iVR: Integrated Vision and Radio Localization with Zero Human Effort

JINGAO XU, School of Software and BNRist, Tsinghua University, P.R. China
HENGJIE CHEN, School of Software and BNRist, Tsinghua University, P.R. China
KUN QIAN, Department of Electrical and Computer Engineering, University of California San Diego, US
ERQUN DONG, School of Software and BNRist, Tsinghua University, P.R. China
MIN SUN, School of Software and BNRist, Tsinghua University, China
CHENSHU WU, Department of Electrical & Computer Engineering, University of Maryland, College Park, US
LI ZHANG, HeFei University of Technology, China
ZHENG YANG*, School of Software and BNRist, Tsinghua University, P.R. China

Smartphone localization is essential to a wide range of applications in shopping malls, museums, office buildings, and other public places. Existing solutions relying on radio fingerprints and/or inertial sensors suffer from large location errors and considerable deployment efforts. We observe an opportunity in the recent trend of increasing numbers of security surveillance cameras installed in indoor spaces to overcome these limitations and revisit the problem of smartphone localization with a fresh perspective. However, fusing vision-based and radio-based systems is non-trivial due to the absence of absolute location, incorrespondence of identification and looseness of sensor fusion. This study proposes iVR, an integrated vision and radio localization system that achieves sub-meter accuracy with indoor semantic maps automatically generated from only two surveillance cameras, superior to precedent systems that require manual map construction or plentiful captured images. iVR employs a particle filter to fuse raw estimates from multiple systems, including vision, radio, and inertial sensor systems. By doing so, iVR outputs enhanced accuracy with zero start-up costs, while overcoming the respective drawbacks of each individual sub-system. We implement iVR on commodity smartphones and validate its performance in five different scenarios. The results show that iVR achieves a remarkable localization accuracy of 0.7m, outperforming the state-of-the-art systems by >70%.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Indoor Localization; Computer Vision; Wireless; Pedestrian Tracking

*Corresponding author

Authors' addresses: Jingao Xu, xujingao13@gmail.com, School of Software and BNRist, Tsinghua University, Room 11-211, East Main Building; 30 Shuangqing Rd, Haidian District, Beijing, 100084, P.R. China; Hengjie Chen, chenhengjie13@gmail.com, School of Software and BNRist, Tsinghua University, Beijing, 100084, P.R. China; Kun Qian, qiank10@gmail.com, Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, 92093, US; Erqun Dong, doneq13@gmail.com, School of Software and BNRist, Tsinghua University, Beijing, 100084, P.R. China; Min Sun, minsun1996@163.com, School of Software and BNRist, Tsinghua University, Beijing, 100084, China; Chenshu Wu, wucs32@gmail.com, Department of Electrical & Computer Engineering, University of Maryland, College Park, Washington DC, MD, 20742, US; Li Zhang, lizhang@hfut.edu.cn, HeFei University of Technology, HeFei, Anhui, China; Zheng Yang, hmilyyz@gmail.com, School of Software and BNRist, Tsinghua University, 30 Shuangqing Rd, Haidian District, Beijing, 100084, P.R. China.

2474-9567/2019/9-ART114 $15.00
https://doi.org/10.1145/3351272

ACM Reference Format:
Jingao Xu, Hengjie Chen, Kun Qian, Erqun Dong, Min Sun, Chenshu Wu, Li Zhang, and Zheng Yang. 2019. iVR: Integrated Vision and Radio Localization with Zero Human Effort. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 114 (September 2019), 22 pages. https://doi.org/10.1145/3351272

1 INTRODUCTION

Accurate and easy-to-deploy indoor localization is a key enabler for many applications on the horizon, such as customer navigation in supermarkets, targeted advertisements in shopping malls, and augmented (virtual) reality in public places. Since the role of localization is essential, minor errors may have significant impact. For example, a few meters of error in location estimate can place a customer in a wrong aisle within a supermarket. Meanwhile, high deployment costs may prevent the pervasive adoption of indoor location service.

Prior proposals achieving sub-meter accuracy usually require physical-layer Channel State Information (CSI) or Ultra-wide Band (UWB) signal that are not available on commercial smartphones [28, 57]. WiFi-based fingerprinting [50, 53, 56, 61] and inertial-based pedestrian dead-reckoning (PDR) [43, 55, 62, 65] are more promising for mobile and pervasive computing. However, these approaches suffer from both large location errors and considerable deployment costs. It is well-known that PDR has intrinsically accumulative errors and can hardly serve as a stand-alone tracking service on smartphones. Received Signal Strength (RSS) fingerprint-based solutions yield meters of location error due to complex indoor multipath environments [45, 54]. In addition, fingerprinting also involves a labor-intensive and time-consuming procedure of site survey to gather the RF signatures for each location. What's worse, such a cumbersome site survey may need to be repeated over time due to environmental dynamics. Finally, as shown in Fig. 1a, indoor semantic maps are desirable to ensure the rationality of the localization results (e.g. , pedestrians should not be located on a table, trajectories should not intersect with walls). Unlike outdoor localization scenarios where we can obtain road network information from map service providers (e.g. , Google Map), in indoor scenarios, the availability and quality of indoor floor-plans cannot always be guaranteed.

Nowadays, surveillance cameras are pervasively deployed in public areas, such as shopping malls, museums, galleries and so on [23]. In many occasions, it is even very common that multiple cameras cover overlapped areas. Researchers realize that these widely installed surveillance cameras could provide complementary advantages to conventional radio-based and IMU-based localization in terms of both accuracy and start-up efforts [7, 35, 49], although vision-based tracking itself could be frequently blocked and fail to locate targets. Furthermore, as computer vision matures, captured images are leveraged for not only pedestrian tracking [9, 33, 41], but also 3D reconstruction of environments [10].

Intuitively, one can fuse the results of vision-based tracking and radio-based localization for improved accuracy. However, translating this intuition into a practical system is non-trivial and faces three significant challenges:

- **Absence of absolute location.** As illustrated in Fig. 1b, vision-based tracking systems are capable of framing pedestrians with bounding boxes in images, however, they cannot obtain absolute locations of pedestrians in world-coordinate as radio-based systems. To solve the problem, many works [35, 49] need manual calibration by users to acquire a projection matrix associating pixels in images with locations on floor plans, which is labor-intensive. In recent years, some systems leverage Structure from Motion (SfM) algorithm to reconstruct the indoor scenes and obtain absolute coordinates of objects [12, 31]. However, these mechanisms usually require hundreds (even thousands) of overlapping images from multi-angle of views and are mainly based on crowdsourced methods [64]. Such a large volume of images cannot be obtained from a handful of ambient surveillance cameras in real-world scenarios.
- **Incorrespondence of identification.** As shown in Fig. 1c, the user ID provided by vision-based approaches (typically labels of pedestrians) cannot be directly associated with the user ID provided by

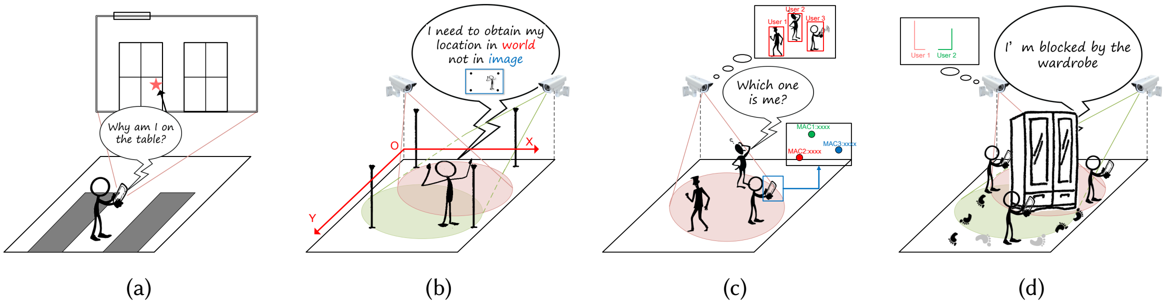


Fig. 1. Challenges of fusing vision-based tracking and radio-based localization systems: (a) Rationality of localization results, pedestrians cannot be localized into inaccessible areas. (b) Absence of absolute location. Vision-based systems cannot obtain absolute locations of pedestrians in world coordinate. (c) Incorrespondence of identification. The user ID provided by vision-based approaches cannot be directly associated with the user ID provided by radio-based approaches. (d) Looseness of sensor fusion. Due to frequent LOS blockages, the generated traces are not smooth and continuous.

radio-based approaches (typically device ID, e.g. , MAC address of NIC). However, this association is a prerequisite to integrate multimodal data.

- **Looseness of sensor fusion.** Existing works [7, 39] are loosely coupled: traces (or tracklets) are directly generated by individual systems independently and then aligned to distinguish users and obtain a fused trajectory. But they rely on an assumption that traces tracklets are accurately constructed from high sampling rates of visual detection, which is hardly tenable in practical scenarios of frequent LOS blockages and erroneous detections, as illustrated in Fig. 1d. Therefore, fusing at tracklet-level may degrade in complicated circumstances, leaving room for improvement.

To tackle the above challenges, we design iVR, a tightly integrated vision and radio localization system that achieves sub-meter accuracy with zero human effort. To obtain the absolute location, we propose an *automatic map construction* algorithm to construct indoor maps and calculate the projection matrix using only a couple of surveillance cameras. However, it is non-trivial to reconstruct a digital map using two cameras. Traditional Binocular Stereo Vision (BSV) [8, 22] requires two identical cameras with parallel optical axes to capture the same scene, which is impractical for surveillance cameras. In iVR, we first utilize the SfM algorithm [27] to determine the relative pose between two cameras from their simultaneously captured images. Then, we generate an equivalent image based on the relative pose according to the imaging principle [6]. The generated image can be treated as captured by a virtual camera whose optical axis is parallel with the other. Finally, we feed these equivalent images into a BSV algorithm and realize automatic map construction and projection matrix acquisition, which was previously only feasible with dedicated binocular stereo cameras. Furthermore, we take full advantage of video frames and construct indoor maps with semantic information to ensure the rationality of localization results.

To associate user identifications and coalesce different data sources in the deep, we devise an *augmented particle filter* to fuse the intermediate results reported by independent systems in early stages, thereby collaboratively nip in the bud the estimated errors. The outputs of our fusion algorithm are (1) location estimates with enhanced accuracy, and (2) associations between pedestrians detected in images and their locations determined by radio-based localization. The proposed particle filter also accounts for location rationality according to semantic maps.

We fully prototype iVR on an Ubuntu server and four different types of smartphones. We conduct extensive experiments in five scenarios, including two different laboratories, an office building, a classroom building and

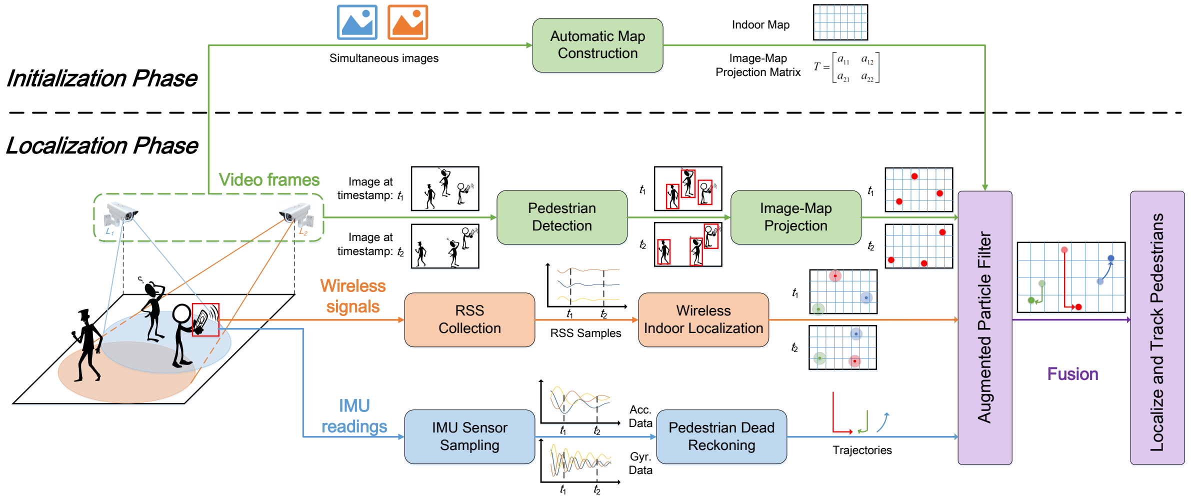

Simultaneous images
Automatic Map Construction
Indoor Map
Image-Map Projection Matrix
$T=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}$
Initialization Phase
Localization Phase
Video frames
L_1
L_2
Image at timestamp: t_1
Image at timestamp: t_2
Pedestrian Detection
t_1
t_2
Image-Map Projection
Wireless signals
RSS Collection
RSS Samples
Wireless Indoor Localization
IMU readings
IMU Sensor Sampling
Acc. Data
Gyr. Data
Pedestrian Dead Reckoning
Trajectories
Augmented Particle Filter
Fusion
Localize and Track Pedestrians

Fig. 2. System Overview

a floor of a shopping mall. The total size of the experimental areas is more than 5,000 m^2. We localize and track pedestrians for more than 20 hours, collecting 20.4k video frames. Evaluation results demonstrate that iVR achieves a mean error of 0.7m and a 90-percentile error of 1.14m for localization, which outperforms the state-of-the-art smartphones-based systems by more than 70%. The tracking success rate is more than 92% in all of the experimental areas, touching 99% in laboratories.

The key contributions are summarized as follows:

- We design an automatic indoor semantic map construction method based on merely a couple of ambient stationary cameras with unparalleled optical axes without manual calibration. To the best of our knowledge, this is the first work that constructs a physical map by using stationary surveillance cameras with unparalleled optical axes.
- We propose a novel augmented particle filter algorithm that tightly couples measurements from multiple orthogonal systems, including vision, radio, and IMU, and jointly estimates a target's location with enhanced accuracy and individual label, making the most of their complementary advantages while overcoming the respective drawbacks of each individual system (e.g. frequent LOS blockages and high frame rate requirement in vision systems, rough localization accuracy and cumbersome site survey in radio systems, and accumulative errors and device placement limitation in inertial system).
- We prototype iVR and conduct extensive experiments in 5 scenarios with the comparison to four state-of-the-art approaches. The evaluation results shown that with zero on-site effort, iVR achieves sub-meter accuracy (0.7m location error in average), outperforming existing systems by 70%.

The rest of this paper is organized as follows. We present an overview in Section 2, followed by detailed presentation about *automatic semantic map construction* in Section 3 and *multimodal localization and tracking* in Section 4. We implement and evaluate iVR in Section 5. We review the start-of-the-art in Section 6 and conclude this work in Section 7.

2 SYSTEM OVERVIEW

Fig. 2 sketches the system architecture of iVR. Multiple cameras continuously monitor public places (e.g. shopping mall, laboratory or office) and stream the recorded videos to the server. Meanwhile, the mobile device (e.g. smartphone) carried by a user logs RSS of wireless signals and IMU sensor readings and then sends them to the server.

During the initialization phase, iVR uses video data recorded from cameras to automatically construct an indoor semantic map of the current monitored area and derive a projection matrix that associates pixels in the images with locations on the map. Thus, iVR does not require external floor plans or any labor-intensive site survey.

During the localization phase, iVR takes as input the three types of data gathered at the server and tracks pedestrians with each type of data. Each type of data has its unique advantages and drawbacks, which are briefly introduced as follows. **Video frames** are processed to detect and localize pedestrians and project them on pre-constructed indoor semantic map. As illustrated in Fig. 2, although it is unable to identify different pedestrians with video frames, the number of pedestrians and their locations can be accurately calculated for further use. **Wireless signals** are used to localize mobile devices held by their users according to signal propagation principle [13]. Compared with video data, it is able to identify different pedestrians, since the MAC addresses of mobile devices are unique. However, simply using wireless RSS can only provide coarse localization results with intolerable large errors, which need to be further refined. **IMU readings** portray relative moving trajectories of pedestrians. When a pedestrian walks, iVR records the readings of the accelerometer and gyroscope on his mobile device and then applies PDR to recognize steps and detect turns. Although lacking global location, once combined with the other two types of data, IMU readings still help distinguish pedestrians and recover global trajectories. For this purpose, iVR further adopts an *Augmented Particle Filter* to fuse intermediate results from the three types of data, differentiate pedestrians and obtain accurate location and trajectory of each individual pedestrian. Finally, iVR sends the tracking results back to each user, where the results might be further used by other location-based or motion-based applications.

Thanks to the design of augmented particle filter, iVR systematically yields enhanced performance in localization accuracy and tracking success rate, meanwhile overcoming respective drawbacks and limitations of each individual system, which are listed below:

- Frequent LOS blockages and high frame rate requirement in vision systems. Vision-based localization and tracking system requires a line-of-sight path between camera and target. Once the target is occluded, it is difficult to re-identify it as the original target [7, 35]. Moreover, traditional visual tracking also requires a high frame rate (e.g. > 15 fps), which is a severe challenge to bandwidth and computational resource.
- Coarse location accuracy and cumbersome site survey in radio system. RSS-range model based localization approaches have been demonstrated coarse-grained and RSS fingerprint based solutions are well-known to suffer from fingerprint spatial ambiguity and temporal instability [53], thus yields meters of location error. What's even worse, cumbersome site-survey which gather RF signatures for each location may need to be repeated over time due to environmental dynamics.
- Accumulative errors and device placement limitation in IMU system. PDR suffers from significant accumulative errors due to the accuracy of the cheap IMU sensors in smartphones and users need to carefully place the phone horizontally in front of their bodies [65], which is user-unfriendly and cannot be fulfilled in practical usage.

In what follows, we first present the design of automatic semantic map construction and augmented particle filter and then explain why iVR yields enhanced performance meanwhile overcoming respective drawbacks.

3 AUTOMATIC SEMANTIC MAP CONSTRUCTION

Automatically constructing indoor map and calculating projection matrix without human intervention is an essential prerequisite to enable vision-based indoor localization with zero human effort. Since surveillance

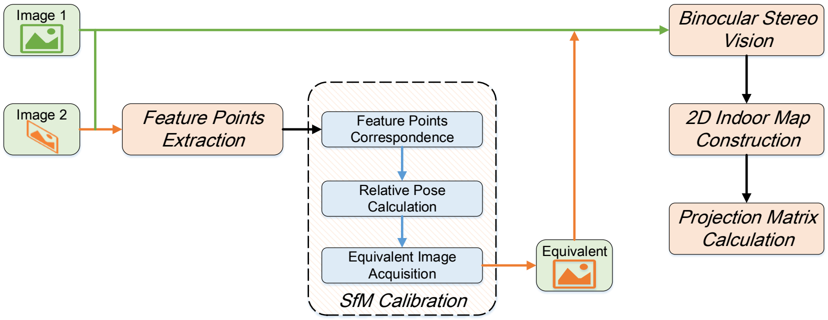


Image 1
Image 2
Feature Points Extraction
Feature Points Correspondence
Relative Pose Calculation
Equivalent Image Acquisition
SfM Calibration
Equivalent
Binocular Stereo Vision
2D Indoor Map Construction
Projection Matrix Calculation

Fig. 3. Workflow of Automatic Map Construction

cameras have been pervasively deployed in public places, such as shopping malls, museums, and supermarkets, it is straightforward to use images captured by these cameras to construct the indoor map.

In recent years, structure from motion (SfM) algorithm [27] has been utilized in many localization systems [12, 31] to reconstruct indoor scenes. However, such a classical mechanism is not suitable in our scenario, hundreds (even thousands) of overlapping images are required for SfM to compute an accurate dense point cloud of a POI [12, 48]. Without no doubt, such volume of images cannot be obtained from two ambient stationary surveillance cameras in our scenario.

Fortunately, we discover that only two ambient cameras still retain the potential to construct indoor map. As in the human visual system [26], two surveillance cameras act as two eyes of a person and form binocular stereo vision (BSV), which can triangulate the 3D geometric information of any objects according to the difference of view angles of the object in two cameras.

Translating this intuition into reality, however, faces a significant challenge: standard BSV requires two identical cameras with parallel optical axes, which cannot be fulfilled by any pair of independently deployed surveillance cameras in practice. To solve the problem, We partially exploit SfM that calculates the relative pose of two cameras from their matched image [58]. Thus, we can calibrate the cameras with their relative pose and emulate images captured by cameras with parallel optical axes.

Fig. 3 shows the logic flow of automatic map construction of iVR. Upon receiving new images from cameras, iVR first extracts feature points from them. We select the SIFT [36] feature due to its highest matching accuracy against other local image features. Then, SfM is carried out to detect correspondences of feature points in two images and calculate the relative pose between two cameras with these correspondences. With the relative pose, iVR virtually rotates the uncalibrated camera (e.g., the camera 2 in Fig. 3) to make its optical axis parallel with the reference camera (e.g., the camera 1), and generate an equivalent image. After calibration, iVR apply BSV to achieve 3D reconstruction and obtain depth information of correspondences of feature points. Finally, the indoor map is constructed by estimating 2D locations of feature points and the semantic floor plan is generated by clustering feature points with pixel-level segmentation information obtained from processing captured image with neural network. The projection matrix from the camera image pixels to the map locations is also calculated by associating feature points in the image and their projections in the map.

The key processes of automatic map construction are further introduced in detail.

3.1 SfM Calibration

iVR exploits the idea of SfM to calibrate a pair of independent surveillance cameras. First, SfM extracts feature points (e.g. SIFT) from pictures and detects matches of feature points across multiple pictures. Then, SfM calculates

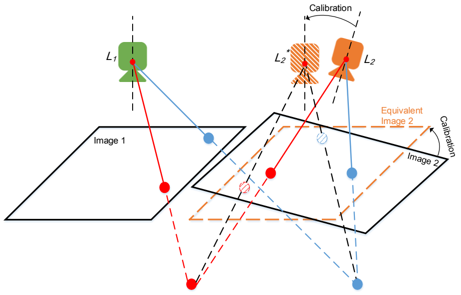


Fig. 4. SfM Calibration

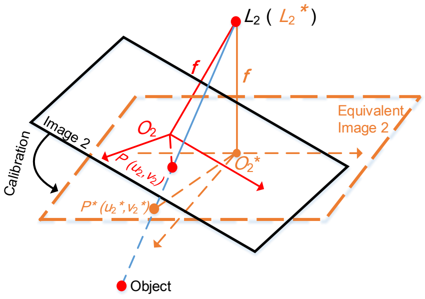


Fig. 5. Equivalent Image Acquisition

relative poses of pairs of cameras when these pictures are taken [58, 59] according to the location difference between matched feature points in different images.

In iVR, surveillance cameras are fixed with location and orientation. Thus, images simultaneously captured by these cameras are used as the input of SfM, and the output is the relative pose between the cameras. Fig. 4 illustrates the process of SfM calibration. For brevity, we denote the relative pose of two cameras as ${}^{L_1}_{L_2}T$, where T represents rotation matrix and L_i the coordinate system of the i-th camera [1]. ${}^{L_1}_{L_2}T$ is computed by solving a so-called Perspective-n-Point (PnP) [30] problem using SfM, as in Argus [58]. Then, the calibration matrix is calculated as:

$$ {}^{L_2^*}_{L_2}T ={}^{L_1}_{L_2} T \cdot {}^{L_2^*}_{L_1} T \tag{1}$$

where ${}^{L_2^*}_{L_1}T$ is the translation transformation matrix from L_1 to L_2^* and can be calculated simply as the location of the cameras is prior knowledge.

Then, iVR projects feature points in the original image with the calibration matrix ${}^{L_2^*}_{L_2}T$ to generate an equivalent image. As shown in Fig. 5, according to imaging principle [6], the object, the corresponding feature point P in the original image, the projection P^* in the equivalent image and the optical center of the lens $L_2(L_2^*)$, are collinear. In the coordinate system L_2:

$$\overrightarrow{L_2P} = (u_2, v_2, -f) \tag{2}$$

and in the coordinate system L_2^*:

$$\overrightarrow{L_2^*P^*} = (u_2^*, v_2^*, -f) \tag{3}$$

where f is the focal length of the camera, (u_2, v_2) is the pixel location of P in the equivalent image, and (u_2^*, v_2^*) is the pixel location of P^* in the equivalent image. With the collinear principle:

$$(u_2^*, v_2^*, -f)^{\mathrm{T}} = \lambda {}^{L_2^*}_{L_2}T(u_2, v_2, -f)^{\mathrm{T}} \tag{4}$$

where λ is unknown scaling factor. Note that u_2^* and v_2^* can be solved from Eq. 4, as there are three variables, u_2^*, v_2^*, and λ, and three equations. After the same process on each pair of matched feature points, we can obtain an equivalent image 2 which can be treated as captured from virtual camera L_2^*, whose optical axes is parallel with reference camera L_1

[1] This convention is brought up by [20], where the left superscript is the reference coordinate system and the left subscript is the objective coordinate system. The multiplication of the rotation matrix requires that the reference coordinate system of the left operand is the same as the objective coordinate system of the right operand, which is canceled with multiplication. For example, ${}^{A}t ={}^{A}_{B} T \cdot {}^{B}t$ means a vector t in coordinate system B multiplied by rotation matrix ${}^{A}_{B}T$, and the result is the t in coordinate system A.

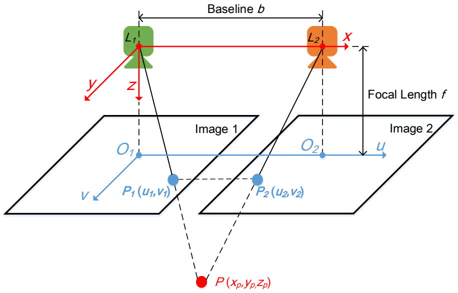


Fig. 6. Binocular Stereo Vision

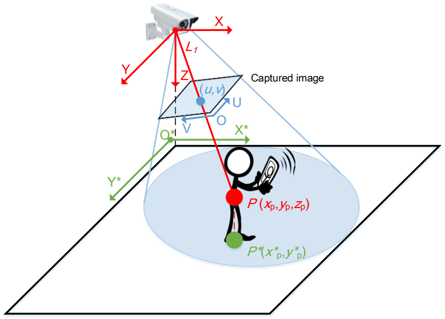


Fig. 7. 2D Indoor Map Construction

3.2 3D Reconstruction with BSV

The approach of acquiring 3D geometric information of objects with BSV is based on visual disparity [38]. Fig. 6 illustrates the schematic diagram of horizontally sighted BSV. The baseline between projective centers of two cameras is denoted as b. The origin of the camera coordinate system L-x-y is at the optical center of the lens of the camera. To simplify the calculation, camera images are drawn in front of the lens by the focal length of f. The origin of the image coordinate system O-u-v is at the intersection of the image plane and the optical axis. The $u-$axis and $v-$axis of the image coordinate system O-u-v are parallel with the $x-$axis and $y-$axis of the camera coordinate system L-x-y respectively.

Suppose that $x-$axes of two cameras of BSV coincides with each other, and a 3D point P has corresponding image points $P_1 : (u_1, v_1)$ and $P_2 : (u_2, v_2)$ on the left and right image plane, then $v_1 = v_2$. According to trigonometry constraints:

$$\begin{aligned} u_1 &= f\frac{x_p}{z_p},\ u_2 = f\frac{x_p - b}{z_p} \\ v_1 &= v_2 = f\frac{y_p}{z_p} \end{aligned} \tag{5}$$

where (x_p, y_p, z_p) are coordinates of P in the left camera coordinate system, f is focal length of the camera. The difference of coordinates P_1 and P_2, termed as visual display, is:

$$d = u_1 - u_2 = f\frac{b}{z_p} \tag{6}$$

With the knowledge of coordinates of points P_1 and P_2 in two images, the 3D coordinate of P is determined as:

$$x_p = \frac{bu_1}{d},\ y_p = \frac{bv_1}{d},\ z_p = \frac{bf}{d} \tag{7}$$

And depth d_p of pixel P_1 on image 1 is:

$$d_p = \sqrt{x_p^2 + y_p^2 + z_p^2} \tag{8}$$

3.3 Semantic Map Construction and Projection Matrix Calculation

As shown in Fig. 7, the projection of P on the indoor map in real world coordinate O^* is determined as:

$$x_p^* = x_p,\ y_p^* = y_p \tag{9}$$

Similarly, iVR localizes the projection of each feature point from image to world coordinate O^*. We further use Mask R-CNN [16] to realize instance semantic segmentation of the captured image at pixel level, after which we will have the semantic information of each feature point. Then, projections with same semantic information are clustered. Finally, we outlining the clusters and mark them as unaccessible areas, generate the indoor semantic map. An example of indoor map is illustrated in Section 5.2.1. In general, clusters of projections of feature points correspond to obstacles in the monitoring area, such as refrigerator, desks and wardrobes.

Finally, with the correspondence between the coordinates of feature points (u, v) in image O and that (x^*, y^*) in world coordinate O^*, the projection matrix T is calculated by solving the optimizing problem:

$$\begin{aligned} \textbf{Minimize}: &\quad \sum_{1 \le i \le N} \sqrt{(x_i^* - u_i^*)^2 + (y_i^* - v_i^*)^2} \\ \textbf{Subject to}: &\quad (u_i^*, v_i^*)^{\mathrm{T}} = T(u_i, v_i)^{\mathrm{T}},\ 1 \le i \le N \end{aligned} \tag{10}$$

where N is the number of feature points extracted from images. The projection matrix is further used to calculate the locations of pedestrians in world coordinate according to their locations detected in the image plane.

4 MULTIMODAL LOCALIZATION AND TRACKING

Typical indoor scenarios are full of multimodal data, such as video frames captured by surveillance cameras, and wireless RSS and IMU readings recorded by mobile devices. While localization with each type of data has been extensively studied in works of literature, the fundamental limitations of these approaches, as discussed in Section 2, are not easy to overcome. To tackle these challenges, iVR exploits advantages of each type of data, design a tightly coupled fusion algorithm and realizes multimodal localization. In this section, we introduce first the preprocessing step of obtaining localization results from each individual type of data, and then the fusion step that merges intermediate results to differentiate pedestrians and improve localization accuracy.

4.1 Unimodal Preprocess

4.1.1 Detection with Vision. The vision part is designed to acquire precise locations of pedestrians, yet without knowing their identities. To achieve this goal, we adopt Mask Region-based CNN (Mask R-CNN) [16] in iVR, which is the state-of-the-art framework for instance recognition and segmentation. In iVR, we use the Mask R-CNN network pre-trained on modified COCO dataset [32], which is dedicated to fast and robust human recognition and segmentation in any scene.

The reason of leveraging Mask R-CNN to detect pedestrians instead of using Category Free Tracking (CFT) algorithms [7, 14, 18] to track pedestrians is CFT algorithms need high frame rate to identify pedestrians. But in practical multi-camera scenarios, the network bandwidth and available computational resources cannot fulfill the requirement. Section 5.2.5 demonstrates that in our experiment, Mask R-CNN can obtain satisfactory result even with a low frame rate (e.g. 1 fps).

When a new video frame is uploaded to the server, iVR uses Mask R-CNN to detect and localize pedestrians in the frame. We denote the coordinates of pedestrians in the image coordinate system as $\{t_i, < u_{t_i}^1, v_{t_k}^1 >, < u_{t_k}^2, v_{t_i}^2 >, \ldots, < u_{t_i}^{C_{t_i}}, v_{t_i}^{C_{t_i}} >\}$, where t_i is the timestamp of the i-th frame and C_{t_i} is the total number of pedestrians detected in the frame. Then the projection matrix T obtained in automatic map construction is applied to calculate the locations of pedestrians $\{t_i, < x_{t_i}^1, y_{t_i}^1 >, < x_{t_i}^2, y_{t_i}^2 >, \ldots, < x_{t_i}^{C_{t_i}}, y_{t_i}^{C_{t_i}} >\}$ in the world coordinate:

$$(x^k_{t_i}, y^k_{t_i})^{\mathbf{T}} = T(u^k_{t_i}, v^k_{t_i})^{\mathbf{T}} \qquad 1 \leq k \leq C_{t_i} \tag{11}$$

4.1.2 Pedestrain Dead-reckoning with IMU. The PDR sub-module is designed to use the accelerate, gyroscope and magnetometer data to perform two key functions. First, it reliably determines whether the pedestrian is walking through step detection. Second, it provides particle filter with a rough estimate of the new step's orientation. For practical usage in the real world scenario, both key functions should be device placement-independent because pedestrians would not always hold their phones horizontally in front of their bodies, as assumed by many related works [62, 65].

For step detection, iVR first uses the smooth filter to filter out noise in accelerometer readings. It then calculates the variance of accelerometer magnitude within a window, which further distinguishes stationary and walking states by an empirical threshold. If the user is in the walking state, it detects steps by searching rising edges with adequate peak values of accelerometer magnitude, as shown in Fig. 8.

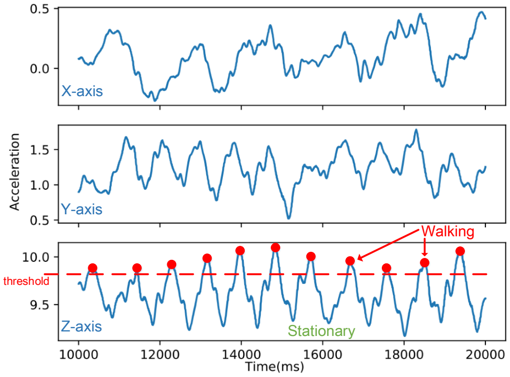


Fig. 8. Acc. data processing for step detection

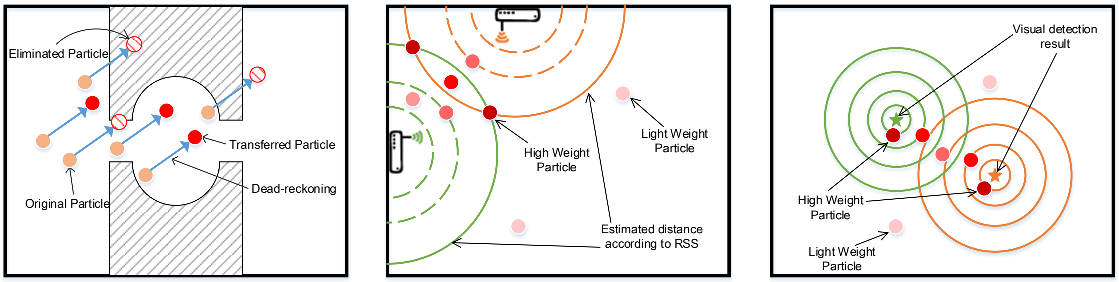
For the orientation measurement of a new step, we adopt the placement-independent algorithm proposed in Zee [43], which takes as input the inertial sensor samples and outputs relative orientation change during one step.

The output of the PDR submodule is a tuple: $< t_j, h_{t_j}, \alpha_{t_j} >$, where t_j is the timestamp, h_{t_j} is the heading direction at time t_j and α_{t_j} is an binary variable, indicating whether there is a step at time t_j.

4.1.3 Localization with Wireless Signal. Although it is theoretically sufficient to only use video and IMU data to localize and identify pedestrians, in some practical cases errors are inevitable. For example, when two pedestrians walk together in a corridor, they may have identical moving patterns that cannot be distinguished by IMU data solely. Therefore, iVR further adopts wireless localization.

In recent years, RSS-fingerprint-based localization algorithms are the mainstream of indoor localization [17], however, these systems also involve a labor-demanding and time-consuming procedure called site-survey. What's even worse, such a cumbersome site-survey may need to be repeated due to environmental dynamics. In iVR, the goal of the wireless sub-module is to help differentiate pedestrians, so we adopt the range-based RSS localization method which is light-weight and fully circumvents labor-intensive site-survey.

By our design, the server collects RSS of wireless signals as well as MAC addresses of mobile devices, and roughly calculate global coordinates of each pedestrian. Specifically, iVR uses the long-distance path loss model,



(a) Particle Movement

(b) Particle Weight Assignment by Wireless

(c) Particle Weight Assignment by Vision

Fig. 9. Particle movement and weight assignment

which describes the relationship between the RSS R_d and the distance d between transceivers:

$$R_d = P_{d_0} - \eta 10lg\frac{d}{d_0} + X_\sigma \tag{12}$$

where P_{d_0} is the reference RSS received at distance d_0, η is the path loss exponent and X_σ is some random noise.

In general, a mobile device sends RSS data $\{t_k, < id_1, R^1_{t_k} >, < id_2, R^2_{t_k} >, \ldots, < id_N, R^N_{t_k} >\}$ to iVR server, where t_k is the timestamp and N is the total number of anchors in the environment. Then, according to Eq. 12, iVR ranges the mobile device and each anchor. The output is $\{t_k, < id_1, d^1_{t_k} >, < id_2, d^2_{t_k} >, \ldots, < id_N, d^N_{t_k} >\}$.

4.2 Tightly Coupled Multimodal Fusion

4.2.1 Pedestrian Association. Before fusing the results from multimodal data, it is necessary to associate visual detection result of each pedestrian with pedestrian dead-reckoning trace during the time interval of two successive video frames.

Suppose at timestamp t_i, the vision detection result of frame i is $\{< x^m_{t_i}, y^m_{t_i} >\}$, $1 \le m \le C_{t_i}$ and the corresponding dead-reckoning result is $< t_i, h_{t_i}, \alpha_{t_i} >$. The association problem is formulated as

$$\begin{aligned} &\arg\min_{\pi_m} \sum_{m=1}^{n} \sqrt{(u^{\pi_m}_{t_{i+1}} - x^m_{t_{i+1}})^2 + (v^{\pi_m}_{t_{i+1}} - y^m_{t_{i+1}})^2} \\ &u^{\pi_m}_{t_{i+1}} = x^{\pi_m}_{t_i} + \Delta\alpha_{t_i} cos(h_{t_i}) \\ &v^{\pi_m}_{t_{i+1}} = y^{\pi_m}_{t_i} + \Delta\alpha_{t_i} sin(h_{t_i}) \end{aligned} \tag{13}$$

where n is the smaller value between C_{t_i} and $C_{t_{i+1}}$, π_m is a permutation of the sequence $(1, 2, \ldots, n)$, and $x_{\pi_i} = (x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n})^T$ is a permutation of the original vector $x = (x_1, x_2, \ldots, x_n)^T$, and Δ is the step length and is set to 0.6m in iVR. Eq. 13 can be viewed as finding best matches between the video detection at time t_{i+1}, and the prediction derived from the video detection and corresponding pedestrain dead-reckoning result. This problem is solved with the Hungarian algorithm [29].

4.2.2 Deep Fusion with Particle Filter. The preprocessing step yields vision detection results $\{t_i, < x^m_{t_i}, y^m_{t_i} >\}$, $1 \le m \le C_{t_i}$, dead-reckoning results $< t_j, h_{t_j}, \alpha_{t_j} >$, and wireless localization results $\{t_k, < id_r, d^r_{t_k} >\}$, $1 \le r \le N$. iVR further combines these intermediate results into an augmented particle filter on a per-step basis. We denote the set of particles as $X = \{X_1, X_2, X_3, \ldots, X_N\}$. The state of each particle $X_i =< x_i, y_i, h_i, w_i >$, where (x_i, y_i) is the current location, $\vec{h_i}$ is the heading direction, and w_i is the weight.

Particle Movement. iVR first uses pedestrian dead-reckoning results to calculate movements of particles. As shown in Fig. 9a, the k^{th} location of the i-th particle is updated as:

$$\begin{aligned} x_i^k &= x_i^{k-1} + \alpha_i^k(\delta_i^k + \Delta)cos(h_i^k + \theta_i^k) \\ y_i^k &= y_i^{k-1} + \alpha_i^k(\delta_i^k + \Delta)sin(h_i^k + \theta_i^k) \end{aligned} \tag{14}$$

To compensate the length variations of walking steps and noises of heading measurements, zero mean Gaussian noises (δ_i^k and θ_i^k) are added. Then, iVR checks whether the movement of each particle violates any indoor environmental constraints (e.g. moving into the wall or furniture) and eliminates outliers. For each valid particle, iVR further updates its weight.

Particle Weight Assignment. Initially, all particles are equally weighted, i.e., $w_i = \frac{1}{N}, i = 1, 2, \ldots, N$. After updating particle locations, particle weights are adjusted according to vision detection results and wireless localization results.

On the one hand, as shown in Fig. 9b, particles whose range to each wireless anchor is similar to the wireless localization results deserve higher weight. Specifically, the weight of the i^{th} particle is updated as:

$$w_i = w_i * \prod_{r=1}^{M} e^{\frac{-(d^r - dis_i^r)^2}{2*\Sigma^2}} \tag{15}$$

where M is the total number of beacons, d^r is the range between the mobile device and the r^{th} anchor calculated from wireless RSS, and h_i^r is the range from the coordinate of the i-th particle. Σ is the normalized parameter and set to 0.9 in iVR.

On the other hand, Fig. 9c illustrates the weight assignment by vision detection results. Particles that are closer to vision detection results will have higher weights, which are additionally updated as:

$$w_i = w_i * e^{\frac{-f_i^2}{2*\Sigma^2}} \tag{16}$$

where

$$f_i = \underset{1 \le a \le C}{\arg\min} \sqrt{(x_i - v_x^a)^2 + (y_i - v_y^a)^2} \tag{17}$$

f_i indicates the nearest distance between the i-th particle and vision detection results $\{(v_x^a, v_y^a),\ 1 \le a \le C\}$.

Particle Re-sampling. After updates of particle weights, iVR re-samples particles according to their new weights. The re-sampling process is as follows:

(1) Normalize weights: $w_i = \frac{w_i}{\sum_{1 \le j \le N} w_j}$, where N is the total number of existing particles.
(2) Randomly generate a number m ranging from [0,1) and determine the index n which make $\sum_{1 \le j < n} w_j \le m < \sum_{1 \le j < n+1} w_j$.
(3) Create a new particle that is the same as X_n ($X_n =< x_n, y_n, h_n, w_n >$).
(4) Repeat the step (2) and (3) until the desired number of re-sampling particles have been generated.

Thus, particles with smaller weights are gradually discarded as disparities of the particles against wireless localization and vision detection results continuously reduce their weights.

Position Decision Strategy. The distribution of particles reflects the likelihood of the real position. Two common approaches are used to determine the target position from particle distribution. One is to set the position of the particle with maximum weight as the target position. The other is to calculate a weighted average of all particles as the target position. Through experiments, we observe that the former approach converges more quickly but has more fluctuations during tracking, while the latter approach needs a longer time to converge but yields stabler position estimation. Thus, iVR combines two approaches by initially selecting the particle with maximum weight, and switching to a weighted average once after convergence. For the weighted average result,

Table 1. Data collection in different scenarios

#	Building type (Areas)	Size(m^2)	#Beacons	#Cameras	#Frames	Duration
1	Laboratory 1 (Whole room)	120	5	4	10.8k	10h in 1 week
2	Laboratory 2 (Whole room)	230	12	3	9k	6h in 2 week
3	Office (Whole floor)	600	20	3	2k	2h in 2 days
4	Classroom (Public areas)	1,360	40	6	2.4k	2h in 3 days
6	Shopping mall (Public areas)	2,130	40	7	3.6k	3h in 1 days

only the top 50% most weighted particles are used. With augmented particle filter, iVR fuses multimodal input and is able to obtain highly accurate and distinguishable trajectories for each pedestrian.

4.2.3 Rationale Behind Tightly Coupled Multimodal Fusion. As aforementioned, the enhanced particle filter, overcoming the defects of each individual submodule, can boost localization accuracy and improve the robustness of tracking. In this subsection, we explain the rationale behind this delighting performance.

First, from a mathematical perspective, particle filters are a realization of the Sequential Monte Carlo method [47]. They estimate system states reliably even when input data are non-Gaussian, nonlinear and even noisy, which is a true profile of our data comprising wireless localization, IMU based dead-reckoning (PDR) and visual recognition bounding box.

Second, in terms of each individual sub-module, the visual system can acquire accurate localization of pedestrians though it cannot distinguish and track pedestrians; PDR and wireless localization system can distinguish and track the users as well as localizing them despite the rough accuracy and accumulative error. Intuitively, if these advantages can be combined together, the system will gain satisfying performance. The design of particle movement and particle weight update strategy in iVR exactly reflects this purpose.

In iVR, the particle movements comply to the orientation and length of each step of PDR, but their weights are only updated according to their proximity with visual localization results. Thus, as shown in Fig. 9a and Fig. 9c, although PDR is imprecise, particles shifted or wrongly positioned have decreasing weights. That said, particles around visual recognition results are assigned higher weight even the visual system is suffering exceptions. For example, two pedestrians gather in one spot such that the visual system cannot distinguish them. In this case, the wireless localization, which keeps tracking of the global coordinates of the particles, plays an important role in distinguishing users by converging particle weights to the true individual pedestrian (as shown in Fig. 9b). This is similar to the cases when pedestrians are occluded from the sight of the camera by huge furniture, or when the visual system wrongly recognizes something else as a pedestrian. All these exceptions of visual results are handled by PDR and wireless localization. They maintain tracking of the true user, though less accurate, but when the visual system resumes its normal state, the pedestrian is accurately localized again.

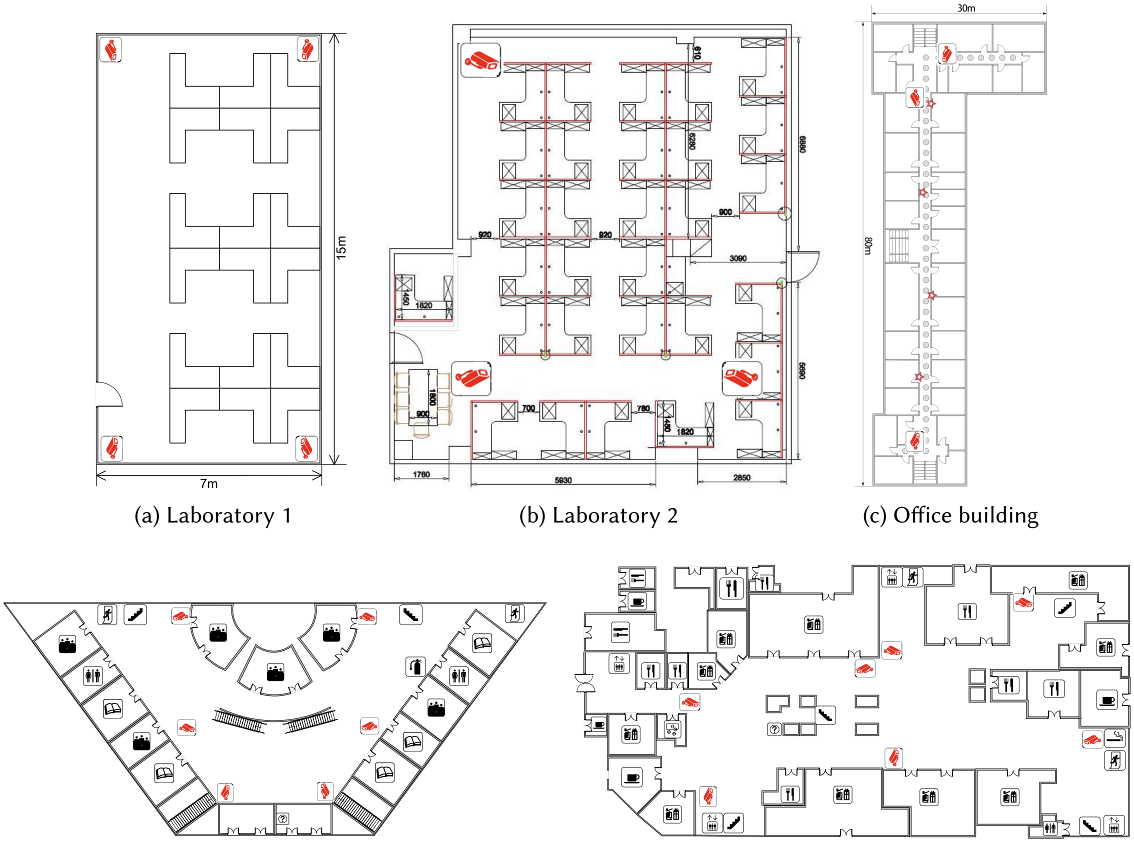
Therefore, comparing to any sub-modules, the whole system improves localization accuracy, distinguishes different pedestrians and overcomes their respective defects.

5 IMPLEMENTATIONS AND EVALUATION

5.1 Experimental Methodology

In this section, we first introduce the experimental settings and then present the detailed evaluation.

5.1.1 Experimental Scenarios. We conducted extensive experiments in two laboratories, a whole floor of an office building, a classroom building and 1st floor of a shopping mall. As shown in Fig. 10, these areas have different floor layouts, diverse wireless environments, and distinct user behavior patterns. In particular, the crowded shopping mall is the most dynamic. There are many people in office buildings in the daytime, which will be



(a) Laboratory 1

(b) Laboratory 2

(c) Office building

(d) Classroom building

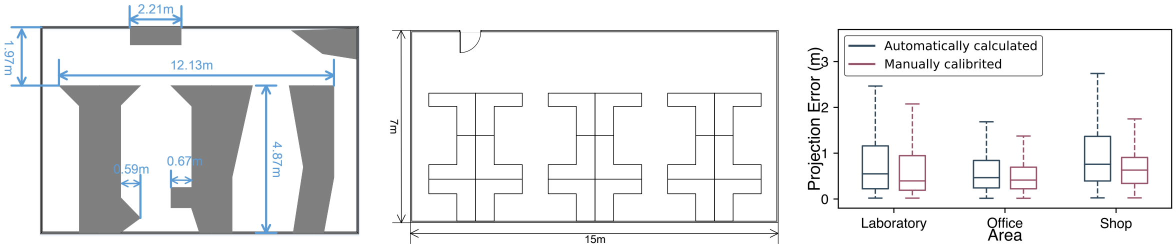
(e) Shopping mall

Fig. 10. Experimental areas

empty in the night. The classroom buildings are crowded or empty to different extents depending on the course schedule. While there are a reasonable number of users in the laboratory most of the time.

The data collection details are summarized in Table 1. We also employ six phones of four different types that are manufactured by different companies for data collection, including two HUAWEI P10, one Lenovo Phab2 pro, two Google Nexus 6p and one Google Pixel, which are equipped with different types of wireless chips and IMU sensors.

5.1.2 Experimental Setup. The client side of iVR is implemented on the Android platform with all of the devices mentioned above, which logs accelerometer and gyroscope readings at 100Hz meanwhile samples Bluetooth ibeacon signal at 60Hz. HIKIVISION-H100 is used as IP cameras to record and continuously stream videos to the server, the size of each frame is 960 × 680 pixels. In each experimental scenarios, there are 2-7 surveillance cameras deployed. The server we use is a Lenovo IdeaPad-Y700 with i7-6700HQ CPU of 2.6GHz main frequency and 16G RAM, runs the Ubuntu 16.0.4 operating system. For Mask R-CNN, the GPU we used is TITAN V with Cuda version 9.1.85 and cudnn-7.05. For SfM, we use and modify Bundler [48], an online open-source SfM project. And we use VisualSFM [52] to validate and visualize the results.



(a) (b) (c)

Fig. 11. Comparison between indoor map automatically constructed and ground-truth floorplan. (a) Indoor map of laboratory 1 automatically constructed by iVR. (b) Floorplan of laboratory 1 provided by the administrator. (c) Vision projection error under automatically constructed and manually calculated matrix.

5.1.3 Ground Truth Acquisition. In order to obtain the ground truth, which is the accurate location of each pedestrian, we recruited 5 volunteers to watch large amounts of surveillance videos, artificially differentiate and track each pedestrian and manually mark their locations on a 2D indoor map. For video frames, volunteers mark a set of tuples (t_i, ID, location) recording each pedestrian's location at timestamp t_i. Totally, our ground truth database has about 35k records [2].

In experiments, we principally test two aspects of performance about iVR: tracking success rate and localization accuracy. The tracking success rate reflects the ability of iVR to distinguish and track different pedestrians. Localization accuracy demonstrates the overall performance of the system including image-map projection error, particle filter fusion deviation, and pedestrian mismatch. Generally speaking, if pedestrian mismatch occurs, it will result in large localization bias.

5.1.4 Comparative Methods. To extensively evaluate the performance of iVR, we additionally implement four different state-of-the-art approaches for comparison, which have been proposed to enhance the primary wireless fingerprinting.

1) **Horus** [63]: A classical probabilistic algorithm that computes the probability distribution of the RSS values at each location as the fingerprint metric, and retrieves the targets of the maximum likelihood as estimated locations.

2) **GIFT** [46]: A metric of binary differential value between RSSs observed at two adjacent locations is exploited as replacements to the original RSS values as fingerprints.

3) **ViViPlus** [60]: Embrace the spatial awareness of RSS values in a novel form of RSS Spatial Gradient (RSG) matrix for enhanced WiFi fingerprints.

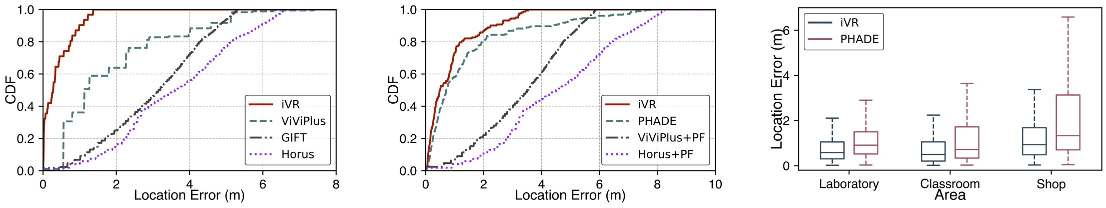
4) **PHADE** [7]: A most related system that extracts human motion features from both video and IMU sensors whereafter fuses the two results to identify and track different users.

iVR can not only localize stationary pedestrians but also track mobile pedestrians. Our experiment with comparative systems includes two parts: localization and mobile tracking. In the first part, we compare iVR with Horus, GIFT, and ViViPlus; In the second part, iVR is compared with PHADE and extended ViViPlus and Horus fusing IMU samples with a traditional particle filter.

5.2 Performance Evaluation

5.2.1 Performance of Automatic Map Construction. As mentioned above, automatic semantic map construction is the key function of iVR. We first examine the effectiveness and accuracy of the function. Fig. 11a and Fig. 11b is the automatically constructed result and ground-truth of laboratory 1. The total width and length of the

[2] Data can be found at https://github.com/xujingao13/iVR.



(a) (b) (c)

Fig. 12. Overall performance comparison with state-of-the-art systems. (a)Different methods localization accuracy. (b)Different methods of tracking accuracy. (c) Comparing iVR with PHADE in different areas.

constructed map are similar to the actual floorplan. Moreover, an automatically constructed semantic map also displays obstacles in the real environment, which is a richer source of information than simply floorplan. iVR use the information to restrict particle's movement as described in Section 4.2.2. Fig. 11c depicts the projection error of automatically calculated and manually calibrated projection matrix in different areas. As shown, the automatically constructed matrix achieves nearly equal performance with manually calibrated matrix in laboratory and office, however, resulting in some large errors (>2m) in the shopping mall. The reason is perspectives of surveillance cameras deployed in shopping malls may enjoy larger differences than other areas, which may lead to *SfM Calibration* function partial invalid. However, compared with manual calibration which is labor-intensive, automatic calculating the matrix is a zero-cost and effective method.

5.2.2 Performance Comparison. Fig. 12a depicts the performance of the proposed iVR as well as three other comparative systems in indoor localization scenarios. As shown iVR achieves the best performance among all comparative systems. The average accuracy of iVR is 0.7m which outperforms Horus by 79.7%, GIFT by 77.2% and exceeds ViViPlus by 67.8%. As for the performance of mobile tracking, as shown in Fig. 12b, the average accuracy of iVR is 0.86m which outperforms PHADE, extended ViViPlus and Horus by 81.2%, 73.6%, and 42.8%. The 95th percentile accuracy outperforms these systems by 60.8%, 40.5%, and 47.2%, respectively.

Furthermore, we meticulously compare iVR with PHADE, which is the most recent system that also based on vision and inertia sensor to differentiate and track pedestrians. As shown in Fig. 12c, iVR significantly outperforms PHADE by at least 25% in all experimental scenarios. iVR and PHADE are both vision-based systems that may suffer from frequently LOS blockages and thus fail to track objects continuously, we further examine the tracking success rate of these systems which is defined as the association accuracy rate between continuous video frames. As shown in Fig. 13, the tracking success rate achieves more than 90% in all areas and outperforms PHADE by more than 10%, traditional vision-based tracking systems like CFT [33] by more than 40%.

The results demonstrate iVR achieves remarkable performance gains based on fusing vision, wireless and inertial sensors. It is worth mentioning that iVR is zero-cost, without human intervention, compared with other systems, iVR doesn't need to conduct site-survey, maintain RSS fingerprint database or calibrate projection matrix manually.

5.2.3 Performance in Different Areas. We evaluate the performance in four different experimental floors as illustrated in Fig. 10, including two laboratories, an office building and a floor in a large shopping mall. Fig. 14 shows the performance of iVR in different areas. As seen, iVR yields an average accuracy of 0.86m in the laboratories, 0.76m in the office building and 1.23m in the classroom, 1.29m in the shopping mall. The corresponding 95th percentile location errors in these three buildings are 3.09m, 2.09m, 3.36m, and 3.41m respectively. The result shows iVR yields similar performance regardless of the environmental difference.

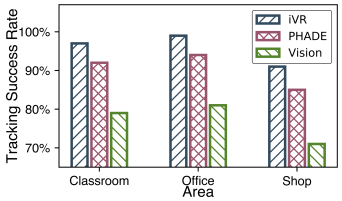


Fig. 13. Tracking success rate in different areas

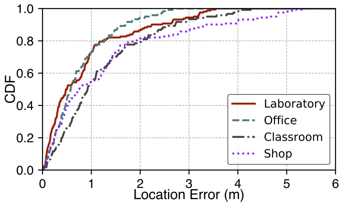


Fig. 14. Different areas

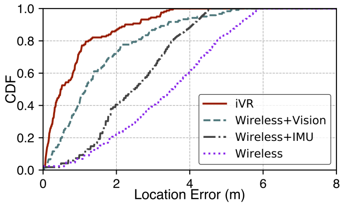


Fig. 15. Effectiveness of fusion

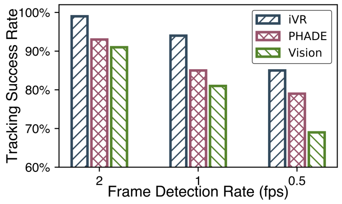


Fig. 16. Different frame detection rate

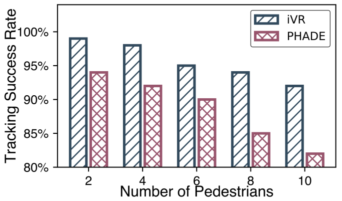


Fig. 17. Different pedestrian number

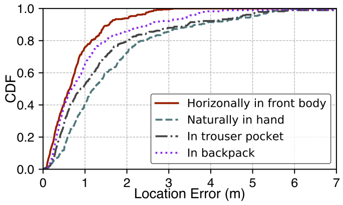


Fig. 18. Different device placement

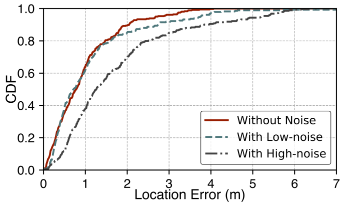


Fig. 19. Different noise strength

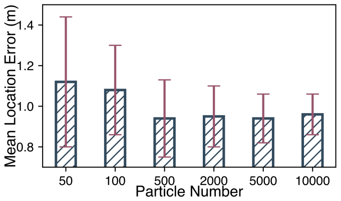


Fig. 20. Different particle number

5.2.4 *Performance of Fusing Sensors and Video.* To demonstrate the effectiveness and accuracy enhanced of fusing multiple sub-systems, we de-couple the three sub-systems in our iVR and evaluate the performance of each module and some combinations of them. Fig. 15 shows the performance of each combination of these modules. As seen, our complete system iVR outperforms Wireless+Vision sub-system by 38%, Wireless+PDR sub-system by 64% and individual Wireless module by more than 80%. The result shows the effectiveness of our system to fuse radio sub-system, vision sub-system, and inertial sub-system. In general, vision sub-system provides high-accuracy but label-free pedestrian localization result, coarse but with an individually pedestrian label is provided by radio sub-system. The corresponding trajectory of a pedestrian, which is calculated by the inertial sub-system will be further used to associate the above two results.

5.2.5 *Impact of Frame Detection Rate.* We also examine the impacts of frame detection rate, which high rate will lead to abundant computational complexity and the lower rate will reduce the tracking success rate. As shown in Fig. 16, the tracking success rate is 99%, 93% and 91% under frame detection rate at 2fps, 1fps and 0.5fps respectively. Compared with PHADE and simple vision-based tracking system, iVR increases the success rate for at least 7.3%, especially at the low frame detection rate, iVR increases for more than 15%.

5.2.6 *Impact of Device Placement.* We evaluate the robustness of iVR on different device placements. We asked a user to carry his smartphone: (1) horizontally in front of body, (2) naturally in his hand, (3) in his trouser pocket, and (4) in his backpack, and walking around for 5 minutes, respectively. We test the localization accuracy of iVR in these cases. As shown in Fig. 18, iVR yields an average accuracy of 0.76m, 1.55m, 1.25m and 1.09m under different device placement respectively.

The results demonstrate that iVR resists to diverse device placement, and the rationale is from two aspects: first, iVR searches rising edges of accelerometer magnitude to detect steps and resorts to PIME algorithm proposed in Zee [43] to calculate relative direction change, both of which have been demonstrated independent of device placement; second, as described in Section 4.2.3, pedestrian dead-reckoning results are further restricted into a reasonable range by vision detection and wireless localization result.

5.2.7 Impact of Multiple Pedestrians. We further examine the impacts of multiple pedestrians using iVR. As shown in Fig. 17, iVR achieves 99%, 98%, 95%, 94%, 92% tracking success rate for 2, 4, 6, 8, 10 users respectively. The performance degrades as the number of users increases because the occlusion between people happens more frequently as there are more users walking in the limited area. This can be mitigated by setting the cameras higher or on the ceiling. Anyhow, iVR outperforms PHADE by more than 5%, and when the number of users increases from 2 to 10, the rate soars to 11%. The enhanced performance lies in the design of augmented particle filter: when the visual system fails to detect a user, wireless localization and PDR will help keep tracking of the global coordinates of the pedestrian, and once the visual system resumes its normal state, the pedestrian is accurately localized again.

5.2.8 Tracking Robustness under Blockage. We evaluate the robustness of iVR by introducing visual blockage deliberately into our vision sub-system by randomly eliminating detection results. For moderate blockage (low strength noise), 20% of the detection results are dropped and for severe blockage (high strength noise), the drop rate increases to 40%. As shown in Fig. 19, the average accuracy of iVR under moderate and severe blocakge maintains 0.96m and 1.15m respectively. The result indicates that although in the real environments, vision-based algorithms may suffer from temporary blockage, the fusion of wireless and IMU module in iVR is helpful to overcome failures of user identification in the vision module and maintain high localization accuracy of iVR.

5.2.9 Impact of Particle Re-sampling Number. In the above, we use 500 particles in our particle filter. Now we dig into the performance of iVR when using different numbers of re-sampling particles ranging from 50 to 10,000. As shown in Fig. 20, the average location error decrease from 1.12m to 0.94m when the number of particles increases from 50 to 500 and maintains the accuracy although the number further increases. The standard deviation decrease from 0.32 to 0.10 when the number of particles increases from 50 to 10,000 but also almost equal to 0.13 when number ranging from 500 to 10,000. So in iVR, using 500 particles is a trade-off between system accuracy and computational complexity.

5.2.10 System Latency. The frame detection rate in iVR is set to 2fps, the rationale is normal pedestrian will move 0.8m during 0.5s which is about 20 pixels reflected in the 960 × 680 video frame. Vision detection is the majority of time-consuming, in iVR, the average detection time using Mask R-CNN framework is 0.32s. During this time, the wireless module and the IMU module will calculate in parallel. For augment particle filter, one round from particle re-sampling to position decision takes average 0.11s at a particle number of 500. So the latency is 0.43s in iVR once after system sample a video frame. In a nutshell, iVR accomplishes pedestrian detection and localization within the system video sampling interval of 0.5s and runs fluently in real-time.

6 RELATED WORKS

Indoor localization has attracted vast research efforts during the past decades. We briefly review the most related latest works in the following.

Easing Deployment. Site survey has been a major bottleneck for fingerprint-based localization, which is time-consuming and labor-intensive. Among various research efforts, recent crowdsourcing-based approach shed promising light in easing the site survey costs [43, 50, 61]. Simultaneous Localization and Mapping (SLAM) techniques are incorporated to avoid the training costs, which result in a set of technique advancements including

WiFiSLAM [11], GraphSLAM [21], and SemanticSLAM [1]. In addition to radio maps, pioneer works including [12, 37, 44, 48] further consider automatic construction of floorplans, which is a prerequisite for any location-aware application. Specifically, JigSaw [12] leverages crowd sensed images captured from mobile users and extracts the position, size and orientation information of individual landmark objects. OPS [37] uses images taken from users to create an approximate 3D structure of the object and camera, and applies mobile phone sensors to scale and rotate the structure to its absolute configuration. These relevant works effectively reconstruct 3D structures of indoor environments, however, require hundreds of overlapping images as input. Differently, iVR exploits the power of binocular stereo vision to automatically construct an indoor semantic map using only two images captured from ambient stationary cameras with unparalleled optical axes, which is demonstrated to be more efficient and requires zero human effort.

Sensor-assisted Indoor Localization. Wireless-RSS-fingerprint-based indoor localization has been demonstrated to suffer from dynamic environment. Some works resort to extra information sources beyond WiFi measurements to gain better accuracy. Ranging via acoustic signals [34] or WiFi Direct [25] among multiple devices are introduced to alleviate fingerprint ambiguity. Fusing inertial sensor data also attracts extensive studies. SurroundSense [4] integrates various sensor hints as multi-modal fingerprints for localization. More commonly, motion information is fused to provide relative locations to improve fingerprint-based localization [62]. Specifically, [19] employs a particle filter fusion technique to combine relative motion information based on step detection with WiFi signal strength measurements. and Zee [43] leverages the inertial sensors to track pedestrians as they traverse an indoor environment. While simultaneously performing WiFi scans, it further utilizes geometric constraints imposed by both mobility information and digital floorplan. Although having gained remarkable accuracy, they generally rely on RSS fingerprint-based localization and IMU sensor-based tracking, the accuracy and practical are facing enormous challenges in the real environment (e.g. fingerprint temporal instability and spatial ambiguity [53], IMU drifting error and device placement restriction). On the contrary, iVR introduce visual localization result to calibrate IMU accumulation deviation and correct wireless based localization result. Furthermore, iVR is device-placement-independent and free of labor-intensive site-survey.

Image-assisted Indoor Localization. Several works exist which utilize a fusion of camera and mobile sensors with a wide variety of applications. Overlay [42] uses a combination of smartphone camera and various sensors to build a geometric representation of an environment to enable augmented reality on the phone. Argus [58] and ClickLoc [59] makes use of visual images to obtain extra position constraints for fingerprinting. Apart from leveraging images captured from smartphone cameras, recent systems also use surveillance cameras to track pedestrians. RAVEL [39] and EV-Loc [49] fuse anonymous visual detections captured by widely available camera infrastructure, with radio readings. PHADE [7] and [24] rely on surveillance cameras to view user's motion patterns and compare the motion with the trajectory calculated from IMU sensors on the user's phone to identify each user and track them. TAR [35] resumes a user trajectory using shopper visual and BLE proximity trace. It leverages a deep-neural-network(DNN)-based visual tracking and person re-identification, which is demonstrated to have high latency and be computationally expensive. Association-Based Tracking (ABT) [3, 5] and Category Free Tracking (CFT) [14, 18] are also open issues and active competitions in computer vision (even in machine learning) field. Although these relevant works achieve high accuracy in pedestrian tracking, it is well known that vision-based systems suffer from frequent LOS blockages in indoor environments. What's even worse, they all rely on high frame rates to identify different pedestrians, which is a waste of network bandwidth and computational resources. Compared with these prior work, iVR leverages the light-weight framework to detect pedestrians in each frame and associates the detection results with radio localization systems to track each user, which is demonstrated to be more robust and works well at low frame rate (e.g. < 2fps).

Dedicated devices based Localization. Recently, several works based on physical layer Channel State Information (CSI), Ultra-wide band (UWB) signal and ultrasonic to localize and track users and achieve decimeter or centimeter level accuracy. SpotFi [28] achieves decimeter-level location accuracy by accurately computing the angle of arrival (AoA) of multipath components using CSI. LiFS [51] leverages the shadowing effect caused by people's blocking line-of-sight paths of Wi-Fi links to achieve passive localization. WiTrack [2] tracks the 3D motion of a user from the UWB signals reflected off her body. Dolphin [15] and BeepBeep [40] present a new design for ultrasonic transmitters and receivers, using range approach to estimate the distance between two cellular phones. While these technologies achieve even centimeter level accuracy, they currently are not available on commodity smartphones for room-level and building-level indoor localization. Specifically, Wi-Fi CSI is only available to some obsolete types of NICs (e.g. Intel 5300 and Atheros 9580) that are not used in up-to-date new brand smartphones. UWB requires specific sensors that are not equipped on current smartphones. Ultrasonic approaches also require specialized audio components. However, smartphones are only equipped with commercial microphones and speakers that supports short-range localization and may generate acoustic noises hearable to sensitive persons.

To conclude, while most of the existing approaches achieve remarkable accuracy for WiFi-based localization, they usually in the meantime introduce additional costs and constraints such as peer cooperation, mobility hints, digital floorplan information, and/or physical layer information, etc, which largely degrades the applicability and ubiquity in practice especial the difficulty of deployment. In the contrary, our proposed approach is a zero cost system without site survey and combines advantages of WiFi, PDR, and vision based on commodity smartphones and pervasively deployed surveillance cameras, thus holding superior potentials for ubiquitous applications.

7 CONCLUSIONS

In this paper, we present iVR, a robust sub-meter accuracy indoor localization system that integrates observations from pervasive surveillance cameras, wireless signals, and mobile sensors. By fusing observations from multiple submodules, iVR successfully overcomes their respective drawbacks and yields great performance that is not achievable by any single submodule alone. We implement iVR on commodity smartphones and conduct extensive experiments in multiple buildings to validate its performance. We believe iVR takes a promising step towards cross-technology system integration to shape a practical smartphone localization service.

ACKNOWLEDGMENTS

This work is supported in part by the National Key Research Plan under grant No. 2016YFC0700100, NSFC under grant 61832010, 61632008, 61672319, 61872081, Microsoft Research Asia, and Tsinghua University Initiative Scientific Research Program.

REFERENCES

[1] H. Abdelnasser, R. Mohamed, A. Elgohary, M. F. Alzantot, H. Wang, S. Sen, R. R. Choudhury, and M. Youssef. 2016. SemanticSLAM: Using Environment Landmarks for Unsupervised Indoor Localization. *IEEE Transactions on Mobile Computing* 15, 7 (July 2016), 1770–1782.

[2] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 2014. 3D tracking via body radio reflections. In *Proceedings of the USENIX NSDI*.

[3] Anton Andriyenko and Konrad Schindler. 2011. Multi-target tracking by continuous energy minimization. In *Proceedings of the IEEE CVPR*.

[4] M. Azizyan, I. Constandache, and R. Roy Choudhury. 2009. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proceedings of the ACM MobiCom*.

[5] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *ECCV*. Springer.

[6] Eric Betzig, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz, and Harald F Hess. 2006. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* 313, 5793 (2006), 1642–1645.

[7] Siyuan Cao and He Wang. 2018. Enabling Public Cameras to Talk to the Public. In *PACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

[8] Steven D Cochran and Gérard Medioni. 1992. 3-D surface description from binocular stereo. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 10 (1992), 981–994.

[9] Piotr Dollár, Ron Appel, and Wolf Kienzle. 2012. Crosstalk cascades for frame-rate pedestrian detection. In *Computer Vision–ECCV 2012*. Springer, 645–659.

[10] Erqun Dong, Jingao Xu, Chenshu Wu, Yunhao Liu, and Zheng Yang. 2019. Pair-Navi: Peer-to-Peer Indoor Navigation with Mobile Visual SLAM. In *Proceedings of the IEEE INFOCOM*.

[11] Brian Ferris, Dieter Fox, and Neil Lawrence. 2007. WiFi-SLAM using Gaussian process latent variable models. In *Proceedings of the IJCAI*.

[12] Ruipeng Gao, Mingmin Zhao, Tao Ye, Fan Ye, Yizhou Wang, Kaigui Bian, Tao Wang, and Xiaoming Li. 2014. Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing. In *Proceedings of ACM MobiCom*.

[13] Julius Goldhirsh and Wolfhard J Vogel. 1998. Handbook of propagation effects for vehicular and personal mobile satellite systems. *NASA Reference Publication* 1274 (1998), 40–67.

[14] Helmut Grabner, Jiri Matas, Luc Van Gool, and Philippe Cattin. 2010. Tracking the invisible: Learning where the object might be. In *Proceedings of the IEEE CVPR*.

[15] Mike Hazas and Andy Ward. 2002. A novel broadband ultrasonic location system. In *Proceedings of the ACM Ubicomp*.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE ICCV*.

[17] S. He and S. H. G. Chan. 2016. Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons. *IEEE Communications Surveys Tutorials* 18, 1 (Firstquarter 2016), 466–490.

[18] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2015. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence* 37, 3 (2015), 583–596.

[19] Sebastian Hilsenbeck, Dmytro Bobkov, Georg Schroth, Robert Huitl, and Eckehard Steinbach. 2014. Graph-based Data Fusion of Pedometer and WiFi Measurements for Mobile Indoor Positioning. In *Proceedings of the ACM UbiComp*.

[20] Robert V Hogg and Allen T Craig. 1995. Introduction to mathematical statistics.(5"" edition). *Englewood Hills, New Jersey* (1995).

[21] J Huang, D Millman, M Quigley, and D Stavens. 2011. Efficient, generalized indoor WiFi GraphSLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

[22] Ramesh Jain, Rangachar Kasturi, and Brian G Schunck. 1995. *Machine vision*. Vol. 5. McGraw-Hill New York.

[23] Niall Jenkins. 2015. 245 million video surveillance cameras installed globally in 2014. *IHS Technology* (2015).

[24] Wenchao Jiang and Zhaozheng Yin. 2017. Combining passive visual cameras and active IMU sensors for persistent pedestrian tracking. *Journal of Visual Communication and Image Representation* 48, 4 (2017), 419–431.

[25] Junghyun Jun, Yu Gu, Long Cheng, Banghui Lu, Jun Sun, Ting Zhu, and Jianwei Niu. 2013. Social-Loc: Improving Indoor Localization with Social Sensing. In *Proceedings of the ACM SenSys*.

[26] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. 1988. Snakes: Active contour models. *International journal of computer vision* 1, 4 (1988), 321–331.

[27] Jan J Koenderink and Andrea J Van Doorn. 1991. Affine structure from motion. *JOSA A* 8, 2 (1991), 377–385.

[28] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. SpotFi:Decimeter Level Localization Using WiFi. In *Proceedings of the ACM SIGCOMM*.

[29] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

[30] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. 2009. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision* 81, 2 (2009), 155.

[31] Liqun Li, Pan Hu, Chunyi Peng, Guobin Shen, and Feng Zhao. 2014. Epsilon: A visible light based positioning system. In *Proceedings of the USENIX NSDI*.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer.

[33] Baiyang Liu, Junzhou Huang, Lin Yang, and Casimir Kulikowsk. 2011. Robust tracking using local sparse appearance model and k-selection. In *Proceedings of the IEEE CVPR*.

[34] Hongbo Liu, Yu Gan, Jie Yang, Simon Sidhom, Yan Wang, Yingying Chen, and Fan Ye. 2012. Push the limit of WiFi based localization for smartphones. In *Proceedings of the ACM MobiCom*.

[35] Xiaochen Liu, Yurong Jiang, Puneet Jain, and Kyu-Han Kim. 2018. TAR: Enabling Fine-Grained Targeted Advertising in Retail Stores. In *Proceedings of the ACM Mobisys*.

[36] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[37] Justin Gregory Manweiler, Puneet Jain, and Romit Roy Choudhury. 2012. Satellites in our pockets: an object positioning system using smartphones. In *Proceedings of the ACM Mobisys*.

[38] Jacques McCoun and Lucien Reeves. 2010. *Binocular vision: development, depth perception and disorders*. Nova Science Publishers, Inc.

[39] Savvas Papaioannou, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2014. Fusion of radio and camera sensor data for accurate indoor positioning. In *Proceedings of the IEEE MASS*.

[40] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the ACM MobiSys*.

[41] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. 2011. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Proceedings of the IEEE CVPR*.

[42] Jain Puneet, Manweiler Justin, and Roy Choudhury. Romit. 2015. Overlay: Practical mobile augmented reality.. In *Proceedings of the 13th Annual International Conference on Mobile Systems*.

[43] Anshul Rai, Krishna Kant Chintalapudi, Venkata N. Padmanabhan, and Rijurekha Sen. 2012. Zee: Zero-effort Crowdsourcing for Indoor Localization. In *Proceedings of the ACM MobiCom*.

[44] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 2011. Fast image-based localization using direct 2d-to-3d matching. In *Proceedings of IEEE ICCV*.

[45] Longfei Shangguan, Zheng Yang, Alex X Liu, Zimu Zhou, and Yunhao Liu. 2016. STPP: Spatial-temporal phase profiling-based method for relative RFID tag localization. *IEEE/ACM Transactions on Networking* 25, 1 (2016), 596–609.

[46] Y. Shu, Y. Huang, J. Zhang, P. Coué, P. Cheng, J. Chen, and K. G. Shin. 2016. Gradient-Based Fingerprinting for Indoor Localization and Tracking. *IEEE Transactions on Industrial Electronics* 63, 4 (April 2016), 2424–2433.

[47] Adrian Smith. 2013. *Sequential Monte Carlo methods in practice*. Springer Science & Business Media.

[48] Noah Snavely, Steven M Seitz, and Richard Szeliski. 2006. Photo tourism: exploring photo collections in 3D. In *ACM Transactions on Graphics (TOG)*.

[49] Jin Teng, Boying Zhang, Junda Zhu, Xinfeng Li, Dong Xuan, and Yuan F Zheng. 2014. EV-Loc: integrating electronic and visual signals for accurate localization. *IEEE/ACM Transactions on Networking (TON)* 22, 4 (2014), 1285–1296.

[50] He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury. 2012. No need to war-drive: unsupervised indoor localization. In *Proceedings of the ACM MobiSys*.

[51] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. 2016. LiFS: Low Human Effort, Device-Free Localization with Fine-Grained Subcarrier Information. In *Proceedings of ACM MobiCom*.

[52] Changchang Wu. 2013. Towards linear-time incremental structure from motion. In *Proceedings of IEEE 3D Vision*.

[53] Chenshu Wu, Jingao Xu, Zheng Yang, Nicholas D. Lane, and Zuwei Yin. 2017. Gain Without Pain: Accurate WiFi-based Localization with Fingerprint Spatial Gradient. In *PACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.

[54] Chenshu Wu, Zheng Yang, and Yunhao Liu. 2015. Smartphones based Crowdsourcing for Indoor Localization. *IEEE Transactions on Mobile Computing* 14, 2 (Feb 2015), 444–457.

[55] Chenshu Wu, Zheng Yang, and Chaowei Xiao. 2018. Automatic Radio Map Adaptation for Indoor Localization using Smartphones. *IEEE Transactions on Mobile Computing* 17, 3 (March 2018), 517–528.

[56] Chenshu Wu, Zheng Yang, Chaowei Xiao, Chaofan Yang, Yunhao Liu, and Mingyan Liu. 2015. Static Power of Mobile Devices: Self-updating Radio Maps for Wireless Indoor Localization. In *Proceedings of the IEEE INFOCOM*.

[57] Jie Xiong and Kyle Jamieson. 2013. ArrayTrack: a fine-grained indoor location system. In *Proceedings of the USENIX NSDI*.

[58] Han Xu, Zheng Yang, Zimu Zhou, Longfei Shangguan, Ke Yi, and Yunhao Liu. 2015. Enhancing Wifi-based Localization with Visual Clues. In *Proceedings of the ACM UbiComp*.

[59] Han Xu, Zheng Yang, Zimu Zhou, Longfei Shangguan, Ke Yi, and Yunhao Liu. 2016. Indoor Localization via Multi-Modal Sensing on Smartphones. In *Proceedings of the ACM UbiComp*.

[60] Jingao Xu, Zheng Yang, Hengjie Chen, Yunhao Liu, Xianchun Zhou, Jinbo Li, and Nicholas Lane. 2018. Embracing Spatial Awareness for Reliable WiFi-Based Indoor Location Systems. In *Proceedings of the IEEE MASS*.

[61] Zheng Yang, Chenshu Wu, and Yunhao Liu. 2012. Locating in Fingerprint Space: Wireless Indoor Localization with Little Human Intervention. In *Proceedings of the ACM MobiCom*.

[62] Zheng Yang, Chenshu Wu, Zimu Zhou, Xinglin Zhang, Xu Wang, and Yunhao Liu. 2015. Mobility Increases Localizability: A Survey on Wireless Indoor Localization Using Inertial Sensors. *Comput. Surveys* 47, 3, Article 54 (April 2015), 34 pages.

[63] Moustafa Youssef and Ashok Agrawala. 2008. The Horus Location Determination System. *Wireless Networks* 14, 3 (June 2008), 357–374.

[64] Xinglin Zhang, Zheng Yang, Yunhao Liu, and Shaohua Tang. 2016. On reliable task assignment for spatial crowdsourcing. *IEEE Transactions on Emerging Topics in Computing* 7, 1 (2016), 174–186.

[65] Pengfei Zhou, Mo Li, and Guobin Shen. 2014. Use it free: Instantly knowing your phone attitude. In *Proceedings of ACM Mobicom*.