IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2021, accepted February 14, 2021, date of publication February 22, 2021, date of current version March 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3060760

3D Point Cloud-Based Indoor Mobile Robot in 6-DoF Pose Localization Using a Wi-Fi-Aided Localization System

MINGCONG SHU, GUOLIANG CHEN, AND ZHENGHUA ZHANG
School of Environment Science and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China

Corresponding author: Mingcong Shu (shumc@cumt.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502105, and in part by the Science Foundation of Jiangsu Province under Grant BK20161181.

ABSTRACT Six-DoF (Six-degree-of-freedom) pose localization based on 3D point clouds is a challenging task for LBSs (localization-based services). This paper proposes a robust and efficient method that uses multimodal information (vision and Wi-Fi signal information) to estimate the 6-DoF pose of an RGBD camera on a robot with respect to complex 3D textured models of the indoor environment that can contain more than 650,000,000 points. Our developed method narrows the search scope, which delimits boundaries initially using the Wi-Fi location system and applies an environment-adaptive approach to determine the radius of the search sphere based on the signal stability of the Wi-Fi location system. In addition, we propose an algorithm for estimating a novel correspondence between local points with a 3D submap by combining 3D points and surface normals to acquire absolute poses from noisy and outlier-contaminated matching point sets for RGBD sensors in dynamic indoor scenes. Then, a novel two-level spatial verification strategy is used to estimate an accurate pose, which includes the use of a RANSAC (Random Sample Consensus) algorithm for identification and a direct least-square method to acquire the pose from the inliers. The proposed method has been implemented and tested extensively in various indoor scenes. The experimental results demonstrate that the Wi-Fi-aided localization system can efficiently localize a mobile robot in a variety of large-scale 3D point cloud datasets to realize stable time consumption and similar performance to state-of-the-art methods.

INDEX TERMS Visual localization, Wi-Fi positioning, 6-DoF pose, multimodal information, computation complexity.

I. INTRODUCTION

Autonomous navigation and localization inside a building are essential capabilities of robotic intelligent systems [1], [2]. The global positioning system (GPS) cannot operate indoors, but visual localization methods can be used in GPS-denied environments. However, human indoor environments change every day. The most challenging problem is efficient self-localization of the robot with sufficient accuracy in a global map without the specification of the initial pose [3]–[7].

Robot localization has been a hot topic for decades due to the difficulties of self-localization in a global map. GPS is usually used for localization in outdoor open areas, and Monte Carlo localization (MCL) [8]–[10] is broadly applied to indoor localization for robots. However, MCL requires the initial pose of the robot, which is a substantial shortcoming. Image-based localization can provide localization for robots anytime in both indoor and outdoor environments. In the following sections, previous studies on image-based localization are reviewed.

A. IMAGE RETRIEVAL-BASED LOCALIZATION

The global localization problem is similar to the recognition problem, which is often approached as an image retrieval problem. The localization of a robot's scene is predicated by efficient image indexing techniques [11]–[18] and can be further improved by spatial reranking [19], distinctive visual information feature selection [20] or feature weighting [21]–[24]. Milford and Wyeth [25] proposed

The associate editor coordinating the review of this manuscript and approving it for publication was Peng Liu.

an approach for place recognition across large perceptual changes that involves linear sequential filtering on image matching. Sünderhauf *et al*. [26] proposed leveraging the robustness of convolutional features with regional proposals for accurate topological localization. Badino and Kanade [27] proposed an approach that fuses LiDAR and image data with a particle filter framework to realize long-term place recognition. Although these approaches have shown impressive results in challenging conditions, they do not provide metric information regarding the 6-DoF pose of the camera. Torii *et al*. [28] used Google Street View images and corresponding depth maps to synthesize virtual views to boost place recognition performance. However, this method can output only an approximate localization with a topological approach, and it cannot determine the exact 6-DoF poses of robots.

B. 3D MAP-BASED VISUAL LOCALIZATION

Recently, techniques for obtaining 6-DoF poses directly have been developed with 3D maps. The maps are usually composed of a 3D point cloud that has been constructed via structure-from-motion (SfM) [29], and the points are related to local image features that are used for triangulation. These approaches establish 2D-3D or 3D-3D correspondences between the query descriptors and the 3D points in the map. The exact 6-DoF poses of robots can be obtained by feature matching and solving a classic perspective-n-point problem or iterative closest point (ICP) [30]–[37]. Outliers are always present in the process of correspondence. The RANSAC algorithm is utilized to cope with outliers. Pose estimation based on local features depends strongly on the quality of the feature matches. Image feature descriptors are not robust during image degradation—e.g., due to blur or viewpoint differences. Moreover, with 3D map extension, the number of points is large, and the time complexity of searching grows rapidly. To solve these problems, McManus *et al*. [38] proposed an approach for learning salient visual elements of a place using a bank of support vector machine (SVM) classifiers. This approach is hybrid, as it uses weak localizers to find the closest topological node in the map and refines the pose using the bank of SVM classifiers according to the place. It realizes submeter localization accuracy and requires 10 MB of storage per place. Kendall [39] proposed directly regressing the camera pose from a monocular image via an end-to-end approach. Kendall and Cipolla [40] showed that modeling the uncertainty in camera pose estimates can lead to higher localization performance. Very recently, Walch *et al*. [41] proposed learning the contextual features of images using spatial LSTMs [42] in combination with the PoseNet architecture to increase the localization accuracy. Compared to approaches that depend on local features, the main limitation of these DL approaches is that the estimated pose tends to have low accuracy and depends strongly on the spatial resolution of the training images.

C. WI-FI FINGERPRINT-BASED LOCALIZATION

Recently, Wi-Fi fingerprint-based indoor localization has become one of the most attractive methods due to the wide deployment and availability of Wi-Fi infrastructure, especially in the research area of pedestrian indoor positioning [43], [44]. Many research institutions and companies, such as Apple and Huawei, have invested in and are exploiting related products. Such schemes depend on the received signal strength indicator (RSSI) of the Wi-Fi signal as the observed value and employ smartphones or other Wi-Fi receptors as clients, which renders them free of extra infrastructure and specialized devices and feasible for indoor scenes. Typically, fingerprint-based approaches consist of two stages: The first stage is offline training, in which a fingerprint database (radio map) is constructed by collecting signals with known location notes. Then, during the online localization stage, the position is determined by matching fingerprint observations against those stored in the database. The training phase is usually completed via a site survey, which is labor intensive and time consuming and leads to a major hurdle for real applications. Recent advances in mobile crowdsourcing have increased the efficiency of fingerprint database construction, thereby rendering Wi-Fi fingerprint-based indoor location methods practical [45]–[48]. However, the established RSSI fingerprint-based location system still suffers from several drawbacks, especially in terms of localization accuracy. The state-of-the-art approaches that depend on fingerprints can realize only meter-level accuracy for 2D positioning (X- and Y-coordinates) without direction information; hence, this system cannot be directly applied in robot localization, which requires much more accurate pose estimation.

D. PROBLEM STATEMENT AND SYSTEM DESCRIPTION

A precise 6-DoF pose estimate is highly important for many multimedia applications, such as unmanned aerial vehicles and humanoid robots. However, several challenges remain in robot localization with RGBD cameras in a 3D point cloud environment. First, self-localization of a robot in a 3D environment is typically a time-consuming task due to the large quantity of point cloud data that must be processed. This is especially critical in robotics tasks in which small and low-cost robots (e.g., aerial robots) are frequently used [49]. Second, in a dynamic indoor environment, visual features, such as 2D image features or 3D point cloud features, can become lost or emerge. For instance, the doors and windows can be opened or closed, books on a table can be opened, and objects can be moved out of a building [31]. Third, indoor environments usually have similar layouts, especially in public indoor scenes, which can cause perceptual aliasing and perceptual variability with visual localization methods [21], [50], [24].

In this paper, we propose an efficient and robust localization system that is based uniquely on multimodal information with a known precomputed 3D textured model and a Wi-Fi

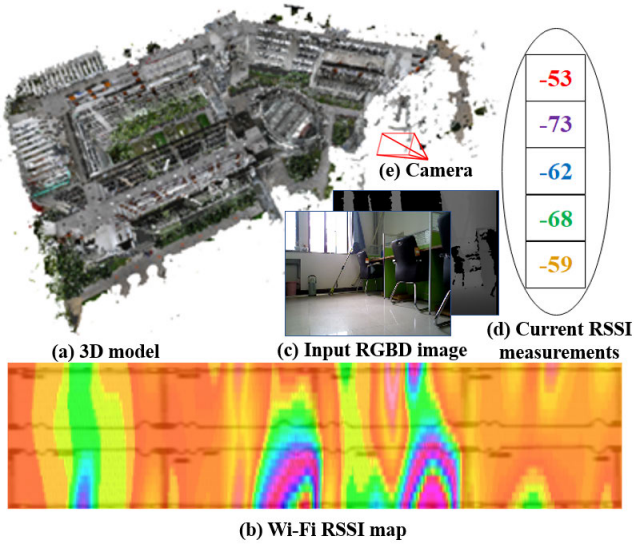


FIGURE 1. Problem definition: In an environment with a known textured 3D model (a) and indoor radio map on calibration points (b), given an input RGBD image (c) and current RSSI measurements (d), the problem is to estimate the pose of the camera that captures the visual and signal information with respect to the environment (e). The main challenge addressed in the paper is to identify the correspondence of points efficiently and reliably for complex 3D models that contain many points. In this example, the model of the CUMT has more than 650,000,000 points.

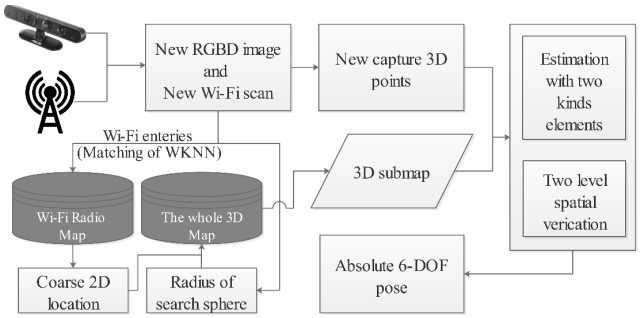


FIGURE 2. Framework of 6-DoF pose estimation with a multimodal location system.

fingerprint map of the indoor environment, as illustrated in Fig. 1. The developed system can efficiently estimate the full pose (translation and rotation) of the RGBD camera on the robot within the map from a single RGBD image and the Wi-Fi received signal strength. No temporal information is used about the previous poses that can constrain the process of estimating the current pose and to where the robot is pointing. While this method significantly increases the complexity of our problem, its output renders the pose estimation robust to issues such as drifting, occlusions of the 3D map, and sudden robot motions.

More precisely, let us assume that the 3D map is composed of n 3D points, and each is associated with a point descriptor that represents its appearance in the process of correspondence. Fig. 2 illustrates the overall scheme of the proposed method. In our case, we must estimate the 3D-3D correspondence between local points and the 3D submap points from the known depth information with the RGBD camera; then, the pose estimate can be acquired. The current m 3D points are compared against all 3D points in the map. However, solving this correspondence problem has a complexity of $O(m \times n)$, which is extremely costly since n can be very large (e.g., 650,000,000 points). Moreover, in dynamic indoor scenes, the n 3D points of the map may change daily—e.g., due to people walking, doors opening, and object movement—which will render the problem more challenging.

E. MAIN CONTRIBUTIONS

We present a novel approach for fast robot 6-DoF pose determination in large-scale indoor environments. We address the three main challenges of robot localization.

(1) Lack of initial pose. Initial pose determination is the first step and a key point for robot global localization in large-scale indoor environments. Typically, researchers use place recognition methods in which the robot attempts to match its scene to previously built maps to identify an initial pose, and image-based retrieval techniques are widely applied to address the problem. However, these topological approaches rely on visual feature landmarks that depend on the path of image capture beforehand, and this method does not satisfy our requirements since it relies on a metric map constructed from 2D image features. To overcome this problem, we built a large-scale indoor 3D map and adapted a coarse-to-fine strategy that aided the Wi-Fi localization system in quickly estimating the RGBD camera's 6-DoF pose of the robot with high accuracy. We utilize Wi-Fi signals as side-channel information to acquire an initial 2D position and fully exploit the Wi-Fi-based global location method, which is robust to dynamic indoor scenes without cumulative error. The multimodal localization system can increase the robustness of the robot pose estimation task in large indoor scenarios.

(2) Computing time and self-similarity in large-scale scenes. Time complexity reduction must be considered to realize the objective of robot self-localization in a large 3D indoor environment more quickly. Indoor environments are often highly self-similar due to many systemic and repetitive elements on large and small scales (e.g., corridors, room, tiles, windows, chairs, and doors). To overcome this problem, we consider the use of features other than images to represent each scene. Visual information can describe many visible features; however, perceptual aliasing is a fatal problem in vision-based localization. In this paper, we propose fusing the Wi-Fi signal feature and visual features to estimate the one-to-one relationship between scene feature expression and geographic position. Moreover, the Wi-Fi-aided localization system can efficiently acquire a coarse 2D position, which can be used to narrow the search scope in global 3D maps within 1 s; then, pose estimation can be conducted via small map executed segmentation. The Wi-Fi signal is used as side-channel information to guide the expensive 3D-3D correspondence. In detail, robots can obtain a small submap by using a Wi-Fi localization system to determine the space

and radius of the submap, which can efficiently reduce the computation time on resource-constrained platforms without the risk of perceptual aliasing or perceptual variability in the visual localization research domain.

(3) Dynamic indoor scenes. Indoor environments are highly dynamic scenes compared to outdoor environments due to the presence of many moveable objects, such as people and furniture, which results in severe occlusion problems in computer vision research when viewed from a close distance. This can cause an incorrect image retrieval result to be obtained, which leads to localization failure. To overcome this problem, we rely on multiple forms of correspondence based on 3D points and surface normal information to realize more accurate and robust pose estimation. Furthermore, in our pose estimation step, a novel correspondence is established between local 3D points and the submap's 3D points. Then, a novel direct least-square algorithm is used to estimate the pose from the inliers obtained by solving RANSAC, which can effectively restrain the influence of dynamic scene factors on the positioning process and improve the positioning accuracy.

II. CONSTRUCTING THE 3D MODEL AND WI-FI FINGERPRINT MAP

Our approach for robot localization assumes that a 3D textured model and Wi-Fi fingerprint map of the indoor scene are available. However, large-scale datasets remain challenging, especially in an indoor environment with various types of features. Datasets that cover larger indoor scenes [3], [4], [51] have so far captured fairly small spaces, such as a single room, and have been constructed from dense-captured sequences of RGBD images. However, they are designed for object retrieval and lack the Wi-Fi fingerprint map; hence, they are not suitable for robot Wi-Fi-aided visual localization.

In this paper, a new large-scale indoor localization dataset with vision and signal maps is introduced for multimodal localization, which includes query images captured from a wide range of viewpoints in various scenes across two floors. To build these precise 3D models of an indoor scene, we use RGBD Mapping [52], which is a SLAM system for continuous image collections. The hardware platform of the mapping system is shown in Fig. 3, which consists of three parts: a motion execution unit, a data acquisition unit, and a processing unit. In our experiment, AICRobo XII, which is a mobile robot developed by ALCRobo, uses a two-wheel differential drive chassis—namely, Kobuki—which was developed by Yujin of Korea. An RGBD camera—namely, ASUS Xtion PRO LIVE, which was developed by Asus—is used as a data acquisition unit, and it is similar to Kinect, which consists of an infrared projector, a CMOS image sensor, and an RGB color camera. Fig. 3 shows the data acquisition unit, and the specifications of the unit are presented in Table 1. The DELL G7 laptop on the robot has an Intel i5-8300 dual-core 3 GHz processor with an NVIDIA GTX1060TI graphics card, 32 GB of RAM, and an attached 802.11 Wi-Fi module, which is used as a processing and testing unit.

FIGURE 3. Testing of the hardware platform in an indoor scene with many AP Macs mounted on the wall.

TABLE 1. Specifications of ASUS Xtion PRO LIVE.

ASUS Xtion PRO LIVE	Specifications
Depth Image Size	VGA (640×480): 30 fps
Resolution	SXGA (1280×1024)
Field of View	58°H, 45°V, 70°D
Distance of Use	Between 0.8 m and 3.5 m
Operation Scene	Indoor
Dimensions	18 × 3.5 × 5 in

Since the error probability increases with increasing mapping area, we use the system to construct only the submap. Then, the complete 3D map can be constructed by using point aligning. By starting from two submap clouds with the largest overlap, a set of point-to-point correspondences between them is manually selected. A rigid transformation is coarsely initialized from the selected correspondences and refined with a generalized ICP algorithm [53]–[55]. The complete indoor 3D model is constructed by using the SLAM mapping system. Then, an indoor Wi-Fi fingerprint map is constructed via a method that we propose in this paper, which is novel and more efficient than previous approaches [56]–[58]. The fingerprint map construction is presented in detail in the following.

To build the indoor fingerprint efficiently and use high-accuracy location information in the vision localization system, we propose a vision-based radio map construction model that utilizes an external Wi-Fi receptor fixed on the robot and a built-in RGBD camera to collect visual information and Wi-Fi fingerprints, as shown in Fig. 4. The visual information includes an RGB color image and a depth image. The Wi-Fi fingerprint includes the media access control (MAC) number of the APs, n, and the corresponding received signal strength value, s. The full indoor Wi-Fi radio map consists of n sites that describe the feature, $f_i^{wifi} = [s_1, s_2, \ldots, s_n]$, with a known location. An overview of the proposed radio map construction method is presented in Fig. 5.

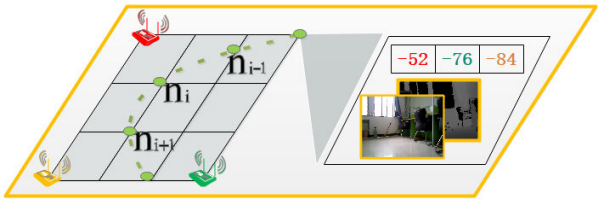


FIGURE 4. Vision-based radio map construction method schematic diagram.

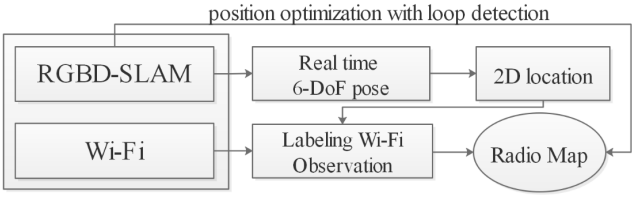


FIGURE 5. System overview of the proposed radio map construction method.

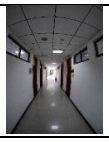
The localization graph in our method is a set of nodes, $n_1, n_2, n_3, \ldots, n_i, \ldots, n_n$(each frame is considered a node) joined by a set of directed edges l_{i-1}, l. Each edge contains odometry data for the 6 DoFs (computed by the RGBD mapping system), which are described by the metric relations between the nodes. The receptor that collects Wi-Fi fingerprints is labeled with a 2D position that is not precise due to system drift in SLAM, and the more accurate location labels of fingerprints can be acquired after the back-end optimization. The trigger frequency of the Wi-Fi receptor depends on the sampling distance, and the empirical value of the distance is set as 0.8 m based on previous studies [59], [60].

For this paper, we select a multistory building 3D model with a variety of indoor scenes, which include offices, corridors, halls and meeting rooms, on the CUMT campus. To unify the map coordinates of the floors into the same coordinate system, we establish a control surveying net that covers these floors by using a Hi-Target ZTS-121 Total Station (Hi-Target Surveying Instrument Co. Ltd., Guangzhou, China) with a 2 s angle error and a 2 mm positioning error every 1000 m. Photographs of the project site are shown in Fig. 6. The whole 3D model is registered on the uniform coordinate system, which belongs to the local coordinate system in geodesy research, and its corresponding elevation datum is on the first floor of our experimental scenes.

Our dataset is composed of 3D maps and Wi-Fi RSSI fingerprints, and the 3D vision maps are constructed from a database of RGBD images geometrically registered to the floor maps. The points in the 3D map consist of the exact location information (x,y,z) including height, while Wi-Fi RSSI fingerprints consist only of exact plane 2D position information with floor descriptions for coarse height information. The database is augmented with a separate set of RGBD query images captured by the same robot to render it suitable for the task of indoor localization. The provided query images are annotated with manually verified ground-truth 6-DoF

FIGURE 6. Hi-Target ZTS-121 Total Station is used as a building control point to unify the maps of various floors (the red boxes in the images are landmarks for more precise surveying).

TABLE 2. Statistics of the dataset.

Scenes	Office	Corridor	Hall	Meeting Room
Environment image				
Dataset size	19,970	41,072	9,716	25,722
Queries	68	103	42	59

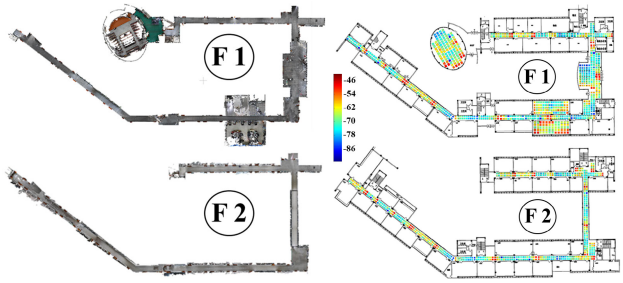


FIGURE 7. Indoor 3D model and Wi-Fi signal maps of the two floors.

camera poses in the global coordinate system of the 3D map. The basic indoor RGBD dataset consists of 96,480 RGBD images obtained by scanning a two-floor lab building at CUMT with the robot. The dataset is divided into four scenes: offices, corridors, halls and meeting rooms. The statistics of the dataset are presented in Table 2.

As in previous visual localization research [2], [61], the whole 3D model is merged manually for high quality, and the fingerprint of the constructing submap overlap areas employs the Kriging interpolation to acquire RSSI values [62], [63]. The overhead view of the whole model is presented in Fig. 7 (left), and Fig. 7 (right) shows the Wi-Fi signal maps of the two floors. In this paper, we use a relational database, the format of which is described in Table 3, to store these data.

III. SIX-DOF POSE ESTIMATION USING THE PROPOSED WI-FI-AIDED LOCALIZATION SYSTEM

The pipeline of the fast 6-DoF pose determination of the robot with the Wi-Fi-aided localization system consists of the following three steps: First, we use the Wi-Fi location

TABLE 3. Storage format of the Wi-Fi database.

RGBD Images	2D Position	AP Number	AP_1	…	AP_m
1	(x_1, y_1)	m1	MAC_1=$RSSI_{11}$	…	MAC_{m1}=$RSSI_{m1}$
2	(x_2, y_2)	m2	MAC_1=$RSSI_{12}$	…	MAC_{m2}=$RSSI_{m2}$
3	(x_3, y_3)	m3	MAC_1=$RSSI_{13}$	…	MAC_{m3}=$RSSI_{m3}$
⋮	⋮	⋮	⋮	⋮	⋮
n	(x_n, y_n)	mn	MAC_1=$RSSI_{1n}$	…	MAC_{mn}=$RSSI_{mn}$

system to identify a coarse 2D position, and advance pose identification is conducted with an adaptive radius of search determined by the stability of the Wi-Fi signal source to ascertain the submap segmentation. Second, we estimate a novel correspondence between local points with a 3D submap by combining 3D points and surface normals for absolute pose acquisition from noisy and outlier-contaminated matching point sets for RGBD sensors in dynamic indoor scenes. Finally, a novel two-level spatial verification strategy for accurate pose estimation is applied, which includes a RANSAC algorithm to identify and a direct least-square method to acquire the pose from the inliers. The three steps are detailed next.

A. SUBMAP ACQUISITION USING WI-FI-AIDED POSITIONING

Most well-known approaches for indoor pedestrian positioning are based on the RSSI of 802.11 transmitted packets, and many studies have demonstrated that satisfactory accuracy can be realized in indoor scenarios by employing fingerprinting techniques [7]. However, the direction information of a robot's pose cannot be estimated from a Wi-Fi-based indoor location. We now utilize the Wi-Fi signal to efficiently search for a 3D submap with no perceptual aliasing and perceptual variability, in contrast to previous place recognition methods and fixed grid submap division. We have explored the relation between Wi-Fi signal stability and accuracy, which determines the size of the 3D submap. The Wi-Fi-based position accuracy is easily influenced by dynamic indoor environments, so a self-adaptive algorithm for determining the size of the submap is proposed. We seek to develop a novel method for evaluating the accuracy and robustness of fingerprint localization. To realize this objective, we formulate a relationship between the received RSSI and location quality. We also use the matching distance value obtained during the RSSI-based Wi-Fi localization to assess the quality. The Wi-Fi data can be automatically collected when the robots move around the scene, and the evaluation method is described in detail as follows.

For the robot receiving Wi-Fi signal strength, we observe the following information at a time frame t in an indoor space:

$$p_i^{(t)} = \{RSSI_1, RSSI_2, \ldots, RSSI_j\}^{(t)} \tag{1}$$

where $p_i^{(t)}$ is the position that corresponds to the time frame t, and RSSI is the received signal strength from AP antenna j. The Wi-Fi fingerprint approach assumes that each position can be uniquely defined by the RSSI signal strength values. For convenience, we usually set the acceptable AP number to be the same, and an infinitesimal value is assigned for the AP RSSI when it can receive no or only a weak signal.

Assuming that the Wi-Fi database of a floor is created, it is essential to efficiently compare the list of Wi-Fi scans to the Wi-Fi scans stored in the database D_{Wi-Fi}. Via this comparison, the most similar fingerprint in D_{Wi-Fi} is identified, and the corresponding position of the fingerprint is obtained.

Basic Wi-Fi fingerprint localization includes deterministic [64] and probabilistic methods [65]. Deterministic algorithms use a similarity metric to differentiate online signal measurement and fingerprint data. Then, the target is located at the closest fingerprint location in signal space [66]. The major advantage of the deterministic methods is their ease of implementation compared to probabilistic algorithms, which usually require some probabilistic assumptions (such as Gaussian noise or probabilistic independence [67]) and more datasets than traditional deterministic algorithms [68]. Traditional deterministic methods can be easily implemented based on k nearest neighbors (k-NN) and the computational complexity is often low [69]. Some other more advanced deterministic algorithms such as support vector machine [70] and linear discriminant analysis [71] show better localization accuracy with higher computational cost. The k-weighted nearest neighbor algorithm is a version of k-NN that has been improved by introducing a weighted distance factor. Since the proposed method constructs a radio map, the RP distribution is sparse or nonuniform. This disadvantage can be overcome by using the k-WNN algorithm with better classification results. Thus, in this paper, we use a k-WNN-based Wi-Fi fingerprint location algorithm, which provides an accurate position with a relatively simple computing method, to acquire a coarse 2D position estimate. The distance that describes the similarity between fingerprints is the Sorensen distance function, which can realize higher accuracy than the Euclidean distance [72]. Therefore, in this paper, we select the Sorensen function for the k-WNN algorithm. The Sorensen distance function is defined as follows:

$$L_{(d,i)} = \frac{\sum_{i=1}^{n} \left| RSSI_{MN}^{(i,n)} - RSSI_{RP}^{(i,n)} \right|}{\sum_{i=1}^{n} \left| RSSI_{MN}^{(i,n)} + RSSI_{RP}^{(i,n)} \right|} \tag{2}$$

where i is a sequence of RPs, n denotes the number of APs with the same MAC between the receptor node and RPs, MN denotes the receptor node, and RP denotes a reference point. Then, the nodes with the k nearest distances are selected as the nearest neighbors. Finally, the distances that correspond to the k neighbors are used to calculate the normalized weight:

$$W_i = \frac{1/L_{(d,i)}}{\sum_{i=1}^{K} (1/L_{(d,i)})} \tag{3}$$

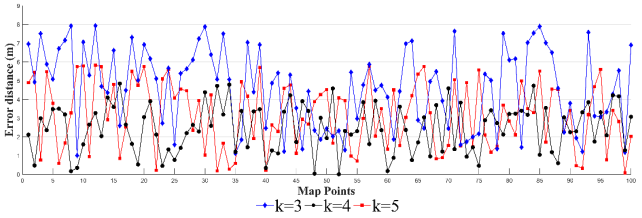


FIGURE 8. Point positioning error curves with various K values.

According to the normalized weight and RP coordinates, a coarse 2D position is acquired using the k-WNN-based Wi-Fi fingerprint via the following formula:

$$P_i(x, y)_{k-WNN} = \{\sum_{i=1}^{K} x_i \cdot W_i, \sum_{i=1}^{K} y_i \cdot W_i\} \tag{4}$$

We choose 100 known positions with distinct K values to calculate the mean square error, and the results are presented in Fig. 8. According to Fig. 8, $k = 4$ is the optimal choice, so we set K to 4 in this paper, and the accuracy of the k-WNN-based Wi-Fi location algorithm is under 6 m.

Due to the uncertainty of positioning using a Wi-Fi signal, only the area of robots can be determined, with no exact location. In this paper, a novel evaluation parameter that we propose is calculated from the ratio test value, R, to indirectly quantitatively evaluate the variability of the Wi-Fi fingerprint. The accuracy of Wi-Fi-based localization depends strongly on the signal RSSI stability [73], [46]. The ratio test value is defined as the significance of the fingerprint similarity, and the value can be calculated using the following formula:

$$R = \frac{L_{(q,i)}^{first}}{L_{(q,i)}^{\sec ond}} < \mu \tag{5}$$

where $L_{(q,i)}^{first}$ is the nearest neighbor between query fingerprint q and reference fingerprint i, and $L_{(q,i)}^{\sec ond}$ is the second nearest neighbor between query fingerprint q and reference fingerprint i. The value μ is compared with the ratio test value, which determines the radius of the rendered submap. In our experiments, two radius values are selected: One radius is set as 3 m, and the other radius is set as 6 m. The ratio value is between 0 and 1, and μ is set as 0.7.

According to Wi-Fi localization, which can provide the initial position and radius of the submaps, robots obtain a coarse localization, and an accurate 6-DoF pose is estimated by identifying a correspondence between location points and submap points. However, previously established point-to-point (P2P) matching approaches would be highly inefficient. We can substantially increase the efficiency by combining 3D points and surface normals, as we describe next.

B. INTEGRATING MULTIPLE FORMS OF CORRESPONDENCE

In this section, two forms of correspondence are considered, as illustrated in Fig. 9. In the 3-3 correspondence, $p_i \in R^3$ and $q_i \in R^3$ represent two sets of 3D points defined in

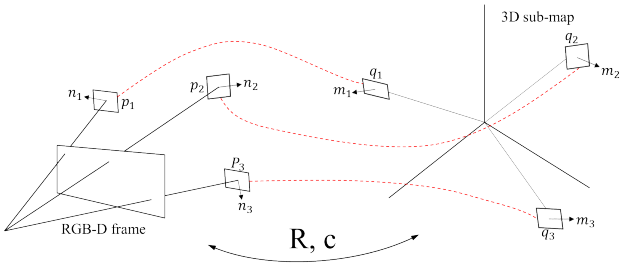


FIGURE 9. Pose estimation using two forms of 3-3 correspondence: ($p_i \Leftrightarrow q_i$) and ($n_i \Leftrightarrow m_i$).

the camera coordinate system C and the world coordinate system W, respectively, and each pair of correspondences can be obtained via rotation and translation—namely, $p_i - R(q_i - c)$. Simultaneously, a pair of 3-3 correspondences can be identified by using the surface normal with surface normal (N-N) correspondence. n_i and m_i are defined as a pair of N-N correspondences in the camera and world coordinate systems, respectively, and are both unit vectors in R^3. The surface normal (N-N) correspondences satisfy the relationship $n_i = Rm_i$, which is independent of the camera center between the pair.

The relationships above with the two forms of correspondence can be incorporated into a single error function of R and c by using least-square estimates. This can be formulated as follows:

$$e^2(R, c) = \frac{1}{|\Lambda_1|} \sum_{i \in \Lambda_1} \|p_i - R(q_i - c)\|^2 + \frac{\psi}{|\Lambda_2|} \sum_{i \in \Lambda_2} \|n_i - Rm_i\|^2 \tag{6}$$

In the above function, we use a weight, ψ, to balance the relative contributions of the two forms of correspondence, and Λ_1, Λ_2 represent the sets of matching pairs with 3D points and their surface normal correspondences, respectively, where $|\Lambda_1|$ is the number of Λ_1 in the i-th set, and $\frac{1}{|\Lambda_1|}$ and $\frac{\psi}{|\Lambda_2|}$ are weights that depend on each importance in the above equation. Outliers are always present and could cause localization to fail. Thus, we propose a novel spatial verification that combines RANSAC and least-square solution to minimize a set of noisy matched pairs that are contaminated with outliers to estimate the 6-DoF poses of robots. This will be described in detail in the next section.

C. POSE ESTIMATION USING TWO-LEVEL SPATIAL VERIFICATION

We now obtain the ultimate robot 6-DoF pose by applying the RANSAC algorithm and least-square solution on all correspondences and removing the noise in a dynamic indoor environment. RANSAC is largely applied in image-based location to identify outliers, and we use the algorithm to delete noisy matched pairs. Two combinations (3-3 or 3-3 and N-N) of each pair are applied using the algorithm. First, the candidate pose can be obtained by using a randomly sampled minimal set of correspondences that are mixtures

of the two forms. The two forms of correspondence provide 5 independent constraints, and there are many possibilities for selecting the minimum set. We choose two forms with a satisfactory mix and available solutions to address the problem. Then, we propose a direct least-square solution that minimizes (6) and estimates R and c for the robots with a set of matching pairs and associated correspondences. We assume the above correspondences are inliers obtained in RANSAC, and a noniterative algorithm is applied, which is based on the direct least-square optimization of 3-3 correspondences proposed by Shinji [74]. In brief, for a specified set of N 3D point correspondences p_i and q_i, the algorithm has three steps:

The translation is eliminated by subtracting the centroid from each set of points: $q_i' = q_i - \bar{q}$ and $p_i' = p_i - \bar{p}$.

The optimized rotation estimate is obtained by $\hat{R} = USV^T$, where USV^T is a singular value decomposition (SVD) of the covariance matrix $Cov = \frac{1}{N}\sum p_i' q_i'^T$, and S is the identity matrix.

The estimate of an optimized camera center in the world coordinate system can be obtained using $\hat{c} = \bar{q} - R^T\bar{p}$.

This algorithm can be extended to include N-N correspondence except for 3D correspondence, and N-N correspondence can be straightforwardly combined with the covariance matrix Cov due to its independence of the camera center. The 3-3 correspondences can be used to identify the camera center in the direct least-square solution, and the rotational constraints are combined with the covariance matrix in the N-N correspondence process. Therefore, the extended covariance matrix is

$$Cov = \frac{1}{|\Lambda_1|}\sum_{i\in\Lambda_1} p_i' q_i'^T + \frac{\psi}{|\Lambda_2|}\sum_{i\in\Lambda_2} n_i m_i^T \qquad (7)$$

The two forms of correspondence (N-N and 3-3) can provide all rotational constraints for estimating the 6-DoF pose, and we set $\psi = \frac{1}{|\Lambda_1|}\sum |p_i'|^2$ to assign them equal roles in determining the rotation. The unit vectors for normalizing p_i' and q_i' can also solve the problem. However, the lost length of the unit vector could decrease the precision in rotation matrix estimation. After using the SVD of Cov to obtain the optimized R in 2 above, the optimized camera center $\hat{c}$ could be estimated. The direct least-square algorithm can obtain $\hat{c}_s$ by optimizing 3-3 correspondences. In the end, the camera center is the mean of $\hat{c}$ and $\hat{c}_s$.

IV. EXPERIMENTAL VERIFICATION

In this section, we describe the experimental setup for evaluating the robot 6-DoF pose estimation performance using our dataset (Section IV.A). The proposed method—namely, ''Wi-Fi-Aid''—is compared with state-of-the-art methods and shows advantages in large-scale indoor scenes (Section IV.B). Finally, we compare the pose estimation accuracies among various indoor scenes and analyze the effect factors for the method (Section IV.C).

A. IMPLEMENTATION DETAILS

Our evaluation dataset was gathered from a two-story building indoor scene. The building was selected for several reasons. First, its size (1479 m^2) allows for a thorough evaluation of our system and has various advantages. Second, it is representative of a practical implementation setting that includes many typical indoor scenarios in an office building with a strong demand for location-based services. Furthermore, the building interior is a complex space filled with vast open meeting rooms, narrow corridors, offices with similar appearances and ever-changing halls, which render it a challenging environment for vision-based methods.

To adequately evaluate the Wi-Fi-aided visual-based localization system, we gauge its performance along two performance axes. The first is the localization time. Six-DoF localization is the most computationally expensive component and requires matching correspondence among abundant points or features. In this paper, we define the localization time as the time from when RGBD data are input into the system to when the 6-DoF pose for the frame is determined via two-level spatial verification. The second aspect we evaluate is the localization accuracy, which is the difference between the system's query pose estimates of the position and orientation and the ground-truth values. We report the position error in meters and the orientation error in degrees.

B. COMPARISON WITH THE STATE-OF-THE-ART METHODS

In this paper, we compare the proposed method with four established methods in terms of both accuracy and processing time. The first compared system utilizes direct 2D-3D matching methods, which are modified versions of state-of-the-art 3D structure-based image localization methods [30], [33]. RootSIFT [15] computes features for whole images in the dataset, and we associate these features with 3D coordinates by using the known geometric information. We extract the features from query images for matching to the dataset 3D point descriptors. No more than five database images are selected to receive the most matches. Then, we obtain poses by utilizing all these matches. Finally, P3P-LO-RANSAC [75] is applied to compute the 6-DoF pose. State-of-the-art image retrieval-based localization is the next compared method, which uses bag-of-visual-words with Hamming embedding [13] to represent images, and a 200K vocabulary is applied in the method described by RootSIFT from our dataset and trained on affine covariant features [21]. The top 100 candidate images are reranked via spatial verification using features [19]. Lowe's ratio test is not applied here, as it would remove too many features that must be retained in indoor scenes. The 6-DoF poses of query images are computed with P3P-LO-RANSAC by using the inliers. We use the same features in the method for fair comparison and P3P-LO-RANSAC for pose estimation as the direct 2D-3D matching algorithm, as described above. Our third method for comparison is based on NetVLAD [76], which is a variation of the above image retrieval-based localization

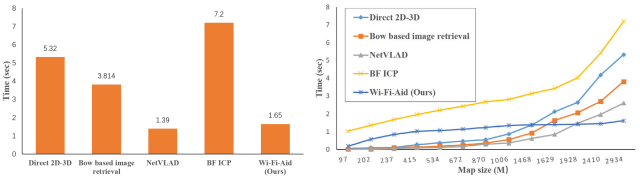


FIGURE 10. Average time consumptions of the five compared methods (right) and their runtimes with increasing map scale (left).

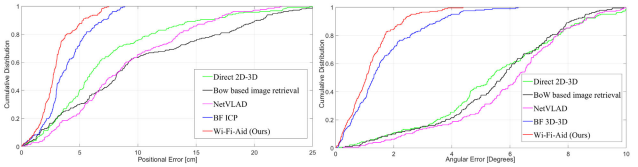


FIGURE 11. CDF plots of the position and orientation errors of direct 2D-3D, Bow-based image retrieval, NetVLAD, BF-ICP and Wi-Fi-Aid (ours).

method. The difference between them is a candidate shortlist, and this method uses NetVLAD for acquisition. The last method we compare is BF (brute-force) matching against 3D points in the whole map and computation of the 6-DoF pose using ICP-RANSAC [77].

1) RUNTIME EVALUATION

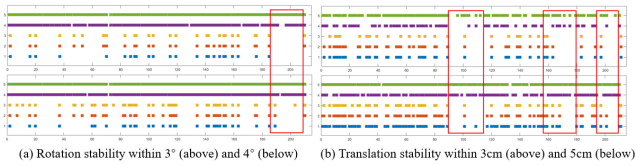
The average runtimes of the five visual localization methods—namely, Direct 2D-3D, Bow-based image retrieval, NetVLAD, Brute Force-ICP (BF-ICP) and Wi-Fi-Aid (ours)—for the whole 3D map are presented in Fig. 10 (left). NetVLAD takes only 1.39 s on average to complete one pose estimation, and the method is the fastest in this experiment. The reason is that NetVLAD, by using a deep learning algorithm, can significantly shorten runtimes during the localization process. However, this comes at the price of lower accuracy, as presented in Fig. 11. Our proposed method, Wi-Fi-Aid, can realize similar time consumption to NetVLAD, and the localization speed of the method is more than 3, 2 and 4 times those of Direct 2D-3D, Bow-based image retrieval and BF-ICP, respectively. However, NetVLAD requires a higher hardware configuration and takes a substantial amount of time to perform model training. Fig. 10 (right) shows the relation between the map of the area (dataset size) and localization runtime for 12 query images that cover only a small map. When positioning with small map size, such as room level, the proposed method needs to increase the time consumption of Wi-Fi positioning. Meanwhile, both the proposed method and the BF-ICP method use additional depth information for matching, which leads to higher complexity compared with other 2D-3D matching methods. When the size of the local map is larger than 1468 M, the radio signals in our method can speed up the selection of matching regions and the size of the region is relatively stable. First, global positioning is carried out based on Wi-Fi positioning, and then fine local positioning is carried out based on image-based localization. The runtime of our approach is stable and varies less with increasing map size, while those of the other four methods vary more with increasing map size, and their time consumptions grow exponentially. Our approach can realize shorter runtime due to the use of radio location as an aid to quickly search and extract submap information; hence, high performance is realized, and the localization system can exhibit stable time consumption with various 3D model sizes in large indoor scenes. Moreover, compared to the BF-ICP method, our method has lower time consumption with all map sizes, which is beneficial to Wi-Fi-aided localization. From Fig. 10, we find that Wi-Fi or other opportunistic radio signals can effectively assist visual localization. This verifies quantitatively that the result in real indoor scenes is the same as previous researchers' conjecture [61], and the adaptive submap segmentation radius algorithm based on signal feature analysis is innovatively proposed for visual localization in section III.A.

TABLE 4. Evaluation of our largescale indoor dataset. The numbers represent the median positional error (cm) and angular error (degrees).

Scene	Direct 2D-3D	BoW-based image	NetVLAD	BF 3D-3D	Wi-Fi-Aid (Ours)
Office	10, 6.24	11, 7.45	10, 6.95	2, 1.32	**2, 1.05**
Corridor	18, 8.37	21, 8.15	17, 8.36	4, 1.49	**3, 1.31**
Hall	13, 7.16	17, 7.84	17, 7.94	3, 1.64	**3, 1.14**
Meeting Room	15, 6.90	19, 7.81	18, 7.43	3, 1.61	**2, 1.07**
Average	14, 7.17	17, 7.81	16, 7.67	3, 1.52	**2.5, 1.14**

2) LOCALIZATION ACCURACY

In Table 4, we present the median position errors and angular errors of the five compared methods for four indoor scenes: office, corridor, meeting room and hall. According to the table, our approach realizes the best position quality in all scenes. Moreover, it is on average one order of magnitude more accurate due to the consideration of geometric information. Compared to pure image feature-based methods, our approach, which uses an RGBD camera, realizes significantly higher pose estimation accuracy for both translation and rotation. Compared to brute force 3D-3D, a slight performance increase in terms of the translation error is realized on the corridor scene dataset, while the gain in terms of the rotation error on all scene datasets is readily observed for the two forms of correspondence we performed. Fig. 11 presents the cumulative distribution functions (CDFs) of Direct 2D-3D, Bow-based image retrieval, NetVLAD, BF-ICP and Wi-Fi-Aid. Our approach yields the best results in terms of both position error and angular error, and it realizes substantially higher accuracy than the visual localization methods—namely, Direct 2D-3D, Bow-based image retrieval and NetVLAD, which use RGB pictures to acquire 6-DoF poses. The RGB images provide only 2D features of indoor scenes and lack depth information, which is important for textureless scenarios. Moreover, depth information can provide more constraint conditions for calculating the pose of the camera. In addition, the accuracy of the method we



(a) Rotation stability within 3° (above) and 4° (below) (b) Translation stability within 3cm (above) and 5cm (below)

FIGURE 12. System stability evaluation of five methods using fixed rotation and translation errors (vertical ordinates 1-4 (Direct 2D-3D, Bow-based image retrieval, NetVLAD and BF-ICP). The results of four established algorithms are compared with those of the proposed method—namely, method 5—which are represented by the green bands. The horizontal ordinates represent query samples, and the red rectangular areas represent poor results in extremely complex environments.

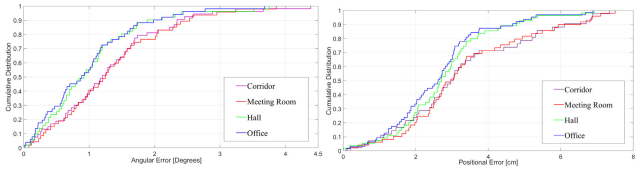


FIGURE 13. CDF plots of the position and orientation errors of Wi-Fi-Aid on four scenes.

FIGURE 14. Challenging test images from all datasets. The images contain structural ambiguities, transparent windows, and severe occlusion.

propose is significantly increased by 10% and 13% in terms of the position error and angular error, respectively, which is the result of integrating multiple forms of correspondence using a surface normal.

3) SYSTEM STABILITY AND EFFICIENCY

In terms of overall system performance, a comparison of the stability of the fixed position accuracy (position and angular errors) between Wi-Fi-Aid and the four competing approaches is reported in Fig. 12. The performance of the positioning system is evaluated by analyzing the fixed 6-DoF pose precision continuity of the entire set of spatially ordered query images. In the experiment, we select angular errors of less than 3° and 4° and position errors of less than 3 cm and 5 cm for system continuity verification and analysis, respectively. Vertical ordinates 1-4 in the figures are the comparison algorithm results, and 5 represents the algorithm we propose. The left panel presents the continuity analysis results of the positioning algorithms under the two fixed angular errors (3° and 4°), and the right panel presents the continuity analysis results of the algorithms under the two fixed position errors (3 cm and 5 cm). In Fig. 12, the continuous stability of the high-quality camera pose estimation algorithm based on 2D picture features is lower, especially in indoor areas with few textures or less light (the area of the red rectangle in Fig. 12), which lead to pose estimation deviation spurts, thereby resulting in an inability to realize high location accuracy. However, the algorithms with depth information constraints (algorithms 4 and 5) can realize higher continuous 6-DoF pose estimation performance under the fixed errors. Therefore, via the acquisition of RGB image information in the process of robot positioning in a large range of indoor scenes, the camera pose can be estimated, while the acquisition of depth information can effectively increase the positioning accuracy and system stability to ensure stable and smooth operation in the whole scene. Compared to algorithm 4—namely, the BF-ICP algorithm—our proposed algorithm 5 can realize superior continuous high-precision positioning performance, which is due mainly to the incorporation of additional point cloud computing vector information into the pose estimation method, which can effectively increase the matching accuracy. In addition, the RANSAC least-square algorithm can effectively eliminate the noise matching error in the dynamic scene, thereby further increasing the pose estimation precision. Regardless of the angular error or position error, the algorithm proposed in this paper can maintain continuous high-precision positioning and realize the robust pose estimation of indoor large-scale dynamic scenes.

C. PERFORMANCE IN VARIOUS INDOOR SCENES

To evaluate the performance of the proposed system in various indoor scenes, we choose four typical indoor environments—namely, a corridor, a meeting room, a hall and an office—to test the robustness of the methods, and 50 points with true 6-DoF poses in each scene are applied in the experiments. As shown in Fig. 13, we also use the error CDF to evaluate the adaptive capacity of the proposed pose estimation system in these four indoor scenes. Our proposed location system performs better in the hall and office than in the corridor and meeting room in terms of both translation errors and rotation errors.

From Fig. 14, we conclude that the two scenes have similar structures and visual features, especially in corridor environments with textureless walls, which lead to vision deviation, thereby resulting in incorrect matches in pose estimation. However, according to Table 4, these four scenes have nearly the same average errors in terms of both position and angle, and the variances of the errors are stable. However, errors are certainly present in the 3D maps of our experimental scenes, and the map quality naturally influences the location accuracy.

V. CONCLUSION

With the objective of improving the imprecise and time-consuming camera 6-DoF pose estimation for robot location within known indoor environments, in this paper, we proposed a novel crowdsourcing pose acquisition algorithm that utilizes a Wi-Fi signal and an RGBD camera. Real-case experiments in large-scale indoor scenes and the results of various state-of-the-art methods demonstrate that the Wi-Fi-aided image-based localization system realized higher accuracy, stability and efficiency than four previously established algorithms (direct 2D-3D, bow-based image

retrieval, NetVLAD, and BF-ICP). In summary, the proposed Wi-Fi-aid image-based localization algorithm for 6-DoF pose estimation with an RGBD camera and a wireless signal receiver placed on a robot has four innovative aspects:

1) To the best of our knowledge, we are the first to deeply apply a Wi-Fi location algorithm to the image-based localization process for 6-DoF pose estimation using an RGBD camera and a Wi-Fi receiver in a large-scale indoor environment that includes various complex scenes, which can inspire researchers in the same field to develop an elegant solution to the image-based localization problem. In addition, SLAM-based Wi-Fi RSSI radio map construction was proposed to efficiently complete the labor-intensive preliminary work in Wi-Fi localization research.
2) We explored the beneficial impacts of the Wi-Fi-aided image-based localization system and found that incorporating Wi-Fi signal location into 6-DoF pose estimation reduces the computational complexity significantly and substantially shortens the localization time consumption in our real-case large-scale indoor scenes. Our approach has a stable time complexity that is independent of the database size (the number of 3D points). Moreover, a novel method for selecting an adaptive radius of search area according to the stability of the Wi-Fi signal source is applied to ascertain the submap segmentation adaptively, and the method we proposed has higher robustness to dynamic indoor scenes without visual aliasing, which can provide a location for each query image with limited errors.
3) We proposed a novel correspondence between local points with a 3D submap by combining 3D points and surface normals to acquire absolute poses from noisy and outlier-contaminated matching point sets for RGBD sensors in dynamic indoor scenes. Furthermore, a novel two-level spatial verification strategy for accurate pose estimation was proposed, which includes a RANSAC algorithm for identifying and a direct least-square method for acquiring the pose from the inliers. The experimental results demonstrate that the accuracy and stability of the proposed algorithm are substantially higher than those of the other four algorithms considered in this paper, and the proposed algorithm outperforms the RGB-image-based algorithm without depth information in terms of both translation error and rotation.
4) We explored the system performance in four indoor scenes (corridor, meeting room, hall and office). The results demonstrate that our method realizes high localization performance in various indoor environments, and the influence of the indoor layout is not readily identifiable in our location system.

For camera 6-DoF pose estimation in indoor scene research, it is ideal to realize optimal performance on large-scale, complex and fast-changing indoor environments; however, there are many issues to be resolved. Our proposed algorithm also has various shortcomings, which will be overcome in our future studies:

First, we should consider combining higher-semantic-level visual information to estimate the camera pose. We performed our experiments under the assumption that light is stable in indoor environments. However, visual localization in large-scale indoor environments is challenging due to complex and dynamic scenes. Moreover, factors that change with time can result in loss of trust in and reliability of the localization map. The captured visual information usually differs with time in map construction. Similar to changes in light intensity, light changes always occur with changes in weather or artificial light, and the cameras are sensitive to lighting and susceptible to degradation in the quality of visual features, which reduce the performance of 6-DoF pose computation in image-based localization. The consideration of geometric and high-level semantic information can increase the robustness of the visual localization system. We will address this aspect in the future by using semantic visual information, which is more robust to lighting changes and has been applied in place recognition, SLAM and self-driving technology [78], [79] to solve dynamic scenes in CV (computer vision) research.

Second, we use an RGBD camera to capture the visual information, whereas general monocular cameras are of lower cost and are more widely used in practice. However, general cameras cannot acquire depth information, and increasingly many smartphones have recently been equipped with lower-cost cameras with TOF (time of flight) technology. We should consider the use of an RGBD camera with TOF on a handheld mobile platform and will perform pose estimation experiments with the latest widely available TOF camera.

Finally, in the experiment, we assume that the 3D map and the radio map have not changed significantly, but sometimes real environments are dynamic and affected by external factors, such as indoor redecoration or temperature variation. Therefore, we should consider map changes and develop an algorithm and scheme for updating significantly changed submaps. This will be our next objective.

REFERENCES

[1] A. Debski, W. Grajewski, W. Zaborowski, and W. Turek, "Open-source localization device for indoor mobile robots," *Procedia Comput. Sci.*, vol. 76, pp. 139–146, 2015, doi: 10.1016/j.procs.2015.12.327.
[2] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-DOF localization in large-scale environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1043–1050.
[3] S. Xun, Y. Xie, L. Pei, and L. Wang, "A dataset for benchmarking image-based localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7436–7444.
[4] J. Valentin, A. Dai, M. Nießner, P. Kohli, P. Torr, S. Izadi, and C. Keskin, "Learning to navigate the energy landscape," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 323–332.
[5] S. Wang, S. Fidler, and R. Urtasun, "Lost shopping! monocular localization in large indoor spaces," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2695–2703.
[6] T. Schmidt, R. Newcombe, and D. Fox, "Self-supervised visual descriptor learning for dense correspondence," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 420–427, Apr. 2017.

[7] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single rgb image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3364–3372.

[8] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte Carlo localization for mobile robots," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 1999, pp. 1322–1328.

[9] D. Fox, W. Burgard, F. Dellaert, and S. Thrun, "Monte Carlo localization: Efficient position estimation for mobile robots," in *Proc. Nat. Conf. Artif. Intell.*, 1999, pp. 1–7.

[10] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, "Robust Monte Carlo localization for mobile robots," *Artif. Intell.*, vol. 128, nos. 1–2, pp. 99–141, May 2001.

[11] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[12] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1578–1585.

[13] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[14] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2161–2168.

[15] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2911–2918.

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[17] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, p. 1470.

[18] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[20] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 889–896.

[21] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1582–1590.

[22] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla, "Learning and calibrating per-location classifiers for visual place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 907–914.

[23] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1169–1176.

[24] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 883–890.

[25] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2012, pp. 1643–1649.

[26] N. Sünderhauf *et al.*, "Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Annu. Conf. Acad. Marketing Sci. (AMS)*. Springer, 2015.

[27] H. Badino, D. Huber, and T. Kanade, "Real-time topometric localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1635–1642.

[28] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1808–1817.

[29] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 72–79.

[30] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1744–1756, Sep. 2017.

[31] A. Irschara, C. Zach, J. M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2599–2606.

[32] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5974–5983.

[33] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2102–2110.

[34] S. Cao and N. Snavely, "Minimal scene descriptions from structure from motion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 461–468.

[35] F. Camposeco, T. Sattler, A. Cohen, A. Geiger, and M. Pollefeys, "Toroidal constraints for two-point localization under high outlier ratios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4545–4553.

[36] S. Choudhary and P. J. Narayanan, *Visibility Probability Structure from SfM Datasets and Applications*. Berlin, Germany: Springer, 2012.

[37] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, *Worldwide Pose Estimation Using 3D Point Clouds*. Berlin, Germany: Springer 2012.

[38] C. McManus, B. Upcroft, and P. Newman, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robot., Sci. Syst. Conf.*, Berkley, CA, USA, Jul. 2014.

[39] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-DOF camera relocalization," *Educ. Inf.*, vol. 31, pp. 2938–2946, 2015.

[40] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2015, pp. 4762–4769.

[41] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization with spatial LSTMs," in *Proc. ICCV*, 2017.

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[43] B. Roberts and K. Pahlavan, "Site-specific RSS signature modeling for WiFi localization," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2009, pp. 1–6.

[44] Q. Chen, G. Huang, and S. Song, "WLAN user location estimation based on receiving signal strength indicator," in *Proc. 5th Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Sep. 2009, pp. 1–4.

[45] M. Xue, W. Sun, H. Yu, H. Tang, and R. Zimmermann, "Locate the mobile device by enhancing the WiFi-based indoor localization model," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8792–8803, Oct. 2019.

[46] W. K. Zegeye, S. B. Amsalu, Y. Astatke, and F. Moazzami, "WiFi RSS fingerprinting indoor localization for mobile devices," in *Proc. IEEE 7th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2016, pp. 1–6.

[47] M. Nowicki and J. Wietrzykowski, *Low-Effort Place Recognition With Wi-Fi Fingerprints Using Deep Learning*. Cham, Switzerland: Springer, 2016.

[48] S. Boonsriwai and A. Apavatjrut, "Indoor WiFi localization on mobile devices," in *Proc. 10th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol.*, May 2013, pp. 1–5.

[49] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 8, 2019, doi: 10.1109/TPAMI.2019.2952114.

[50] R. Arandjelovi and A. Zisserman, "DisLocation: Scalable descriptor distinctiveness for location recognition," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 188–204.

[51] E. Wijmans and Y. Furukawa, "Exploiting 2D floorplan for building-scale panorama RGBD alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 308–316.

[52] A. Concha and J. Civera, "RGBDTAM: A cost-effective and accurate RGB-D tracking and mapping system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 6756–6763.

[53] G. K. Tam, Z.-Q. Cheng, Y.-K. Lai, F. C. Langbein, Y. Liu, D. Marshall, R. R. Martin, X.-F. Sun, and P. L. Rosin, "Registration of 3D point clouds and meshes: A survey from rigid to nonrigid," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 7, pp. 1199–1217, 2013.

[54] M. Magnusson, A. Lilienthal, and T. Duckett, "Scan registration for autonomous mining vehicles using 3D-NDT," *J. Field Robot.*, vol. 24, no. 10, pp. 803–827, 2007.

[55] N. J. Mitra, N. Gelfand, H. Pottmann, and L. Guibas, "Registration of point cloud data from a geometric optimization perspective," in *Proc. Eurographics/ACM SIGGRAPH Symp. Geometry Process.*, 2004, pp. 22–31.

[56] P. Chen, Y. B. Xu, L. Chen, and Z. A. Deng, "Survey of WLAN fingerprinting positioning system," *Appl. Mech. Mater.*, vols. 380–384, pp. 2499–2505, Aug. 2013.

[57] A. Eleryan, M. Elsabagh, and M. Youssef, "Synthetic generation of radio maps for device-free passive localization," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2011, pp. 1–5.
[58] M. B. Kjrgaard, "A taxonomy for radio location fingerprinting," in *Proc. Int. Symp. Location-Context-Awareness*, 2007, pp. 139–156.
[59] S. Dai, L. He, and X. Zhang, "Autonomous WiFi fingerprinting for indoor localization," in *Proc. ACM/IEEE 11th Int. Conf. Cyber-Phys. Syst. (ICCPS)*. New York, NY, USA: ACM, 2020.
[60] B. A. Labinghisa and D. M. Lee, "Indoor localization algorithm based on behavior-driven predictive learning in crowdsourced Wi-Fi environments," *Mod. Phys. Lett. B*, vol. 33, nos. 14–15, May 2019, Art. no. 1940036.
[61] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg, "Wide area localization on mobile phones," in *Proc. 8th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2009, pp. 73–82.
[62] S.-S. Jan, S.-J. Yeh, and Y.-W. Liu, "Received signal strength database interpolation by kriging for a Wi-Fi indoor positioning system," *Sensors*, vol. 15, no. 9, pp. 21377–21393, Aug. 2015.
[63] H. Zhao, B. Huang, and B. Jia, "Applying kriging interpolation for WiFi fingerprinting based indoor positioning systems," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2016, pp. 1–6.
[64] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Proc. IEEE Conf. Comput. Commun. 19th Annu. Joint Conf. IEEE Comput. Commun. Societies (INFOCOM)*, Mar. 2000, pp. 775–784.
[65] M. Youssef and A. Agrawala, "The horus location determination system," *Wireless Netw.*, vol. 14, no. 3, pp. 357–374, Jun. 2008.
[66] D. Han, S. Jung, M. Lee, and G. Yoon, "Building a practical Wi-Fi-based indoor navigation system," *IEEE Pervasive Comput.*, vol. 13, no. 2, pp. 72–79, Apr./Jun. 2014.
[67] W. Sun, J. Liu, C. Wu, Z. Yang, X. Zhang, and Y. Liu, "MoLoc: On distinguishing fingerprint twins," in *Proc. IEEE 33rd Int. Conf. Distrib. Comput. Syst.*, Jul. 2013, pp. 226–235.
[68] J. Seitz, T. Vaupel, J. Jahn, S. Meyer, J. G. Boronat, and J. Thielecke, "A hidden Markov model for urban navigation based on fingerprinting and pedestrian dead reckoning," in *Proc. 13th Int. Conf. Inf. Fusion*, Jul. 2010, pp. 1–8.
[69] S. He and S.-H.-G. Chan, "Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 466–490, 1st Quart., 2016.
[70] C.-L. Wu, L.-C. Fu, and F.-L. Lian, "WLAN location determination in e-home via support vector classification," in *Proc. IEEE Int. Conf. Netw., Sens. Control*, Mar. 2004, pp. 1026–1031.
[71] G. Nuño-Barrau, *A New Location Estimation System for Wireless Networks Based on Linear Discriminant Functions and Hidden Markov Models*. London, U.K.: Hindawi, 2006.
[72] J. Torres-Sospedra, R. Montoliu, A. Martinez-Uso, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Oct. 2014, pp. 261–270.
[73] J. Bi, Y. Wang, X. Li, H. Qi, H. Cao, and S. Xu, "An adaptive weighted KNN positioning method based on omnidirectional fingerprint database and twice affinity propagation clustering," *Sensors*, vol. 18, no. 8, p. 2502, Aug. 2018.
[74] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, Apr. 1991.
[75] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized RANSAC," in *Proc. BMVC*, 2012.
[76] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018.
[77] Y. Guo, Y. Gu, and Y. Zhang, "Abstract: Invariant Feature Point based ICP with the RANSAC for 3D Registration," *Inf. Technol. J.*, vol. 10, no. 2, pp. 276–284, 2011.
[78] G. A. Kumar, J. H. Lee, J. Hwang, J. Park, S. H. Youn, and S. Kwon, "LiDAR and camera fusion approach for object distance estimation in self-driving vehicles," *Symmetry*, vol. 12, no. 2, p. 324, Feb. 2020.
[79] H. Zhang, G. Chen, Z. Wang, Z. Wang, and L. Sun, "Dense 3D mapping for indoor environment based on feature-point SLAM method," in *Proc. 4th Int. Conf. Innov. Artif. Intell.*, 2020, pp. 42–46.

MINGCONG SHU is currently pursuing the Ph.D. degree with the School of Environment Science and Spatial Informatics, China University of Mining and Technology. His research interests include indoor positioning and navigation, visual localization, and computer vision.

GUOLIANG CHEN is currently a Professor and a Ph.D. Supervisor with the School of Environment Science and Spatial Informatics, China University of Mining and Technology. He is also the Director of the Key Laboratory of the National Administration of Surveying, Mapping and Geographical Information for Land, Environment and Disaster Monitoring. He has published more than 20 academic articles, has been awarded eight patents, and has copyrights on ten software programs. His research interests include indoor-outdoor positioning technology and geographic location-based services. He is a two-time National Science and Technology Progress Prize laureate. He has been awarded more than two provincial or ministry-level prizes.

ZHENGHUA ZHANG is currently pursuing the Ph.D. degree with the School of Environment Science and Spatial Informatics, China University of Mining and Technology. His main research interests include 3D deep learning and 3D reconstruction.