Enhancing Camera-Based Multimodal Indoor Localization With Device-Free Movement Measurement Using WiFi

Yanchao Zhao, *Member, IEEE*, Jing Xu, *Student Member, IEEE*, Jie Wu, *Fellow, IEEE*, Jie Hao, *Member, IEEE*, and Hongyan Qian, *Member, IEEE*

***Abstract*—Indoor localization is of great significance to a wide range of applications in the era of mobile computing. The maturity of the computer vision techniques and the ubiquity of embedded sensors in commercial off-the-shelf (COTS) smartphones shed the light on the submeter localization services for indoor environment. The state-of-the-art indoor localization works suffer from high-cost deployment and inaccurate results due to the coarse readings from internal measurement units (IMUs) sensors in the smartphones. In this article, we mainly innovate in introducing the WiFi-sensing technology to extract the distance information in a low-cost and device-free manner. Along with the computer vision technology, we model and implement an accurate and easy-to-deploy system for indoor localization. This system enhances indoor localization with multimodal sensing via two images, IMU sensors reading and CSI of WiFi signal. Specifically, we first model and design camera-based, sensor and WiFi-assisted indoor localization and propose several algorithms in this model. We then implement a prototype with smartphones and commercial WiFi devices and evaluate it in several distinct indoor environments. The experimental results show that 92-percentile error is within 0.2 m for indoor targets which sheds light on submeter indoor localization.**

***Index Terms*—Indoor localization, multimodal sensing, smartphone.**

I. Introduction

A. Motivation

INDOOR localization has served as an indispensable part for a wide range of applications and services, such as customer navigation in shopping malls [1], object localization and tracking in airports [2], and routing robots in an automated factory [3]. The essence of these applications lies in the measurements of distance and angle for indoor localization. However, most state-of-the-art multimodal-based localization schemes, such as Argus [4] and ClickLoc [5], are mainly flawed in measurement accuracy or cost, although with tremendous related attempts, accurate and robust indoor localization remains unsolved. Meanwhile, recent progress in the hardware ability of mobile smartphones has shed light on and the need for vision-based indoor localization. As the camera is essentially a more powerful sensor capable of harvesting the environment information in high dimensions, the vision-based method [5]–[8] is a promising direction to improve the indoor localization to a new level. However, the cameras, especially the smartphone equipped ones, are unstable and noisy, thus also pose great challenges to efficient and accurate localization. As the most camera-taken images are lacking the depth information of the objects, it is essential to propose a multimodal mechanism fully utilizing the IMU and other possible means to help the smartphone cameras accurately extracting the distance information in the images.

Manuscript received July 5, 2019; revised September 30, 2019; accepted October 15, 2019. Date of publication October 22, 2019; date of current version February 11, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0802300; in part by the National Natural Science Foundation of China under Grant 61602238 and Grant 61602242; and in part by NSF under Grant CNS 1824440, Grant CNS 1828363, Grant CNS 1757533, Grant CNS 1618398, Grant CNS 1651947, and Grant CNS 156412. *(Corresponding authors: Yanchao Zhao; Jie Hao.)*

Y. Zhao, J. Xu, J. Hao, and H. Qian are with the College of Computer Sciences and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210046, China, and also with the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China (e-mail: yczhao@nuaa.edu.cn; haojie@nuaa.edu.cn).

J. Wu is with the Center for Networked Computing, Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA.

Digital Object Identifier 10.1109/JIOT.2019.2948605

B. Proposed Approach

In this article, we propose a nonintrusive and high-accurate solution for vision-based indoor localization. We mainly innovate in introducing the multimodal data from commercial off-the-shelf (COTS) WiFi devices, internal measurement units (IMUs) sensors, and the monocular camera of smartphones together to derive the distance and direction. The basic idea is as follows. We match similar scenes from two photographs collected by the monocular camera and identify the target objects in the images for further distance extraction. Based on the detection results, we further extract the geometric information of the image space, where the inertial sensors data and channel state information (CSI) values are continuously collected during the user's movements, with which we manage to extract the moving distance and direction. Then, we fuse this information with the map and get a geometric relationship of the image space to the inertial space, supplying the actual distance and orientation in the physical space to achieve the users' localization.

C. Challenges and Solutions

To implement our multimodal localization as a practical system, three technical challenges need to be addressed. The first challenge is how to extract accurate distances and directions between different photographs. The indoor localization results in this article can be severely affected by the distance and angle measured by inertial sensors. Owing to the subtle changes of the user's gesture during localization, the collected data of inertial sensors may be inconsistent with the previous measurements and result in attitude drifts. The distance and orientation obtained from IMU sensors could also be erroneous. Hence, it is necessary to eliminate the influences of attitude drifts and ensure the reliability of sensor data. As a solution, we basically introduce the fusion data from both the IMU sensors and the WiFi passive sensing technology and get the improved results of distance and orientations.

The second challenge is how to obtain the geometric relationship between two images and map it to inertial space. Images collected by the user provide scene information, visual clues, and geometric relationships for the point of interests (PoIs). Therefore, effective methods to extract sufficient information of the indoor scenario are essential. On the one hand, we consider sensor data to divide regions for feature matching of objects in images, which reduces the complexity of matching. On the other hand, we adopt a multitarget detection framework to detect objects in two images, and output the coordinate information of each target. We can further obtain detailed geometric relationships through calculations of coordinate data, acquiring the positions of target objects in images. Moreover, considering the obtained distance and direction, we can complete coordinate mapping by utilizing the correspondence of the same measurement in different spaces.

The third challenge is how to fuse the multimodal sensing information, especially the vision and the WiFi sensing, for better measurement performance. For visual clues collected by the monocular camera, the related module outputs coordinate data for the proposed localization model and the distance ratio of the target object to the camera's visual boundary in the image space. For the user's movement and distance measurement in the inertial space, we take into account both the IMU sensors and WiFi CSI, and conquer the drawbacks of attitude drift, error accumulations of IMU sensors, and multipath effect on CSI values. Finally, we manage to design an approach taking advantage of both methods. Basically, it first obtains time constant from first-order system equations. Then, both distance estimation from WiFi Fresnel Zone model and the orientation from IMU sensors are combined and output accurate distance estimations.

The fourth challenge is how to apply our system into a multiple person environment. Although WiFi CSI-based moving distance measuring is much more accurate than the IMU sensors, it could be easily compromised by the disturbance from the other moving persons in the environment. We manage to do this by introducing the virtual samples mechanism and extract the distance information with the other movement interference in such environment.

D. Contributions

We make four contributions in this article that can be summarized as follows.

1) We demonstrate the feasibility of enhancing indoor localization based on multimodal sensing via camera and sensors of smartphones and design the system that can be used to assist the reconstruction of building interior view and indoor navigation further.
2) We propose an innovative algorithm for multimodal distance and orientation estimation in indoor environments, making full use of scene information from images and extracting the user's motion information effectively.
3) We propose a framework to enable the WiFi-based device-free moving distance derivation to be performed in multiple person scenarios, so that our localization algorithm could be applied in real environment settings.
4) We implement a proof-of-concept localization prototype and evaluate it in various indoor environments. The experimental results show that the 92-percentile error is within 0.2 m for indoor targets which makes our solution achieving submeter accuracy overall.

II. Related Work

The rich inertial sensors in smartphones and the widespread use of the CV technology have attracted extensive research focusing on using one or multiple sensing modalities to determine the indoor location, including heading direction, movement distance, and walking trajectory of the pedestrian. Many approaches have been proposed in the localization system, including utilizing wireless signal [9], [10], multiple sensors [11], [12], and images [8], [13]. The closely related work can be roughly divided into the following three categories.

1) *WiFi-Based Indoor Localization:* Wu *et al.* [14] proposed a mechanism analyzing WiFi signal features through 2-D Fresnel model to determine the walking distance and direction of users indoor, which detects both centimeter-scale and decimeter activities with high accuracy. Yu *et al.* [15] built a method that uses CSI values provided by COTS WiFi devices to measure the movement distance and heading direction of human hands. Vasisht *et al.* [16] utilized a novel algorithm that computes distances between antennas and the client with a MIMO access point through multiplying the time-of-flight with the speed of light to achieve decimeter-level localization accuracy. Recent work [17] even tried to use the learning-based method to enhance the outdoor localization. In summary, WiFi-based methods perform well in terms of accuracy but they are greatly affected due to multipath interference in the complex indoor environments. In this article, multimodal sensing indoors can help weaken the measurement error caused by multipath effect.
2) *Image-Based Indoor Localization:* Gao *et al.* [18] utilized CV and crowdsourcing to reconstruct a floor plan by extracting direction and position information from images, and acquiring the spatial relation from

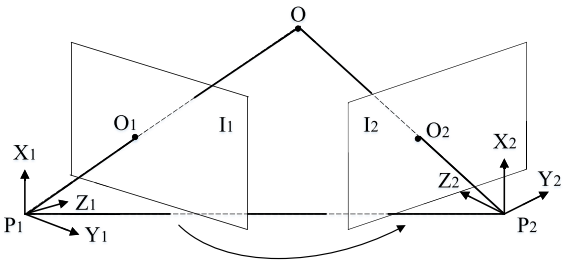


Fig. 1. Target localization model based on two images.

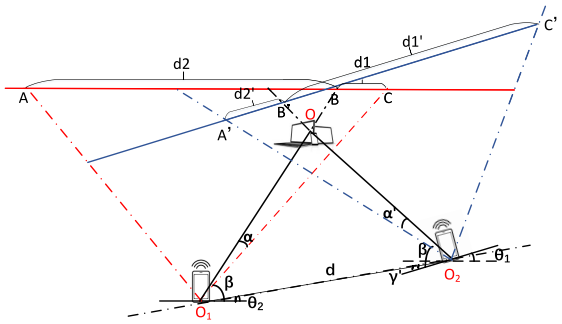


Fig. 2. Multimodal indoor localization model.

embedded sensors. Chen *et al.* [19] proposed a mechanism that jointly leverages images taken with smartphones and sensor data to reconstruct indoor skeleton. Zheng *et al.* [20] proposed a vision-based navigation system Travi-Navi that collects high-quality images on the trajectory of pedestrian, which packs visual clues and sensor data for accurate position measurements. Most image-based localization methods require massive images to construct the back-end database. In this article, our aims include reducing the overhead of image database construction, and not relying on the construction of building structures. The authors in [6], [7], and [21] used the positioning method of image database with image descriptors and efficient indexing. Meanwhile, Vedadi and Valaee [22] automatically constructed the image database capable of synthesis with any indoor image-based localization method. Then, Werner *et al.* [23] proposed three mode positioning with the first mode based on an image similar to [6], [7], and [21] and the other two based on the same video streaming as in [24]. Recent papers such as [25], and [26] further improved in the form of a single image and 2-D floor map localization.

3) *Multimodal Indoor Localization:* Xu *et al.* [4] proposed an enhanced WiFi-based localization approach by extracting geometric constraints from crowdsourced images collected by the back camera of smartphones and reduced fingerprint ambiguity by mapping constraints against fingerprint spaces. Xu *et al.* [5] proposed a method that is rooted in extracting semantic information from a few images and combining it with sensor data after optimization. Dong *et al.* [27] leveraged WiFi fingerprints to select partitions for building 3-D models from the crowdsourced 2-D photographs collected by smartphones, which meshes the paths recognized from the user motion and compiles a trajectory navigation for the pedestrian. Fusion algorithms commonly achieve higher indoor localization accuracy but with high cost of calculation and more complex deployments generally.

III. Preliminary

In theory, localization can be performed given as few as two images shooting at different positions. The localization principle is shown in Figs. 1 and 2. Two photographs I_1 and I_2 of a PoI (we use target and PoI interchangeably hereafter) O are taken at two positions P_1 and P_2. Let O_1 and O_2 denote the corresponding projections. We have

$$\begin{cases} \frac{\sin(\gamma-\alpha)}{\sin\alpha} = \frac{d_2}{d_1} \\ \frac{\sin\alpha'}{\sin(\gamma-\alpha')} = \frac{d_2'}{d_1'} \end{cases} \tag{1}$$

where γ is the maximum shooting angle for the camera, $\alpha = \angle OO_1C$, $\alpha' = \angle OO_2B$, d_2/d_1, d_2'/d_1', respectively, represent the distance ratio of the target to the boundaries in the two images. Thus, the orientation and distance of the camera from the target in the image can be calculated as follows:

$$\begin{cases} \angle OO_1O_2 = \alpha + \beta - \theta_2 \\ \angle OO_2O_1 = \alpha' + \beta - \theta_1 + \theta_2 \\ OO_1 = \sin\left(\alpha' + \beta - \theta_1 + \theta_2\right) * d / \sin\angle O_1OO_2 \\ OO_2 = \sin(\alpha + \beta - \theta_2) * OO_1 / \sin\left(\alpha' + \beta - \theta_1 + \theta_2\right) \end{cases} \tag{2}$$

where θ_1 and θ_2, respectively, represent the rotation angle of the smartphone, d is the camera moving distance from P_1 to P_2, and $\beta = (1/2)(180^\circ - \gamma)$.

In this way, if projections O_1 and O_2 are detected in the two images, and the rotation angles θ_1, θ_2 and the distance d are known, we can derive the position of the camera via (2). In the following section, we will describe how to obtain the required information for localization.

IV. Multimodal Localization

A. System Overview

As explained in the last section, we require two photographs taken in two shooting positions and we need to detect the PoIs in the two photographs, obtain the rotation angels θ_1, θ_2 of the smartphone, and calculate the distance d between two shooting positions. Therefore, we divide the proposed multimodal indoor localization system into four components as shown in Fig. 3.

1) *Multimodal Data Collection:* As depicted in Fig. 4, the data collection phase begins with taking one photograph of the PoIs in an indoor environment. Then, the user pushes the smartphone to take the second photograph of the same PoIs. During the movement, the smartphone continuously collects the IMU data and the WiFi CSI signal data.

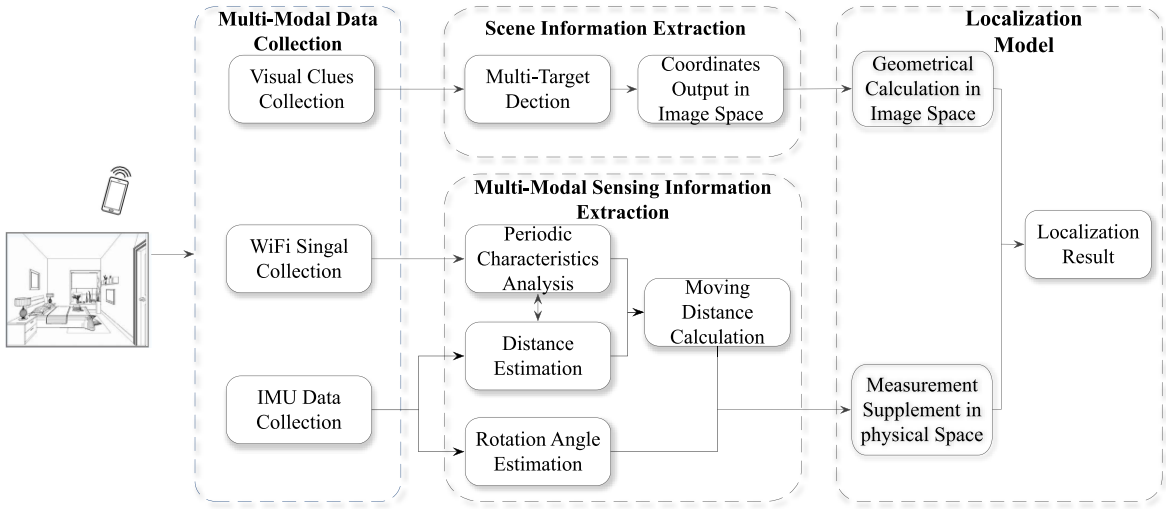


Fig. 3. System architecture.

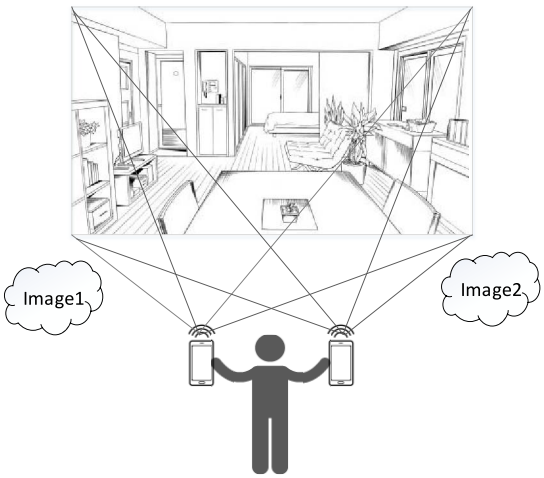


Fig. 4. Illustration of PoIs photography taking.

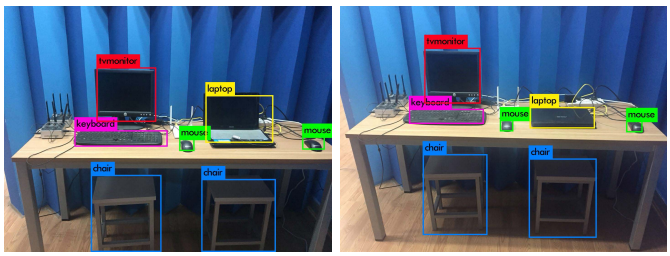


Fig. 5. Target detection in images.

2) *Scene Information Extraction:* This module is used to obtain the geometrical relationship between the two photographs in the image space, i.e., d_1/d_2 and d_1'/d_2' as shown in (1). Different from the traditional computer vision technique which calculates the correspondence between two photographs at the pixel level, we directly utilize a target (PoI), such as a keyboard or a laptop as shown in Fig. 5 to calculate the correspondence. Moreover, in order to avoid useless PoI detection based on low-quality images, we first assess the image quality. In particular, we first integrate the collected IMU data to roughly estimate the moving distance between the two images, determining the detection range accordingly, and perform feature matching in the detection area. Then, the scale-invariant feature transform (SIFT) [28] algorithm is utilized to match the feature points in two images to confirm whether the quality of the two photographs is sufficient. If the correspondence degree of the two photographs is relatively low (the number of correspondences is lower than a threshold $a = 300$), the proposed system will guide the user to retake another two photographs.

Consequently, we adopt a multiobject detection framework based on YOLO V2 [29] to perform target detection, which mainly uses a joint training method for target classification that ensures the detection accuracy in real time. YOLO V2 identifies the targets in two images and outputs their coordinates in the image space. The detection results in two images are shown in Fig. 5. The working procedure is described in Algorithm 1. As a result, we can calculate the geometric relationship between the images.

3) *Multimodal Sensing Information Extraction:* We integrate the IMU data and CSI data to accurately measure the rotation angles θ_1 and θ_2 and the accurate moving distance d of the smartphone. As the orientation measurement of IMU has sufficient accuracy, we directly use it to obtain the smartphone orientation, i.e., θ_1 and θ_2 in (2). The accelerometer built in the smartphone can be used to estimate its movement distance, which however often has a large error from the ground truth. Fortunately, multisensor data fusion combining the IMU sensor and WiFi signal can effectively improve the accuracy. Based on the accelerometer sensor measurement as the raw data, the gyroscope is used to measure the angular velocity of the smartphone, and the rotation vector sensor is used to convert the acceleration value from

Algorithm 1 Target Detection

Input:

Image set I=$\{I_1, I_2\}$; Movement distance between two images d collected by IMU; A preset correspondence degree threshold a.

Output:

Distance ratios d_2/d_1 and d'_2/d'_1 used in Eq. (1).

Divide the two images into multiple regions for feature matching according to d;
Calculate the correspondence degree α of the two images by SIFT;
while $\alpha < a$ **then**
    Take another photos;
Detect target objects utilizing YOLO V2;
for each detected target **do**
    Record the distance ratios d_2/d_1 and d'_2/d'_1 from visual boundaries.
return the average d_2/d_1 and d'_2/d'_1

Algorithm 2 Localization by Multimodal Data

Input:

IMU data; two images; WiFi CSI data;

Output:

Global position of the smartphone (the user);

Geometric Relations Acquisition:
Utilize the IMU data to determine the detection ranges in two images;
Identify multiple targets in two images;
Get the coordinates of the targets in the image space;
Obtain the distance ratios;
Moving Distance and Rotation Angle Estimation:
Process multisensor data for primary distance measurement according to Eq. (3);
Obtain the rotation angle from the rotary vector sensor;
Calculate the distance via CSI signal for complementary measurement;
Fuse both distance measurements according to Eq. (4) to obtain a final distance value d;
Localization:
Performing localization according to Eq. (2);

the smartphone coordinate system to the inertial coordinate system. As for the WiFi signal, when the user pushes the smartphone, the phase of the dynamic component of CSI will change accordingly, leading to the fluctuation of waveform. Hence, the moving distance can be estimated based on the Fresnel zone model [30]. When the smartphone moves between adjacent ellipses in the Fresnel zone, the reflection path changes by half of the wavelength. We can infer the moving distance by calculating the number of ellipses the smartphone passes through. Then, we further combine the measurements of wireless CSI signal to complete the distance calculation.

4) *Indoor Localization via Multimodal Sensing:* Based on the obtained geometrical relationship between the two photographs, distance, and orientation of the smartphone, localization is finally realized via (2). The working procedure is summarized in Algorithm 2. In the following, in detail, we illustrate how we could infer accurate d based on the IMU data and WiFi signal.

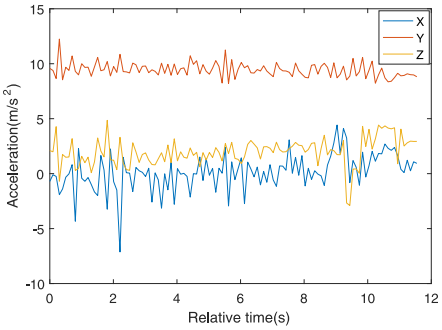


Fig. 6. Original acceleration data.

B. Accurate Distance Inference Based on IMU and WiFi Data

1) Rough Estimation via IMU Sensors: In indoor environments, the data of sensors and wireless channel are affected by activities people perform indoors. As aforementioned, we first perform primary moving distance and rotation angle estimation via multisensor data. As shown in Fig. 6, during the user's movement, accelerometers are susceptible to external disturbances, incurring large fluctuations and many high frequency components in data. Hence, the Butterworth low-pass filter is a natural choice which removes high-frequency noises. In order to eliminate the influence of static gravitational acceleration, the original data need to be dehomogenized. The waveform after gravity effect elimination and Butterworth filter is shown in Fig. 7. Due to drifts caused by movement, the data acquired by the accelerometer are not continuous with the previous moment, which leads to angle deviations. Therefore, the quaternion space coordinate conversion algorithm is considered to map the collected accelerometer data from the smartphones coordinate system to the actual inertial coordinate system in order to reduce the effects of drifts. The formula for coordinate conversion is expressed as follows:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} ABC \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3}$$

$$A = \begin{bmatrix} m_1^2 + m_2^2 - m_3^2 - m_4^2 \\ 2(m_2 m_3 - m_1 m_4) \\ 2(m_2 m_4 + m_1 m_3) \end{bmatrix} \tag{4}$$

$$B = \begin{bmatrix} 2(m_2 m_3 + m_1 m_4) \\ m_1^2 + m_3^2 - m_2^2 - m_4^2 \\ 2(m_3 m_4 - m_1 m_2) \end{bmatrix} \tag{5}$$

$$C = \begin{bmatrix} 2(m_2 m_4 - m_1 m_3) \\ 2(m_2 m_3 + m_1 m_4) \\ m_1^2 + m_4^2 - m_2^2 - m_3^2 \end{bmatrix} \tag{6}$$

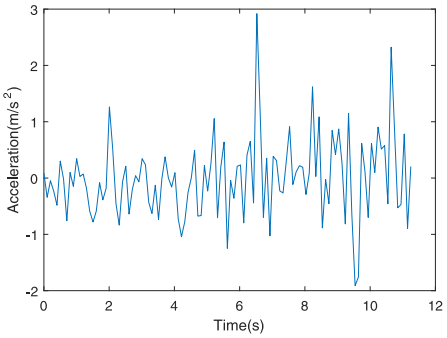


Fig. 7. Acceleration data processed with Butterworth filter.

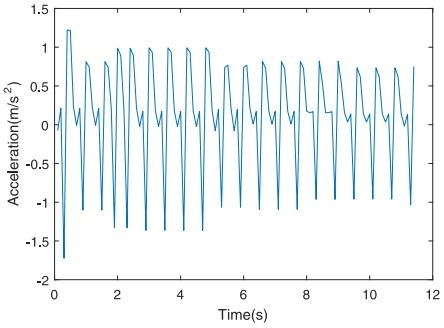


Fig. 8. Acceleration after conversion.

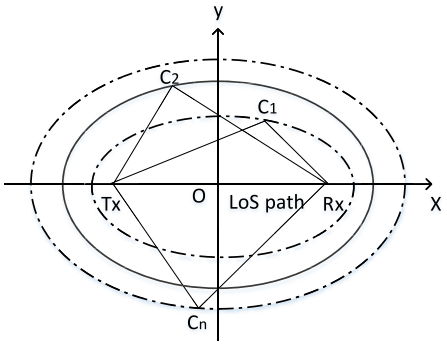


Fig. 9. Fresnel zone model.

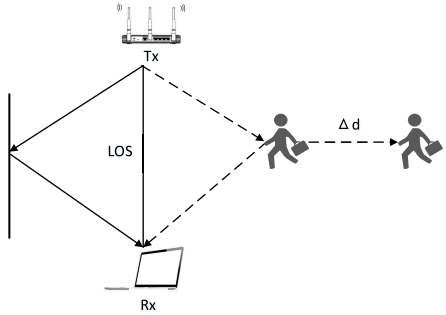


Fig. 10. Indoor multipath environment.

where $[x, y, z]^T$ represents the smartphone coordinate, $[x', y', z']^T$ represents the inertial coordinate, m_1 represents the numeric part of rotation vectors, m_2 represents the rotation vectors along the x-axis, so as m_3 for y-axis, and m_4 for z-axis.

Fig. 8 shows the accelerometer data distribution after the coordinate conversion and Butterworth filter. Moreover, the gyroscope provides angular velocity information for the calculation of moving distance. Then, we combine the angular velocity and acceleration information to obtain the moving distance as follows:

$$s = \int_0^t \left[a' \cos(\omega t) + \upsilon_0\right] \mathrm{d}t \tag{7}$$

where ω is the angular velocity, υ_0 is the initial speed, a' is the accelerometer measurement, and s is the moving distance.

2) Calibration via CSI of WiFi: We further calibrate the distance measurement based on the WiFi CSI signals integrating with the Fresnel model. As described in Section IV-A, after the first photograph was taken, the user is required to push the smartphone to take the second photograph. When the smartphone moves in the Fresnel zones, the radio signal travels from the transmitter to the receiver through the direct path and reflected path. The Fresnel zone demonstrates the relationship of reflector's location and continuously marks the positions in which channel frequency response (CFR) power is enhanced or degraded [30]. As shown in Fig. 9, when a smartphone's location (C_1) is at the first Fresnel zone boundary, the reflected path is $\lambda/2$ longer than the direct path, where λ is the wavelength. This adds the phase shift by π because the two signals are in the same phase and result in constructive interference. Similarly, the smartphone (C_2) located at the second Fresnel zone boundary leads to destructive interference.

In order to characterize the case of CSI signal in the presence of moving objects indoors, we have studied the typical setup of WiFi devices indoor shown in Fig. 10, where the signal is transmitted indoors through multiple paths to the receiving end. These paths can be divided into static and dynamic paths. The received signal $H(f, t)$ can be expressed by the following equation:

$$H(f, t) = H_s(f) + H_d(f, t) = H_s(f) + a(f, t)e^{-\frac{j2\pi d(t)}{\lambda}} \tag{8}$$

where $H_s(f)$ is the static vector representing the sum of the signals from the static path; $H_d(f, t)$ is the dynamic vector, introduced by the moving object; $a(f, t)$ is the amplitude of the dynamic path; $e^{-(j2\pi d(t)/\lambda)}$ is a complex value representation of the initial phase offset; and $d(t)$ is the dynamic path length. It can be seen from the formula that when the length of the reflected signal changes by λ, its phase shift is 2π. Therefore, the received signal $H(f, t)$ has a time-varying amplitude in the complex plane

$$|H(f, \theta)|^2 = |H_s(f)|^2 + |H_d(f)|^2 + 2|H_s(f)||H_d(f)| \cos\theta \tag{9}$$

where θ is the phase difference between the static vector and the dynamic vector. This model is essential for extracting the movement and gesture information from WiFi CSI data [31]. The static vector is the combined direct signal and the reflect signal from static object, while the dynamic vector is the one from moving objects. Basically, the dynamic vector could be obtained by subtracting the current signal from the signal in the static environment. In our system, we only need the phase difference between two dynamic vectors by subtracting them,

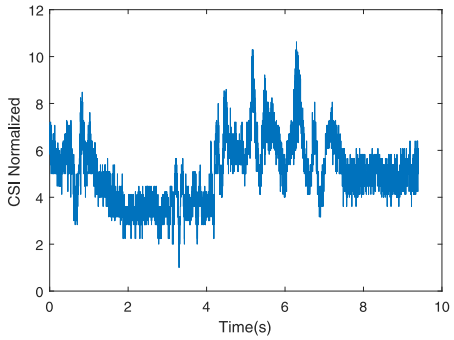


Fig. 11. Original CSI data.

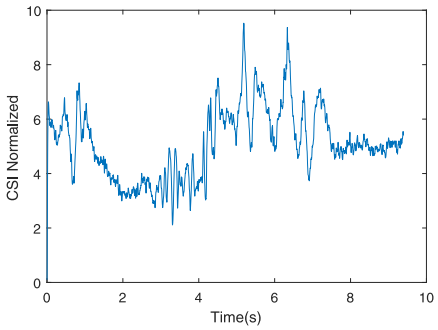


Fig. 12. CSI data processed with Butterworth filter.

as the static vectors stay the same. Therefore, when a smartphone in the room passes through several WiFi Fresnel areas, the amplitude of the CSI varies with the peak and trough. We can infer the time of the smartphone crossing the boundaries of the Fresnel zone by observing signal fluctuations to measure moving distance.

As shown in Fig. 11, the original CSI data contain many redundant information incurred by carrier frequency offset (CFO). Since the frequency of common human activities is often within 200 Hz, we adopt the Butterworth low-pass filter with a cutoff frequency of 200 Hz. As depicted in Fig. 12, it is obvious to find that high-frequency noises are removed the after Butterworth filter. However, the noise between 1 and 200 Hz cannot be estimated. As shown in Fig. 13, we use principal component analysis (PCA) to reduce the full-frequency noise further. As discussed above, we can acquire distance information by counting the regions the smartphone passes through. In detail, each subcarrier corresponds to a Fresnel zone since the wavelength of that is different. We filter each subcarrier to smooth out the signal and count the number of fluctuation periods in frequency domain to obtain the moving distance in the Fresnel zone. Note that there will exist multiple WiFi signals in the environment, thus, the unintended WiFi signal also has impact on our Fresnel zone. Mostly, the unintended signal will be harmful to our system, especially for those having strong signal strength. However, as our system only relies on detection of the peak of the signal to build the Fresnel zone model, our system is robust to small interference.

3) Data Fusion for Moving Distance: Considering the accumulative error of IMU data and the instantaneous error in the periodic characteristics of the CSI signal, we combine both

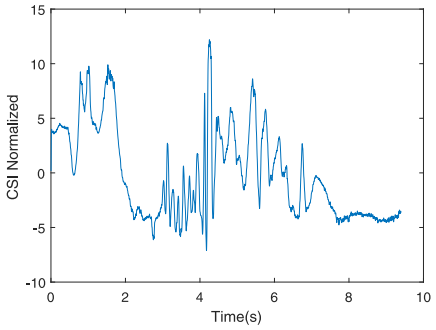


Fig. 13. PCA-based denoising.

measurements to obtain a final distance measurement. Let d_1 denote the distance measurement by IMU and d_2 by CSI signals. To compensate for the drift and dynamic error caused by IMU, we consider that wireless CSI data are more reliable in a short time. Hence, d_1 has a small weight coefficient to weaken the influence of the abnormal data while d_2 has a larger weight to suppress the impact of frequency offsets. We can calculate the associated weight coefficients through determining the time constant, which is an important descriptive measurement of dynamic performance in the first-order systems. The differential equation and transfer function of the first-order system are expressed as follows:

$$\frac{dX(t)}{dt} + cX(t) = u(t) \tag{10}$$

$$R(s) = \frac{X(s)}{u(s)} = \frac{1}{s+c} \tag{11}$$

where $X(t)$ is the system output, $u(t)$ is the system input, $R(s)$ is the transfer function, and c is a constant. If the substituting unit step input $u(t)$ into solution, $u(t)$ can be given by the following equation:

$$u(t) = \begin{cases} 1 & \tau \geq 0 \\ 0 & \tau < 0. \end{cases}$$

Then, the system output is represented by

$$X(\tau) = 1 - e^{-c\tau}. \tag{12}$$

The time constant τ is defined as $1/c$, and its response $X(1/c)$= 0.632. So, we can obtain weight coefficient from the following equation:

$$\eta = \tau/(\tau + t). \tag{13}$$

Therefore, the accurate moving distance d can be finally calculated by

$$d = \frac{\tau}{\tau + t} \times d_1 + \left(1 - \frac{\tau}{\tau + t}\right) \times d_2. \tag{14}$$

With the distance to multiple PoIs and the position of these PoIs, we can easily use this distance to localize the photograph shooter's location.

V. Localization in Multiple User Environment

In this section, we further extend the proposed multimodal localization to the scenarios with multiple users.

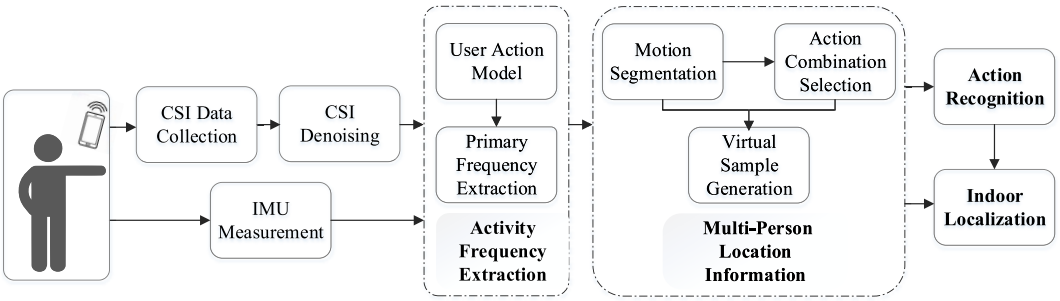


Fig. 14. Procedure of multiperson indoor localization.

A. Basic Idea

As mentioned in Section IV-B, we rely on CSI signal to obtain a precise distance measurement. Although with very high precision, the CSI-based distance could be easily compromised with multiple moving objects nearby. Specifically, multiple (mobile) users performing various actions (including sitting down, walking, and so on) will cause big inference in the CSI signal, which will bring a big challenge in extracting the moving distance of the target user.

To address this challenge, we first assume that, during a short time period, only a single target user (who wants to localize him/herself) performs the *push* action to measure the distance with the CSI signal. Then, we need to identify the push action from the mixing CSI signal influenced by multiple actions. Basically, we introduce the virtual action sample to resolve this problem. Here, the virtual action samples refer to a possible CSI signal incurred by different action combinations. With these virtual action samples, we could try to extract the push action from the other overlapping CSI signals by subtracting the real received signal to the generated virtual samples. Moreover, the virtual sample is advantageous in reducing the training cost. We do not require multiple users to provide training samples for all possible combinations of predetermined actions. Instead, we utilize the real sample of each action to generate a virtual sample for any desired combination of actions. In this way, only a single user is required to provide training samples for each action.

The workflow of the localization in multiperson scenario is as follows: 1) we detect whether there are multiple users performing actions simultaneously indoors; 2) we determine the number of simultaneous actions performed in the detected CSI signal; 3) we identify the time range where multiple actions are performed and we further separate the start and end time of each action; 4) we filter out reasonable action combinations and generate a virtual user action sample; 5) we compare the detected action sample with the generated virtual samples to identify the types of action combinations; and 6) we determine the sequence of actions that contain the target action and feature model, from which we calculate the orientation and distance. In summary, as shown in Fig. 14, the solution is divided into three main modules: 1) data acquisition and processing; 2) frequency extraction; and 3) action recognition and indoor localization. The detailed design of each component is presented as follows.

B. Data Acquisition and Processing

This module collects the CSI data from commercial wireless devices. As described above, the transmitter has two antennas and the receiver has three antennas. Thus, it takes $2 \times 3 \times 30 = 180$ CSI streams as the input and converts each CFR value in each stream to CFR power through multiplying by its complex conjugate. Specifically, (8) is transformed into

$$H(f,t) = e^{-j2\pi\Delta ft}\left(H_s(f) + \sum_{\forall k \in \cup_{i=1}^{n_u}} a_k(f,t)e^{-\frac{j2\pi d_k(t)}{\lambda}}\right) \quad (15)$$

where n_u is the number of actions performed simultaneously and k is the number of dynamic paths. The CFR power can be further calculated by the following formula:

$$\begin{aligned}|H(f,t)|^2 = {} & |H_s(f)|^2 + \sum_{\forall k \in \cup_{i=1}^{n_u}\varphi_i} |a_k(f,t)|^2 \\ & + \sum_{\forall k,p \in \cup_{i=1}^{n_u}\varphi_i; k \neq p} 2|a_k(f,t)a_p(f,t)| \\ & \cos\left(\frac{2\pi(\upsilon_k - \upsilon_p)t}{\lambda} + \frac{2\pi\left(d_k(0) - d_p(0)\right)}{\lambda} + \phi_{kp}\right) \\ & + \sum_{\forall k \in \cup_{i=1}^{n_u}\varphi_i} 2|H_s(f)a_k(f,t)| \\ & \cos\left(\frac{2\pi\upsilon_k t}{\lambda} + \frac{2\pi d_k(0)}{\lambda} + \phi_{sk}\right) \end{aligned} \quad (16)$$

where $[(2\pi(d_k(0) - d_p(0)))/\lambda] + \phi_{kp}$ and $[(2\pi d_k(0))/\lambda] + \phi_{sk}$ represent the initial constants of different indoor paths.

Next, similarly, we use the wireless CSI processing method mentioned above, i.e., the Butterworth low-pass filtering and PCA processing. Since the principal components after PCA have correlation, the signals of human actions captured in a certain principal component can also be acquired in the other principal components. So the CSI component with the highest signal-to-noise ratio among the principal components is selected as the primary source of data for subsequent modules.

C. Action Frequency Extraction

1) User Action Modeling: We establish the feature model $V_{n,k}^t$ for each action which includes four types of characteristics: time, moving distance, orientation, and speed of the movement. Let $V_{n,k}^t$ represent the feature vector generated

from the kth training sample of the nth predefined action, and the feature vector is defined as

$$V_{n,k}^{t} = \left\{ \text{time},\ \text{distance},\ \text{orientation},\ \frac{\text{distance}}{\text{time}} \right\} \tag{17}$$

where *time* is the action duration; *distance* is the action span in the Fresnel region, characterized by the number of fluctuation periods in the frequency domain; and *orientation* is the direction of the action, which can be expressed by the distance ratio on the two coordinate axes in the 2-D Fresnel region mentioned above. It can be characterized using the discrete wavelet transform (DWT) to calculate the phase delay between different subcarriers; *distance*/*time* is the speed of the movement. Since the duration of different training samples will be different, the feature model needs to be normalized by the number of window steps of the frequency to quantify the frequency introduced by actions in the CFR power. The normalization is as follows:

$$\overline{V_{n,k}^{t}} = \frac{V_{n,k}^{t}}{(\text{time} - w)/m_n} \tag{18}$$

where w is the width of the sliding window and m is the number of training samples for the nth predefined action.

2) Primary Frequency Extraction: We take the denoised flow of the training samples for any given action (obtained from the output of the data acquisition and processing module) as input and slide it over a short time width window in a certain step size. First, we collect the CSI data under the static background (that is, no active users) in the indoor environment. Then, we obtain the mean variance of the collected wireless CSI data after PCA and DWT processing, and the value is set to a threshold T for judging whether the indoor user is in an active state. In the active state, it is necessary to further determine the number of total frequencies of the CSI due to users' activities. All paths of CSI in the indoor environment include the static paths and the dynamic paths. The number of total paths is represented by N, which can be measured by the number of peaks greater than threshold T of all frequencies. When the user performs activities indoors, if the number of paths whose path length changes at different rates is s, that is, the number of primary frequencies which are introduced by human actions is s, and the number of secondary frequencies is $\binom{s}{2}$, where $N = s + \binom{s}{2}$. Specifically, we consider the binary search method to determine the number of primary and secondary frequencies and then distinguish which frequencies are the primary frequencies. We calculate the pairwise difference between the s frequency values and figure out the distance between the vectors for the $\binom{s}{2}$ difference value and the remaining $\binom{s}{2}$ frequencies. Finally, we select the set of vectors with the shortest vector distance as the primary and secondary frequencies, respectively. Since the module needs to search for the primary and secondary frequencies in turn, the calculation cost is relatively high compared with the other modules. But this module is only performed once during the training phase, which does not have a great impact on the running speed during localization.

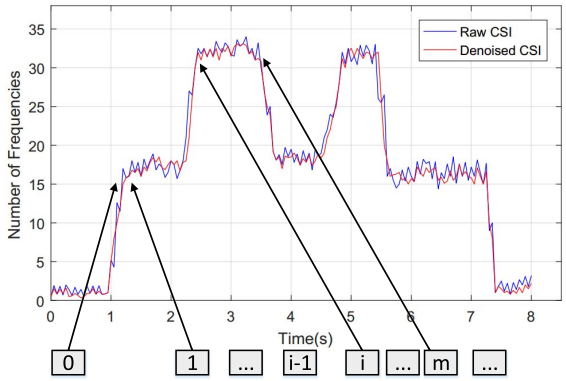


Fig. 15. Segmentation of combination and action.

D. Action Recognition and User Localization

1) Action Segmentation: We take the wireless CSI stream after denoising of samples as input and continuously slide a window of short duration on it. For the case where multiple users simultaneously perform actions, it can be observed that when the variance of the wireless CSI data in the continuous window is greater than the static threshold T, it can be determined that one or a group of actions is performed by the users; otherwise, it can be determined that one or a group of actions is terminated by indoor users.

When multiple users perform actions simultaneously, the segmentation method of action combination is inconspicuous for each action division. For the beginning and end of each action in a combination, we consider making a judgment by processing the frequency vector of samples in each window step. In the first window step, the number of frequency vectors N_{f0} is only introduced by the first action. Then, we compare the number of frequency vectors $N_{f\tau}$ with the exponential weighted average $\mu_{f(\tau-1)}$ of frequency vectors, where

$$\begin{aligned} \mu_{f(\tau-1)} &= \frac{\mu_{(\tau-2)} + N_{f(\tau-1)}}{2} \\ \mu_{f0} &= N_{f0}. \end{aligned} \tag{19}$$

If $N_{f\tau} - \mu_{f(\tau-1)} > \alpha^{i} N_{f0}$, it means that the number of frequency vectors increases rapidly at time τ, which indicates that another actions are performed by the other users, where i is the number of actions, and α is a constant. Correspondingly, $N_{f\tau} - \mu_{f(\tau-1)} < \alpha^{i} N_{f0}$ represents that the number of frequencies is rapidly reduced, which means that an action is terminated at this time. Considering that when a user introduces a new action in wireless environment, the existing action is being executed by other users, so α must be less than 1, and we take 0.9 as its initial value in the algorithm design. When the system detects that the number of actions in the beginning is greater than that in the end, the value of α increases by a step of 0.02. Otherwise, α decreases by a step of 0.02. When a new action is detected, the feature model $V_{n,k}^{t1}$ is established for the action; when an action is ended, we establish the feature model $V_{n,k}^{t2}$. Then, we pair $V_{n,k}^{t1}$ with $V_{n,k}^{t2}$, and return a set of time series (t_1, t_2). As illustrated in Fig. 15, with the highest matching degree of feature model, so as to match the start and end of an action in combinations.

2) Action Combination Selection: With n_A as the number of predefined actions and n_u as the number of simultaneously performed actions, we have ${n_u}^{n_A}$ action combinations. The number of candidates is large, which makes the action recognition too complicated. Therefore, we need to reduce the number of candidate actions and filter out reasonable action combination. The specific method is: 1) since the frequency length caused by human actions does not exceed 300 Hz, we remove the candidate actions whose amplitude difference between any frequencies is greater than $\sqrt{300}$ and 2) we do not consider the combination of predefined action samples that match the detected feature model by less than 50%. Based on the start and end time of each action, we filter the possible action combinations and their execution time pairs. The action combination and time pair are used as the input of the virtual action sample generation module.

3) Virtual Sample Generation: Based on the action combination, we generate the virtual samples in the form of a binary matrix. Before generating virtual samples for any given action combination, we need to determine whether the duration between randomly selected training samples matches the filtered ones. Basically, if the duration T_{combine} of filtered action combinations is greater than the duration T_{train} of the training samples, the $(T_{\text{combine}} - T_{\text{train}})/T_{\text{train}}$ primary frequency sets are supplemented and renumbered in chronological order; if the duration T_{combine} of filtered action combinations is less than the duration T_{train} of the training samples, the $(T_{\text{train}} - T_{\text{combine}})/T_{\text{train}}$ primary frequency sets are truncated and renumbered in chronological order.

Specifically, first, we retrieve the sets of indoor user activity frequencies in the action training samples associated with the time series pair (t_1, t_2); then, the entries in the matrix corresponding to primary frequencies are set to 1, which is equivalent to inserting the characteristics of primary frequencies in virtual samples; moreover, we insert the characteristics of secondary frequencies into virtual samples, and calculate the pairwise difference between primary and secondary frequencies. We set the entries in the matrix that correspond to the calculated values of primary and secondary frequency differences to 1, and insert the characteristics of secondary frequencies into virtual samples; finally, we complete the generation of virtual samples by supplementing the features of the primary and secondary frequencies.

4) Localization of Multiple Users: The action recognition of multiple users in an indoor environment is essentially comparing the similarity between the binary matrix M_j of detected samples and the binary matrix M_v of virtual samples. We take virtual samples, and denoised wireless CSI stream from the start of first detected action to the end of the last action as input. We first use frequency vectors obtained from the above work as detected sample matrices. When the frequency is greater than the static threshold, the user is not stationary, and the action is performed. So we set the corresponding entry in the matrix to 1; otherwise, we set it to 0. Thus, the detection sample matrix M_j is generated. Furthermore, we confirm the similarity between the detected samples and virtual samples by calculating the Jaccard coefficients of M_j and M_v, and identify a plurality of actions performed by indoor users.

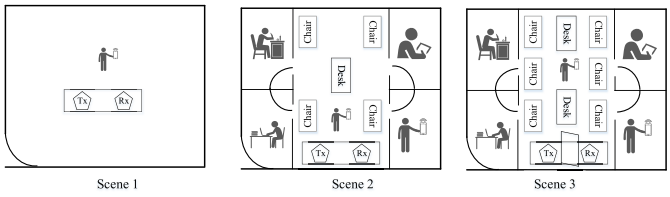


Fig. 16. Experimental settings in different environments.

In the end, the moving distance of the smartphone is finally calculated. We can infer the localization result based on (2) in the same way as in Section IV.

VI. Performance Evaluation

In this section, we will introduce the experimental settings and evaluation setup of the system, then we will evaluate the performance of the proposed method in terms of accuracy and efficiency.

A. Experimental Implementation and Setup

1) Prototype Implementation: The prototype consists of a Google Nexus 5X as the monocular camera and IMU sensor data acquisition device, an Intel NUC D54250WYKH computer with an Intel 5300 NIC as the WiFi signal receiver, and a mini R1C wireless router as the transmitter. The implementation is conducted in the 5-GHz frequency band with 20-MHz band with channels. In addition, the transmitter has two antennas and the receiver has three antennas. We sample the IMU data at a rate of 50 samples/s and CSI singles at 2500 samples/s. In each scenario in indoor environments, we choose 10 PoIs and take two photographs for each PoI from different angles. In total, we have collected 1200 samples from eight volunteers.

2) Experiments in Single User Environment: We first design several experimental settings when the user varies in arm movement and the movement of arm and body during the process of taking photographs. Then, the effect on localization accuracy caused by different smartphone movement patterns is explored. In this experiment, we use the real distance value as the baseline. Furthermore, we compare the effect of the acquired image number on the results. On the other hand, we compare the existing indoor localization system Argus [4], ClickLoc [5] to our solution on the same dataset. The extensive experiments are conducted in the following environments: an empty room, a laboratory with scattered tables and chairs, and the same laboratory with more tables and chairs and also a metal plate is placed between the user and the receiver, as illustrated in Fig. 16.

3) Experiments in Multiple Users Environment: For the multiperson environments, we also consider three indoor scenarios as shown in Fig. 16. The data collection process is as follows. Basically, we collect training samples of five actions from eight volunteers, including walking (w), pushing (p), sitting down (s), falling down (f), and running (r). We first ask each volunteer to provide 20 samples for each action in a random indoor scenario, and then take 16 sets of action combinations consisting of 2, 3, and 5 actions each. In total, we

TABLE I
ACTION COMBINATION IN THE EXPERIMENTAL DATA SET

ID	Number of Actions	User ID & Actions
1	2	$\nu_1, w\ \nu_4, w$
2	2	$\nu_1, p\ \nu_4, w$
3	2	$\nu_1, p\ \nu_4, r$
4	2	$\nu_1, p\ \nu_2, p$
5	3	$\nu_3, w\ \nu_6, p\ \nu_7, s$
6	3	$\nu_5, w\ \nu_8, p\ \nu_2, s$
7	3	$\nu_5, p\ \nu_8, w\ \nu_2, r$
8	3	$\nu_4, f\ \nu_1, p\ \nu_2, s$
9	4	$\nu_2, w\ \nu_3, p\ \nu_7, f\ \nu_8, s$
10	4	$\nu_2, r\ \nu_3, w\ \nu_7, s\ \nu_8, p$
11	4	$\nu_1, r\ \nu_2, w\ \nu_4, s\ \nu_6, p$
12	4	$\nu_1, p\ \nu_3, p\ \nu_6, w\ \nu_8, f$
13	5	$\nu_1, p\ \nu_3, r\ \nu_5, w\ \nu_6, f\ \nu_8, s$
14	5	$\nu_1, p\ \nu_2, r\ \nu_4, w\ \nu_5, f\ \nu_7, s$
15	5	$\nu_1, w\ \nu_3, p\ \nu_5, r\ \nu_6, p\ \nu_8, s$
16	5	$\nu_2, f\ \nu_3, w\ \nu_4, p\ \nu_7, s\ \nu_8, f$

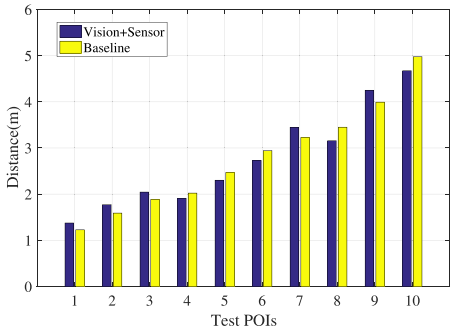


Fig. 17. Overall distance measurement performance.

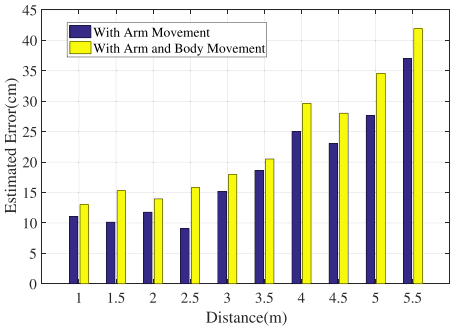


Fig. 18. Impact of activity diversity.

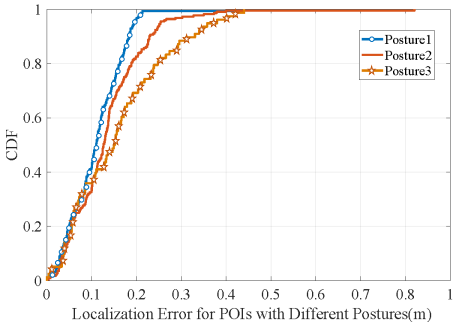


Fig. 19. Impact of different movement pattern.

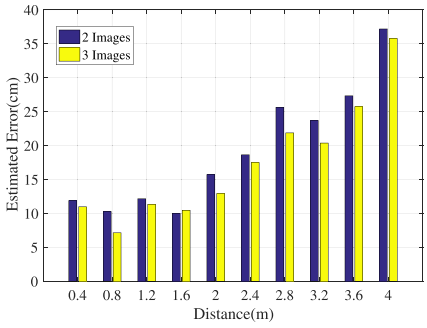


Fig. 20. Performance with different image numbers.

have collected 1760 action samples, including 800 samples of a single action and 960 samples of action combinations, 60 samples for each action combination. Table I summarizes the collected samples and the numbering of involved volunteers with corresponding action IDs.

B. Experiment Results

1) Single-User Environment: We present the performance evaluation result for our solution in single-user environment.

Overall Performance: First, we evaluate the accuracy of the prototype using the aforementioned settings. We basically conduct experiments testing the distance to 10 different PoIs at the laboratory and the meeting hall. Fig. 17 summarizes the performance of our solution, where the results are compared with the "baseline" (ground-truth distance). We can see that the 92-percentile localization errors are 0.2 m.

Arm Movement Versus Arm and Body Movement: Fig. 18 presents the localization accuracy with the user's arm movement and with the user's arm and body movement during localization. It shows that the 92-percentile errors of a single movement and a compound movement are 0.19 and 0.23 m, respectively. Multiple actions affect the IMU data and wireless data. However, the movement of the user's arm and body can be distinguished by the analysis of multimodal data. Thus, the accuracy is not greatly reduced.

Different Movement Pattern: When moving the smartphone between two shooting positions, we consider, including left and right translations, forward and backward translations, and oblique movements that can be decomposed into translation and rotation. As shown in Fig. 19, the 92-percentile estimated errors of the above postures are 0.18, 0.2, and 0.21 m, respectively. The reason for this difference is that the first two movement methods do not need to consider the measurements of rotation angles, which reduces the localization errors to some extent.

Two Photographs Versus Three Photographs: Fig. 20 illustrates the localization accuracy of taking two photographs and three photographs in the same scene. As shown in the figure, collecting three images reduces the estimated error. The reason is that the third image outputs an additional distance and direction from the user to the target objects in images, which can help correct the previous localization result.

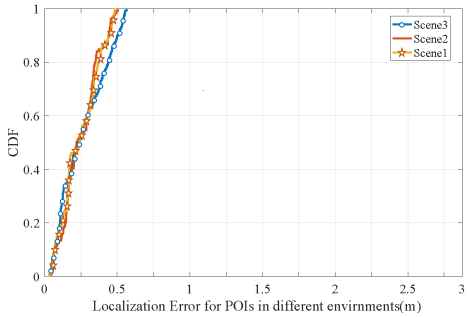


Fig. 21. Performance with different environments.

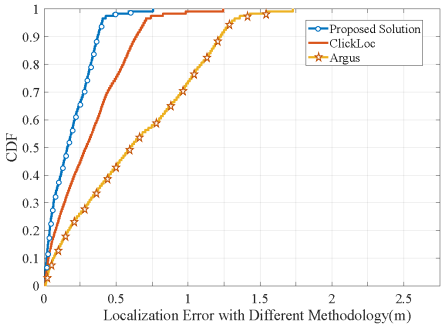


Fig. 22. Performance with different solutions.

Distinct Environments: From Fig. 21, we can see that the 92-percentile localization error in scene 1 is about 0.17 m, and that in scenes 2 and 3, the errors are 0.2 and 0.21 m, respectively. It demonstrates that the localization system performs well in all three environments.

Comparison of the State-of-the-Art: Then, we compare our system with the existing methodology. Argus is an indoor localization system that estimates the user's distance and direction via combining WiFi with visual clues, which extracts geometric constraints in the image space and takes joint methods to map the constraints to the fingerprint space. ClickLoc is a system for indoor localization through multimodal measurements on smartphones. It uses the core technology of image-based semantic information extraction and sensor-based data fusion. For a fair comparison, we evaluate Argus, ClickLoc, and our proposed solution. We further provide the same condition of sampling and measurement in the same dataset. Fig. 22 shows the comparison results. We can observe from Fig. 22 that our scheme is superior to Argus by about 36% and ClickLoc by about 19%.

2) Multiple Users Environment: In Figs. 23 and 24, we evaluate the accuracy of action recognition under the combinations of two to five actions, where multiple users sequentially execute these action combinations. It can be seen that the proposed method can achieve an average recognition accuracy of 92.87% and 94.05% when the actions are simultaneously and sequentially executed, respectively.

Under the premise that three actions are performed indoors at the same time, we investigate the influence of the distance between the users on the localization accuracy. We set the distance between the users to 65, 80, 95, 110, and 125 cm to evaluate the system performance. As shown in Fig. 25, as

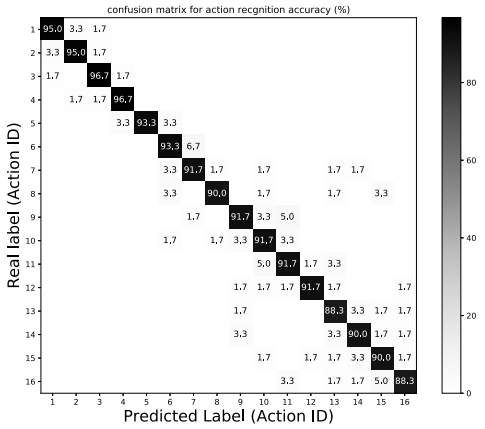


Fig. 23. Recognition result of the actions performed simultaneously.

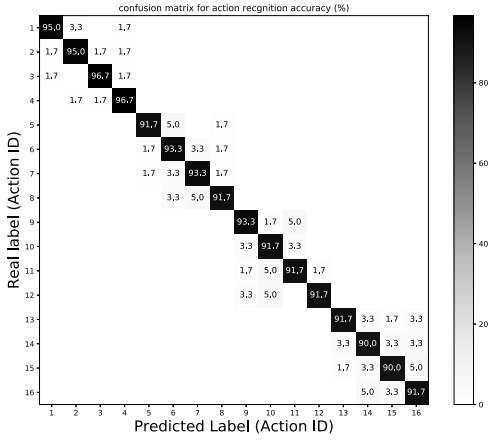


Fig. 24. Recognition result of the actions performed in sequence.

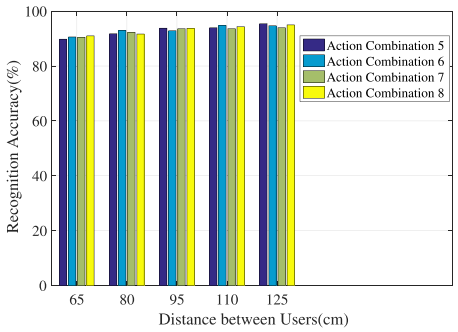


Fig. 25. Impact of the distance between the users.

the distance between users increases, the recognition accuracy of multiperson motion increases. The reason is that when the users are relatively close, the user's dynamic signal reflection is hindered to a certain extent. Furthermore, in this situation, the user may not be able to perform certain actions completely.

In addition, we evaluate the results of multiperson motion detection in three scenes. As can be seen from Fig. 26, the solution proposed in this article performs well in all scenes, even when the reflection path in the indoor environment increases and the line-of-sight path is blocked.

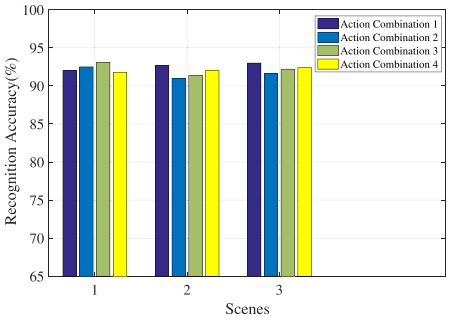


Fig. 26. Impact in different indoor environments.

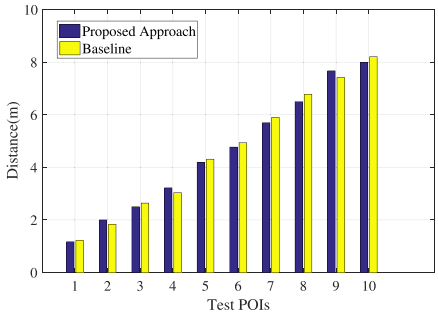


Fig. 27. Performance in multiperson environment.

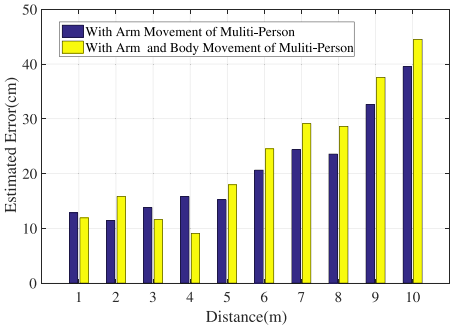


Fig. 28. Impact of activity diversity of multiperson.

Furthermore, we evaluate the overall localization performance of the target user in a multiperson environment. As shown in Fig. 27, the experimental results show that the localization error in multiperson environment is not much different from that of only one active user. The distance error of the 92-percentile is 0.22 m. Compared with the localization for a single user, the measurement error of multiperson is increased by 0.02 m, indicating that our localization algorithm is robust to multiperson interference.

Similarly, we evaluate the performance of localization in the indoor environment where multiple people perform actions simultaneously, considering the case in which the user only performs the arm movement and the simultaneous movement of the arm body. As shown in Fig. 28, the 92-percentile distance error under a single-arm movement is 0.22 m, and the 92-percentile distance error of the compound movement is 0.23 m. We verify the performance of indoor localization with three movement patterns in multiperson environment as well.

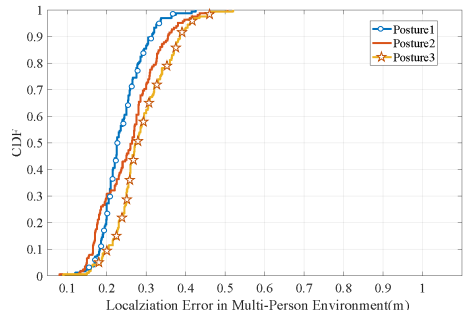


Fig. 29. Impact of movement patterns of multiperson.

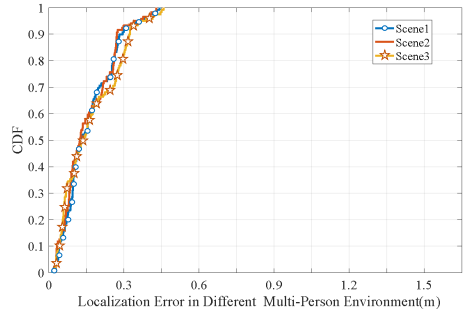


Fig. 30. Localization performance in different environments.

We consider that users perform actions with left and right translations, forward and backward translations, and oblique movements. In Fig. 29, it shows that the 92-percentile estimated errors of above postures are 0.2, 0.21, and 0.23 m. We further analyze the performance of multiperson indoor localization in three scenes. As shown in Fig. 30, the 92nd percentile errors are 0.19, 0.21, and 0.22 m. In general, our proposed indoor localization solution of multiple people has been experimentally verified to perform well.

C. Results Summary

By analyzing the experimental results above, we can summarize the following.

1) The localization accuracy of moving the smartphone with a single action is more accurate than that with compound action. However, the decrease is acceptable by further actions recognition.
2) The movement pattern of the smartphones while taking photographs has a small impact on the localization results, which demonstrates that our system is robust to different photograph-taking postures.
3) With more images taken for localization, the estimated errors are decreased.
4) The measurements of distance and orientation have low dependence on the environments.
5) Our proposed algorithm could accurately extract the push action of the photograph from multiple user environment where all the others are conducting different actions. All in all, the prototype performs well overall in different indoor environments.

VII. Conclusion

With the trends toward widespread use of the CV technique, improved sensor accuracy, and enhanced wireless connectivity, we envisioned the user-friendly and extensible computation localization service. In this article, we proposed a multimodal approach to enhance indoor localization with camera and WiFi signal. The core techniques are rooted in the mapping model from image space to physical space, the algorithm of distance, and orientation measurements. We conducted comprehensive theoretical studies and the experimental results showed that the 92-percentile error was within 0.2 m for indoor PoIs within 5 m. Our estimation could localize the user with only one PoI in two pictures taking in the same location. Furthermore, our method is robust in case of insufficient data, so it can also be applied to the indoor localization systems with sparse information.

References

[1] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "SpotFi: Decimeter level localization using WiFi," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 269–282, 2015.

[2] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu, "Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices," in *Proc. ACM MobiCom*, 2014, pp. 237–248.

[3] J. Wang, F. Adib, R. Knepper, D. Katabi, and D. Rus, "RF-Compass: Robot object manipulation using RFIDs," in *Proc. ACM MobiCom*, Miami, Florida, USA, 2013, pp. 3–14.

[4] H. Xu, Z. Yang, Z. Zhou, L. Shangguan, K. Yi, and Y. Liu, "Enhancing WiFi-based localization with visual clues," in *Proc. ACM UbiComp*, Osaka, Japan, 2015, pp. 963–974.

[5] H. Xu, Z. Yang, Z. Zhou, L. Shangguan, K. Yi, and Y. Liu, "Indoor localization via multi-modal sensing on smartphones," in *Proc. ACM UbiComp*, Heidelberg, Germany, 2016, pp. 208–219.

[6] H. Taira *et al.*, "InLoc: Indoor visual localization with dense matching and view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7199–7209.

[7] Y. Xia, C. Xiu, and D. Yang, "Visual indoor positioning method using image database," in *Proc. Ubiquitous Position. Indoor Navig. Location Based Services (UPINLBS)*, Mar. 2018, pp. 1–8.

[8] Z. Liu, L. Cheng, A. Liu, L. Zhang, X. He, and R. Zimmermann, "Multiview and multimodal pervasive indoor localization," in *Proc. ACM Multimedia (MM)*, Mountain View, CA, USA, 2017, pp. 109–117.

[9] D. Carrillo, V. Moreno, B. Úbeda, and A. F. Skarmeta, "Magicfinger: 3D magnetic fingerprints for indoor location," *Sensors*, vol. 15, no. 7, pp. 17168–17194, 2015.

[10] J. Xiong and K. Jamieson, "Arraytrack: A fine-grained indoor location system," in *Proc. USENIX NSDI*, Lombard, IL, USA, 2013, pp. 71–84.

[11] J. G. Manweiler, P. Jain, and R. R. Choudhury, "Satellites in our pockets: An object positioning system using smartphones," in *Proc. ACM MobiSys*, 2012, pp. 211–224.

[12] Y. Tian *et al.*, "Towards ubiquitous indoor localization service leveraging environmental physical features," in *Proc. IEEE INFOCOM*, Toronto, ON, Canada, 2014, pp. 55–63.

[13] S. Papaioannou, H. Wen, A. Markham, and N. Trigoni, "Fusion of radio and camera sensor data for accurate indoor positioning," in *Proc. IEEE MASS*, Philadelphia, PA, USA, 2015, pp. 109–117.

[14] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, "WiDir: Walking direction estimation using wireless signals," in *Proc. ACM UbiComp*, Heidelberg, Germany, 2016, pp. 351–362.

[15] N. Yu, W. Wang, A. X. Liu, and L. Kong, "QGesture: Quantifying gesture distance and direction with WiFi signals," in *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, 2018, pp. 1–23.

[16] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-level localization with a single WiFi access point," in *Proc. USENIX NSDI*, Santa Clara, CA, USA, 2016, pp. 165–178.

[17] J. Wang, J. Luo, S. J. Pan, and A. Sun, "Learning-based outdoor localization exploiting crowd-labeled WiFi hotspots," *IEEE Trans. Mobile Comput.*, vol. 18, no. 4, pp. 896–909, Apr. 2019.

[18] R. Gao *et al.*, "Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing," in *Proc. ACM MobiCom*, 2014, pp. 249–260.

[19] S. Chen, M. Li, K. Ren, X. Fu, and C. Qiao, "Rise of the indoor crowd: Reconstruction of building interior view via mobile crowdsourcing," in *Proc. ACM SenSys*, Seoul, South Korea, 2015, pp. 59–71.

[20] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao, "Travi-Navi: Self-deployable indoor navigation system," in *Proc. ACM MobiCom*, 2014, pp. 471–482.

[21] Y. Chen, R. Chen, M. Liu, A. Xiao, D. Wu, and S. Zhao, "Indoor visual positioning aided by CNN-based image retrieval: Training-free, 3D modeling-free," *Sensors*, vol. 18, no. 8, p. 2692, 2018.

[22] F. Vedadi and S. Valaee, "Automatic visual fingerprinting for indoor image-based localization applications," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[23] M. Werner, M. Kessel, and C. Marouane, "Indoor positioning using smartphone camera," in *Proc. IEEE Int. Conf. Indoor Position. Indoor Navig.*, 2011, pp. 1–6.

[24] M. Li, N. Liu, Q. Niu, C. Liu, S.-H. G. Chan, and C. Gao, "SweepLoc: Automatic video-based indoor localization by camera sweeping," in *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, Sep. 2018, pp. 1–25.

[25] S. Wang, S. Fidler, and R. Urtasun, "Lost shopping! Monocular localization in large indoor spaces," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2695–2703.

[26] Z. Huang, N. Gu, J. Hao, and J. Shen, "3DLoc: 3D features for accurate indoor positioning," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 4, p. 141, 2018.

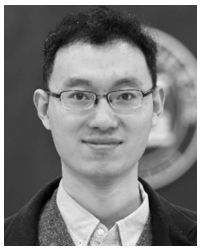
[27] J. Dong, Y. Xiao, M. Noreikis, Z. Ou, and A. Ylä-Jääski, "iMoon: Using smartphones for image-based indoor navigation," in *Proc. ACM SenSys*, Seoul, South Korea, 2015, pp. 449–450.

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

[29] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517–6525.

[30] D. Zhang, H. Wang, and D. Wu, "Toward centimeter-scale human activity sensing with Wi-Fi signals," *Computer*, vol. 50, no. 1, pp. 48–57, 2017.

[31] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Paris, France, 2015, pp. 65–76.



Yanchao Zhao (M'15) received the B.S. and Ph.D. degrees in computer science from Nanjing University, Nanjing, China, in 2007 and 2015, respectively.

He is currently an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing. Since 2011, he has been a visiting student with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. He is also a member of the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing. His current research interests include wireless network, mobile computing, edge computing, and device-free sensing.

Jing Xu (S'18) received the B.S. degree from the Nanjing University of Information Science and Technology, Nanjing, China. She is currently pursuing the M.S. degree with the Collage of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing.

Her current research interests include indoor localization and mobile computing.

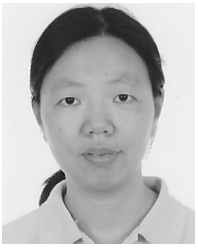
Jie Wu (F'09) received the B.S. degree in computer engineering and the M.S. degree in computer science from the Shanghai University of Science and Technology (currently, Shanghai University), Shanghai, China, in July 1982 and in July 1985, respectively, and the Ph.D. degree in computer engineering from Florida Atlantic University, Boca Raton, FL, USA, in August 1989.

He was a Program Director with the National Science Foundation and a Distinguished Professor with Florida Atlantic University, Boca Raton, FL, USA. He is the Director of the Center for Networked Computing and a Laura H. Carnell Professor with Temple University, Philadelphia, PA, USA, where he also serves as the Director of International Affairs with the College of Science and Technology. He served as the Chair of the Department of Computer and Information Sciences from summer 2009 to summer 2016 and an Associate Vice Provost for International Affairs from fall 2015 to summer 2017. He regularly publishes in scholarly journals, conference proceedings, and books. His current research interests include mobile computing and wireless networks, routing protocols, cloud and green computing, network trust and security, and social network applications.

Dr. Wu was a recipient of the 2011 China Computer Federation (CCF) overseas Outstanding Achievement Award. He serves on several editorial boards, including the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON SERVICE COMPUTING, the *Journal of Parallel and Distributed Computing*, and the *Journal of Computer Science and Technology*. He was the General Co-Chair for IEEE MASS 2006, IEEE IPDPS 2008, IEEE ICDCS 2013, ACM MobiHoc 2014, ICPP 2016, and IEEE CNS 2016, as well as the Program Co-Chair for IEEE INFOCOM 2011 and CCF CNCC 2013. He was an IEEE Computer Society Distinguished Visotor, the ACM Distinguished Speaker, and the Chair for the IEEE Technical Committee on Distributed Processing. He is a CCF Distinguished Speaker.

Jie Hao (M'14) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, in 2014.

From 2014 to 2015, she was a Post-Doctoral Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore. She is currently a Lecture with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. She is also a member of the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing. Her current research interests include visible light sensing and Internet of Things.



Hongyan Qian (M'16) born in 1973. She received the Ph.D. degree in computer application from the Nanjing University of Aeronautics and Astronautics, Nanjing, China.

She is an Associate Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. She is also a member of the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing. Her current research interests include computer network wireless communication and information security.