This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

Efficient Locality Classification for Indoor Fingerprint-based Systems

Ka-Ho Chow Suining He Jiajie Tan S.-H. Gary Chan

Abstract—Locality classification is an important component to enable location-based services. It entails two sequential queries: 1) whether a target is within the site or not, i.e., inside/outside region decision, and 2) if so, which area in the region the target is located, i.e., area classification. Locality classification is hence more coarse-grained and efficient as compared with pinpointing the exact target location in the region. The classification problem is challenging, because fingerprints may not exist outside the region for training. Furthermore, the target may sample an incomplete RSSI vector due to, say, random signal noise, momentary occulsion or scanning duration. The algorithm also has to be computationally efficient.

We propose INOA, a scalable and practical locality classification overcoming the above challenges. INOA may serve as a plug-in before any fingerprint-based localization, and can be incrementally extended to cover new areas or regions for large-scale deployment. Its preprocessor cherry-picks only those discriminating access points, which greatly enhances computational efficiency and accuracy. By formulating a "one-class" classifier using ensemble learning, INOA accurately decides whether the target is within the region or not. Extensive experimental trials in different sites validate the high efficiency and accuracy of INOA, without the need of full RSSI vectors collected at the target.

Index Terms—Fingerprinting, locality classification, area classification, inside/outside region decision, context-awareness.

1 Introduction

In indoor fingerprint-based localization [1], [2], the RSSI (received signal strength indication) vectors of wireless signal are first collected at various locations of a region by professional surveyors or through crowdsourcing [3]. These RSSI vectors and their collected locations, the so-called *fingerprints*, are then stored offline in a database. In the online query phase, a user (or target) first samples an RSSI vector at her/his spot. Given the fingerprint database, the vector is then mapped to a location via a certain measure (say, the minimum Euclidean distance of signal vectors [1]). Due to its wide applicability in complex indoor environments without making any assumption on the path loss model [4], fingerprinting is practical and promising to deploy.

Beyond pinpointing exact locations of a target in a fingerprint region, the region may be partitioned into areas according to layout or functionalities (accompanied by the region boundaries), for example, floors, food court, zones, and the like. In order to support fingerprint-based indoor location-based service (ILBS), we hence need efficient *locality classification*, which consists of the following two sequential queries:

1) *In/Out region decision:* Given any RSSI vector, is the target within the boundary of the fingerprint region or not?
2) *Area classification:* If the target is within the region, in which area (say, floors or zones) is he/she at?

Efficient locality classification leads to better ILBS. It enables, for example, a context-aware service where indoor and outdoor maps can be seamlessly switched for a roaming user. Knowing the area the target is at, the search space of fingerprints can be greatly reduced, leading to efficient pinpointing of the target location. Yet another application is indoor user analytics, where user density in a certain area or shop can be studied.

Devising efficient and ubiquitously deployable locality classification is challenging. This is because fingerprints are often collected only *inside* a region of interest, with few or no training fingerprints in the outside (exterior) region to differentiate the interior ones. This is in stark contrast to traditional multi-class classification algorithms, which usually require a substantial amount of training data from both classes (i.e., the presence of fingerprints both inside and outside the region).

Besides the absence of exterior fingerprints, the locality classification algorithm must achieve a performance robust enough to handle partially missing or noisy data. In a fingerprint system, a target often cannot collect the full RSSI vector at its location due to random missing signals arising from, for example, signal noise, obstruction of access points, scanning duration, and so on. Traditional one-class classification algorithms [5], [6] cannot easily accommodate such incompleteness in data, as they often assume a complete vector. For a good user experience, the locality classification should work despite missing signals, and vectors which cannot be classified with high confidence (e.g., being at the boundary of adjacent areas) should be rejected.

Computational efficiency is also an important concern for locality classification. Previous localization approaches are fine-grained by nature, searching over all fingerprints to pinpoint a user location. This is not scalable to a large geographical region. The in-out decision and area classification, due to their partitioning or coarse-grained nature, should be computationally efficient, scalable to a large region, and incrementally extensible to new regions or areas. It may

- *The authors are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China.*
  E-mail: {khchowad, sheaa, jtanad, gchan}@cse.ust.hk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

also serve as a precursor to the fine-grained localization to narrow the search space of fingerprints.

We propose **INOA**, a highly robust and efficient locality classification system for **in**side/**o**utside region decision and **a**rea classification overcoming the above challenges. It is applicable to various mobile devices (given automatic or crowdsourced device calibration [6]), complementary to any existing fine-grained localization system (including state-of-the-art [1], [2], [3], [4]), and may serve as a universal plug-in to enhance their localization efficiency. INOA consists of the following novel and salient features:

- *Preprocessing for site adaptivity, classification accuracy & computation efficiency*: INOA consists of a preprocessing module which is adaptive to different sites or indoor structures. Via AP signal analysis, it conducts clustering and signal compression to minimize redundant or correlated APs. By retaining only the discriminating APs, locality classification becomes more accurate and computationally efficient.
- *Robustness against missing signals and incrementally extensible to new areas*: INOA is robust against missing signals in the collected RSSI vector. We formulate the in-out decision as a *data description* problem [5], which classifies the signal patterns outside the region as outliers (*novelties*) versus those inside (*regulars*). We use ensemble learning as a backbone design, where multiple one-class classifiers are embedded in order to capture the signal patterns in different aspects of learning. Our formulation is incrementally extensible to new regions or areas, i.e., the region can be extended by including additional sites without the need to retrain the entire system or classification model.
- *Rejection of unclassifiable signals*: INOA consists of an efficient rejection module which denies a vector if the sampled (noisy) signals do not provide location results with sufficient high confidence. In this way, INOA identifies the credibility of signals for improved ILBS result presentation and user experience.

We show in Figure 1 the major modules in INOA. The modules are divided into *offline* and *online* phases. Each area in the region is assigned a unique ID (indicating its area or locality). In the offline phase, given the fingerprints and their corresponding area IDs, we first perform *fingerprint preprocessing* and construct a new signal space by eliminating non-discriminative and redundant APs. Then, we train classification models for the in-out decision and area classification, and store their model parameters in the database.

In the online phase, the user samples an RSSI vector and then requests his/her area. The *unclassifiable signal rejection* module rejects the vector which cannot lead to a good result with high confidence (due to excessive missing or noisy signals). Otherwise, the *inside/outside region decision* module first checks whether the user is inside the fingerprint region or not. If so, the vector is then fed into the *area classification* module to find and return the area the user is at.

We have implemented INOA, and conducted extensive experimental trials in various sites (our university campus, a premium shopping mall named Hong Kong Harbor City, and a business building named Hong Kong Cyberport). For concreteness, we consider Wi-Fi RSSIs in our expositions and experiments, though INOA is applicable to any existing or emerging fingerprint signals such as Bluetooth [7] or

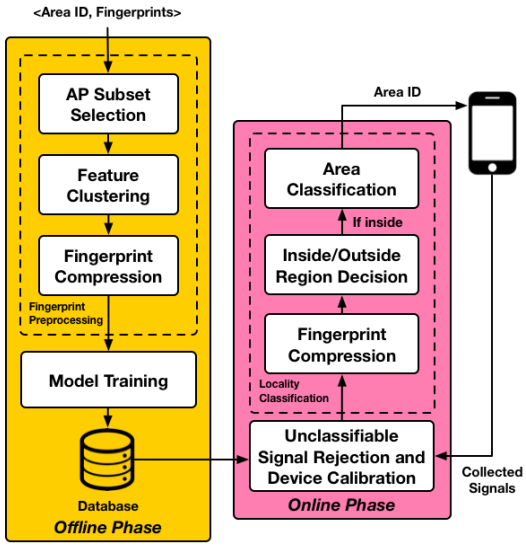


Fig. 1: System modules and work flow of INOA.

channel state information [8]. Our study shows that INOA outperforms state-of-the-art algorithms substantially (more than 20% in most of our study) despite many incomplete signals. Moreover, INOA performs far more efficiently than our state-of-the-art comparison schemes [6], [9] and hence can scale to a large site in its practical deployment.

The remainder of this paper is organized as follows. After briefly reviewing related works in Section 2, we discuss the fingerprint preprocessing module of INOA in Section 3. Section 4 presents our locality classification algorithms in terms of inside/outside region decision and area classification. We discuss unclassifiable signal rejection and device calibration in Section 5. Illustrative experimental results are presented in Section 6, followed by concluding remarks in Section 7.

2 Related Work

Fingerprint-based algorithms for indoor localization have been extensively studied for decades, pioneered by RADAR [1], Horus [2] and PCA [10]. The works either utilize some similarity metrics between reference point (RP) fingerprints and target signals, or assume a certain probabilistic model on the RSSI. Though the results are impressive, they mainly focus on the fine-grained pinpointing of user location, and have not studied large-scale and more effective coarse-grained locality classification. Our paper investigates such a problem and proposes an algorithm serving as a plug-in for the above state-of-the-art systems to improve their computation efficiency and accuracy.

(1) Inside/outside Region Decision: Locality classification begins with identifying the targets inside the fingerprint region (i.e., inside/outside region decision). As INOA is the first work addressing the generic in-out decision problem in ILBS, we review the indoor/outdoor detection problem, a common application of the in-out decision, instead.

- *Signal thresholding*: A straightforward way is to set thresholds as decision boundaries. The work in [11] considers the target is outdoor if more than 3 satellites are received in the GPS signals, while [12] uses the accuracy drop of GPS indoors as an indicator. [13] and [14] enhances the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

detection accuracy by using the fusion of multiple sensors (inertial sensors, light detectors and magnetometers) to monitor abrupt environmental changes. They define several thresholds (e.g., the light intensity) in different scenarios to make the decision. Nevertheless, time (daytime or nighttime), weather (cloudy or rainy) and nearby skyscrapers may affect GPS detection and sensor readings (e.g., the sun is blocked by the clouds). To capture the transition from outdoors to indoors, or vice versa, the work in [15] installs iBeacon near the entrances of the building and keeps monitoring the number of beacons received by the target. Then, they determine whether it is indoor or outdoor by jointly considering the GPS signals. However, installing iBeacon introduces extra effort/costs and may not be feasible in many spacious sites.

- *Advanced learning techniques*: Due to the simplicity and imperfection of thresholding, more advanced learning models have been studied. The scheme in [16] leverages LTE Measurement Data to train a supervised learning model for indoor/outdoor classification. Training data from both environments is required, and hence is labor-intensive. The work in [17] proposes a semi-supervised learning model using readings from multiple sensors. It still requires labeled data from both environments to achieve a good performance. The work in [18] uses cameras to detect the scenes (indoor/outdoor) from a set of training images using machine learning algorithms. However, continuously enabling the camera significantly and rapidly consumes power in mobile devices.

INOA is more versatile than all the existing works, applicable to scenarios beyond simply indoor/outdoor. It makes in-out decisions in any context (e.g., room, floor, building, etc.) as it differentiates between inside and outside regions with only radio frequency (RF) signal patterns and does not rely on abrupt environmental changes. Also, INOA does not require fingerprints outside the region, which is an essential feature to serve as a plug-in for any existing fingerprint-based ILBS where only inside fingerprints are available.

(2) One-class Classification with Missing Data: We formulate INOA as a one-class classification problem, making it applicable to any context beyond indoor/outdoor. As briefly mentioned, missing data is often a serious problem for one-class classifications. In fingerprint-based ILBS, missing data (signals) is inevitable due to measurement imperfection and incompleteness. Hence, simply applying existing one-class classification techniques does not work. Here, we briefly review some recent works on handling missing data from a machine learning perspective.

- *Data imputation*: The typical way to address missing data is to replace it by a default value [19] (e.g., the weakest signal intensity in our case), so-called data imputation. Recently, some approaches have attempted to model the non-missing data to predict the missing ones based on some distributions. The authors in [20] leverage Bayesian Multiple Imputation, while [21] introduces a combination of K-nearest neighbor and self-organizing map. However, the unique characteristics of RF signals make data imputation difficult. RF signals can be missing either at random or deterministically (e.g., if the source is too far away from the target). If we improperly impute data into the signals which are not missing at random, the resultant signal pattern will be contaminated.
- *Feature transformation*: The works in [22] and [23] leverage the negligible change in (dis)similarity measures when a few data (say, dimensions or features) are missing. To make an in-out decision, one-class classification is conducted in a similarity space by transforming RP fingerprints and target signals using a certain similarity measure. The transformation process is tedious as every RSSI vector should be compared against all RP fingerprints. Hence, in large-scale deployment, its offline training and online querying complexities are intractable. Also, similarity-based classification is not scalable as the system should be retrained from scratch whenever an area is added or removed.

In contrast, INOA does not require data imputation, and it addresses missing data from a new perspective. It employs ensemble learning and creates multiple simple models in 2-D space. INOA handles missing signals effectively and efficiently, and is scalable without the need of retraining the entire system.

(3) Area Classification: Given the user is within the region, we conduct area classification to locate the target. Here, we overview the recent approaches focusing on either floor/room localizations (the most common application) or general area classifications.

- *Signal thresholding and pattern recognition*: The works in [24] and [25] utilize an accelerometer to achieve floor localization by detecting walking patterns, while [26] and [27] differentiate floors by significant air pressure differences. They require initial location input and constant calibration over time as the distance errors accumulate with the accelerometer, and barometer readings may be affected by thermal changes. On the other hand, room-level classification using sound [28] and light [29] has attracted much attention. In general, the intensity change or signal reflection by dividing walls [30] is utilized to differentiate rooms. However, tedious calibration or modification over smartphones for the above signals is often inevitable. None of the above works are applicable to general areas without explicit wall partitions or a large difference in elevation.
- *Advanced learning techniques*: The schemes in [31] and [32] respectively leverage a probabilistic model and clustering techniques to identify areas. While the above works seem promising, they do not consider the impact of missing data due to measurement error [33]. The work in [34] leverages K-means and Kohonen layer to compress the fingerprint database, and an artificial neural network to estimate the floor. Despite its extendability to area classification, the system is not scalable as offline computation time in training the Kohonen layer and neural network is usually heavy due to the tedious training process and random parameter initialization. Also, the parameters (e.g., number of neurons) should be learned from scratch when an area is added to or removed from the system. The work in [35] proposes a probabilistic framework and some heuristics to find the significant APs. However, it works only for floor detection.
- *Nearest neighbor search*: The scheme in [36] searches against the entire Wi-Fi fingerprint database of different rooms and buildings to find the nearest one, which is identical to fine-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

grained localization and hence computationally expensive for large-scale deployment. The work in [37] proposes a K-nearest neighbor approach. Despite its efficiency given the database construction scheme, it assumes that signal sources are uniformly distributed, and is highly sensitive to the number of sources available and the quality of signal measurements.

Different from the above schemes, area classification in INOA is not restricted to floor/room localization scenarios. It is applicable to the classification of any area even without explicit partitions or a large difference in terms of elevation. INOA also proposes a preprocessing module that not only reduces the impact of missing signals in area classification and target localization, but also offline training and online querying time.

A preliminary version of this work has been reported in [6]. We advance it in the following major ways: 1) The previous work has not comprehensively studied inside/outside region decision. The problem is challenging, and we propose an algorithm under the paradigm of novel ensemble learning, which achieves an excellent performance; 2) The previous work has not considered the case of incomplete signals due to missing AP detection. We propose an algorithm which is more robust to signal loss; 3) We present effective fingerprint preprocessing to reduce the time (computation) and space (storage) complexity; 4) We conduct more extensive experimental studies to validate the performance of INOA.

3 Offline Fingerprint Preprocessing

We consider formulating locality classification into a standard classification problem in machine learning. Each AP is viewed as a feature, and the total number of APs determines the size of the feature (signal) space. Such a number may be very large in reality. For example, on our university campus, we have detected a total of $1,498$ APs in only two out of thirteen floors of a building. Having a huge feature space often leads to heavy computation and overfitting [38] in practice. Due to such high redundancy, it is intuitive to expect that not all of the APs should be kept for locality classification. To improve computational efficiency and accuracy for both offline training and online querying, the preprocessing module retains only those discriminating APs.

This section is organized as follows. We first introduce the preliminaries of this work in Section 3.1. Then, we discuss how to select the discriminating AP subset in Section 3.2 based on information theory, and how to cluster their salient features in Section 3.3. After that, the process to compress the RSSI vectors from the original signal space to the new feature space is presented in Section 3.4.

3.1 Preliminaries

The major symbols used in this paper are summarized in Table 1. The site survey may be conducted at predefined reference points (RPs). Each fingerprint is assigned with an area ID, which can be given by simple region characteristics, say the floor, zone or the building.

Let N be the number of RPs, and L be the total number of detectable APs. To mitigate the random effect, we collect multiple RSSI samples to form an RSSI vector at each RP. The mean RSSI $\bar{f}_l^{(n)}$ at RP n from AP l is given by

$$\bar{f}_l^{(n)} = \frac{1}{S_l^{(n)}} \sum_{s=1}^{S_l^{(n)}} f_l^{(n),s}, \tag{1}$$

where $f_l^{(n),s}$ is the s-th RSSI sample (in dBm) at RP n from AP l, and $S_l^{(n)}$ is the number of RSSI samples collected. The RSSI vector at RP n is defined as

$$\mathbf{F}^{(n)} = [\bar{f}_1^{(n)}, \bar{f}_2^{(n)}, \ldots, \bar{f}_L^{(n)}]. \tag{2}$$

By definition, if AP l is not detected at RP n (i.e., $S_l^{(n)} = 0$), the signal strength is stored as 0. The measured RSSI vector of the target is similarly denoted as $\mathbf{T}$:

$$\mathbf{T} = [\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_L], \tag{3}$$

where $\bar{t}_l$ is the mean RSSI from AP l.

Let $\boldsymbol{\mathcal{A}}$ be the set of area IDs to be classified (also known as classes in machine learning). For each $\mathbf{F}^{(n)}$, we have the corresponding area ID $r^{(n)} \in \boldsymbol{\mathcal{A}}$. Similarly, for each target RSSI vector $\mathbf{T}$ measured inside the fingerprint region, we denote its area ID as $y \in \boldsymbol{\mathcal{A}}$, which is to be identified in our area classification study.

TABLE 1: Major symbols used in INOA.

Notation	Definition
N	Number of RPs in fingerprint database
L	Number of APs
$\bar{f}_l^{(n)}$	Mean RSSI at RP n from AP l (dBm)
$f_l^{(n),s}$	s-th RSSI sample at RP n from AP l (dBm)
$S_l^{(n)}$	Number of RSSI samples collected at RP n from AP l
$\mathbf{F}^{(n)}$	RSSI vector received at RP n
$\mathbf{T}$	RSSI vector received at target
$\boldsymbol{\mathcal{A}}$	Set of area IDs
μ	Percentage of APs extract in AP subset selection
$\mathcal{C}_k$	k-th AP cluster
ϵ	Threshold of minimum intra-cluster correlation
$\boldsymbol{\mathcal{C}}$	Set of AP clusters obtained by feature clustering
$\hat{\mathbf{F}}^{(n)}$	Preprocessed RSSI vector at RP n
$\hat{f}_k^{(n)}$	Mean RSSI from APs in the k-th cluster at RP n (dBm)
$\hat{\mathbf{T}}$	Preprocessed RSSI vector at target
Γ_{ij}	Base learner for clusters $\mathcal{C}_i$ and $\mathcal{C}_j$
ν	Upper bound of error fraction in training dataset
$\hat{\mathbf{T}}_{ij}$	Tuple of RSSIs from clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ at target
$\boldsymbol{\Gamma}_{\hat{\mathbf{T}}}$	Subset of base learners involved in online query
κ	Threshold of minimum target vector length
β	Threshold of minimum area classification probability
φ	Threshold of area classification probability difference
Y	Optimal size of the new space in PCADD

3.2 AP Subset Selection

In locality classification, we prefer APs whose signals are the most useful for discriminating between locations. To cherry-pick the APs with such a property, we exploit an information theory approach [39] in the first preprocessing strategy, namely *AP subset selection*, which retains only the important APs with high information gain.

The information gain of AP ℓ, denoted by IG_ℓ, measures its discriminative power across the entire fingerprint region and is given by

$$IG_\ell = H(RP) - H(RP|AP_\ell), \tag{4}$$

where $H(RP) = -\sum_{n=1}^{N} P(RP_n) \log P(RP_n)$ is the entropy of the RPs when the RSSI from AP ℓ is unknown. $H(RP|AP_\ell) = -\sum_v \sum_{n=1}^{N} P(RP_n, AP_\ell =$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

$v) \log P(RP_n|AP_\ell = v)$ computes the conditional entropy of RPs given AP ℓ's RSSI values, and v is one possible value of RSSI from AP ℓ. Here, we denote RP_n and $AP_\ell = v$ as the events of locating at RP n and receiving an RSSI value v from AP ℓ respectively. By assuming a user can be equally likely to be at any location (i.e., $P(RP_n)$ is a constant for all n), we can further simplify Equation (4) and use only the negative of the conditional entropy as the discriminative power.

In the AP subset selection, we compute the discriminative power for each AP, and the top μ% of APs with the highest value are selected. Our experimental results show that this strategy improves not only the computational efficiency, but also the classification accuracy, since the signals considered in the following modules originate from the APs which are the best at providing spatial discrimination.

3.3 Feature Clustering

After AP subset selection, clustering is performed to further reduce dimensionality by combining highly correlated features. Figures 2a and 2b visualize the signal heatmaps of two APs on the campus, while Figure 2c illustrates the relationships of their signals by a scatter plot. These plots show that some APs resemble each other, implying redundancy within the feature space. One potential reason is that some signal sources are co-located within a small region, and signals broadcast from them should be highly correlated. In our experiments where signals from public APs are utilized, we observe some APs (e.g., the two in Figure 2) are almost identical. This is because setting multiple virtual APs in the same physical device (AP) is a common practice. Their signals originate from one exact location and therefore are roughly the same.

Duplicated information does not provide extra spatial discrimination, and introduces unnecessary computation. To eliminate redundant computation and reduce model complexity, we propose a *clustering as preprocessing* technique to transform groups of highly correlated APs into new features. In other words, each new feature is defined as a group of correlated APs. Specifically, we assign AP l to cluster $\mathcal{C}_k$ if the intra-cluster correlation satisfies

$$\frac{1}{|\mathcal{C}_k|}\sum_{j\in\mathcal{C}_k}\frac{\sum_n(\bar{f}_l^{(n)}-\bar{\mathcal{H}}_l)(\bar{f}_j^{(n)}-\bar{\mathcal{H}}_j)}{\sqrt{\sum_n(\bar{f}_l^{(n)}-\bar{\mathcal{H}}_l)^2}\sqrt{\sum_n(\bar{f}_j^{(n)}-\bar{\mathcal{H}}_j)^2}} \geq \epsilon, \quad (5)$$

where $\bar{\mathcal{H}}_l = \frac{1}{N}\sum_{n=1}^{N}\bar{f}_l^{(n)}$ and ϵ is the minimum required correlation where an AP belongs to a particular cluster. All clusters $\boldsymbol{\mathcal{C}}$ subsequently constitute a new feature space, which is concise and comprehensive.

3.4 Fingerprint Compression

After obtaining clusters of APs $\boldsymbol{\mathcal{C}}$, the original raw fingerprints should be projected into the new space of fewer features so as to reduce model complexity. For each RP fingerprint, we take the average RSSI from the APs in the same cluster to be the feature value, and hence the fingerprint $\mathbf{F}^{(n)}$ is transformed into $\hat{\mathbf{F}}^{(n)} = [\hat{f}_1^{(n)}, \hat{f}_2^{(n)}, \ldots, \hat{f}_{|\boldsymbol{\mathcal{C}}|}^{(n)}]$ where

$$\hat{f}_l^{(n)} = \frac{1}{\sum_{j\in\mathcal{C}_l} S_j^{(n)}}\sum_{j\in\mathcal{C}_l}\sum_{s=1}^{S_j^{(n)}} f_j^{(n),s}. \quad (6)$$

Similarly, the target RSSI vector should also be converted to $\hat{\mathbf{T}}$. After fingerprint compression, the physical meaning of a new "feature" becomes the average signal strength of a group of highly correlated APs. As our experimental trials show that $|\boldsymbol{\mathcal{C}}| \ll L$, the new space is much more compact and representative of the signal patterns inside the region.

As indicated in Equation (6), a feature value in the new space comprises one or more AP. As long as any AP in the same cluster is detected, the corresponding signal value can be derived, making the feature much less likely to be null. This reduces the chance of experiencing missing signals which arise from measurement randomness or obstruction. As a result, its impact on area classification and fine-grained target localization can be reduced by our preprocessing module. We further show its benefit in the experimental evaluation. In the following discussion, we assume the RSSI vectors are preprocessed unless otherwise stated, and the terms cluster and feature are used interchangeably.

4 Online Locality Classification

Beyond the fingerprint preprocessing, in this section we present how to classify the locality of the target given the signals collected. We first verify that they are being measured inside the fingerprint region by using the inside/outside region decision algorithm (Section 4.1). Then, area classification (Section 4.2) is conducted to determine in which area the target is.

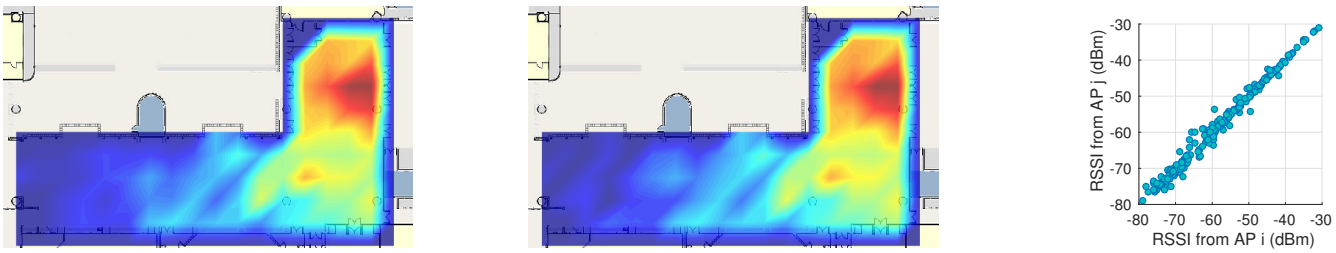
4.1 Inside/Outside Region Decision

To support the in-out decision, we formulate it into a one-class classification or the so-called data description problem, as we are usually not given fingerprints outside the region due to large survey costs. The goal of one-class classification is to distinguish between a set of known objects and all other possible ones. In our context, given a set of inside-region fingerprints, we find a data description to model them. If the target RSSI vector resembles the data description, it is likely to belong to the inside-region, and we can conduct area classification to localize the target. Otherwise, it is likely to be an outlier or novelty sampled outside the region.

In contrast to traditional approaches using a single data description to model the entire feature space which is vulnerable to missing values (signals), INOA employs a set of base learners under the ensemble learning paradigm. Each base learner is a data description which captures the patterns of a feature pair. We devise a hierarchical combination scheme to determine the novelty of the target RSSI vector given the classification results of the base learners. The rationale is that, if the target RSSI vector is sampled inside the region, the non-missing signal pairs should resemble the patterns captured by the corresponding base learners. Breaking a single data description problem into multiple ones ensures that all base learners contributing to the final in-out decision are given complete inputs. Otherwise, they will not be invoked.

Specifically, each cluster pair $\mathcal{C}_i$ and $\mathcal{C}_j$ is associated with a base learner Γ_{ij}, which distinguishes their patterns sampled outside the region from those inside. Each base learner

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing



(a) Signal heatmap for AP i (F4-0F-1B-93-A4-F3) (b) Signal heatmap for AP j (F4-0F-1B-93-A4-F5) (c) RSSI plot for APs i and j

Fig. 2: Visualization of two highly correlated APs discovered on the campus.

Γ_{ij} is formulated as a ν-support vector data description (ν-SVDD) [40] with the learning problem:

$$\begin{aligned} \underset{R_{ij}}{\text{minimize}} \quad & R_{ij}^2 + \frac{1}{\nu N}\sum_{n=1}^{N}\xi^{(n)}, \\ \text{subject to} \quad & ||\Phi(\hat{\mathbf{F}}_{ij}^{(n)}) - \mathbf{a}_{ij}||^2 \le R_{ij}^2 + \xi^{(n)}, \ \ \forall n, \end{aligned} \tag{7}$$

where $\hat{\mathbf{F}}_{ij}^{(n)} = (\hat{f}_i^{(n)}, \hat{f}_j^{(n)})$, $\xi^{(n)}$ is the non-negative slack variable for the n-th fingerprint; $\Phi(\cdot)$ is a feature map which can be computed by the RBF kernel; while $\mathbf{a}_{ij}$ and R_{ij} are the center and the radius of the hypersphere respectively. ν-SVDD finds, in principle, a small hypersphere that encloses as many training samples as possible. As illustrated in Figure 3, the base learner Γ_{ij} is a highly interpretable 2-D problem capturing all valid patterns of $\mathcal{C}_i$ and $\mathcal{C}_j$. Target signals $\hat{\mathbf{T}}_{ij} = (\hat{t}_i, \hat{t}_j)$ located outside the decision boundary (i.e., the solid line) will be regarded as a novelty. The upper bound of the error fraction ν for all the base learners helps to exclude some noisy data in the training process. In this way, we may get a more robust classifier.

After solving the above optimization problem with the Lagrange technique, we obtain the model parameter $\boldsymbol{\theta}_{ij}$ which should be stored in the database for online queries. Then, the base learner can be represented as

$$\Gamma_{ij}(\hat{\mathbf{T}}_{ij}) = \begin{cases} 1, & \text{if } g(\hat{\mathbf{T}}_{ij}|\boldsymbol{\theta}_{ij}) < 0, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $g(\cdot|\boldsymbol{\theta}_{ij}) = 0$ is the decision boundary for Γ_{ij}, and the base learner set

$$\boldsymbol{\Gamma} = \{\Gamma_{ij} \mid i, j \in \{1, 2, \ldots, |\boldsymbol{\mathcal{C}}|\}, i \ge j\} \tag{9}$$

will be used throughout the online classification.

Given a target RSSI vector, only a subset of base learners, $\boldsymbol{\Gamma}_{\hat{\mathbf{T}}} = \{\Gamma_{ij} \mid \hat{t}_i, \hat{t}_j \neq 0\}$, participate in predicting the novelty of $\hat{\mathbf{T}}$. An illustrative example of an online query is presented in Figure 4. We design a two-layer combination scheme. After getting a prediction from the base learners, we estimate the novelty of each cluster $\mathcal{C}_i$. This can be done by using the predictions associated with cluster $\mathcal{C}_i$ and all clusters $\mathcal{C}_j$ received in $\hat{\mathbf{T}}$. Instead of using simple and uniform majority voting, we use a *weighted* voting scheme which prefers prediction from clusters with a strong RSSI in $\hat{\mathbf{T}}$. The intuition is that a strong intensity in RSSI is more likely to be observed at locations around the APs than those far away [41], showing a potentially high confidence in pinpointing the locations. Specifically, in the first layer, to calculate the novelty of cluster $\mathcal{C}_i$, we assign weight w_j to the base learner Γ_{ij}. In the second layer, we determine the novelty score ς of the target vector by performing a similar weighted voting process on the results from the first layer. We assign a higher weight to the intermediate voting result (say, cluster $\mathcal{C}_i$) from the first layer with a stronger RSSI $\hat{t}_i$.

The novelty score ς given the preprocessed target RSSI vector $\hat{\mathbf{T}}$ is formulated as

$$\varsigma = \frac{1}{\sum_{i=1}^{|\boldsymbol{\mathcal{C}}|} w_i} \sum_{i=1}^{|\boldsymbol{\mathcal{C}}|} w_i \left[\frac{1}{\sum_{j=1}^{|\boldsymbol{\mathcal{C}}|} w_j} \sum_{j=1}^{|\boldsymbol{\mathcal{C}}|} w_j \Gamma_{ij}(\hat{\mathbf{T}}_{ij}) \right], \tag{10}$$

where w_i is the weight assigned based on the RSSI $\hat{t}_i$ from cluster $\mathcal{C}_i$ and is given by

$$w_i = \begin{cases} (\frac{1}{\hat{t}_i})^2, & \text{if } \hat{t}_i \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Then, the inside/outside region decision algorithm concludes that the target is inside the fingerprint region if $\varsigma \le 0.5$ (i.e., the majority are not novelties).

4.2 Area Classification

After verifying that the target RSSI vector is measured inside the fingerprint region, we can proceed and determine in which area the target is. In one-class classification, the model (e.g., the solid line in Figure 3) sometimes suffers from over-fitting or under-fitting since our training dataset is imbalanced (e.g., only inside-region fingerprints are provided in the in-out problem), and hence we have a lack of knowledge to get the optimal decision boundary separating two classes (i.e., regulars and outliers). Different from the inside/outside region decision, the existing fingerprint systems naturally provide fingerprints in different areas. Hence, instead of using only fingerprints in an area to train a single data description for each of them, we can leverage all fingerprints in the entire site to learn a more robust model by using binary-class classification techniques.

We formulate the problem into a well-known "one-against-all" form. For each area, we find a probabilistic support vector machine (SVM) model [42] to distinguish its unique signal patterns from the others, which are together considered to be a single class [43]. Specifically, the SVM model for, say, area i is to compute the probability $p_i(x) = \sigma(w_i^T x + b_i)$ given an RSSI vector x where $\sigma(\cdot)$ is the sigmoid function. The parameters can be trained by the optimization problem:

$$\begin{aligned} \underset{w_i}{\text{minimize}} \quad & \frac{1}{2}||w_i||^2 + C\sum_{n=1}^{N}\xi_i^{(n)}, \\ \text{subject to} \quad & \tilde{r}^{(n)}(w_i^T\hat{\mathbf{F}}^{(n)} + b_i) \ge 1 - \xi_i^{(n)}, \forall n, \end{aligned} \tag{12}$$

where C is a regularization parameter, $\xi_i^{(n)}$ is the non-negative slack variable for the n-th fingerprint at area i, and $\tilde{r}^{(n)} = 1$ if $r^{(n)} = i$ and $\tilde{r}^{(n)} = -1$ otherwise. In total, $|\boldsymbol{\mathcal{A}}|$ SVM models are trained for the site with $|\boldsymbol{\mathcal{A}}|$ areas and their parameters (i.e., $\{(w_i, b_i)\}_{i=1}^{|\boldsymbol{\mathcal{A}}|}$) are stored in the database.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

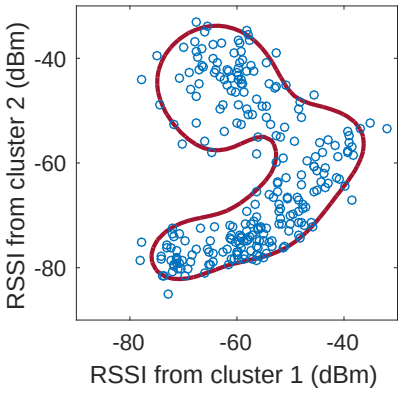


Fig. 3: Illustration of a base learner $\Gamma_{1,2}$.

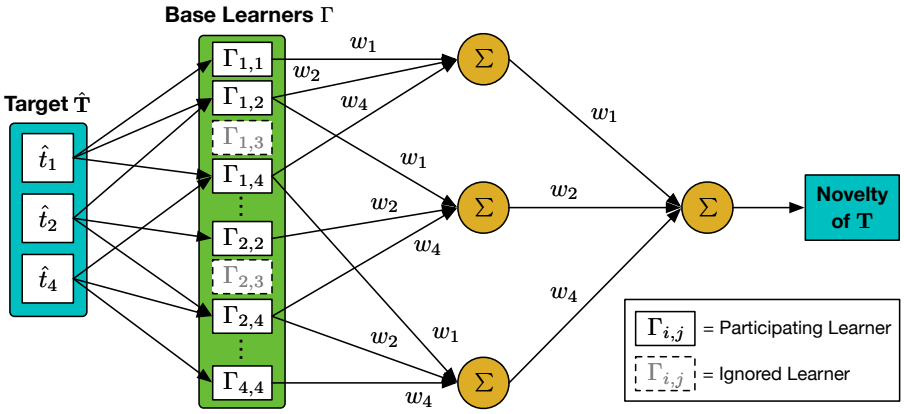


Fig. 4: Simple example of inside/outside region decision with 3 received RSSIs.

During an online query, each SVM model finds the probability that the target RSSI vector is in this area against the others. Given $|\mathcal{A}|$ probabilities, we then find the area with maximum probability and return its ID y to the user:

$$y = \underset{i=1,\ldots,|\mathcal{A}|}{\text{argmax}}\, p_i(\hat{\mathbf{T}}) = \underset{i=1,\ldots,|\mathcal{A}|}{\text{argmax}}\, \sigma(w_i^T \hat{\mathbf{T}} + b_i). \quad (13)$$

Besides that, the probabilities can be further used for signal rejection or probabilistic localization.

5 Signal Rejection and Device Calibration

To further improve accuracy, in this section we discuss how to reject unclassifiable signals (Section 5.1) and calibrate heterogeneous devices (Section 5.2).

5.1 Signal Rejection

As noted before, some target RSSI vectors may not be classifiable (due to high measurement noise, significant missing signals or target at area junction). We hence design a signal rejection scheme to reject the vector which does not lead to good classification results with sufficiently high confidence.

Specifically, we first define $\mathbf{T}$ is classifiable only if

$$||\mathbf{T}||_0 \geq \kappa \cdot \min(||\mathbf{F}^{(1)}||_0, \ldots, ||\mathbf{F}^{(N)}||_0), \quad (14)$$

where $||\mathbf{x}||_0$ is the l_0-norm counting the total number of non-zero elements in $\mathbf{x}$, and κ is the predefined threshold parameter. Target RSSI vectors violating the above fingerprint length requirement, are likely to be collected far from the fingerprint region. INOA labels them as unclassifiable *outliers* and rejects them without further classification.

In addition, due to measurement noise, RSSI vectors may be similar across different areas which would lead to incorrect results. Through our deployment experience, we have observed that the decision probabilities of the area classification algorithm may be similar in different areas. If the decision is highly uncertain (unclassifiable), rejecting this signal can prevent misclassification and improve the user experience. During area classification, we reject $\hat{\mathbf{T}}$ if all of the probabilities $\{p_a(\hat{\mathbf{T}}) \mid a \in \mathcal{A}\}$ calculated are less than a predefined threshold, or the difference between the largest $(p(\hat{\mathbf{T}})^{(1)})$ and the second largest $(p(\hat{\mathbf{T}})^{(2)})$ probability is less than a certain value. Specifically, we reject the target RSSI vector $\hat{\mathbf{T}}$ if

$$\forall a, p_a(\hat{\mathbf{T}}) \leq \beta, \text{or } p(\hat{\mathbf{T}})^{(1)} - p(\hat{\mathbf{T}})^{(2)} \leq \varphi p(\hat{\mathbf{T}})^{(1)}. \quad (15)$$

The rejection parameters β and φ can be determined via empirical studies over offline-collected data.

For those ILBSs capable of identifying the target (e.g., client-based navigation), we can leverage the historical localities to predict the label instead of discarding the unclassifiable target vector directly. In the experimental evaluation, we demonstrate the effectiveness of a history-based INOA using a simple sliding window.

5.2 Device RSSI Calibration

For the same RSSI, different smartphones may have different measurement values due to their Wi-Fi network interface differences [44]. For each target RSSI $\bar{t}_l$ from AP l, a linear shift d from the true RSSI $\tilde{t}_l$ (i.e., the RSSI captured by the RP fingerprints) using a different device [45] has been reported, i.e., $\bar{t}_l = \tilde{t}_l + d$. We consider a scalable online calibration in order to reduce offline manual effort. We first calculate the similarity between the target RSSI vector and each RP fingerprint. After that, the RPs with similar signal vectors can be leveraged for online signal calibration in order to get d. RSSI vector comparison uses cosine similarity [46], denoted as $cos(\mathbf{T}, \mathbf{F}^{(n)})$, between $\mathbf{T}$ and $\mathbf{F}^{(n)}$. The cosine similarity compares the relative signal trends of different APs rather than the absolute RSSI values. For each $\bar{t}_l \in \mathbf{T}$, we find the corresponding $\bar{f}_l^{(n)}$'s at RPs from AP l. Given pairs of $[\bar{t}_l, \bar{f}_l^{(n)}]$, we conduct the linear regression [47] and obtain the corresponding offset d for target RSSI $\bar{t}_l$. To mitigate the effect of random noise, we find the top several RPs with $cos(\mathbf{T}, \mathbf{F}^{(n)}) > \zeta$ (say, $\zeta = 0.95$ in our experiment) for linear RSSI calibration [48].

The device calibration can be conducted in a crowdsourced manner. We can leverage the ILBS user data for calibration and store those d's for different phone models. At the beginning, given the MACs and Wi-Fi interface vendors, some smartphones are calibrated online and their RSSI offsets are stored in the database. The same smartphone models of later users can then benefit from those crowdsourced parameters.

6 Experimental Evaluation and Illustrative Results

We have conducted extensive experimental trials in several spacious sites to evaluate INOA. In this section, we first present the experimental settings (Section 6.1), followed by the evaluation of the inside/outside region decision

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

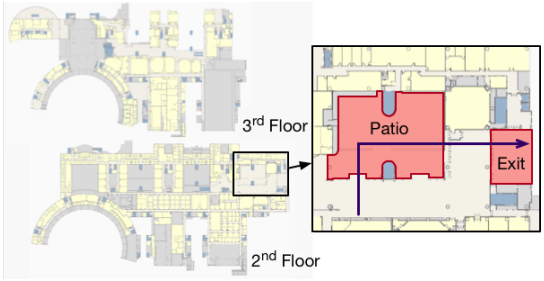


Fig. 5: Maps of HKUST - The campus and a validation trajectory.

(Section 6.2) and area classification (Section 6.3). Then, we analyze the complexity of different schemes (Section 6.4) and discuss the deployment of INOA (Section 6.5).

6.1 Setting and Metrics

We conduct trials in the following venues, spanning a wide range of characteristics:

- HKUST: The university campus (Figure 5) where we collect $3,874$ fingerprints (3m grid size) from the public areas of two floors (more than a $4,000\text{m}^2$ area). A patio (outdoor) and an exit are located on the 2nd floor. This serves as a case study where the site consists of both indoor and outdoor environments. Since outdoor fingerprints are usually unavailable due to survey cost limitations, we have to first ensure the target is indoors by making the in-out decision. Then, area classification is conducted to deduce which floor the target is on.
- Harbor City (HKHC): The premium shopping mall (Figure 6) where we collect $8,044$ fingerprints (3m grid size) from three floors (more than $10,000\text{m}^2$ area). This site mainly consists of corridors, and we consider that fingerprints are unavailable in one of the premium stores on the 2nd floor. If a target is estimated to be outside of the store, we then determine which floor the target is on by the area classification.
- Cyberport (HKCP): The premium business building (Figure 7) where 826 fingerprints (4m grid size) are collected from two floors (more than $50,000\text{m}^2$ area). This site has a multi-storey lobby, which is a popular design in many modern buildings and makes locality classification and target localization difficult. We consider that the ILBS does not cover the 4th floor and therefore the fingerprint regions are on the 2nd and the 3rd floors. If the target is estimated to be outside of the 4th floor where ILBS is not provided, we then conduct the area classification to determine the floor.

The above venues are representative of commonly visited ILBS sites (i.e., campuses, shopping malls and business buildings) with various typical structures (i.e., a mixture of indoors and outdoors, corridors and often a multi-storey lobby). During the site survey, we utilize different smartphones, including HTC One X, Coolpad F1, Lenovo A680, Samsung S3 and Mi for RP fingerprint and target data collection. At each RP, we collect fingerprints from four different directions (north, south, west and east) to reduce the effect of human bodies on signals, with a scanning duration of 15 seconds to ensure all receivable signals are collected. Before the site survey, we have no knowledge of the AP locations

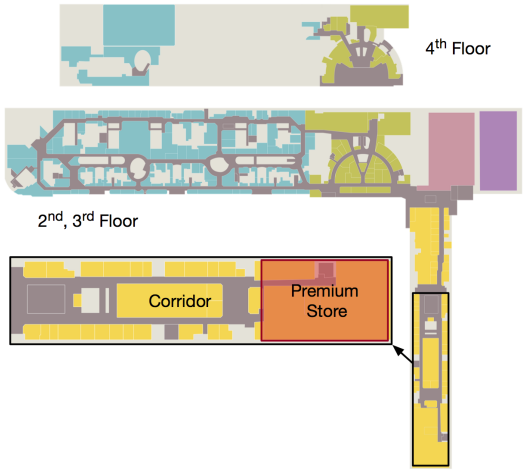


Fig. 6: Maps of HKHC - Mall.

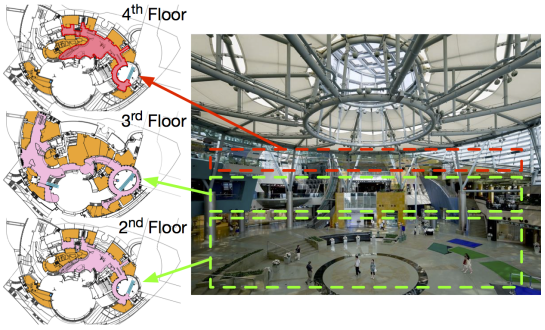


Fig. 7: Maps and photos of Hong Kong Cyperport - Business.

and their exact number. Wi-Fi fingerprint and target data collection is conducted during working hours and therefore we do not exclude the cases when there are people (crowds) nearby.

We evaluate the inside/outside region decision and area classification based on the following performance metrics:

- *True positive rate (TPR)*: evaluates the ability to correctly identify the target inside the fingerprint region. TPR is given by $TPR = TP/(TP + FN)$. TP denotes the number of true positives while FN denotes the number of false negatives.
- *True negative rate (TNR)*: evaluates the ability to reject the target outside the fingerprint region. TNR is given by $TNR = TN/(TN + FP)$. TN denotes the number of true negatives while FP denotes the number of false positives.
- *True rate (TR)*: gives a performance measure to overview the robustness of the in-out decision, and is given by $TR = (TPR + TNR)/2$. An unbiased one-class classifier should have high TPR and TNR such that TR ideally approaches 1. $TR = 0.5$ means that the biased classifier always chooses either a regular or an outlier to be the prediction.
- *Classification accuracy*: is the number of correct area classifications over that of all the classifiable samples. It characterizes the robustness under different building structures and signals noise levels.
- *Online querying time*: calculates the mean prediction time of each target. The less time the calculation takes, the less energy the mobile device consumes and the shorter time the users have to wait.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

We utilize *TPR*, *TNR* and *TR* for the in-out decision where "positive" corresponds to inside signals, while "negative" corresponds to signals outside the region. To evaluate the in-out decision algorithm fairly, we collect the same number of targets inside and outside the region to form the test dataset. 120 targets are sampled on the campus, while 112 and 88 targets are obtained in the mall and the business building. Similarly, we also collect the same number of targets in each area to test the performance of area classification. For each area inside the campus, the mall and the business building, 59, 89 and 50 targets are sampled respectively. All targets in the test dataset are for performance evaluation only and not used to conduct parameter tuning.

We evaluate the preprocessing scheme by comparing it with the Fisher criterion- (FC) [49] and online optimization-based (OOPT) [50] approaches. For the in-out decision, we compare it with the support vector data description (SVDD) [5] and the PCA data description (PCADD) [6], which perform comparatively well in our previous study. Also, (dis)similarity-based data description (SBDD) [23], which aims at classifying instances with missing data, is included. For area classification, we compare it with artificial neural network (ANN) [6], signal heuristic classification (SHC) [9], nearest neighbor (NN) [36] and deep belief network (DBN) [51] to evaluate our area localization ability in different sites after filtering away targets outside the fingerprint region.

We have empirically studied the parameters for INOA, and the detailed settings are summarized in Table 2. They are used in all sites without being individually fine-tuned. In the in-out decision, at most 10% of the training data can be noisy ($\nu = 0.1$), and the selected RBF kernel width ($\gamma = 0.005$) makes the model flexible enough for different sites. For the comparison schemes, although we assume signals outside the region are unavailable, we still collect them to find the optimal parameters using grid search for each site. In SVDD, the upper bound of the error fraction and the choice of kernel are selected. We have to select different parameters for different sites as they are highly sensitive to the problem dimensionality (i.e., AP number). In PCADD, the optimal size Y of the new space is first selected, followed by finding the optimal threshold of the reconstruction error. The magnitude of the reconstruction error highly depends on Y, which is site-dependent. In SBDD, the similarity measure is chosen. The parameters and the characteristics of SBDD are similar to those of SVDD as SBDD trains an SVDD model after transforming the RSSI vectors into the similarity space. All RP fingerprints are used to train the in-out decision models. In area classification, we conduct 5-fold cross validation to train the optimal model parameters. We construct a typical four-layer ANN. The first hidden layer has 20 neurons, while the second has 3. In DBN, the number of training epochs is set to $1,000$, with 3 hidden layers. The sigmoid function is applied in its neural network classification.

6.2 Inside/Outside Region Decision

In a practical deployment of fingerprint-based ILBS, the number of detected APs in the targets is usually less than those in RP fingerprints, which is shown in Figure 8. A smartphone needs to provide an instant scanning result of the detected APs, and hence the scanning duration for a target RSSI vector is limited. Figure 9 further shows the fraction of an RSSI vector collected with different scanning durations. Based on our deployment experience, we can afford at most one-second scanning duration to ensure that the ILBS is real-time and user-friendly. As a result, only about 60% of the APs are detected and form the target RSSI vector. In practice, many more signals are missing due to other reasons such as infrastructure change or signal obstruction. This accounts for the observation that the maximum number of detected APs in the targets (around 59 APs) is less than that in the RP fingerprints (around 81 APs) in Figure 8. Based on the above, by default (baseline scenario) we use only half of the complete target RSSI vector (i.e., 50% RSSIs are randomly removed) in the following evaluation.

To evaluate the classification robustness in the presence of missing signals, we randomly remove the RSSIs from the target vectors and use the model which is trained based upon complete (i.e., originally collected) RP fingerprints for online locality queries. For each target, we repeat RSSI removal and locality classification 100 times to simulate the randomness of missing signals and get the average performance. The *TR* using different preprocessing approaches for the proposed in-out decision algorithm is illustrated in Figure 10 where R-INOA represents the proposed scheme without performing fingerprint preprocessing. The smaller the fraction of the target RSSI vector required for a correct decision, the more robust the classifier. In general, INOA achieves higher *TR* than FC and OOPT. We find that FC cannot effectively classify targets outside the region since evaluating the Fisher criterion of APs over anchor points in the region of interest may not be reliable in this case. Although OOPT achieves better *TR* than R-INOA, it has a heavy computational burden during online queries. On the contrary, INOA requires only offline preprocessing and significantly reduces the computational effort. Based on the

TABLE 2: Baseline parameters of INOA used in all trials. [9]

Phases	Parameter	Empirical Value
Preprocess (Section 3)	μ	80
	ϵ	0.85
Base Learner (Section 4)	ν	0.1
	γ	0.005
Rejection (Section 5)	β	0.6
	φ	0.3
	κ	0.3

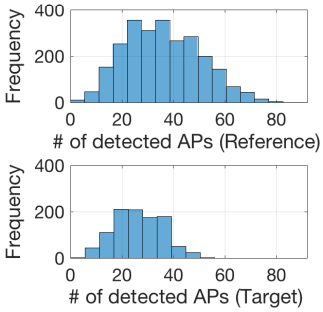


Fig. 8: Histogram of detected AP number at RPs and targets (mall).

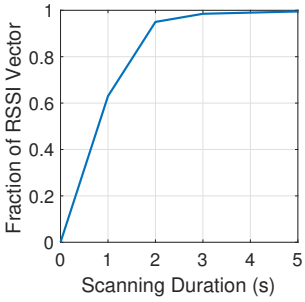


Fig. 9: Fraction of an RSSI vector detected versus scanning duration (mall).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

above, we only consider INOA with preprocessing in the following evaluation.

Figure 11 shows the *TR* of different algorithms in all sites. The *TR* of all algorithms generally increases with the fraction of the target vector. When the target vector is complete, it conforms more to the patterns captured by the model. Hence, the performance is usually better if the scanning duration is longer, resulting in a more complete target vector. Specifically, INOA converges at a high *TR* much faster than the others and it only needs a surprisingly small portion (0.3 to 0.4) of the complete target vector to get a high *TR* (over 0.85). This explains why we define $\kappa = 0.3$ in the signal rejection module. The main reason, which has led to such a significant improvement, is that we classify the target vector by examining the novelty of the non-missing feature pair individually. This strategy ensures that our algorithm is remarkably robust against missing signals and the curse of dimensionality. Moreover, the preprocessing module can effectively filter out non-discriminative APs and cluster salient features, leading to a performance boost. The experimental results validate the inherent redundancy where we only need to use a small portion of an RSSI vector to make our decision converge. Extra RSSIs do not actually alter the final voting result.

To the contrary, other classification schemes are not robust enough against signal missing, especially on the campus and in the mall. They require a sufficient amount of RSSIs in order to achieve a high *TR* though most of them can handle conditions when a few of the signals are missing. In SVDD, the location of the target vector in the feature space changes significantly when some RSSIs are missing. Even if the RSSI vector is collected inside the fingerprint region, the location estimation may fall outside the decision boundary since the model expects a complete target vector. In SBDD, although signal transformation into similarity space can reduce the impact of missing signals, the similarity score changes significantly when a large number of signals are missing. Moreover, in an open space without an explicit wall partition, the signal patterns become similar. Regulars transformed into similarity space may be similar to those collected outside the region. This accounts for the unsatisfactory performance of SBDD in the mall where no explicit partition has been provided between the corridor and the stores. In PCADD, which depends on a reconstruction error threshold, the missing signals make it inappropriate to differentiate outliers from regulars.

Experimental results also show that the accuracy improvement of INOA over the others is much more compelling in the sites where a large number of APs are discoverable (high dimensionality). In the campus ($1,498$ APs) and the mall ($3,083$ APs), INOA can accurately predict the novelty of target RSSI vectors with only a tiny portion (about 0.3) of the full vector, while other algorithms always bias towards one of the outcomes until more than half of the target vector is received. In order to differentiate the signals from one another, many feature values should be available, especially in high dimensional signal space. Otherwise, misclassification occurs easily due to the signal ambiguity.

Figure 12 shows the *TPR* and *TNR* of different in-out decision algorithms on the campus and in the mall given half target vectors. This shows that INOA is stable overall and obtains both high *TPR* and *TNR* in a practical scenario (i.e., lots of signals are missing). Other algorithms tend to bias towards one of the outcomes. As illustrated in the experiment, INOA maintains a similar performance until almost all RSSIs are missing (say, more than 70% in our case).

We also conduct an experiment on the campus to evaluate the online device calibration scheme. Figure 13 shows the *TR* of INOA using different devices before and after applying the calibration. The RP fingerprints used in this experiment are collected using Mi. We can observe a higher *TR* among different devices after the proposed RSSI calibration. This is mainly because the proposed scheme successfully adjusts the RSSIs among different devices, enhancing the scalability in practical deployment.

Providing a stable performance under different scenarios is important for practical deployment. Figure 14a shows the *TR* of different schemes with a different portion of deployed APs. We remove APs randomly in each site and repeat the experiment 100 times to get the average performance. INOA and other algorithms are shown to be stable with different portions of deployed APs, which is a good property for fast ILBS deployment. Figure 14b shows the *TR* versus the fingerprint survey grid size. The performance of INOA and SVDD is overall stable when the grid size is within a reasonable range, because they retrieve support vectors from the remaining signals after fingerprint reduction, which is less sensitive to the training sample number. When the survey grid size is too large, the RP fingerprints cannot capture all signal patterns inside the region and therefore all algorithms eventually approach the 0.5 *TR* (i.e., random guessing). Figure 14c shows the performance of INOA with different base learner size k. Only having 1-D in each base learner is not appropriate because the model ignores all relationships between APs. Although increasing the base learner size can theoretically encode the patterns from more APs, the results show that the improvement is not significant. To balance between classification accuracy and computation efficiency, we choose $k = 2$ in our settings.

6.3 Area Classification

To evaluate the area classification after fingerprint preprocessing, Figure 15a shows the accuracy of all area classification algorithms given different fractions of target RSSI vectors. SHC suffers from missing signals as it largely relies on a strong RSSI to differentiate the areas. The complex structure of the hidden layer in ANN successfully encodes the key features of the signal patterns. Missing some input in the hidden units does not alter the activated output of the hidden layer significantly. Multiple stacked restricted Boltzmann machines in DBN introduce a deep structure and extract good features from the fingerprints, making it less sensitive to signal loss. Through optimization, the support vectors in SVM can preserve the differentiation of areas after signal removal.

As discussed in the inside/outside region decision, the robustness of an algorithm is important. Figure 15b presents area classification accuracy versus survey grid size. All algorithms degrade in accuracy when the survey grid size

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

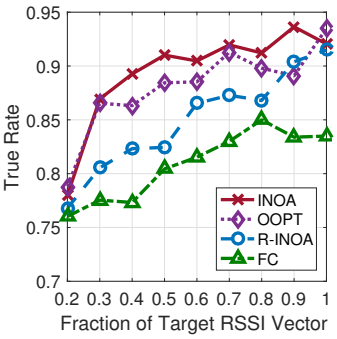
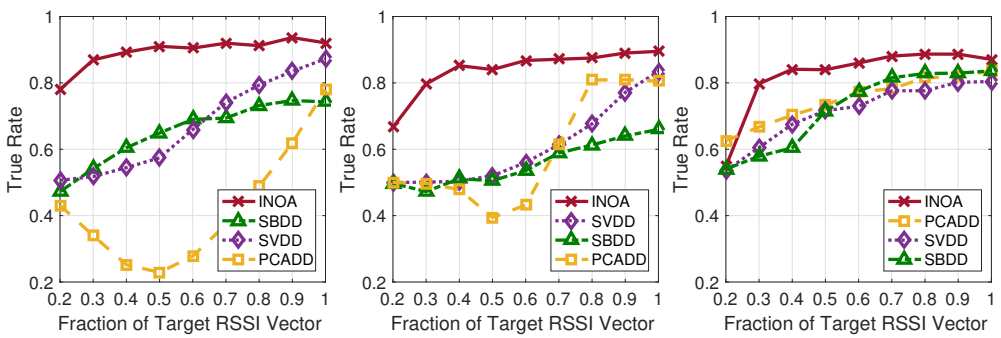
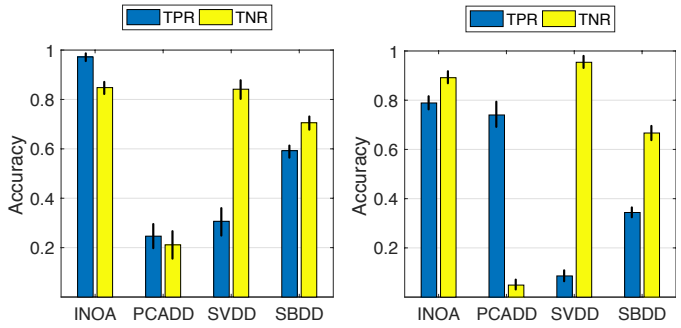


Fig. 10: *TR* of different preprocessing algorithms (campus).



(a) Campus (b) Mall (c) Business

Fig. 11: *TR* of inside/outside region decision algorithms given different fraction of target RSSI vector.



(a) Campus (b) Mall

Fig. 12: *TPR* and *TNR* of different in-out decision algorithms with standard deviation given half target RSSI vectors.

increases. The accuracy of SHC degrades under a large grid size, because some strong signal measurements are lost under RP removals. Overall, SVM performs with better accuracy than other algorithms. Similar to SVDD, SVM retrieves support vectors from the remaining signals. DBN decreases in classification accuracy as the trained belief network largely relies on sufficient signal patterns in order to encode the fingerprint map. Figure 15c shows the area classification accuracy of all algorithms given a different portion of deployed APs. Similar to the in-out decision, we remove APs randomly and repeat 100 times to get the average performance. The results indicate that the area classification algorithms generally do not deteriorate significantly. SHC is less accurate because it depends on finding the APs with strong RSSIs.

Figure 16 summarizes the area classification accuracies using different algorithms. We observe that they all achieve high accuracy on the campus, mainly because there are many wall partitions in the building to differentiate the signals, which matches the observations in [9], [52], [53]. However, some algorithms achieve lower accuracy in the mall and the business building because of noisy signals there. SVM achieves robust performance in all sites, as it finds the support vectors which can effectively differentiate the RSSI vectors in the sites.

6.4 Complexity Analysis

The offline training and online querying time complexities of different algorithms are summarized in Table 3. To be an efficient in-out decision algorithm, its online querying time complexity must be independent of the RP fingerprint number N and AP number L. Otherwise, the computation time increases with the scope of the ILBS. SBDD and PCADD depend on the above factors as the target vector should be compared with every training sample and is therefore inappropriate for making responsive decisions. SVDD is the most efficient because the time complexity is only proportional to the number of support vectors $\bar{Q}_{\text{SVDD}}$ in the model. However, due to its simplicity, the performance is unsatisfactory as discussed previously. For INOA, it depends on the number of APs detected by the target and the average number of support vectors $\bar{Q}_{\text{INOA}}$ in the base learners. In practice, $|\mathbf{T}|$ is always limited (the averages in our experiments are 27 on the campus, 26 in the mall and 47 in the business building). Also, we find that $\bar{Q}_{\text{INOA}}$ and $\bar{Q}_{\text{SVDD}}$ are often much less than L. Therefore, the in-out decision algorithm in INOA is highly effective, efficient and applicable for large-scale deployment.

TABLE 3: Offline training and online querying time complexities for in-out decision algorithms.

Algorithm	Offline Training	Online Querying
INOA	$\mathcal{O}(N^2L + N^2\lvert\mathcal{C}\rvert^2)$	$\mathcal{O}(\lvert\mathbf{T}\rvert + \lvert\mathbf{T}\rvert^2\bar{Q}_{\text{INOA}})$
SVDD	$\mathcal{O}(N^2L)$	$\mathcal{O}(\bar{Q}_{\text{SVDD}})$
SBDD	$\mathcal{O}(N^2L^2)$	$\mathcal{O}(NL^2)$
PCADD	$\mathcal{O}(Y^3)$	$\mathcal{O}(L + NY^2)$

To empirically study the efficiency of INOA, we first show in Figure 17 the running time reduction of locality classification before and after applying fingerprint preprocessing. The running time on the campus is reduced by half because AP subset selection and feature clustering significantly reduce the number of APs from $1,498$ to 846 (about 56%). In the business building, 72 out of 280 APs (about 26%) are removed by preprocessing. In the mall, the improvement is not as much as the others because only 652 out of $3,038$ APs (about 21%) are removed. Then, Figure 18 presents the online querying time of different algorithms on the campus. This site is challenging in terms of efficiency due to the large fingerprint database ($3,874$ RPs). The quantitative results validate that our scheme offers an accurate but rapid response to the users. For PCADD and SBDD, we can observe that the online querying time is unacceptably long. Computation time is expected to increase when more areas (e.g., classrooms) are added to the ILBS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2839112, IEEE Transactions on Mobile Computing

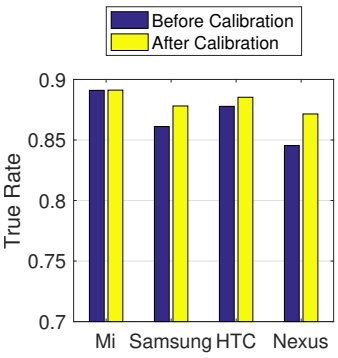
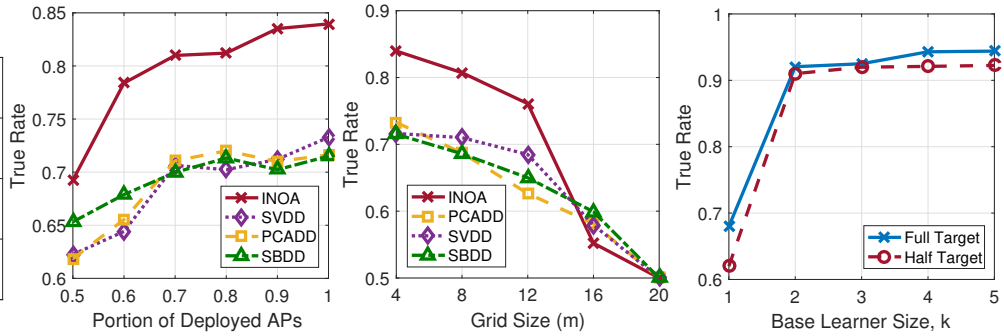
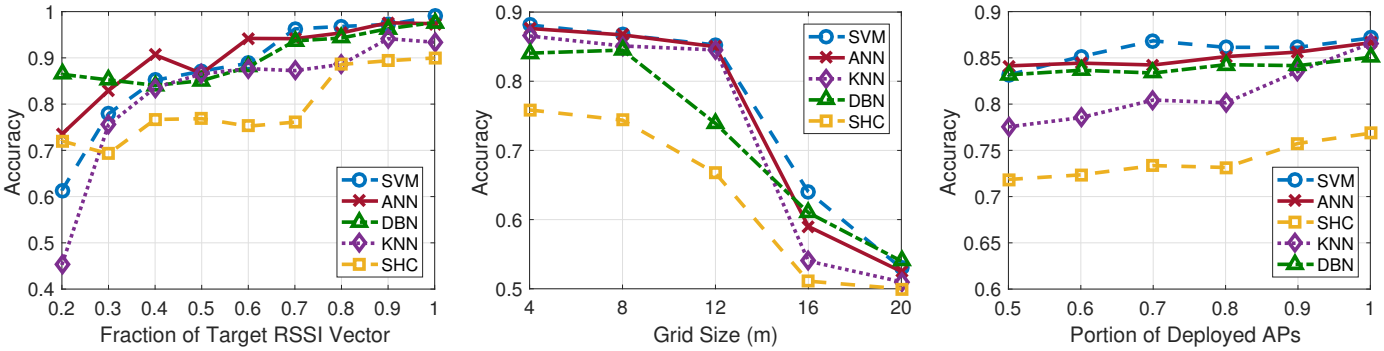


Fig. 13: *TR* of INOA using different devices (campus).



(a) Deployed APs (business) (b) Survey grid size (business) (c) Base learner size (campus)

Fig. 14: Performance (*TR*) comparison of different inside/outside region decision algorithms with different settings.



(a) Accuracy versus fraction of target RSSI vector (mall).

(b) Accuracy versus survey grid size (business).

(c) Accuracy versus portion of deployed APs (mall).

Fig. 15: Performance comparison of different area classification algorithms with different site settings.

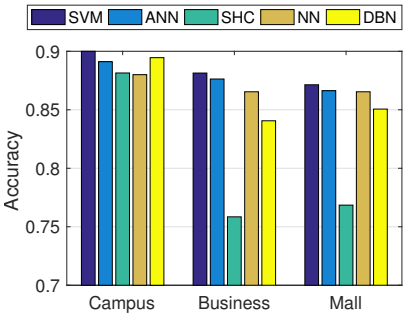


Fig. 16: Area classification accuracy of all algorithms in different sites.

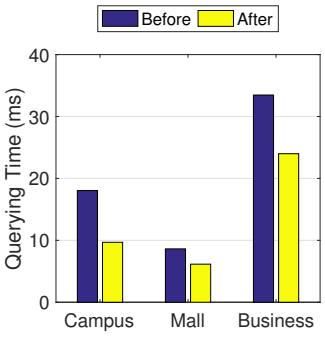


Fig. 17: Online querying time before and after pre-processing.

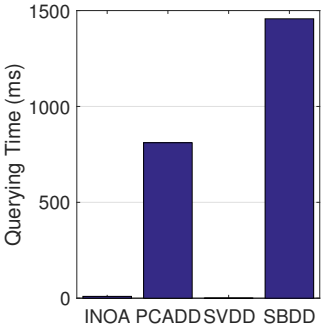


Fig. 18: Online querying time on the campus.

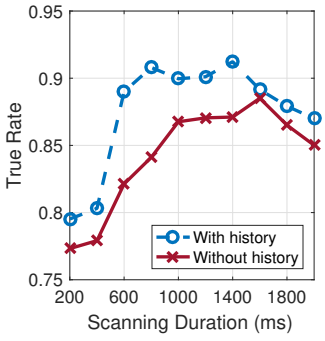


Fig. 19: *TR* of INOA with/without considering historical results (campus).

6.5 Deployment Discussion

Historical localities of each user, if available, can be used to further improve the performance and avoid signal rejection. We have implemented a sliding window of size 3 to smooth out the results by outputting the majority of the last three decisions. Figure 19 summarizes the *TR* of the in-out decision of the target vectors collected along the trajectory on the campus depicted in Figure 5 with different scanning durations. At first, the *TR* of the history-based INOA is significantly better than the one without considering historical localities. In addition, we observe that the *TR* starts declining when scanning duration takes longer than $1,600$ milliseconds. Given a long scanning duration, the target can move from one location to another, and the RSSI vector formed can therefore be ambiguous. For area classification, the improvement is similar, and we omit the results. Hence, if INOA is applied to the applications (e.g., client-based navigation) where we can memorize the localities of each user, we recommend using histories to provide a better user experience.

The current trend of indoor localization is to keep crowdsourcing the signals from the users walking freely inside the target site and improving the localization performance over time [54]. Crowdsourcing is beneficial to INOA as new RP fingerprints, which replace the deteriorated ones, can be used to update the base learners periodically. At the beginning (cold start) stage of INOA, the empirical parameters can be applied to train a coarse model. The base learners with APs received in the crowdsourced signals (both regulars and outliers) are enhanced by re-training them using the crowdsourced signals and RP fingerprints.

To extend INOA to new areas or regions, we can add

or re-train only those base learners with APs detected in the newly collected RP fingerprints. This avoids re-training the entire system when only a set of RP fingerprints is added, which is a practical and common scenario. Moreover, we highly recommend further reducing the computation through parallel programming because our design is extremely easy to parallelize. The predictions of base learners in the in-out decision and the predictions of SVM models in the area classification are all independent and can be delegated to different processors.

7 Conclusion

We design the locality classification for indoor fingerprint-based systems. Such a mechanism entails two queries, inside/outside region decision and area classification. Due to the measurement noise, some signals may not be received by the target, and hence an RSSI vector is always incomplete in terms of feature numbers compared with the RP fingerprint. However, traditional one-class classification algorithms for the in-out decision often suffer severely from missing values. To address this problem, we propose INOA, a highly efficient and scalable locality classification algorithm.

INOA is capable of handling missing or noisy signals and can be incrementally extended to new regions or areas. It may also serve as a plug-in to any indoor localization system for efficiency enhancement. INOA exploits the paradigm of ensemble learning and decomposes the in-out problem into multiple subproblems in the form of base learners. We utilize them to formulate an algorithm which significantly mitigates the impact arising from missing signals. We further present a novel fingerprint preprocessing module. It removes redundant information and combines salient features to speed up the computation. We have implemented INOA and conducted extensive experimental trials in several different sites. Compared with other existing approaches, INOA is shown to be efficient and remarkably robust against incomplete target RSSI vectors. Only a small portion (about 30%) of the RSSIs in the target vector are needed to offer a highly accurate prediction.

References

[1] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Proc. IEEE INFOCOM*, vol. 2, 2000, pp. 775–784.

[2] M. Youssef and A. Agrawala, "The Horus location determination system," *Wireless Netw.*, vol. 14, no. 3, pp. 357–374, 2008.

[3] C. Luo, L. Cheng, M. C. Chan, Y. Gu, J. Li, and Z. Ming, "Pallas: Self-bootstrapping fine-grained passive indoor localization using WiFi monitors," *IEEE Trans. Mobile Computing*, vol. 16, no. 2, pp. 466–481, Feb 2017.

[4] M. Tachikawa, T. Maekawa, and Y. Matsushita, "Predicting location semantics combining active and passive sensing with environment-independent classifier," in *Proc. ACM UbiComp*, 2016, pp. 220–231.

[5] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[6] S. He, J. Tan, and S.-H. G. Chan, "Towards area classification for large-scale fingerprint-based system," in *Proc. ACM UbiComp*, 2016, pp. 232–243.

[7] R. Faragher and R. Harle, "Location fingerprinting with Bluetooth low energy beacons," *IEEE JSAC*, vol. 33, no. 11, pp. 2418–2428, Nov 2015.

[8] K. Wu, J. Xiao, Y. Yi, M. Gao, and L. M. Ni, "FILA: Fine-grained indoor localization," in *Proc. IEEE INFOCOM*, March 2012, pp. 2210–2218.

[9] P. Bhargava, S. Krishnamoorthy, A. K. Nakshathri, M. Mah, and A. Agrawala, "Locus: An indoor localization, tracking and navigation system for multi-story buildings using heuristics derived from Wi-Fi signal strength," in *Proc. MobiQuitous*. Springer, 2012, pp. 212–223.

[10] S.-H. Fang and T. Lin, "Principal component localization in indoor wlan environments," *IEEE Trans. Mobile Comput.*, vol. 11, no. 1, pp. 100–110, 2012.

[11] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, "No need to war-drive: unsupervised indoor localization," in *Proc. ACM MobiSys*, 2012, pp. 197–210.

[12] K. Tabata, H. Konno, K. Tsung, W. Morioka, A. Nishino, M. Nakajima, and N. Kohtake, "The design of selective hybrid positioning by utilizing accuracy information for indoor-outdoor seamless positioning and verification in tokyo station," *Proc. IEEE IPIN*, 2015.

[13] M. Li, P. Zhou, Y. Zheng, Z. Li, and G. Shen, "Iodetector: A generic service for indoor/outdoor detection," *ACM Trans. Sensor Netw.*, vol. 11, no. 2, p. 28, 2015.

[14] M. Jia, Y. Yang, L. Kuang, W. Xu, T. Chu, and H. Song, "An indoor and outdoor seamless positioning system based on android platform," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, 2016, pp. 1114–1120.

[15] H. Zou, H. Jiang, Y. Luo, J. Zhu, X. Lu, and L. Xie, "Bluedetect: An ibeacon-enabled scheme for accurate and energy-efficient indoor-outdoor detection and seamless location-based service," *MDPI Sensors*, vol. 16, no. 2, p. 268, 2016.

[16] A. Ray, S. Deb, and P. Monogioudis, "Localization of lte measurement records with missing information," in *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[17] V. Radu, P. Katsikouli, R. Sarkar, and M. K. Marina, "A semi-supervised learning approach for robust indoor-outdoor detection with smartphones," in *Proc. ACM SenSys*, 2014, pp. 280–294.

[18] C. Chen, Y. Ren, and C.-C. J. Kuo, "Indoor/outdoor classification with multiple experts," in *Big Visual Data Analysis*. Springer, 2016, pp. 23–63.

[19] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[20] S. S. Khan, J. Hoey, and D. Lizotte, "Bayesian multiple imputation approaches for one-class classification," in *Canadian Conference on Artificial Intelligence*. Springer, 2012, pp. 331–336.

[21] Z.-g. Liu, Q. Pan, J. Dezert, and A. Martin, "Adaptive imputation of missing values for incomplete pattern classification," *Pattern Recognition*, vol. 52, pp. 85–95, 2016.

[22] P. Juszczak and R. P. Duin, "Combining one-class classifiers to classify missing data," in *International Workshop on Multiple Classifier Systems*. Springer, 2004, pp. 92–101.

[23] M. Millán-Giraldo, R. P. Duin, and J. Sánchez, "Dissimilarity-based classification of data with missing attributes," in *Proc. IEEE CIP*, 2010, pp. 293–298.

[24] H. Ye, T. Gu, X. Tao, and J. Lu, "F-loc: Floor localization via crowdsourcing," in *Proc. IEEE ICPADS*, 2014, pp. 47–54.

[25] K. Khaoampai, K. N. Nakorn, and K. Rojviboonchai, "FloorLoc-SL: floor localization system with fingerprint self-learning mechanism," *Int. Journ. Distrib. Sens. Netw.*, 2015.

[26] J. Ying, C. Ren, and K. Pahlavan, "A barometer-assisted method to evaluate 3d patient geolocation inside hospital," in *Proc. IEEE ISMICT*, 2016, pp. 1–4.

[27] X. Shen, Y. Chen, J. Zhang, L. Wang, G. Dai, and T. He, "Barfi: Barometer-aided Wi-Fi floor localization using crowdsourcing," in *Proc. IEEE MASS*, 2015, pp. 416–424.

[28] Y.-C. Tung and K. G. Shin, "Echotag: accurate infrastructure-free indoor location tagging with smartphones," in *Proc. ACM MobiCom*, 2015, pp. 525–536.

[29] Z. Yang, Z. Wang, J. Zhang, C. Huang, and Q. Zhang, "Wearables can afford: Light-weight indoor positioning with visible light," in *Proc. ACM MobiSys*, 2015, pp. 317–330.

[30] M. Azizyan, I. Constandache, and R. Roy Choudhury, "Surroundsense: mobile phone localization via ambience fingerprinting," in *Proc. ACM MobiCom*, 2009, pp. 261–272.

[31] S. Hotta, Y. Hada, and Y. Yaginuma, "A robust room-level localization method based on transition probability for indoor environments," in *Proc. IEEE IPIN*, 2012, pp. 1–8.

[32] Y. Jiang, X. Pan, K. Li, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang, "Ariel: Automatic Wi-Fi based room fingerprinting for indoor localization," in *Proc. ACM UbiComp*, 2012, pp. 441–450.

[33] N. T. Nguyen, R. Zheng, and Z. Han, "UMLI: An unsupervised mobile locations extraction approach with incomplete data," in *Proc. IEEE WCNC*, 2013, pp. 2119–2124.
[34] R. S. Campos, L. Lovisolo, and M. L. R. de Campos, "Wi-fi multi-floor indoor positioning considering architectural aspects and controlled computational complexity," *Expert systems with applications*, vol. 41, no. 14, pp. 6211–6223, 2014.
[35] M. A. A. Rahman, M. Dashti, and J. Zhang, "Floor determination for positioning in multi-story building," in *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*. IEEE, 2014, pp. 2540–2545.
[36] H.-X. Liu, B.-A. Chen, P.-H. Tseng, K.-T. Feng, and T.-S. Wang, "Enhanced area estimation algorithms for indoor wireless localization," in *Proc. IEEE ICCE-TW*, 2016, pp. 1–2.
[37] P. Gupta, S. Bharadwaj, S. Ramakrishnan, and J. Balakrishnan, "Robust floor determination for indoor positioning," in *Communications (NCC), 2014 Twentieth National Conference on*. IEEE, 2014, pp. 1–6.
[38] E. P. Xing, M. I. Jordan, R. M. Karp *et al.*, "Feature selection for high-dimensional genomic microarray data," in *ICML*, vol. 1. Citeseer, 2001, pp. 601–608.
[39] Y. Chen, Q. Yang, J. Yin, and X. Chai, "Power-efficient access-point selection for indoor location estimation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 7, pp. 877–888, 2006.
[40] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, J. C. Platt *et al.*, "Support vector method for novelty detection." in *NIPS*, vol. 12. Citeseer, 1999, pp. 582–588.
[41] T. S. Rappaport *et al., Wireless communications: principles and practice*. Prentice Hall PTR New Jersey, 1996, vol. 2.
[42] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, vol. 2, pp. 27:1–27:27, 2011.
[43] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.
[44] A. M. Hossain, Y. Jin, W.-S. Soh, and H. N. Van, "Ssd: A robust rf location fingerprint addressing mobile devices' heterogeneity," *IEEE Transactions on Mobile Computing*, vol. 12, no. 1, pp. 65–77, 2013.
[45] L. Li, G. Shen, C. Zhao, T. Moscibroda, J.-H. Lin, and F. Zhao, "Experiencing and handling the diversity in data density and environmental locality in an indoor positioning service," in *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 2014, pp. 459–470.
[46] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*. Cambridge University Press, 2010.
[47] C. Laoudias, D. Zeinalipour-Yazti, and C. G. Panayiotou, "Crowd-sourced indoor localization for diverse devices through radiomap fusion," in *Indoor Positioning and Indoor Navigation (IPIN), 2013 International Conference on*. IEEE, 2013, pp. 1–7.
[48] S. He, T. Hu, and S.-H. G. Chan, "Contour-based trilateration for indoor fingerprinting localization," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2015, pp. 225–238.
[49] A. Kushki, K. N. Plataniotis, and A. N. Venetsanopoulos, "Intelligent dynamic radio tracking in indoor wireless local area networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 3, pp. 405–419, 2010.
[50] ——, "Kernel-based positioning in wireless local area networks," *IEEE transactions on mobile computing*, vol. 6, no. 6, pp. 689–705, 2007.
[51] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[52] H.-X. Liu, B.-A. Chen, P.-H. Tseng, K.-T. Feng, and T.-S. Wang, "Map-aware indoor area estimation with shortest path based on RSS fingerprinting," in *Proc. IEEE VTC Spring*, 2015, pp. 1–5.
[53] A. Varshavsky, A. LaMarca, J. Hightower, and E. De Lara, "The skyloc floor localization system," in *Proc. IEEE PerCom*, 2007, pp. 125–134.
[54] S. He, W. Lin, and S.-H. G. Chan, "Indoor localization and automatic fingerprint update with altered ap signals," *IEEE Transactions on Mobile Computing*, vol. 16, no. 7, pp. 1897–1910, 2017.

Ka-Ho Chow received his Bachelor of Engineering degree in Computer Science from the Hong Kong University of Science and Technology (HKUST) in 2016. He is now working towards the Master of Philosophy degree at the Department of Computer Science and Engineering, HKUST, under the supervision of Prof. S.-H. Gary Chan. His research interest includes spatiotemporal data mining, big data, indoor localization and mobile computing.

Suining He received the Ph.D. degree at the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST) in 2016. He is currently working as a postdoctoral research fellow at the Real-Time Computing Lab (RTCL), Department of Electrical Engineering and Computer Science, the University of Michigan, Ann Arbor, MI. His research interest includes indoor localization, smartphone sensing and mobile computing.

Jiajie Tan received the Bachelor of Engineering degree from Zhejiang University, Hangzhou, Zhejiang, China, in 2012, and is pursuing the Ph.D. degree in the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology (HKUST), Hong Kong, China. His research interest includes indoor localization, people sensing and mobile computing.

S.-H. Gary Chan (S89-M98-SM03) is currently Professor in the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong. He is also the Director of Entrepreneurship Center, and Chair of the Committee on Entrepreneurship Education Program, Center for Education Innovation, HKUST. He received MSE and PhD degrees in Electrical Engineering from Stanford University (Stanford, CA) in 1994 and 1999, respectively, with a Minor in Business Administration. He obtained his B.S.E. degree (highest honor) in Electrical Engineering from Princeton University (Princeton, NJ) in 1993, with certificates in Applied and Computational Mathematics, Engineering Physics, and Engineering and Management Systems. His research interest includes multimedia networking, mobile computing, data analytics and IT entrepreneurship.

Professor Chan has been an Associate Editor of IEEE Transactions on Multimedia (2006-11), and a Vice-Chair of Peer-to-Peer Networking and Communications Technical Sub-Committee of IEEE Comsoc Emerging Technologies Committee (2006-13). He is and has been Guest Editor of Elsevier Computer Networks (2017), ACM Transactions on Multimedia Computing, Communications and Applications (2016), IEEE Transactions on Multimedia (2011), IEEE Signal Processing Magazine (2011), IEEE Communication Magazine (2007), and Springer Multimedia Tools and Applications (2007). He was the TPC chair of IEEE Consumer Communications and Networking Conference (IEEE CCNC) 2010, Multimedia symposium of IEEE Globecom (2007 and 2006), IEEE ICC (2007 and 2005), and Workshop on Advances in Peer-to-Peer Multimedia Streaming in ACM Multimedia Conference (2005).

Professor Chan has co-founded several startups deploying his research results. Due to their innovations and commercial impacts, his projects have received local and international ICT awards (2012-2015). He is the recipient of Google Mobile 2014 Award (2010 and 2011) and Silver Award of Boeing Research and Technology (2009). He has been a visiting professor and researcher in Microsoft Research (2000-11), Princeton University (2009), Stanford University (2008-09), and University of California at Davis (1998-1999). He was Undergraduate Programs Coordinator in Department of Computer Science and Engineering (2013-15), Director of Sino Software Research Institute (2012-15), Codirector of Risk Management and Business Intelligence program (2011- 2013), and Director of Computer Engineering Program (2006-2008) at HKUST. He was a William and Leila Fellow at Stanford University (1993-94), and the recipient of the Charles Ira Young Memorial Tablet and Medal, and the POEM Newport Award of Excellence at Princeton (1993). He is a Chartered Fellow of The Chartered Institute of Logistics and Transport Hong Kong (CILTHK), and a member of honor societies Tau Beta Pi, Sigma Xi and Phi Beta Kappa.