length of the foot step of each particular person is considered. The foot counter gives the positioning result of 5m with the deviation of 3m for the route length of 120m. In our test, the route length outside camera view is only about 10m. In addition, the bias for foot counter is accumulated from time to time, so in 10m this deviation will be 0.8m (equivalent to 8% of the route length). This makes the deviation of 8cm per one meter labeled in the dataset in comparison with the truth positions.

After the step of synchronization between WiFi and vi-

sual data, the interpolation method is applied to calculate the person positions that are outside the camera field of views.

6.2 Evaluation metrics

In order to evaluate the performance of vision-based tracking, the metrics of Multi Object Tracking Precision (MOTP) [18], Global Multiple Object Tracking Accuracy (GMOTA) [19], and CMC (Cumulative Match Curve) are utilized.

Assuming that for each time step t, a multi-person tracker outputs a set of hypotheses $\{h_1, .., h_m\}$ for a set of visible people $\{u_1, .., u_n\}$. MOTP measures the positioning error for all matched pairs of person and tracker hypothesis on all frames. This metric is defined by:

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t} \tag{22}$$

where $d_{i,t}$ is Euclidean distance between ground truth and tracker hypothesis values for the person i^{th} at time frame t. In this work, it is Euclidean distance between ground truth and tracker hypothesis of FootPoint positions. The element c_t indicates the number of matched pairs at time step t.

GMOTA is an extension of MOTA (Multiple Object Tracking Accuracy) [18]. MOTA measures the number of errors the tracker made in terms of false negatives (missed detections), false positives (wrong detections), mismatches and failure to recover tracks. This score is computed as follows:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + ID_t)}{\sum_t g_t} \tag{23}$$

where FN_t is false negatives, FP_t is false positive, ID_t shows the number of instantaneous identity switches, and g_t denote the number of ground truth detections at time frame t. In GMOTA score, the ID_t is replaced by global ID_t (gID_t). This means that gID_t presents the performance of the tracker in preservation of person identity assignments in a global manner instead of instantaneous identity assignments of MOTA.

$$GMOTA = 1 - \frac{\sum_t (FN_t + FP_t + gID_t)}{\sum_t g_t} \tag{24}$$

The CMC is employed as the performance evaluation metric for vision-based person Re-ID. The CMC curve presents the expectation of finding correct match in the top n matches.

The accuracy of the WiFi-based localization system is evaluated by the statistical values of maximal error, error average, and error at reliability of 90%. Maximal error is the maximum distance deviation in meter between the positions determined by the system and the ground truth positions. The error average refer to the average distance deviation in meter between the positions determined by the system and the ground truth positions. Error at reliability of 90% indicates the distance deviation value in meter in which 90% of the testing times are smaller than this value.

The performance of fusion method is evaluated in this work by the metric of GMOTA.

6.3 Experimental results

In vision-based person localization, at each camera FOV, person identification is done by a so-called process of identification by tracking. This means a trajectory which belongs to an individual in the current frame is linked to the corresponding one from the previous frame based on an optimal assignment of Euclidean distances between them. However, this results in ID switches when people switch to each others.

The proposed method in 3.2 for person Re-ID helps to solve not only person identification in each camera FOV, but also person Re-ID among multiple cameras by using a robust appearance-based descriptor built on each detected human ROI at each FootPoint position. This allows to perform tracking by identification. However, person identification and Re-ID performance still need to be improved, especially in case of inter-person occlusions and people have similar appearances.

The proposed fusion algorithm allows adding clearer ID information of WiFi adapter for performing tracking by identification.

In the following sections, the testing results for WiFi-based localization, vision-based localization and Re-ID, fusion-based tracking are shown.

6.3.1 WiFi-based localization results

The system parameters of the WiFi-based localization model are calculated, then based on these, the positioning results are given out.

Firstly, the training process using GA algorithm is set up with the configuration provided in Table 1. Using these data, the optimal parameters are produced as in Table 2.

Parameter	Value	Parameter	Value
Population size	20	Tolerance	10^{-6}
Elite count	5	Selection	Uniform
Crossover fraction	0.5	Crossover	Scattered
Time limit	No	Mutation	Uniform
Maximal generations	No	Creation population	Uniform

Table 1: Genetic algorithm configuration.

Parameter	Values for the first scenario	Values for the second scenario
P_0	-41 dBm	-36.1757 dBm
n	1.1	2.2029
k_σ	1.0035 m^{-1}	5.3147 m^{-1}
r_0	5 m	2.5117 m
k_d	49.23 $dBm{\cdot}m^{-1}$	5.1311 $dBm{\cdot}m^{-1}$

Table 2: Optimized system parameters for the first and the second scenarios of testing environments.

Fingerprint Feature	Maximal error (m)	Average error (m)	Error at reliability of 90% (m)
RSSI	6.3	1.86	2.99
Distance	6.27	1.89	2.98

Table 3: Evaluations for distance and RSSI features in case of using coefficient λ.

Fingerprint Feature	Maximal error (m)	Average error (m)	Error at reliability of 90% (m)
RSSI	6.06	1.76	3.55
Distance	6.5	1.59	2.9

Table 4: Localization results using different features of distance and RSSI, without using coefficient λ.

Secondly, the weights of different values of θ based on dissimilarity are given out (see Figure 11). Different values of λ are presented in Figure 12, with $\lambda = 0.5 \times 10^{-6}$, the influence is reduced by 3 when fingerprints is scanned from 1 month since the testing time (roughly 2.6×10^6 seconds). Similarly, when fingerprints is taken from 2 months since the testing time, the influence takes only 10% compared with that of new fingerprints. In this work, we choose $k = 9$, $\theta = 1.1$ and $\lambda = 2 \times 10^{-6}$.

The radio maps and fingerprint locations in the testing environment are shown in Figure 13a and Figure 13b. The regions with deep pink color indicate that more APs are available than the regions with light pink color.

The localization experiments are conducted by using fingerprinting method with distance features calculated by the proposed probabilistic propagation model. The comparative results are also given out for using fingerprinting method with RSSI features. Additionally, the stability and reliability of radio map with distance features is also confirmed by the evaluations with coefficient λ.

Figures 14, 15, 16 show the comparative results when the coefficient λ is taken into account. The localization results, distribution of the localization results compared to the real locations, and the reliability of the localization result as a function of the localization error are shown correspondingly in these figures. The details for these results are shown in Table 3. It can be seen from the experiments that the positioning errors at reliability of 90% when using distance features are a little bit higher than using RSSI features. However, without using λ, the localization reliability for RSSI features decreases, whilst it is stable for distance features. The results for this are shown in Figures 17, 18, 21, and in Table 4, with the error at reliability of 90% is 3.55m for RSSI features, but it is 2.9m for distance features. The above experiments show that using distance features for fingerprint data will result in more stable and reliable radio maps in comparison with using RSSI features. Moreover, this also brings lower cost for updating fingerprint data, which is considered as one of the most challenging problem of fingerprinting method in WiFi-based localization.

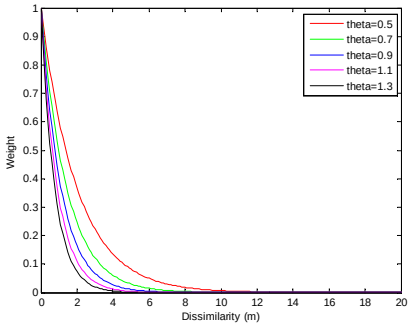


Figure 11: Weights of different values of θ based on dissimilarity.

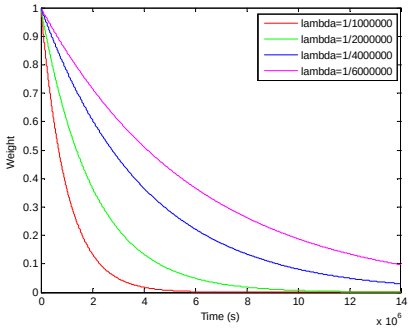
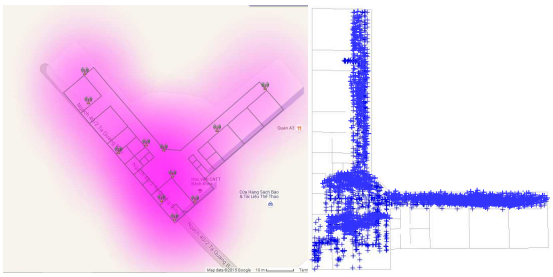


Figure 12: Weights of different values of λ based on dissimilarity.



(a) (b)

Figure 13: (a) the radio map, with (b) 2000 fingerprint locations collected in the testing environment.

	Vision-based evaluations		The proposed fusion algorithm	
	Hallway (Cam 1)	Showroom (Cam 3)	Hallway (Cam 1)	Showroom (Cam 3)
MOTP (cm)	24.3	21.3	24.3	21.3
FN (%)	17.1	26.4	7.6	12.6
FP (%)	22.7	18.3	3.4	2.1
gID	28.3	11.6	4.9	2.3
GMOTA (%)	31.2	52.6	83.9	85.7

Table 5: The comparative results of the proposed fusion algorithm against the vision-based evaluations on testing data of Script 1.

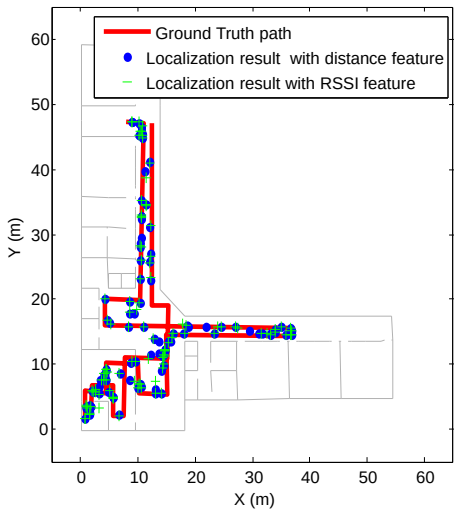


Figure 14: Localization results with distance and RSSI features when using coefficient λ.

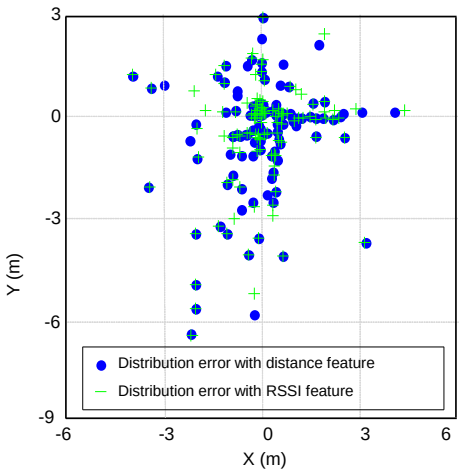


Figure 15: Distribution of localization error for distance and RSSI features when using coefficient λ.

6.3.2 Experimental results for vision and fusion-based tracking.

The performance of vision-based person localization and Re-ID is evaluated on Script 1 and Script 2 databases. In addition, the comparative results gained from fusion system of camera and WiFi are also indicated on these.

Firstly, vision-based person Re-ID evaluations are done on Script 1 data. The human ROIs are manually extracted from the frames captured by three non-overlapping cameras: Cam 1 (hallway), Cam 2 (lobby) and Cam 3 (showroom). The human ROIs from Cam 2 are used for training phase (see Figure 19) and the human ROIs from Cam 1 and Cam 3 for testing phase (see Figure 20).

We train the system with totally 10 people, including two testing ones, by the images of human ROI extracted from Cam 2. Figure 22 shows person recognition rates for this experiment, with Rank 1 is 51.1%.

Table 5 shows the results for vision-based localization, with two scenarios of Hallway (Cam 1) and Showroom (Cam 3) are considered. MOTP evaluated on the vision-based localization system with 24.3cm and 21.3cm for Hallway and Showroom scenarios respectively. These values are retained for the fusion model of camera and WiFi. GMOTA ratio for Hallway is better than Showroom, with correspondingly 31.2% compared to 52%. However, by being integrated with WiFi, these values increase incredibly to 83.9% for Hallway and 85.7% for Showroom. This resulted from the sharply decreases in the rates of FN, FP and gID in both scenarios. Additionally, in comparison with the perfect case of manual human detection in vision-based Re-ID, the performance of person tracking by identification is not as good as the results from the proposed fusion algorithm.

Secondly, further evaluations for the proposed fusion algorithm, the experiments are done on the data of Script 2. This dataset is very challenging compared to Script 1, because of severe occlusions and the similarity in human appearance. Moreover, people moving together in the same route is also a challenge for WiFi-based localization.

In the experiments with this data, we use the ground truth data of FootPoint positions and the corresponding human ROIs for testing evaluations. The parameter gID in GMOTA metric now indicates the performance of tracker in maintaining the person ID when he/she moves from one camera FOV to others or re-appears in one camera FOV. Table 6 shows the comparative results of GMOTA when applying the fusion algorithm and Rank 1 for person Re-ID. It should be noted that FN and FP are not included

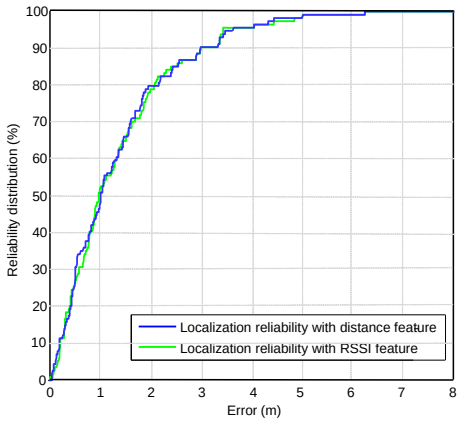


Figure 16: Localization reliability for distance and RSSI features when using coefficient λ.

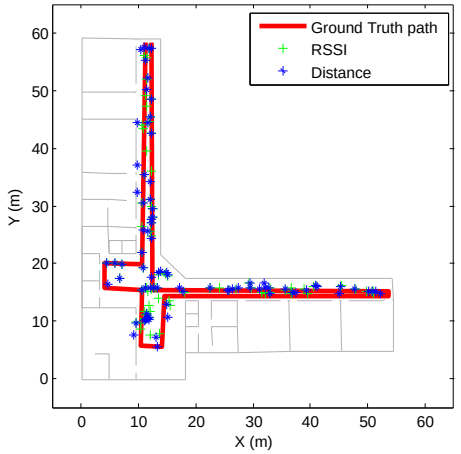


Figure 17: Localization results for distance and RSSI features, without using coefficient λ.

in the testing evaluations of GMOTA because we use the ground truth data of FootPoint positions and human ROIs. In this case, only gID is taken into account. This means performance of maintenance person ID in tracking now depends only on the performance of WiFi-based person localization. In comparison with GMOTA values from Script 1, GMOTA figures from Script 2 are much lower. It is only 31.7% for the scenario of two moving people, 16.5% and 11.2% for scenarios of three and five moving people, respectively. This can be explained that data of Script 2 is much challenging than Script 1. People moving together in very close proximity is not only a burden for vision-based person localization and identification, but also for WiFi-based person localization because of noisy WiFi data when people are close to each other.

However, in comparison with person Re-ID by kernel descriptor, these results are much higher. In this experiments, besides the number of testing people, we train the system with 20 other people at check-in gate for person Re-ID. The recognition rate at Rank 1 is only 12.6% for scenario of two moving people, which is 19.1% lower than

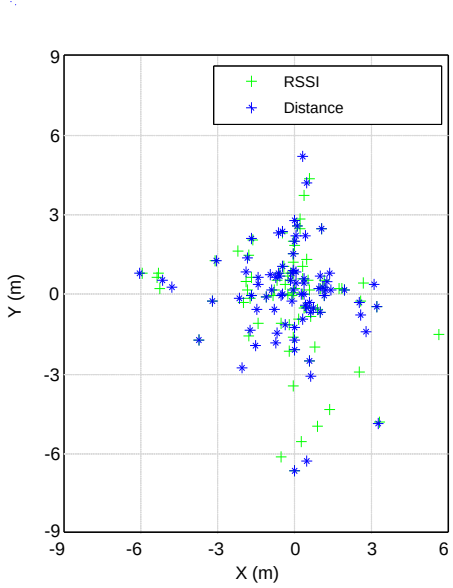


Figure 18: Distribution of localization error for distance and RSSI features, without using coefficient λ.

	Two people	Three people	Five people
GMOTA (%)	31.7	16.5	11.2
Rank 1 (%)	12.6	8.9	5.6

Table 6: The experimental results for person tracking and person Re-ID with Script 2 dataset.

fusion-based method. Rank 1 figures for scenarios of three and five moving people is 8.9% and 5.6%. Clearly, performance of person Re-ID based on kernel descriptor will be degraded in case of the similar human appearance.

From the above comparative evaluations, we can see that by using the proposed fusion algorithm, the performance of person tracking by identification and person Re-ID is improved significantly. The vision-based person localization with high accuracy, together with the clear ID information from WiFi-enable device are integrated into each detected FootPoint position. This allows to do tracking by identification at each camera FOV, and based on this, the person Re-ID in non-overlapping camera networks can be solved more effectively than applying only vision-based method.

7 Conclusion

In this work, person localization and Re-ID in surveillance regions covered by WiFi signals and disjointed FOV cameras are improved by a fusion algorithm based on Kalman filter and optimal assignment technique. This algorithm is executed with the position observations on 2D floor map achieved from each single system of camera or WiFi.

Evaluation on the multimodal dataset shows outperforming results when the proposed fusion algorithm is applied. The high positioning accuracy of vision-based system is maintained in multimodal person localization system. Ad-

Figure 19: Training examples of manually-extracted human ROIs from Cam 2 for person 1 (images on the left) and person 2 (images on the right).

(a)

(b)

Figure 20: Testing examples of manually-extracted human ROIs from Cam 1 (images on the left column) and Cam 3 (images on the right column) for (a) person 1 and (b) person 2.

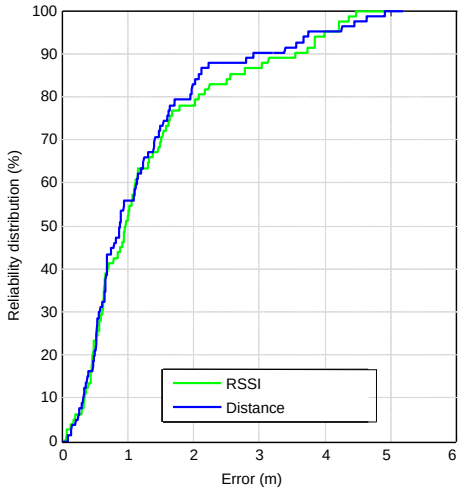


Figure 21: Localization reliability for distance and RSSI features, without using coefficient λ.

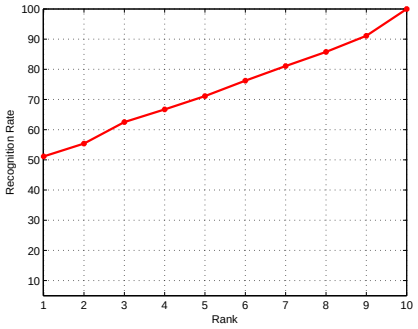


Figure 22: Person Re-ID evaluations on testing data of two moving people.

ditionally, the fusion algorithm allows tracking by identification and based on this person Re-ID in non-overlapping cameras is done with clear identity information taken from the WiFi-based system.

In the future works, some other localization techniques, such as RFID or UWB, can be integrated into a multimodal system in order to improve the positioning accuracy and person Re-ID. The fusion algorithm for person localization and Re-ID is also correspondingly broaden to adapt this addition.

Acknowledgement

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2013.32.

References

[1] Van den Berghe, Sam and Weyn, Maarten and Spruyt, Vincent and Ledda, Alessandro (2011) Combining wireless and visual tracking for an indoor environment, *International Conference on Indoor Positioning and Indoor Navigation (IPIN-2011)*.

[2] MIYAKI, Takashi, YAMASAKI, Toshihiko, et AIZAWA, Kiyoharu (2007) Visual tracking of pedestrians jointly using wi-fi location system on distributed camera network, *2007 IEEE International Conference on Multimedia and Expo*, IEEE, 2007. p. 1762–1765.

[3] Rekimoto, Jun and Shionozaki, Atsushi and Sueyoshi, Takahiko and Miyaki, Takashi (2006) PlaceEngine: a WiFi location platform based on realworld folksonomy *Internet conference*, p. 95–104.

[4] Cheng, Yu-Chung and Chawathe, Yatin and LaMarca, Anthony and Krumm, John (2005) Accuracy characterization for metropolitan-scale Wi-Fi localization, *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, ACM, p. 233–245.

[5] Alahi, Alexandre and Haque, Albert and Fei-Fei, Li (2015) RGB-W: When Vision Meets Wireless, *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, p. 3289–3297.

[6] Pham, T. T. T., Le, T. L., Vu, H., and Dao, T. K. (2017) Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method, *Image and Vision Computing*, Elsevier, p. 44-62.

[7] Kuhn, Harold W (1955) *Naval research logistics quarterly*, Wiley Online Library, p. 83–97.

[8] Zhang, Zhengyou (2000) A flexible new technique for camera calibration, *Pattern Analysis and Machine Intelligence*, IEEE, p. 1330–1334.

[9] Thi Thanh Thuy Pham, Anh Tuan Pham, Hai Vu (2015) A new technique for linking person trajectories in surveillance camera network, *Conference on Fundamental and Applied IT Research (FAIR)*, p. 8–15.

[10] Bo, Liefeng and Ren, Xiaofeng and Fox, Dieter (2010) Kernel descriptors for visual recognition, *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, p. 244–252.

[11] Dao, Trung-Kien and Pham, Thanh-Thuy and Castelli, Eric (2013) A robust WLAN positioning system based on probabilistic propagation model, *9th International Conference on Intelligent Environments (IE)*, IEEE, p. 24–29.

[12] Goldsmith, A. (2005), Wireless communications, Cambridge university press.

[13] Roberts B. and Pahlavan K. (2009) Site-specific rss signature modeling for wifi localization, *In Global Telecommunications Conference*, IEEE, p. 1–6.

[14] Munoz D., Lara F.B., Vargas C., and Enriquez-Caldera R. (2009), *Position location techniques and applications*, Academic Press.

[15] Haupt, Randy L and Haupt, Sue Ellen (2004) *Practical genetic algorithms*, John Wiley & Sons.

[16] Jungmin So, Joo-Yub Lee, Cheal-Hwan Yoon, Hyunjae Park (2013) An Improved Location Estimation Method for Wifi Fingerprint-based Indoor Localization, *International Journal of Software Engineering and Its Applications*.

[17] Arsham Farshad, Jiwei Li, Mahesh K. Marina, Francisco J. Garcia (2013) A Microscopic Look at WiFi Fingerprinting for Indoor Mobile Phone Localization in Diverse Environments, *International Conference on Indoor Positioning and Indoor Navigation*.

[18] Bernardin, Keni and Stiefelhagen, Rainer (2008) Evaluating multiple object tracking performance: the CLEAR MOT metrics, *EURASIP Journal on Image and Video Processing*, Springer, p. 1–10.

[19] Ben Shitrit, Horesh and Berclaz, Jerome and Fleuret, François and Fua, Pascal (2013) Tracklet-based Multi-Commodity Network Flow for Tracking Multiple People, *No. EPFL-PATENT-186751*, WO.

[20] Kothari, Nisarg and Kannan, Balajee and Glasgwow, Evan D and Dias, M Bernardine (2012) Robust indoor localization on a commercial smart phone, *Procedia computer science*, Elsevier, p. 1114–1120.