Enabling Surveillance Cameras to Navigate

Liang Dong*, Jingao Xu*, Guoxuan Chi*, Danyang Li*, Xinglin Zhang†, Jianbo Li‡, Qiang Ma* and Zheng Yang*
*School of Software and BNRist, Tsinghua University
†School of Computer Science and Engineering, South China University of Technology
‡Qingdao University

***Abstract*—Smartphone localization is essential to a wide spectrum of applications in the era of mobile computing. The ubiquity of smartphone *mobile cameras* and surveillance *ambient cameras* holds promise for offering sub-meter accuracy localization services thanks to the maturity of computer vision techniques. In general, *ambient-camera-based* solutions are able to localize pedestrians in video frames at fine-grained, but the tracking performance under dynamic environments remains unreliable. On the contrary, *mobile-camera-based* solutions are capable of continuously tracking pedestrians, however, they usually involve constructing a large volume of image database, a labor-intensive overhead for practical deployment. We observe an opportunity of integrating these two most promising approaches to overcome above limitations and revisit the problem of smartphone localization with a fresh perspective. However, fusing *mobile-camera-based* and *ambient-camera-based* systems is non-trivial due to disparity of camera in terms of perspectives, parameters and incorrespondence of localization results. In this paper, we propose iMAC, an integrated mobile cameras and ambient cameras based localization system that achieves sub-meter accuracy and enhanced robustness with zero-human start-up effort. The key innovation of iMAC is a well-designed fusing frame to eliminate disparity of cameras including a *construction of projection map function* to automatically calibrate ambient cameras, an *instant crowd fingerprints model* to describe user motion patterns, and a *confidence-aware matching* algorithm to associate results from two sub-systems. We fully implement iMAC on commodity smartphones and validate its performance in five different scenarios. The results show that iMAC achieves a remarkable localization accuracy of 0.68m, outperforming the state-of-the-art systems by $> 75\%$.**

***Index Terms*—Indoor Localization, Crowdsourcing, Wireless, Pedestrians Tracking, Map Construction**

I. Introduction

The popularity of mobile and pervasive computing has stimulated extensive interests in indoor applications, such as customer navigation in museums, targeted advertisements in shopping malls, and personnel emergency rescue in factories. Therein, accurate and easy-to-deploy indoor localization is a key enabler for these services on the horizon. During the past decades, crowdsourced WiFi-based fingerprinting [1]–[3] and inertial-based pedestrian dead-reckoning (PDR) [4] hit the mainstream. However, it is well known that PDR has intrinsically accumulative errors [5], and WiFi fingerprint suffers from temporal instability and spatial ambiguity [6], [7], which make these methods yield meter-level accuracy. While meter-level accuracy can roughly localize or navigate a customer within a shopping mall, sub-meter level accuracy is helpful to determine which aisle he/she is facing within a particular store, to provide detailed information when a customer stands in front of a painting in a museum, and to guide a rescuer to find trapped workers in a race against time.

Recently, as computer vision techniques mature, two arising trends may overcome the above limitations and underpin a practical solution to push the limit of wireless localization: First, surveillance cameras are pervasively deployed in public areas, such as shopping malls, museums, and galleries. Researchers realize that these widely installed ambient cameras could provide complementary advantages to conventional wireless localizations in terms of accuracy. Specifically, these **ambient-camera-based** approaches [8]–[12] rely on surveillance cameras and radio sub-systems to extract user's motion patterns (traces or tracklets) from continuous video frames and wireless signals respectively. Then, different motion patterns are aligned to differentiate users and obtain a fused trajectory with enhanced accuracy. However, the visual tracking performance may degrade in complicated circumstances due to frequent LOS blockages and erroneous detections. Moreover, the pedestrian's motion patterns depicted by wireless system are coarse-grained due to localization bias and accumulative errors [7], [11].

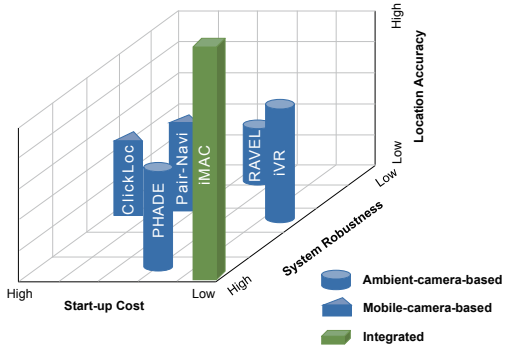


Fig. 1. Comparison of the state-of-the-art works.

Second, vision capability has become more powerful on mobile devices. Images captured by mobile are leveraged to assist localization and navigation. Among **mobile-camera-based** approaches, simultaneous localization and mapping (SLAM) and structure from motion (SfM) technologies have made rapid progress and been widely deployed [13]–[17]. These approaches are capable of precisely tracking mobile cameras' location and pose, but involve a labor-intensive and time-consuming site survey to gather images (or keyframes) about

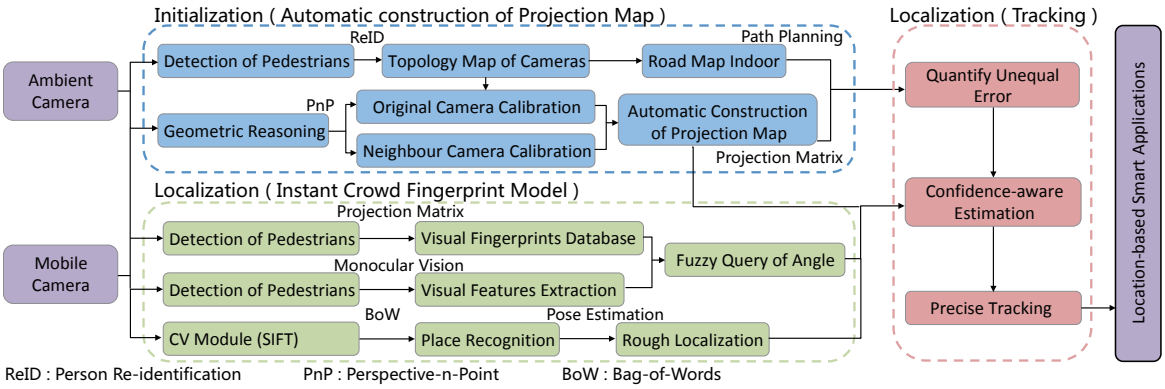


Fig. 2. System Overview of iMAC

landmarks. What's worse, due to frequent LOS blockages by crowds and environmental dynamics, such a cumbersome site survey needs to be repeated over time.

Albeit inspiring, as illustrated in Fig. 1, none of previous studies achieve enhanced localization accuracy and robustness, meanwhile, ease start-up efforts. Intuitively, since *mobile-camera-based* and *ambient-camera-based* methods enjoy their unique advantages, can we fuse these two arising trends together to push the limit of indoor localization and achieve all three goals simultaneously? The integration will improve the precision and robustness of localization, as the leverage of *mobile-camera-based* methods could provide a more fine-grained user motion pattern than wireless systems. On the other hand, deployment costs will be reduced: frames captures by surveillance cameras can be served as image database for mobile systems. However, translating this intuition into a practical system is non-trivial and faces three significant challenges:

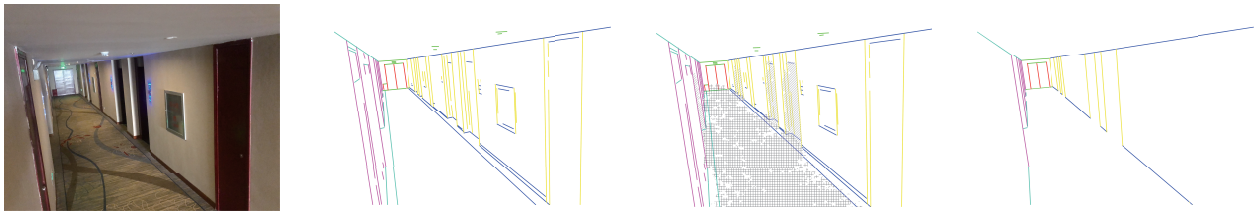
- **Absence of absolute location.** Ambient-camera-based systems are capable of detecting pedestrians in video frames, however, they cannot obtain absolute locations of pedestrians in world coordinate (or in floor plan). To solve the problem, previous works [8]–[10] need manual calibration of the camera to acquire a projection matrix, which is labor-intensive. The most recent work iVR [11] leverages SfM algorithm to automatically calibrate cameras, however, it requires multi-cameras viewing overlapping areas, thus merely fulfill a part of scenarios.
- **Incorrespondence of identification.** The user IDs provided by vision-based approaches are typically the labels of pedestrians. However, the sequence of labels individually acquired from ambient-camera-based and mobile-camera-based systems are unordered and mismatched. This association is a prerequisite to integrate results from each sub-system.
- **Disparity of camera perspective.** Although mobile cameras and surveillance cameras view the same area, the perspective and contents they obtain would vary a lot. Specially, public ambient cameras are stationary and view the area from a top-view, compared with horizon-view from mobile cameras. It is impractical to directly match their visual features using current computer vision techniques.

To tackle all challenges above, we propose iMAC, an **i**ntegrated **M**obile and **A**mbient **C**ameras based localization that achieves sub-meter accuracy and enhanced robustness with zero start-up efforts. To acquire absolute location, we propose an *automatic construction of projection map* frame to calibrate all the ambient cameras and acquire their projection matrices without human intervention. To associate user identifications from two sub-systems, we propose an *instant crowd fingerprints model* (ICFM), a real-time visual description of user motion patterns. Different from WiFi fingerprint, ICFM exploits moving pedestrians as instant beacons to describe user features, which is demonstrated to be more efficient and timely. Meanwhile, we analyze the disparity of camera perspective to find the same estimation error in location will correspond to unequal errors in angle. In some critical areas, a small variation in the location could introduce an extremely large angle estimation error which seriously interferes the result of localization. We mathematically quantify this unequal measurement error and purposely adopt a confidence-aware factor to analyze the similarity of visual features between mobile cameras and ambient cameras.

We fully prototype iMAC on three different types of smartphones and an Ubuntu server and conduct extensive experiments in five typical public scenarios with a practical ambient camera system, including a floor of an office building, a teaching building, a holiday hotel, an art museum and a shopping mall. Evaluation demonstrates that iMAC achieves a mean error of 0.68m and a 80-percentile error of 1.0m in all scenarios, which outperforms state-of-the-art smartphone-based systems by 76.2%. The tracking success rate is more than 90% in all scenarios, including sophisticated scenarios with multiple static pedestrians, where previous methods all malfunction.

The key contributions are summarized as follows:

- We propose a novel system to fuse *ambient-camera-based* and *mobile-camera-based* approaches, making the most of their complementary advantages while overcoming the drawback about labor-intensive start-up efforts. To the best of our knowledge, this is the first work that integrates



(a) Original Image (b) Geometric Reasoning (c) Pedestrian Trajectory (d) After Processing

Fig. 3. Indoor Geometric Reasoning by Line Segments with Crowd Mask

ambient camera and mobile camera together and achieves enhanced localization accuracy.

- We design an automatic ambient camera calibration algorithm without the prior knowledge of camera poses and human intervention, compared with recent works.
- We fully prototype iMAC and conduct extensive experiments in 5 different scenarios with 4 state-of-the-art approaches. The evaluation results show that with zero start-up efforts, iMAC achieves sub-meter accuracy (0.68m location error on average), outperforming existing works by 76.2%.

In the rest of this paper, we first present an overview in Section 2, followed by *automatic construction of projection map* in Section 3. *Instant crowd fingerprint model* is presented in Section 4. Section 5 explains how we achieve *precise localization and tracking with confidence-aware estimation*. We introduce the settings of experiments in section 6 and make detailed evaluations in Section 7. In the end, we review the related work in Section 8 and conclude the proposed work in Section 9.

II. System Overview

Fig.2 sketches the system architecture of iMAC. Multiple ambient cameras continuously monitor public areas and stream the recorded videos to the server. Meanwhile, the mobile camera carried by a user logs visual clues and streams the processed features to the server.

A. Workflow from the user perspective

In iMAC, the user records the surrounding environment with its monocular camera and sends them to iMAC server. In return, iMAC server will send a location tag to the user on the floor plan. During navigation, iMAC is compatible with both visual targets (e.g a picture of Starbuck or a suspect) and semantic location (e.g Room 211) as destinations. Finally, the user will receive the optimum path and visual instructions to achieve there.

B. Workflow from the server perspective

In the initialization stage (in Fig.2), iMAC server automatically calibrates all the ambient cameras and obtains their projection matrices with zero effort.

In the localization stage, a user sends a query (including images of the environment and description of the destination) to iMAC server. First, a rough location is estimated by a place recognition system called FAB-MAP [18]. Afterwards, to achieve precise localization, we put forward Instant Crowd Fingerprint Model which identifies the user appearing in the candidate areas. During matching period, we mathematically quantify unequal estimation between ambient cameras and mobile cameras, and achieve precise tracking by confidence-aware estimation. After locking the user and obtain his location, iMAC sends the optimum path and visual instructions to the user.

III. Automatic construction of projection map

Automatically acquiring projection matrix is an indispensable prerequisite to enable ambient-camera-based navigation to acquire absolute location without human intervention. Most previous works depend on manual measurement to calibrate ambient cameras, which is a labor-intensive and time-consuming process. Existing techniques including SfM and visual SLAM require hundreds of overlapping images from different perspectives to reconstruct 3D model of objects, which is unaccessible towards sparse distributed ambient cameras. Most recent work iVR [11] constructs semantic map requiring two ambient cameras to view same area, which is a strong assumption and invalid in most cases. We design a scheme combining floor plan to automatically calibrate ambient cameras and acquire their projection matrix with no assumption and other prior information.

A. Original Camera Pose Estimation

iMAC combines the idea of SfM and crowd trajectory to calibrate the first batch of ambient cameras which monitor corridors, coners and doors (Fig.3a). To calculate the map relationship between image-generated 3D point cloud and absolute location, we adopt Indoor Geometric Reasoning [19] which assumes that indoor environments satisfy the Manhattan World assumption and recognize the three dimensional structure of the interior of a building from a collection of line segments automatically extracted from single indoor image. However, merge and filter operations [15] fail to effectively extract building structure from line segments (Fig.3b) due to clutter of various objects in complex indoor scenarios. Inspired by crowdsourcing strategy, we capture the trajectory of pedestrian movements and generate a crowd mask (Fig.3c) through particle filter algorithm. Assuming the appear and disappear centers of the crowd as the doors or coners, we effectively remove redundant line segments and extract building structure (Fig.3d) corresponding to physical scale deriving from the floor plan.

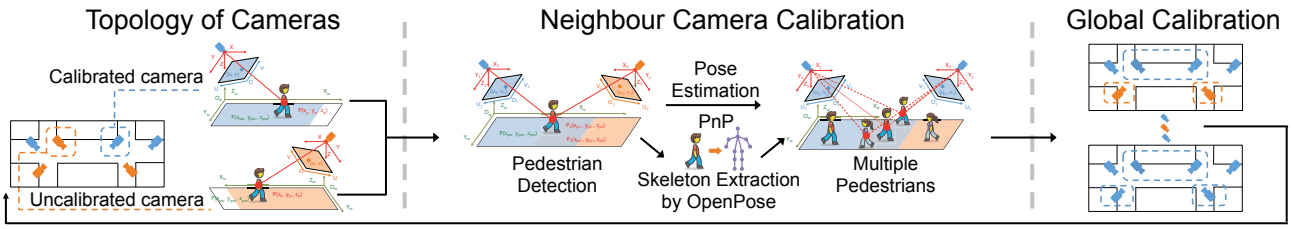


Fig. 4. Workflow of Automatic Camera Calibration

Then, iMAC exploits the idea of Perspective-n-Point [20] (PnP) to calibrate camera external parameters. Concretely, after Indoor Geometric Reasoning we acquire a set of points correspondences, each composed of a 3D reference point $\mathbf{P_i} = (X_i, Y_i, Z_i)^{\mathrm{T}}, i = 1, \ldots, n, n \geqslant 4$ expressed in world coordinates and its 2D projection $\mathbf{p_i} = (u_i, v_i, 1)^{\mathrm{T}}, i = 1, \ldots, n, n \geqslant 4$ expressed in image coordinates. $\mathbf{T}$ is the transformation matrix with which we can acquire the absolute location of the points on image. Then it comes to solving an optimizing problem to estimate the transformation matrix $\mathbf{T}$:

$$\mathbf{T} = \arg\min_{\mathbf{T}} e = \arg\min_{\mathbf{T}} \frac{1}{2} \sum_{i=1}^{n} \left\| \mathbf{p_i} - \frac{1}{s_i} \mathbf{K T P_i} \right\|_2^2, \quad (1)$$

where e is the cost function of reprojection error, s_i is the depth of point $\mathbf{P_i}$, $\mathbf{K}$ is the intrinsic matrix which assumed easy to known from factory defaults.

B. Neighbour Camera Pose Estimation

Although we acquire satisfied pose estimation of some original cameras, more ambient cameras whose monitoring areas unmatching the condition have to be calibrated automatically. Fortunately, for security reasons, ambient cameras systems are required to cover public space [21] which means overlap exists between neighbour cameras. However, these narrow overlapping areas can not support the SfM algorithm to extract enough corresponding feature points.

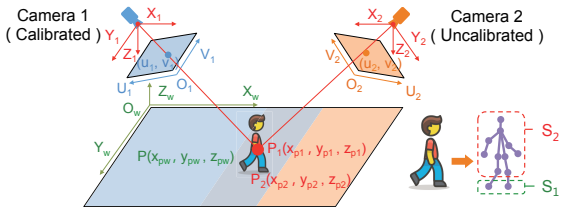


Fig. 5. Calibration of Neighbour Cameras

Thanks to astonishing progress of pedestrian detection in recent years, iMAC can calibrates neighbour cameras through keypoints extracted from the same pedestrian appearing in the overlapping area. Fig.4 illustrates the process of neighbour cameras pose estimation.

First, iMAC topologizes the ambient cameras by their neighbouring relations and selects a pair of know-unknown cameras. Then, iMAC recognizes the same pedestrian in the overlapping area through ReID (Pedestrian Re-Identification) technique [22], [23] which performs well under tight spatio-temporal constraint. To this pedestrian, iMAC adopts OpenPose [24] (a realtime approach to detect the 2D pose of multiple people in an image) to extract his skelecton and select his arthrosis as feature points in neighbour images. Afterwards, iMAC exploits these corresponding points to calculate camera pose estimation.

As shown in Fig.5, $\mathbf{P}$ is a pedestrian recognized in the overlapping area of a pose-estimated camera 1 and a pose-unestimated camera 2. $\mathbf{S_1}$ containing the foot keypoints on the floor plane where $Z = 0$ and the rest keypoints are contained in $\mathbf{S_2}$. According to the pinhole model, we get pixel coordinates $\mathbf{p_1} = (u_1, v_1, 1)^{\mathrm{T}}$ and $\mathbf{p_2} = (u_2, v_2, 1)^{\mathrm{T}}$ on image planes, which are corresponding points of point $\mathbf{P} = (X, Y, Z)^{\mathrm{T}}$:

$$\begin{cases} s_1 \mathbf{p_1} = \mathbf{K_1}(\mathbf{R_1 P} + \mathbf{t_1}) \\ s_2 \mathbf{p_2} = \mathbf{K_2}(\mathbf{R_2 P} + \mathbf{t_2}) \end{cases}, \quad (2)$$

where $\mathbf{K_1}$, $\mathbf{R_1}$, $\mathbf{t_1}$ are known parameters of calibrated camera 1 and $\mathbf{K_2}$, $\mathbf{R_2}$, $\mathbf{t_2}$ are unknown parameters of uncalibrated camera 2. Using PnP algorithm [20], we can obtain $\mathbf{K_2}$, $\mathbf{R_2}$, $\mathbf{t_2}$ and acquire the projection matrix of camera 2.

Finally, we calibrate all ambient cameras and obtain their projection matrices, which enable iMAC to acquire absolute location of detected objects in world coordinates.

IV. Instant Crowd Fingerprint Model

Mobile-camera-based navigation depends on high-quality recognition of the landmark, which suffers from environment fluctuations and frequent LOS blockages of crowds. Although ambient camera offers instant information of environment, it is unworkable to directly match images from the mobile camera and the ambient camera since perspective disparity. Conversely thinking, the crowd not only leads to LOS blockages but also offers a unique description of pedestrian location and motion pattern. iMAC proposes a brand new model called Instant Crowd Fingerprint Model to discern different pedestrians based on the description of crowds.

Fig.6 illustrates this process. First, iMAC sever uses MobileNetV3 (a class of efficient models for mobile vision applications) [25] to detective pedestrians appearing in candidate areas and acquire their absolute locations. Afterwards, we calculate the geometric estimation of each pedestrian $\mathbf{P_i}, i = 1, \ldots, n, n \geqslant 3$ to distinguish each potential user. Concretely speaking, each pedestrain $\mathbf{P_i}$ has a series of angles $\boldsymbol{\alpha_i} = (\alpha_{i1}, \ldots, \alpha_{ij}, \ldots \alpha_{im}), j = 1, \ldots, m$, which engendered with the rest m pedestrians in sight:

$$\alpha_{ij} = \arccos \frac{\overrightarrow{P_i P_j} \cdot \overrightarrow{P_i P_{j+1}}}{\left\| \overrightarrow{P_i P_j} \right\| * \left\| \overrightarrow{P_i P_{j+1}} \right\|} \quad (3)$$

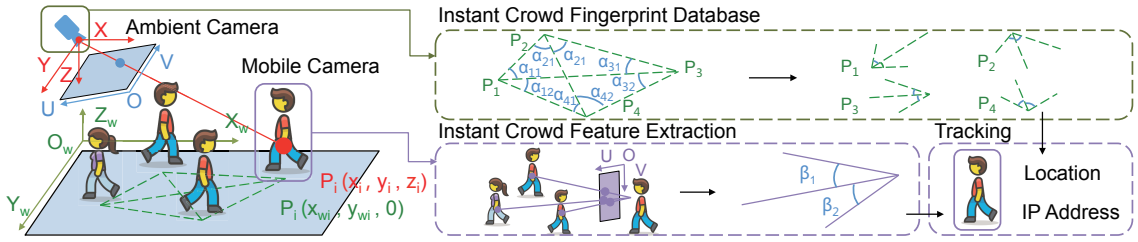


Fig. 6. Workflow of Instant Crowd Fingerprint Model

Up to now, iMAC sets up an instant fingerprint database of candidate pedestrians. However,it becomes difficult to estimate geometric relationship for mobile cameras due to scale ambiguity of monocular vision system.

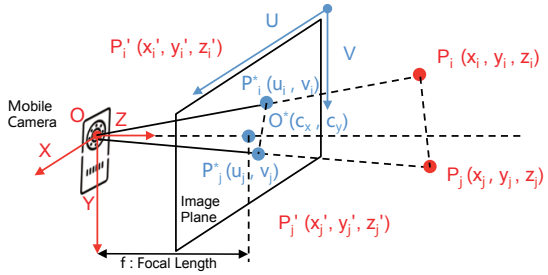


Fig. 7. Extraction of Instant Crowd Feature from Mobile Camera Based on Monocular Vision

Fortunately, we find it still accessible to obtain angle informations in (Fig.7). $\mathbf{P_i} = (x_i, y_i, z_i)$ and $\mathbf{P_j} = (x_j, y_j, z_j)$ are 3D world coordinates of two objects and $\mathbf{P'_i} = (x'_i, y'_i, z'_i)$ and $\mathbf{P'_j} = (x'_j, y'_j, z'_j)$ are their projection on the image plane where $Z = f$ (f is ***focal length***). $\mathbf{P^*_i} = (u_i, v_i)$ and $\mathbf{P^*_j} = (u_j, v_j)$ are their 2D pixel coordinates in the image. According to trigonometric constraints:

$$\angle P_i O P_j = \angle P'_i O P'_j, \tag{4}$$

our aim equals to calculate $\angle P'_i O P'_j$:

$$\angle P'_i O P'_j = \arccos \frac{(x'_i, y'_i, z'_i) \cdot (x'_j, y'_j, z'_j)}{\|(x'_i, y'_i, z'_i)\| * \|(x'_j, y'_j, z'_j)\|} \tag{5}$$

Afterwards, iMAC obtains $\boldsymbol{\beta} = (\beta_1, \ldots \beta_i \ldots \beta_s), i = 1, \ldots s, s \geqslant 2$ as an instant fingerprint on the mobile side, which will be uploaded to iMAC sever and compared with other fingerprints in ICFM database.

V. Precise Localization and Tracking with Confidence-aware Estimation

However, it is quite unwise to directly compare the similarities of geometric features between mobile camera and ambient camera. Since each side of them has a different function of error, among which the error of ambient camera depends on location error, but the error of mobile camera comes from angle error.

As shown in Fig.8.a, θ is an estimation error of angle from the mobile camera, L is the corresponding location error from the ambient camera, d is the unit distance from a candidate pedestrian to reference pedestrian:

$$L = 2(M-1)d \sin\frac{\theta}{2} \tag{6}$$

When θ is set to a constant, L becomes a linear increasing function of M. That is to say, to each candidate pedestrian, the farther a reference pedestrian stands away, the more confidence this reference pedestrian has.

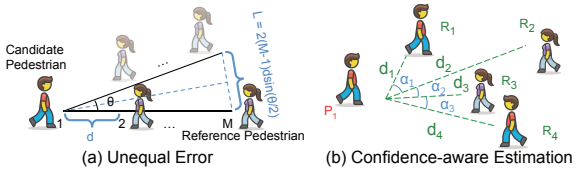


Fig. 8. Unequal estimation

To eliminate the unequal error, we set different confidence to angles in the fingerprint database. For example (in Fig.8.b), P_1 is a candidate pedestrian, R_1, R_2, R_3 and R_4 are its reference pedestrians, d_1, d_2, d_3 and d_4 ($d_2 \geqslant d_4 \geqslant d_3 \geqslant d_1$)are distance between them. α_1, α_2 and α_3 are fingerprints of P_1. According to Eq. (6), we first set the confidence of the farthest reference pedestrian R_2 to 1, and the rest R_1, R_3, R_4 to $\frac{d_1}{d_2}$, $\frac{d_3}{d_2}$, $\frac{d_4}{d_2}$ respectively. Then we set different confidence factor of fingerprints according to the influence of two sides of the angle:

$$\begin{cases} F_1 = \frac{d_1}{d_2} \cdot 1 \\ F_2 = 1 \cdot \frac{d_3}{d_2} \\ F_3 = \frac{d_3}{d_2} \cdot \frac{d_4}{d_2}, \end{cases}, \tag{7}$$

where F_1, F_2 and F_3 are the confidence of α_1, α_2 and α_3 respectively. Meanwhile, these confidence factors will be used to calibrate the rough comparison during query process, which means each likelihood of angles will multiply its correspond confidence factor to get the last value of likelihood.

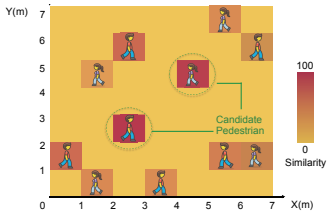


Fig. 9. Query result of one frame by confidence-aware estimation

Fig.9 illustrates the estimation result of a frame. Usually, we keep several candidate pedestrians a time and repeat the same execution until the right pedestrian is locked.

TABLE I
DIFFERENT REPRESENTATIVE SCENARIOS OF EXPERIMENTS

Scenario	Size(m^2)	Original Cameras	Neighbour Cameras	Frames	Duration
Office building	600	3	6	20.3k	1h(office hours)&1h(rush hours)
Teaching building	1360	4	8	28.4k	2h(break hours)
Art museum	860	3	6	13.4k	1h(weekday)&1h(weekend)
Holiday hotel	1120	3	6	14.6k	2h(intermittently in 5days)
Shopping mall	2130	4	8	26.4k	1h(weekday)&1h(weekend)

VI. EXPERIMENT SETTINGS

A. Implementation Setup

We prototype iMAC front-end on three phones of different types, including a Google Pixel, a HUAWEI P30 and an iPhone X, which are equipped with different types of mobile cameras and computing resources. Images are processed on the phone and uploaded to a server, which is a desktop computer with i7-9700F CPU of 4.7GHz main frequency and 16G RAM, runs the Ubuntu 16.04 operation system. The ambient camera we use is HIKIVISION-C3A ,which continuously stream recorded videos to the server. We use Bundler [26] for SfM, EPnP [27] for PnP. We also use VisualSFM [28] to validate and visualize our results.

B. Implementation Scenarios

We implement experiments in five different typical public areas, including a floor of an office building, a teaching building, a holiday hotel, an art museum and a shopping mall. In each scenario, We collect video data during different periods of the day to guarantee the cover of different crowd flows situations. The summarize of collected videos are listed in Table I.

C. Ground truth Acquisition

To acquire the ground truth of cameras pose, we manually measure the location and orientation of each ambient camera in the scenarios. Then we use the measurements to calculate projection as ground truth. In total, we collect 49 calibration results of ambient cameras.

To acquire the ground truth of localization and tracking, we invite 3 volunteers to label the video. They manually recognize the user and localize the user through measured projection matrices. Specifically, each user on each frame will have a tuple (UID, Loc, t_i), where UID is the ID of users, Loc is the ground truth location and t_i represents the timestamp of each frame. Overall, our label collection contains 45K records.

VII. PERFORMANCE EVALUATION

A. Evaluation Methods

We evaluate the performance of iMAC in three fields.

First we evaluate the self-calibration performance of ambient cameras. Since original cameras and neighbour cameras are calibrated through different approaches, they are analyzed separately. We use the classic precise chessboard calibration method in [29] as the control group. We contrast calibration error of rotation and translation respectively.

Then we test overall localization accuracy of iMAC and compare its performance with three different representative indoor localization fusing surveillance cameras observation or mobile camera observation:

- **RAVEL** [12]: RAVEL (Radio And Vision Enhanced Localization) is a generic vision+radio tracking framework, which fuse visual signals from surveillance cameras and WiFi radio signals and is the first paper that proposes a practical solution of radio-aided visual tracking.
- **PHADE** [9]: PHADE is a recent vision+sensor tracking framework, which relies on surveillance cameras viewing users motion patterns, and compares the uniqueness of these patterns with the patterns extracted from user's IMU data.
- **iVR** [11]: iVR is a most recent vision+radio+sensor tracking framework, which combines observations from surveillance cameras, WiFi radio signals and IMU data and outperform the state-of-the-art system.
- **ClickLoc** [15]:ClickLoc is a typical high accurate localization system integrating mobile cameras and IMU signals from the smartphone.

In the end, we focus on evaluating tracking success rate. Since tracking success rate is the main influence factors of localization based on ambient cameras. If tracking successfully, the localization accuracy depends on projection accuracy of ambient cameras, which depends on calibration accuracy of ambient cameras. If tracking incorrectly, it will result in a large bias in localization. During this period, we introduce a classic vision-based object tracking system [30] as a contrast, which is a robust collaborative model accounting for drastic appearance change especially occlusion problem.

- **SDC&SGM** [30]: A robust appearance model that exploits both holistic templates and local representations, which develops a sparsity-based discriminative classifier (SDC) and a sparsity-based generative model (SGM).

B. Performance of Pose Estimation

As mentioned before, automatically acquiring camera external parameters without human intervention is a basic premise of all localization schemes based on ambient cameras. We first test the calibration accuracy of original cameras. We choose 8 ambient cameras in each scenario and calibrate 40 original cameras automantically in total.

1) Original Cameras Calibration: Fig.10.c illustrates that our method achieves similar accuracy in total compared with Zhang's [29] standard result. Concretely speaking, our method achieves better performance in rotation calibration (Fig.10.b)

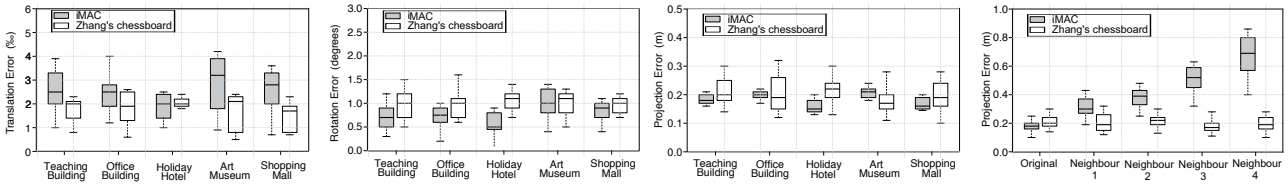


Fig. 10. Evaluation of Automatic Camera Calibration

in most scenarios and outperform Zhang's output by 0.5 degrees on average. Meanwhile, it achieves worse performance in translation calibration (Fig.10.a) in most scenarios. It is because the corresponding points we take in our method is far from the original camera and far apart from each other. But the corresponding points on Zhang's standard chessboard are much closer to the original camera and close to each other. Although the error of world coordinates of our corresponding points is larger than that of Zhang's corresponding points. It induces worse performance in the translation calibration but little influence in the rotation calibration.

Meanwhile, our method performs better accuracy in the holiday hotel. Since the holiday hotel has more regular texture especially in the area of guest corridors which offers more corresponding points for calibration.

Eventually, our method achieves roughly the same performance in average projection accuracy compared with Zhang's (Fig.10.c). Moreover, our result is more stable in each scenario since it has a smaller range of waving. The rationale behind is rotation accuracy becomes more influential to projection than translation accuracy, when the distance between point and camera grows. And as expected, our method performs the best in the holiday hotel, which even outstands Zhang's by nearly 40%.

2) Neighbour Cameras Calibration: Afterwards, we evaluate the performance of neighbour cameras calibration. Compared with original camera (Fig.10.d), the average projection error of neighbour camera is about 0.1m larger since the cumulative error. Since pedestrians offer the key points which play a decisive role in calibration, Fig.11 analyzes the relationship between projection accuracy and the number of pedestrians. Although the projection produces large errors at the first three pedestrians, it becomes narrow and stable with the increase of pedestrians and ultimately stabilizes after 10 pedestrians. As a result, we only list neighbour cameras in a white list after being calibrated by more than 20 pedestrians.

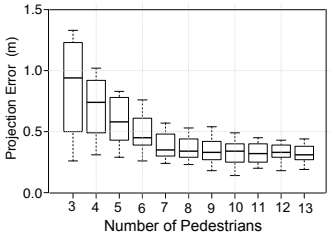


Fig. 11. Relationship between projection accuracy and number of pedestrians

Although we obtain precise calibration results on original cameras, cumulative error will be transmitted to every next neighbour camera. Fig.10.d analyzes this cumulation through increasing layers of neighbour cameras. According to observation, the projection error is linearly proportional to the layers of neighbour cameras in topology structure. And the fourth layer of nighbour camera still has an acceptable projection error in 1m. In practise, an uncalibrated camera may connect to different original camera through distinct routines. Thus, this camera will engender multiple calibration results derived from different original cameras. Fortunately, this regularity directs us to adopt projection result from neighbour camera which is more closer to an original camera in topological relationship.

On the whole, we accomplish a reliable solution to self-calibrate the global cameras which has similar accuracy to Zhang's standard results. Although Zhang's method has been a flexible and convenient calibration method, it still costs us about half an hour and two professional volunteers to calibrate each camera on average. Since Zhang's method only offers calibration result in coordinate of chessboard. It induces additional labor and bias to manually calibrate the location of chessboard. By comparison, our method leverages a zero-cost and effective method to calibrate ambient cameras and help construct the indoor projection map.

C. Performance of Localization

1) Overall Comparison: Compared with three other state-of-the-art indoor localization systems, iMAC achieves the best performance in overall accuracy(in Fig.12.a). The average localization accuracy of iMAC is 0.68m, which surpasses iVR by 34.7%, PHADE by 76.2%, ClickLoc by 77.4%, and RAVEL by 83.4%.

In ambient-camera-based systems, it is noteworthy that WiFi, IMU and vision make different contribution to the final performance. Basically, WiFi plays a fundamental role to offer a rough localization, which obtains 3-5m precision and avoids excessive outliers. IMU plays a definitive role in distinguishing pedestrians in proximate space, which differ in the shape of trajectory. That is why RAVEL has better result(6m) in maximum error than PHADE(7m), although it performs worse in average accuracy. iVR integrates the advantages of both IMU and WiFi to achieve an overall better localization system. On this basis, iMAC replaces WiFi fingerprints with instant visual geometry fingerprints, which performs more accurate, realtime and low-cost.

In mobile-camera-based systems, location accuracy depends on visual recognition of landmark. Once fails in recognition, ClickLoc will degenerate into WiFi-based localization. Thus,

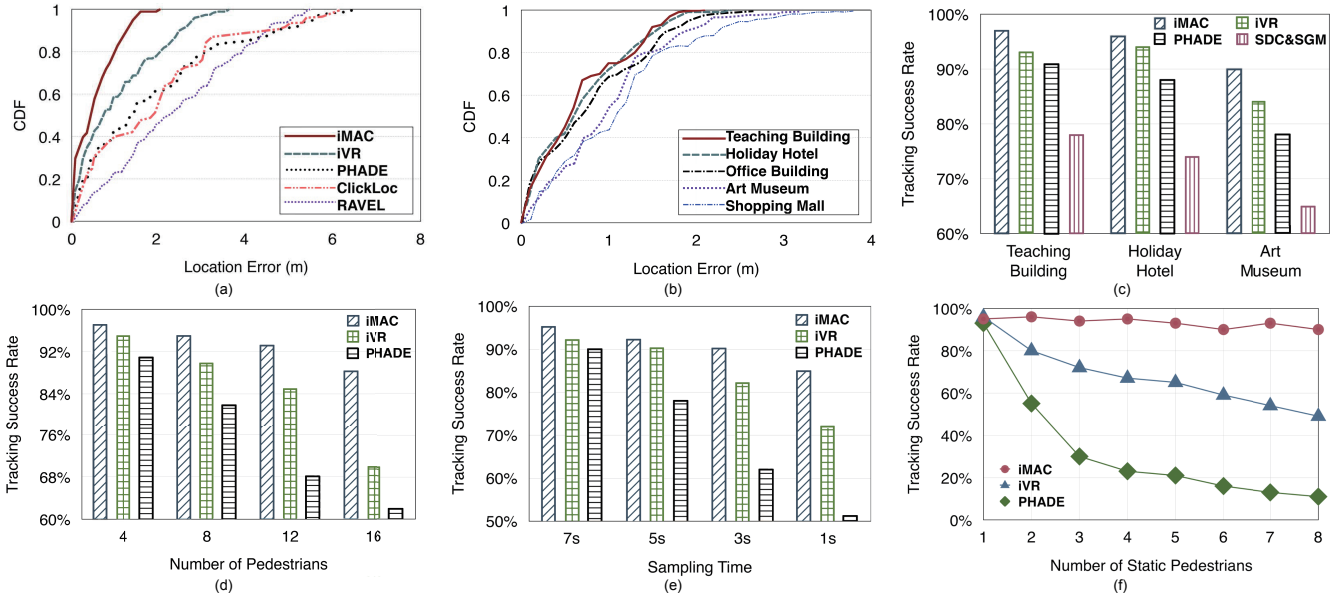


Fig. 12. Evaluation of localization and tracking: (a) Overall location accuracy comparison with state-of-the-art systems. (b) Location accuracy in different public scenarios. (c) Comparison of overall tracking success rate in different public scenarios. (d) Influence to tracking success rate from number of pedestrians. (e) Performance comparison of different sampling time for retracking. (f) Performance comparison in complex scenarios with multiple static pedestrians.

ClickLoc has a better performance than RAVEL in average accuracy but performs as bad as PHADE in maximum error. iMAC leverages ambient cameras to enhance accuracy and significantly surpasses ClickLoc in average accuracy and worst accuracy.

For above reasons, iMAC attains outstanding location performance through fusing ambient camera and mobile camera and acquires better accuracy without human efforts.

2) Performance in Different Scenarios: To meticulously evaluate iMAC, Fig.12.b depicts location accuracy in different scenarios. The 80-percentile error in each scenario is within 2m, meaning iMAC has better performance in different environments. Among them, teaching building, holiday hotel and office building contribute better average accuracy, which are 0.65m, 0.68m and 0.74m respectively. However, art museum and shopping mall contribute slightly worse average accuracy, which are 1.22m and 1.31m. According to our observation, visual occlusion is the primary cause of the drop in accuracy. Complexity of crowd and environment still makes negative influence to some degree.

Basically, iMAC resists the striking disparity between diverse scenarios and achieves an acceptable accuracy in all scenarios.

D. Performance of Tracking

1) Overall Success Rate: Since iMAC, iVR, PHADE are recent indoor localization systems drawing in surveillance cameras, we further analyze the success rate of tracking. For better understanding, we add a classical visual tracking algorithm (**SDC&SGM** [30]) into comparison. Fig.12.c depicts the comparison in 3 distinct scenarios. Significantly, iMAC, iVR and PHADE all achieve better success rate and higher robustness than **SDC&SGM**, which proves combining surveillance video and mobile sensors is a promising way to enhance and promote indoor localization and tracking.

Moreover, iMAC achieves the highest rate in each scenario and shows high robustness, keeping more than 90% success rate regardless of environments. Meanwhile, iVR gains slightly inferior success rate (in 4%) in teaching building and holiday hotel, which is slightly superior (in 6%) than that of PHADE. However, both iVR and PHADE have a more significant drop of success rate in art museum than that of iMAC. Although all these visual tracking algorithms suffer from visual occlusion, iMAC still wins a relative robustness in complex environment by adopting instant geometry features.

Thus, ICFM is demonstrated to have better performance and robustness than using WiFi fingerprints in tracking people in real scenarios. It is remarkable that iMAC gets rid of human intervening in map construction, collecting radio fingerprints and calibrating ambient cameras.

2) Number of Pedestrians: Obviously, the number of pedestrians influences the visual processing and disturbs tracking scheme. We further test the influence of multiple pedestrians in iMAC, which is shown in Fig.12.d.

iMAC achieves the best success rate when there are different number of pedestrians under the camera, which is 97%,95%,93%,88% for 4, 8, 12, 16 pedestrians respectively. iVR also shows high accuracy over 90% within 8 pedestrians, which precipitately drops down to 70% when there are 16 pedestrians. Worse still, PHADE drops down to less than 60% when the number of pedestrians is 16. Although all three methods degenerate with the increase of pedestrians, which induces visual occlusion. iMAC shows significantly better resistance against scenarios with more pedestrians.

Previous solutions depend on WiFi and IMU datas will lose efficacy when more pedestrians appear with similar trajectory

and close location. Thus, their accuracy will deteriorate precipitately, which put those location-based indoor smart application out of commission. However, the increase of pedestrians meanwhile brings more complex geometry features of crowds, which offers rich discrimination for ICFM module and benefits iMAC in crowded public areas.

As a result, iMAC achieves better advantages in scenarios with multiple pedestrians, which are insurmountable for all previous methods.

3) ReTracking Delay: We also concern the time cost for one-time tracking, since it has quite an influence on retracking and relocation. As shown in Fig.12.e, iMAC shows stably excellence success rates, which is 95%, 92%, 90% and 85% when sampling time is set to 7s, 5s, 3s and 1s respectively. iVR performs an equally better rate, which is 92% and 90% in 7s and 5s. However, iVR drops down to 82% and 72% when sampling time reduces to 3s and 1s. Worse still, PHADE faces this drop (78%) even earlier when sampling time reduces to 5s. Eventually, PHADE reaps an unacceptable success rate (51%) when sample time is compressed to 1s.

The above results verify that iMAC keeps an effective performance in each short sampling. Thus it proves ICFM is a highly discriminable real-time model compared with IMU driven model. Since the latter depends on trajectory difference over a period of time, which costs more time to achieve a high accuracy as stable as iMAC. iVR alleviates this shortage by fusing WiFi signals into consideration. However, the improvement of fusing WiFi signals is also limited to a short time slice. For instance, normal human walks about 1.2m per second in a relaxed state [31], which is within the location error engendered by WiFi signals.

iMAC performs extraordinary speed in retracking, which enables users to gain precise location as soon as they pick up their smartphones.

4) Static Pedestrian: Pedestrians regularly slow down or halt their steps in public areas like art museum or shopping mall. We evaluate the performance of iMAC in scenarios with multiple static pedestrians. Concretely, we set different number of static pedestrians ,and set one of them as a user. The result is shown in Fig.12.f.

iMAC shows overwhelming advantages in distinguishing different static pedestrians, achieving over 90% regardless of numbers of pedestrians. Although iMAC, iVR and PHADE all achieve high success rate when there is only one static pedestrian. iVR suffers from multiple static pedestrians, which soon linearly decreases to below 49% when there are 8 static pedestrians. Worst of all, PHADE performs like a random selection algorithm. The rational is IMU module becomes completely out of action when there are multiple static pedestrians. Due to the same reason, iVR performs relatively better since WiFi module still makes efforts to offer a rough difference in location.

Compared with state-of-the-art systems, iMAC overcomes difficulties in dealing with scenarios with multiple static pedestrians, which is very common in public areas.

VIII. Related work

iMAC is the first work to combine mobile cameras and ambient cameras. Here we list most recent works related to our work.

A. Mobile-camera-based Localization

Vision has higher resolution than WiFi kind of radio signals and IMU signals, several existing works leverage mobile camera to improve performance of location service.

OPS [32] integrates GPS, inertial sensors and multiple images of a same object to furnish an outdoor object localization system. Sextant [33] leverages environmental physical features from inertial sensors and mobile cameras to triangulate user locations using at least 3 photos. ClickLoc [15] fuses the advantages of mobile cameras, WiFi fingerprints and IMU signals to achieve an easy-to-use image-based indoor localization system with multi-modal sensing. Travi-Navi [34] and Pair-Navi [17] both provide trace-driven navigation on smartphone. Travi-Navi records high-quality images and sensor readings during a guider's walk on the navigation paths. The followers track the navigation trace, get prompt visual instructions and image tips. Pair-Navi exploits visual SLAM based on mobile cameras to achieve a real-time P2P navigation without help of other sensors in smartphone.

B. Ambient-camera-based Localization

Researchers integrate images from ambient cameras, radio signals and IMU signals to achieve higher accuracy.

RAVEL [12] and EV-Loc fuses visual signals from surveillance cameras with WiFi radio signals for higher location accuracy. [35] combines visual signals from surveillance cameras and sensors signals from IMU to achieve robust pedestrian tracking. Shortly afterwards, PHADE [9] extracts uniqueness patterns of users in surveillance cameras and compares these patterns with user's IMU data to discern different users. Most recently, iVR [11] designs a tightly coupled fusion algorithm to exploit advantages of visual signals, IMU signals and WiFi signals, which outperforms previous systems in accuracy and performs more robust in multi-pedestrian scenario.

C. Easing start-up effort

Indoor floor plan construction has been a major bottleneck for image-based localization, which is time-consuming and labor-intensive. Tango [36] reconstructs 3D indoor structure in real time fusing a depth camera and extra motion capture sensors. Jigsaw [37] using SfM to construct 2D floor plan with commodity smartphones by carefully designed 'Click-Walk-Click' model. IndoorCrowd2D [38] integrates mobile cameras and inertial measurements to construct building interior skeleton. ClickLoc [15] reduces the overhead of image database by correlating image-generated relative models to physical coordinates. iVR [11] further reduces human intervention leveraging two ambient cameras to construct an indoor semantic map

IX. Conclusions

In this paper, we present iMAC, a robust sub-meter accuracy indoor localization and navigation system which fuses observation from mobile cameras and ambient cameras. By integrating observation from two sub-modules, iMAC finally overcomes their respective bottlenecks of heavy start-up efforts and calibration efforts and achieves enhanced accuracy and robustness. iMAC is implemented on several commercial smartphones in different scenarios to validate its performance. The result demonstrates that iMAC shows the light of offering universal indoor location service and becoming a practical indoor navigation system without human efforts.

X. Acknowledgement

This work is supported in part by NSFC under grant 61832010, 61632008, 61672319, 61872081, 61632013.

References

[1] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proceedings of the 18th annual international conference on Mobile computing and networking*, 2012, pp. 269–280.

[2] Z. Yang, Z. Zhou, and Y. Liu, "From rssi to csi: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, pp. 1–32, 2013.

[3] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 313–325.

[4] F. Li, C. Zhao, G. Ding, J. Gong, C. Liu, and F. Zhao, "A reliable and accurate indoor localization method using phone inertial sensors," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012.

[5] Z. Yang, C. Wu, Z. Zhou, X. Zhang, X. Wang, and Y. Liu, "Mobility increases localizability: A survey on wireless indoor localization using inertial sensors," *ACM Computing Surveys*, 2015.

[6] C. Wu, J. Xu, Z. Yang, N. D. Lane, and Z. Yin, "Gain without pain: Accurate wifi-based localization with fingerprint spatial gradient," in *PACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2017.

[7] J. Xu, Z. Yang, H. Chen, Y. Liu, X. Zhou, J. Li, and N. Lane, "Embracing spatial awareness for reliable wifi-based indoor location systems," in *Proceedings of the IEEE MASS*, 2018.

[8] J. Teng, B. Zhang, J. Zhu, X. Li, D. Xuan, and Y. F. Zheng, "Ev-loc: integrating electronic and visual signals for accurate localization," *IEEE/ACM Transactions on Networking (TON)*, 2014.

[9] S. Cao and H. Wang, "Enabling public cameras to talk to the public," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018.

[10] X. Liu, Y. Jiang, P. Jain, and K.-H. Kim, "Tar: Enabling fine-grained targeted advertising in retail stores," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018.

[11] J. Xu, H. Chen, K. Qian, E. Dong, M. Sun, C. Wu, L. Zhang, and Z. Yang, "ivr: Integrated vision and radio localization with zero human effort," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019.

[12] S. Papaioannou, H. Wen, A. Markham, and N. Trigoni, "Fusion of radio and camera sensor data for accurate indoor positioning," in *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*, 2014.

[13] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, 2017.

[14] H. Xu, Z. Yang, Z. Zhou, L. Shangguan, K. Yi, and Y. Liu, "Enhancing wifi-based localization with visual clues," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015.

[15] ——, "Indoor localization via multi-modal sensing on smartphones," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.

[16] Z. Yin, C. Wu, Z. Yang, and Y. Liu, "Peer-to-peer indoor navigation using smartphones," *IEEE Journal on Selected Areas in Communications*, 2017.

[17] E. Dong, J. Xu, C. Wu, Y. Liu, and Z. Yang, "Pair-navi: Peer-to-peer indoor navigation with mobile visual slam," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2019.

[18] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, 2008.

[19] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[20] S. Li, C. Xu, and M. Xie, "A robust o (n) solution to the perspective-n-point problem," *IEEE transactions on pattern analysis and machine intelligence*, 2012.

[21] J. Ratcliffe, *Video surveillance of public places*, 2006.

[22] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[23] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, 2014.

[24] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[25] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[26] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM transactions on graphics (TOG)*, 2006.

[27] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, 2009.

[28] C. Wu, "Towards linear-time incremental structure from motion," in *2013 International Conference on 3D Vision-3DV 2013*, 2013.

[29] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, 2000.

[30] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *2012 IEEE Conference on Computer vision and pattern recognition*, 2012.

[31] F. C. Anderson and M. G. Pandy, "Dynamic optimization of human walking," *Journal of biomechanical engineering*, 2001.

[32] J. G. Manweiler, P. Jain, and R. Roy Choudhury, "Satellites in our pockets: an object positioning system using smartphones," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*, 2012.

[33] Y. Tian, R. Gao, K. Bian, F. Ye, T. Wang, Y. Wang, and X. Li, "Towards ubiquitous indoor localization service leveraging environmental physical features," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, 2014.

[34] Y. Zheng, G. Shen, L. Li, C. Zhao, M. Li, and F. Zhao, "Travi-navi: Self-deployable indoor navigation system," *IEEE/ACM transactions on networking*, 2017.

[35] W. Jiang and Z. Yin, "Combining passive visual cameras and active imu sensors for persistent pedestrian tracking," *Journal of Visual Communication and Image Representation*, 2017.

[36] E. Marder-Eppstein, "Project tango," in *ACM SIGGRAPH 2016 Real-Time Live!*, 2016.

[37] R. Gao, M. Zhao, T. Ye, F. Ye, Y. Wang, K. Bian, T. Wang, and X. Li, "Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014.

[38] 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 *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 2015.