IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 30, 2020, accepted October 23, 2020, date of publication November 16, 2020, date of current version February 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3034654

An Accurate and Real-Time Commercial Indoor Localization System in LTE Networks

WEI FANG[1], (Member, IEEE), CHANGJUN XIE[1], (Member, IEEE), AND BIN RAN[2]
[1]School of Automation, Wuhan University of Technology, Wuhan 430070, China
[2]Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI 53706, USA

Corresponding author: Wei Fang (fangwei49@whut.edu.cn)

ABSTRACT In this paper, a commercial high-accuracy, low-cost and real-time indoor building-level localization system is proposed, which is applicable for locating the Minimization of Drive-Tests (MDT) data in the long-term-evolution (LTE) cellular communication network system. The system utilizes MDT data containing Global Navigation Satellite Systems (GNSS) information which is easy to collect and low cost to assist indoor localization, instead of using indoor drive test (DT) data which needs high manual collection and maintenance costs. In order to compensate for the loss of location accuracy, this paper innovatively divide the online process into two phases: indoor and outdoor (IO) identification phase and indoor localization phase. A real-time and precise GMM-based unsupervised algorithm is applied to identifying if the non-GNSS MDT data is in indoor environment in IO identification phase. Then, a multi-class classification algorithm based on Bayesian classifier is used to locate indoor MDT data to the specific building. The results of experiments conducted in an in-service LTE network using more than 100 LTE base stations demonstrate that the proposed technique yields a IO identification accuracy of 90% and an indoor location accuracy of 49.3m(@67%) respectively, which provides at least 30.2% enhancement in location accuracy compared to traditional technology without DT fingerprint database. This proposed indoor localization technique is applicable in network optimization and Operation and Maintenance (O&M) to assist communication service providers to reduce their operating expense (OPEX) by locating those MDT data without GNSS information.

INDEX TERMS Long-term-evolution, indoor localization, indoor and outdoor identification, Bayesian classifier, Gaussian mixture model, penetration loss.

I. INTRODUCTION

A. CSP MDT INDOOR LOCALIZATION SCENE

Recently, the indoor and outdoor localization technology based on Base Station (BS) signal has been gaining attention again. Benefiting from the rapid development of big data storage and processing technology [1]–[4], as well as 3rd Generation Partnership Project (3GPP) R9's support for user equipment (UE) reporting Minimization of Drive-Tests (MDT) measurement reports [5], then communication service providers (CSPs) are able to collect, store and locating massive MDT data of the whole network, and then utilize these located MDT data to replace traditional drive test (DT) data to improve the efficiency of network optimization and reduce the cost of network operation and maintenance (O&M) [6]. In addition, with the advantage of wide coverage, the whole located MDT data can also be utilized in public security, such as crowd impact analysis for epidemic prevention assistance, suspect tracking, etc. If there is no localization system based on BS signal to locating these MDT data, all applications mentioned above will not be able to use, and these data will not be able to generate great value. 3GPP is also aware of the importance of wireless signal based localization, and the location management function (LMF) network element is specially defined in New Radio (NR) release 16 protocol to provide a variety of localization technologies based on BS wireless signal [7].

Traditional localization scene can be categorized into two groups based on request initiator and data sets used in the methods: UE and CSP MDT localization scene respectively [8]. For UE localization scene, localization methods running in Google Maps, Amap or other navigation applications (Apps), are mainly initiated by UE. With the authorization of UE, the localization module in these Apps can access various data, such as Global Navigation Satellite Systems (GNSS) data [9], motion sensor data [10], image sensor data [12], [12], Wi-Fi data [13]–[15] and Base Station signal data [16]–[19], to provide high-accuracy location services for UEs, and the location accuracy is usually better than 5 meters. However, CSP MDT Localization Scene is different. Only GNSS data and BS signal data of UE can be uploaded to the BS along with MDT data and collected by CSPs [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco Rafael Marques Lima.

To ensure GNSS information can be reported in MDT, UE must turn on the GNSS service switch and perform navigation services. Under this limitation, the proportion of GNSS MDT in the whole MDT data is generally less than 20%, which comes from our statistical results of 50 million MDT data from more than 10 thousand LTE cells. That is, more than 80% of the MDT data need to be located. Obviously, the reduction of available data sources and the huge distance between BS will inevitably lead to the increase of location error of this part of MDT data. However, MDT has the characteristics of 24-hour uninterrupted reporting and being collected when the data service rate is very poor, which make the location service based on MDT data have irreplaceable role and value for CSPs.

Indoor localization technology is playing a more and more important role in CSP MDT localization scene, since more than 80% of the current users spend their time indoors [20]. CSPs have taken the evaluation of indoor network coverage quality as one of their key tasks to enhance user stickiness to the network. With the help of indoor localization technology based on MDT data, CSPs can quickly know the signal quality coverage of each building, so as to carry out network optimization work in time, which can effectively improve user satisfaction and reduce operating expense (OPEX).

B. RELATED INDOOR LOCALIZATION TECHNOLOGY

Indoor UE can access two types of BS, one is the micro BS deployed indoors, the other is the macro BS deployed outdoors. Since the coverage radius of indoor micro BS is small (generally 30 meters), we default that the UEs are in the indoor environment when accessing a micro BS. The above scene is similar to the traditional indoor localization scene based on sensor networks [21]–[23], that is, almost all terminals connected to these signal transmitters are located indoors. However, the situation becomes complicated when UE is connecting to the macro BS, because we can't directly identify whether UE is in indoor or outdoor environment. Therefore, there is an urgent need for an indoor location technology to locating the UEs in the indoor environment to the corresponding building in this case. It is noted that if there is no additional explanation, the indoor localization research in this paper is working in this case.

Traditional indoor localization technology based on BS signal can be categorized into three groups: time based [26], [27], angle based [28], [29] and fingerprint based algorithm [30], [31]. Time-based and angle-based algorithms have low commercial cost, while they are usually greatly affected by the physical environment and channel quality. Error estimates of location related parameters in these two algorithms will occur when the signal arrives at UE from non-Line-of-Sight (NLoS) multipath or there is penetration loss, which resulting in poor location accuracy. In this case, indoor MDT data is likely to be located outdoors, and vice versa.

Fingerprint-based algorithm constructs a fingerprint database (FD) representing the unique combination of BS signal measurements with respect to a indoor or outdoor geographic location [30], [31], which is less affected by environmental factors. Usually, a building area will be marked by many fixed sampling locations, and the adjacent sampling locations are generally 0.5m to 2m apart. Then, some test UEs collect BS signal at each sampling point for a while, then these collected data will be used to build the indoor FD of each sampling location. Finally, UE will be located in the building containing the most suitable sampling points, where the most suitable sampling point is the one which is most similar to the signal characteristics of UE. Of course, indoor localization systems with this method have high location accuracy. However, it is very time-consuming to mark the sampling locations and collect the corresponding BS signal. Moreover, FD of a building should be updated regularly to ensure the accuracy, which means that the commercial cost will be further increased. These shortcomings make it difficult for these localization techniques to be commercialized on a large scale.

Some other fingerprint based localization systems take into account the commercial costs, such as CellSense [19], NBL [24], ARTLoc [8] and DeepLoc [25]. These systems utilize GNSS MDT data to generate the FD in offline phase, and then complete the locating of non-GNSS data in online phase. Since traditional DT data is replaced by GNSS MDT data to build FD, localization techniques used in these systems possess the characteristics of low commercial cost, which makes these systems applicable on a large scale. However, due to the severe deterioration of GNSS location performance in indoor environment, the FD built by GNSS MDT data is mainly distributed outdoors [8], [19], which will result in indoor MDT data being located outdoors. In this case, the indoor location accuracy obtained by these systems will be seriously reduced, which makes many indoor applications unable to be used.

C. CONTRIBUTIONS OF THIS PAPER

Our later research is aimed at a commercial indoor localization system, and the image of this system is deployed in CSPs in many countries around the world. Our localization system is implemented on the distributed platform Spark and Storm, and performs more than 10 billion location requests every day. The locatioin results of this system are widely used in indoor and outdoor network O&M and CSP location-based services, including crowd impact analysis for epidemic prevention assistance, suspect tracking, user portraits, city security services, flow analysis, intelligent advertising push, etc. Therefore, location accuracy, location efficiency and commercial cost are all important for this indoor localization system.

However, according to our research, current indoor localization systems mentioned in related literature do have defects in such commercial scenarios. As we mentioned above, although the fingerprint-based algorithms used in those systems have good location accuracy, the commercial cost is too high to build the indoor DT FD. On the contrary, systems with time-based, angle-based and GNSS MDT

fingerprint-based algorithm have low commercial cost, but they will locate the original indoor UE to outdoors, resulting in big location error. A low-cost, high-precision indoor localization system is urgently needed to assist CSPs in large-scale indoor network O&M and monetizing their data.

To this end, this paper proposes a commercial high-accuracy, low-cost and real-time indoor building-level localization system to address above challenges. The localization technology in the system does not utilize DT data to build the indoor FD, so the commercial cost is low. In order to compensate for the loss of location accuracy, we innovatively divide the online localization process into two phases: indoor and outdoor (IO) identification phase and indoor localization phase, to improve location accuracy.

In the IO identification phase, the non-GNSS MDT data will be identified as indoor or outdoor to ensure that the MDT data originally staying indoors can be located in the building. A real-time, unsupervised and Gaussian mixture model (GMM)-based [32] IO identification algorithm is proposed in this phase, and its design principle comes from the fact that the main difference between indoor and outdoor UE in similar location with the same serving cell is reflected in the outdoor-to-indoor (O2I) penetration loss of reference signal receiving power (RSRP). Then, our algorithm will learn the distribution characteristics of indoor and outdoor RSRP, which are used to assist in the identification of indoor and outdoor UE.

In the indoor localization phase, MDT data identified as indoor environment will be located in a specific building. A high-accuracy, low-cost and real-time indoor building-level localization algorithm is proposed in this phase. Firstly, GNSS MDT data around the building are extracted, and the O2I penetration loss compensation will work on these data, which will be labeled by corresponding building ID. Then, we model the indoor localization problem as a multi-class model, where building ID is the label data and signal characteristics are extracted from these MDT data. Finally, a Bayesian classifier [33], [34] is trained base on these data, which will assist in locating non-GNSS MDT data to the corresponding building in online phase.

To summarize, the contribution of this paper is three-fold:

i) We introduce a commercial high-accuracy, low-cost and real-time indoor building-level localization system, which extends the online phase to the IO identification phase and indoor localization phase for the first time.

ii) A real-time, unsupervised and GMM-based IO identification algorithm is proposed, which ensures that the original indoor data can be located in the building without using the indoor DT FD.

iii) In order to locate indoor MDT data to specific building, we model the indoor localization problem as a multi-classification problem and then introduce a high-accuracy, low-cost and real-time indoor building-level localization algorithm based on Bayesian classifier.

The rest of paper is structured as follows. In Section II, our proposed localization system is introduced. In section III, we present the IO identification algorithm based on GMM in our system. Then the indoor localization algorithm in indoor localization phase is presented in section IV. Section V presents performance evaluation of our system. Finally, Section VI concludes the paper and gives directions for future work.

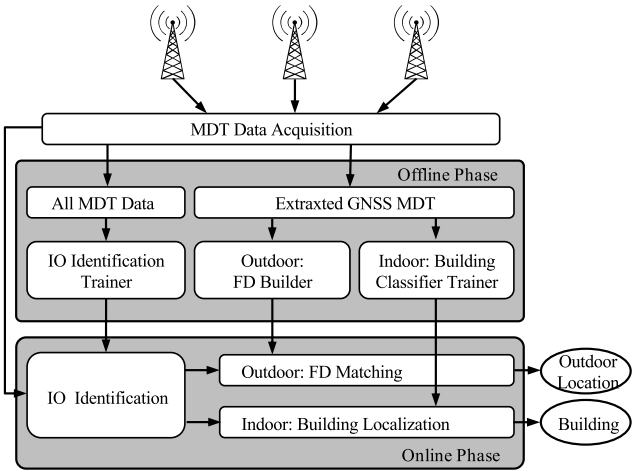

FIGURE 1. Functional block diagram of proposed localization system.

II. PROPOSED LOCALIZATION SYSTEM

In this section, we introduce the background of our proposed commercial building-level indoor localization system, including functional block diagram, measurements required for localization and the the theoretical basis of later sections.

A. PROPOSED LOCALIZATION SYSTEM

The proposed localization system consists of two parts: offline phase and online phase. Figure 1 shows the functional block diagram of our proposed localization system in this paper.

In the offline phase, the MDT data is utilized to train three functional modules. The IO identification model is trained based on the whole MDT data using GMM technology. Part of the whole MDT data contains GNSS information, which we called GNSS MDT. Comparing GNSS with building map, most of the GNSS MDT data are in outdoor environment, which are used to build outdoor FD. The specific process is introduced in [8]. Due to the severe deterioration of GNSS location performance in the indoor environment, there is little GNSS MDT data in indoor buildings. In order to train the indoor building classifier with low commercial cost, we extract all the GNSS MDT data around the building and compensate O2I penetration loss on these data. Then, the corresponding building ID will be used as a label for these data, which is available to train the indoor building classifier. The offline phase is a batch computing model.

In the online phase, non-GNSS MDT will be located to either an outdoor location or a building. Firstly, the location scene feature(indoor or outdoor) can be determined by IO identification model trained offline. Then, the classified outdoor MDT will be matched with the appropriate grid in FD and located to an outdoor location. The identified indoor

MDT will be sorted to a appropriate building. The online location module is a kind of streaming computing model, which supports a distributed streaming computing system. Under the support of the high-accuracy, low-cost and real-time and full-scene localization system, the requirements of most commercial scenarios can be met.

It is worth mentioning that the outdoor localization system has been described and verified in detail in [8], so the following contents of this paper will mainly focus on the indoor building-level localization system.

B. DATA PREPARATION

3GPP protocol has introduced many localization algorithms, including Observed Time Difference of Arrival (OTDOA) for LTE and Downlink (DL)-TDOA, DL-Angle of Departure (AoD), Uplink (UL)-TDOA, UL-Angle of Arrival (AoA), Multi-Round Trip Time (RTT) for 5G NR [7]. The required downlink measurements are based on DL-Positioning Reference Signal (PRS), and the uplink measurements are based on UL-Sounding Reference Signal (SRS). These measurements and localization algorithm both require additional computational overhead for BS and UE. Although it is possible to migrate the calculation to mobile edge computing (MEC) server, there are still not enough computing resources to locate the whole MDT data utilizing such algorithms. Thus, in our proposed localization system, we only use the information in MDT reports that can be collected directly from the BS, which is necessary for the BS operation and does not require additional measurement or calculation.

The MDT report carries vital information required for location, including GNSS(if available), Reference Signal Receiving Power (RSRP) of serving cell and neighbor cells, and Timing Advance (TA) of serving cell. Latest 3GPP protocol supports active or passive GNSS acquisition, however, the GNSS MDT is less than 20% of the whole MDT. The remaining 80% non-GNSS MDT needs to be located.

RSRP is a key indicator related to UE location. Theoretically, the farther the UE is from the antenna, the smaller the measured RSRP is. But affected by multipath effect and shadowing effect, RSRP usually has poor representation of distance.

TA is a more relevant indicator of UE location. In LTE, the size of various fields in the time domain is expressed in time units $T_s = 1/(\Delta f_{\text{ref}} \cdot N_{\text{f,ref}})$, $\Delta f_{\text{ref}} = 15 \cdot 10^3$Hz and $N_{\text{f,ref}} = 2048$. T_s presents a distance of 4.88m between UE and base station. In NR, with $\Delta f_{\max} = 480 \cdot 10^3$Hz and $N_f = 4096$, the time unit T_c is much smaller. T_c presents a distance of 0.076m. For the convenience of subsequent explanation, TA is expressed in $16T_s$ granularity. One TA in $16T_s$ granularity presents a distance of 78.125m.

It is worth mentioning that the proposed localization system is effective for both LTE and NR. Due to the smaller time unit of NR, it can achieve more accurate TA measurements. The location accuracy of NR is expected to be better. But there is no NR MDT data so far, so we use LTE data to verify the algorithms.

TABLE 1. MDT information needed for location.

Information	Definition
$p(x, y)$	GNSS longitude and latitude
r_0	Serving cell RSRP(dBm)
t_0	Serving cell TA(in 16 T_s)
r_i	Neighbor cell RSRPs(dBm)

TABLE 2. Material penetration losses.

Material	Penetration loss(in dB, f is in GHz)
Standard	$L_{\text{glass}} = 2 + 0.2f$
IRR glass	$L_{\text{IIRglass}} = 23 + 0.3f$
Concrete	$L_{\text{concrete}} = 5 + 4f$

TABLE 3. O2I building penetration loss model.

Loss model	Path loss through external wall: PL_{tw} (in dB)
Low-loss model	$5 - 10\log_{10}\left(0.3 \cdot 10^{-L_{\text{glass}}/10} + 0.7 \cdot 10^{-L_{\text{concrete}}/10}\right)$
High-loss model	$5 - 10\log_{10}\left(0.7 \cdot 10^{-L_{\text{IIRglass}}/10} + 0.3 \cdot 10^{-L_{\text{concrete}}/10}\right)$

All the MDT information needed for location is shown in Table 1.

C. O2I PENETRATION LOSS

Obviously, in indoor scene and outdoor scene, there is a big difference in RSRP from the same serving cell measured at similar locations, which is called O2I penetration loss.

In 3GPP TS 38.901 [35], the pathloss incorporating O2I building penetration loss is modeled as

$$\text{PL} = \text{PL}_b + \text{PL}_{tw} + \text{PL}_{in} + N\left(0, \sigma_P^2\right) \tag{1}$$

where PL_b is the basic outdoor path loss, PL_{tw} is the building penetration loss through the external wall, PL_{in} is the inside loss dependent on the depth into the building, and σ_P is the standard deviation for the penetration loss. The last three terms of the formula are all related to the O2I building penetration loss. For convenience of estimation, we only consider the most important external wall penetration loss PL_{tw} which is characterized as

$$\text{PL}_{tw} = \text{PL}_{npi} - 10\log_{10}\sum_{i=1}^{N}\left(p_i \times 10^{\frac{L_{material\_i}}{-10}}\right) \tag{2}$$

where PL_{npi} is an additional loss added to the external wall loss to account for non-perpendicular incidence; $L_{material\_i} = a_{material\_i} + b_{material\_i} \cdot f$, is the penetration loss of material i, example values of which can be found in Table 2; p_i is proportion of i-th materials, where $\sum_{i=1}^{N} p_i = 1$; and N is the number of materials.

The standard deviation for the penetration loss is expressed as σ_P. Table 3. gives PL_{tw} for two O2I penetration loss models.

Assuming a typical frequency of $f = 1.9$GHz, the O2I penetration loss of low-loss model is 12dB, and that of high-loss model is 22dB. In practice, due to the different building materials, the O2I penetration loss will be different, generally in [12, 22]dB. In Section III, we put forward the

IO identification algorithm based on the principle of O2I penetration loss.

D. MODEL OF BUILDING LOCALIZATION PROBLEM

Traditional LTE or NR based indoor localization system is aimed at a small location area, usually just a single building, to locate the UE inside the specified building. However, in commercial scenarios, the location area is much larger than a single building, and sometimes even all buildings in the entire city, as many as hundreds of thousands.

Therefore, it is necessary to determine the building where the UE is located precisely and efficiently before performing indoor localization. Essentially this is a supervised multi-classification problem. The signal measurements are the features, including RSRP and TA. The building where the UE is located is the category to be estimated. By matching GNSS MDT around the building, we can label part of the GNSS MDT with building, which can be used as training data.

III. IO IDENTIFICATION PHASE

In this section, a real-time, unsupervised and Gaussian Mixture Model (GMM)-based IO identification algorithm is proposed, and the non-GNSS MDT data will be identified as indoor or outdoor by this algorithm to ensure that the MDT data originally staying indoors can be located in the building. We assume that all the MDT data used here are connected to at least one neighbor cell. It should be noted that the IO identification algorithm is mainly applied to the UE accessing the outdoor macro BS in the indoor environment. For the UE accessing to the indoor micro BS, we default that this UE is in the indoor environment.

As we mentioned above, according to the O2I penetration loss of BS wireless signal, the distribution of indoor RSRP and outdoor RSRP is quite different in an region with similar distance and direction to the BS. According to this fact, our algorithm will learn the distribution parameters of indoor and outdoor RSRP, and use these parameters to perform IO identification. In the actual operation of our system, IO identification algorithm is divided into three processes, namely characteristics extracted from MDT data, offline phase and online phase. Then, we will elaborate these three processes in detail.

A. CHARACTERISTICS EXTRACTED FROM MDT DATA

It is worth mentioning that before locating, how do we know that UEs are in the region with similar distance and direction to the serving cell. To solve this problem, we assume that if eNodeBID(gNodebID in NR), cellID, timing advance (TA) and the Physical Cell Identifier (PCI) corresponding to the strongest RSRP of neighbor cells, which is referred to as the SNPCI (Strongest Neighbor PCI), are the same, then the two MDT data can be considered to be located in close proximity. One interpretation is that the same SNPCI and TA respectively indicate that the direction and distance between UE and serving cell are similar. In this case, the RSRP of indoor and outdoor serving cell will show obvious O2I penetration loss.

TABLE 4. Characteristics extracted from MDT data.

Characteristics	Definition
eNodeBID	ID of eNodeB or BBU
cellID	ID of RRU
NSPCI	Neighbor Strongest PCI
TA	Timing Advance(in 16 T_s)
RSRP	Serving cell RSRP(dBm)

The characteristics extracted from MDT data by IO identification algorithm are shown in Table 4, where BBU and RRU refer to building base band unite and remove radio unit, respectively.

B. OFFLINE PHASE

After grouping the total data based on eNodeBID, cellID, SNPCI and TA, the distribution of RSRP set in each group $\Phi_i (i = 1, 2, \ldots)$ can be considered as the sum of indoor RSRP distribution and outdoor RSRP distribution. Suppose that indoor RSRP and outdoor RSRP follow Gaussian distribution of $\phi(x|\mu_0, \sigma_0^2)$ and $\phi(x|\mu_1, \sigma_1^2)$, respectively. $\mu_i (i = 0, 1)$ is the mean value and $\sigma_i^2 (i = 0, 1)$ is variance value of Gaussian function. Then, we use p_0 and p_1 represents the probability that a MDT data in set Φ_i belongs to indoor and outdoor environment respectively, and $p_0 + p_1 = 1$, $p_0 \geq 0$, $p_1 \geq 0$. Then the distribution of a RSRP value in Φ_i can be expressed as:

$$\begin{aligned} P(y|\boldsymbol{\theta}) &= \sum_{z=0}^{1} P(y, z|\boldsymbol{\theta}) = \sum_{z=0}^{1} P(y|z, \boldsymbol{\theta}) P(z|\boldsymbol{\theta}) \\ &= p_0 \phi(y|\mu_0, \sigma_0^2) + p_1 \phi(y|\mu_1, \sigma_1^2) \end{aligned} \tag{3}$$

where y is the observed RSRP value of serving cell, $\boldsymbol{\theta} = (\mu_0, \mu_1, \sigma_0^2, \sigma_1^2, p_0, p_1)^T$ is the parameter vector to be estimated and $P(y|\boldsymbol{\theta})$ is the conditional probability distribution of observed RSRP. The variable z indicates that the MDT data is in indoor($z = 0$) or outdoor($z = 1$) environment, which is a latent variable that cannot be directly observed. Obviously, formula (3) is a Gaussian mixture model.

We use vector $\mathbf{Y} = (y_1, y_2, \ldots, y_N)^T$ to represent a set of independent RSRP observations in Φ_i, then

$$P(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[p_0 \phi(y_i|\mu_0, \sigma_0^2) + p_1 \phi(y_i|\mu_1, \sigma_1^2) \right] \tag{4}$$

Based on Maximum Likelihood criterion, the optimal parameter vector $\boldsymbol{\theta}'$ can be expressed as:

$$\boldsymbol{\theta}' = \arg\max_{\boldsymbol{\theta}} P(\mathbf{Y}|\boldsymbol{\theta}) \tag{5}$$

However, there is no closed form solution to above formula, so we use the Expectation Maximization (EM) algorithm to estimate the optimal parameter $\boldsymbol{\theta}$. EM algorithm is an iterative algorithm. In each iteration process, the algorithm will solve the maximum value of the lower bound function, so as to gradually approach the maximum of likelihood function, and obtain the optimal parameters $\boldsymbol{\theta}'$. The specific solution process of EM algorithm has been very general,

so this paper only lists the results of some key steps:

$$\hat{\gamma}_{jk}^{(i)} = E[\gamma_{jk}|\mathbf{Y}, \boldsymbol{\theta}^{(i)}] \quad j = 1, ...N$$
$$= \frac{\alpha_k^{(i-1)}\phi(y_j|\mu_k^{(i-1)}, \sigma_k^{2(i-1)})}{\sum_{k=0}^{1}\alpha_k^{(i-1)}\phi(y_j|\mu_k^{(i-1)}, \sigma_k^{2(i-1)})}, \quad k = 0, 1 \quad (6)$$

$$\mu_k^{(i)} = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}^{(i)} y_j}{\sum_{j=1}^{N}\hat{\gamma}_{jk}^{(i)}}, \quad k = 0, 1 \quad (7)$$

$$\sigma_k^{2(i)} = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}^{(i)}\left(y_j - \mu_k^{(i-1)}\right)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}^{(i)}}, \quad k = 0, 1 \quad (8)$$

$$\alpha_k^{(i)} = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}^{(i)}}{N}, \quad k = 0, 1 \quad (9)$$

where superscript (i) denotes *ith* iteration, subscript j represents the value of *jth* MDT data. Subscript k is *kth* Gaussian distribution model, $k = 0$ and $k = 1$ denote indoor environment and outdoor environment, respectively. $\gamma_{jk}^{(i)}$ is

$$\gamma_{jk}^{(i)} = \begin{cases} 1, & j\text{th observed data is from } k\text{th model} \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

We define $\boldsymbol{\theta}' = (\mu_0', \mu_1', \sigma'^2_0, \sigma'^2_1, p_0', p'_1)^T$ as the GMM parameter vector solved by EM algorithm. When $\mu_0' < \mu'_1$, the distribution of indoor and outdoor RSRP can be expressed as $\phi(x|\mu'_0, \sigma'^2_0)$ and $\phi(x|\mu'_1, \sigma'^2_1)$, respectively, and vice versa. The interpretation is that the indoor RSRP is always less than the outdoor RSRP because of the O2I penetration loss. For each data set grouped by eNodeBID, cellID, SNPCI and TA, we will obtain a series of parameter vector $\boldsymbol{\theta}'$ of GMM by EM algorithm. Then, these vectors are used to construct the indoor and outdoor parameters database (IOPD), which will be accessed by IO identification algorithm in online phase.

It should be noted that if $\left|\mu_0' - \mu_1'\right| < T$, which means that the distinction between indoor and outdoor RSRP distribution is not obvious, there may be only indoor or outdoor environment in this case, and this is not the focus of the article. Here, T is also called the available threshold of IO identification algorithm.

C. ONLINE PHASE

The characteristics to be extracted from MDT data in online phase are the same as those in Table 1. Then, we utilize characteristics, such as eNodeBID, cellID, SNPCI and TA, to match the IOPD trained in the offline phase, and finally obtain the most matching IO RSRP distribution parameter vector $\boldsymbol{\theta} = (\mu_0', \mu_1', \sigma'^2_0, \sigma'^2_1, p_0', p'_1)^T$. Here, we assume that $\boldsymbol{\theta}_0 = (\mu_0', \sigma'^2_0, p'_0)^T$ and $\boldsymbol{\theta}_1 = (\mu'_1, \sigma'^2_1, p'_1)^T$ represents indoor and outdoor parameter vector respectively. So the joint probability distribution can be expressed as

$$P(y, indoor|\boldsymbol{\theta}_0) = P(y|indoor, \boldsymbol{\theta}_0)P(indoor|\boldsymbol{\theta}_0) \quad (11)$$
$$= p_0' N(y|\mu_0', \sigma'^2_0)$$
$$P(y, outdoor|\boldsymbol{\theta}_1) = p_1' N(y|\mu_1', \sigma'^2_1) \quad (12)$$

where $P(y, indoor|\boldsymbol{\theta}_0)$ represents the probability that the observed RSRP y is in the indoor environment when indoor parameters $\boldsymbol{\theta}_0$ are known, and the meaning of $P(y, outdoor|\boldsymbol{\theta}_1)$ is similar.

Then, we can obtain

$$z = \begin{cases} 0, & \text{if } p(y, indoor|\boldsymbol{\theta}_0) > p(y, outdoor|\boldsymbol{\theta}_1) \\ 1, & \text{otherwise} \end{cases} \quad (13)$$

where $z = 0$ and $z = 1$ indicate that the MDT data is identified as indoor or outdoor environment respectively.

Note that in the online phase, we only need to match with the IOPD and compare the joint probability according to the matching parameters, which makes the online phase can be deployed in the streaming system and applicable in real-time scenarios. In addition, for the UEs accessing the indoor micro station, we will default that they are in the indoor environment. The IO identification algorithm described in this section is mainly used for UEs which access the outdoor serving cell but stay in the indoor environment.

IV. BUILDING LOCALIZATION ALGORITHM BASED ON BAYESIAN CLASSIFIER

In this section, we propose a building localization algorithm based on Bayesian classifier which can precisely, efficiently and cost effectively determine the building where the UE is located. Indoor MDT data will be located in the specified building by this algorithm after IO identification phase. In order to reduce the commercial cost and ensure that the system running the algorithm can be commercialized on a large scale, we use the GNSS MDT data which can be easily collected instead of using indoor DT data to train our model. In order to ensure the location accuracy, we extract all the GNSS MDT data around the building and compensate O2I penetration loss on these data. Then, the corresponding building ID will be used as a label for these data, which is available to train a supervised classification model. Finally, we model the building localization problem as a supervised multi-class classification problem, and use Bayesian classifier to achieve high-accuracy, low-cost and real-time indoor building-level localization.

Multi-class classification algorithms can be divided into two categories: natural multi-class classifier (e.g. neural network, SVM, decision tree, Bayesian classifier, etc.) and multiple output units (e.g. one vs. all or one vs. one). In this paper, taking into account the high location efficiency requirements of commercial localization system, we only consider using a simple natural multi-class classifier, such as decision tree or Bayesian classifier. In such problem of building classification, the characteristics we use are not strictly independent. Decision tree ignores the correlation

between attributes. Therefore, we decide to use Bayesian classifier, assuming the appropriate distribution to fit the relevant features.

The signal measurements feature vector $\mathbf{x} \in \mathbf{R}^{D\times 1}$ can be expressed as

$$\mathbf{x} = (x_1, x_2, x_3, \cdots, x_D)^{\mathrm{T}} = (t_0, r_0, r_1, \cdots, x_m)^{\mathrm{T}} \quad (14)$$

where $[\cdot]^{\mathrm{T}}$ is the matrix transposition operator, $x_1 = t_0$ is serving cell c_0 TA, $x_2 = r_0$ is serving cell c_0 RSRP, $x_{i+2} = r_i, i \in [1, m]$ is neighbor cell c_i RSRP, the dimension of signal measurements feature vector is $D = m + 2$.

At different plane locations or floors in a building, UE may access different serving cells, and the detected neighbor cells may also be different. Only a small part of cells can be accessed as a serving cell in a building. In order to improve the efficiency of building localization, we take the same serving cell as the necessary condition to extract the candidate building set, which can significantly reduce the amount of calculation.

Suppose there are M cells in total. For cell c_i $(i = 1, \cdots, M)$, in the training set, it has been accessed as serving cell in K_i buildings. The building set $\mathbf{y}_i \in \mathbf{R}^{K_i\times 1}$ can be expressed as

$$\mathrm{y}_i = \left(y_i^1, \cdots, y_i^j, \cdots, y_i^{K_i}\right)^{\mathrm{T}} \quad (15)$$

where y_i^j is the j-th building accessed to serving cell c_i.

In order to better fit the correlation between features, we treat all the features as continuous variables. According to Bayesian theory, given the feature vector $\mathbf{x}$ accessed to serving cell c_i, its located building can be predicted as

$$y_i^j = \underset{y_i^j \in \mathrm{y}_i}{\arg\max}\, p_{c_i}\left(y_i^j|\mathbf{x}\right) \quad (16)$$

where y_i^j is the located building, $p_{c_i}\left(y_i^j|\mathrm{x}\right)$ is the posterior probability density of building y_i^j. The above formula means that, under the premise of a known feature vector $\mathbf{x}$ accessed to serving cell c_i, the probability density of $\mathbf{x}$ belonging to $y_i^1, \cdots, y_i^{K_i}$ is calculated respectively. Then the building with the largest probability density is selected as the located building of feature vector $\mathbf{x}$.

Based on Bayes formula, $p_{c_i}\left(y_i^j|\mathrm{x}\right)$, the posterior probability density of y_i can be expressed as

$$\begin{aligned} p_{c_i}\left(y_i^j|\mathrm{x}\right) &= \frac{p_{c_i}\left(\mathrm{x}|y_i^j\right) P_{c_i}\left(y_i^j\right)}{p_{c_i}\left(\mathrm{x}\right)} \\ &= \frac{p_{c_i}\left(\mathrm{x}|y_i^j\right) P_{c_i}\left(y_i^j\right)}{\int_{y_i\in\mathbf{y}_i} p_{c_i}\left(\mathrm{x}|y_i\right) P_{c_i}\left(y_i\right) dy_i} \end{aligned} \quad (17)$$

where $p_{c_i}\left(\mathrm{x}|y_i^j\right)$ is the posterior probability density of x, $P_{c_i}\left(y_i^j\right)$ is the prior probability of building y_i^j, $p\left(\mathrm{x}\right)$ is the total probability density of x which is a constant for any buildings. Therefore, to get the maximum value of $p_{c_i}\left(y_i^j|\mathrm{x}\right)$, we only need to calculate the maximum value of $p_{c_i}\left(\mathrm{x}|y_i^j\right) P_{c_i}\left(y_i^j\right)$.

$P_{c_i}\left(y_i^j\right)$ can be calculated from training set as

$$P_{c_i}\left(y_i^j\right) = \frac{\mathrm{n}\left(\mathbf{S}_i^j\right)}{\mathrm{n}\left(\mathbf{S}_i\right)} \quad (18)$$

where $\mathrm{n}\left(\cdot\right)$ is the element number operator, $\mathbf{S}_i$ is training set accessed to serving cell c_i, $\mathbf{S}_i^j$ is a subset of $\mathbf{S}_i$ in building y_i^j.

Next, we will introduce the calculation method of $p_{c_i}\left(\mathrm{x}|y_i^j\right)$ in details. Features are continuous and correlated, so we assume that the signal measurements feature vectors obey multivariate normal distribution. The probability density function of $p_{c_i}\left(\mathrm{x}|y_i^j\right)$ can be expressed as

$$\begin{aligned} p_{c_i}(\mathbf{x}|y_i^j) &= N_{c_i,y_i^j}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right] \end{aligned} \quad (19)$$

where $\mathbf{x} \in \mathbf{R}^{D\times 1}$ is vector of multi-variables, in this case, feature vector. $\boldsymbol{\mu} \in \mathbf{R}^{D\times 1}$ and $\boldsymbol{\Sigma} \in \mathbf{R}^{D\times D}$ are mean vector and covariance matrix of multivariate normal distribution, respectively.

In offline phase, suppose a training set $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)^{\mathrm{T}}$, each of which is independently sampled from a multivariate normal distribution. We use Maximum Likelihood Estimation (MLE) to estimate the probability density function. The log-likelihood equation $\ln L\left(p\right)$ can be expressed as follows.

$$\begin{aligned} \ln L\left(p\right) &= \ln L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\right) \\ &= -\frac{ND}{2}\ln\left(2\pi\right) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| \\ &\quad -\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_n-\boldsymbol{\mu}\right) \end{aligned} \quad (20)$$

To maximize the likelihood function, we let

$$\begin{cases} \dfrac{\partial \ln L\left(p\right)}{\partial \boldsymbol{\mu}} = 0 \\ \dfrac{\partial \ln L\left(p\right)}{\partial \boldsymbol{\Sigma}} = 0 \end{cases} \quad (21)$$

the estimated parameters are as follows

$$\begin{cases} \bar{\boldsymbol{\mu}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\mathbf{x}_n \\ \bar{\boldsymbol{\Sigma}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\left(\mathbf{x}_n-\boldsymbol{\mu}\right)\left(\mathbf{x}_n-\boldsymbol{\mu}\right)^T \end{cases} \quad (22)$$

So far, we have estimated the parameters of $p_{c_i}(\mathrm{x}|y_i^j)$ when accessing a specific cell c_i in a specific building y_i^j. We will fit a distribution for each situation where UE accesses different serving cells in each building.

In online real-time location phase, firstly determine the candidate buildings according to the serving cell accessed. Then, for each candidate building, the building prior probability $P_{c_i}(y_i^j)$ have been calculated offline, and $p_{c_i}(\mathrm{x}|y_i^j)$

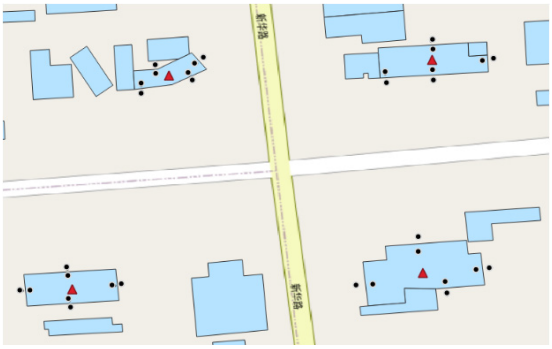

FIGURE 2. The bird's-eye view map of the DT.

of each building is calculated according to online feature vector $\mathbf{x}$ and the distributions $N_{c_i,y_i^j}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ estimated offline. Last the building with the largest probability density $p_{c_i}\left(\mathrm{x}|y_i^j\right) P_{c_i}\left(y_i^j\right)$ is selected as the estimated building of online feature vector $\mathbf{x}$.

The above proposed algorithm can achieve high-accuracy and high-efficiency building localization and complement the full-scene localization system.

V. MEASUREMENT CAMPAIGNS AND LOCALIZATION PERFORMANCE

In order to test and verify our proposed indoor localization system, we make an experiment in a real LTE system. In this section, we will introduce the experiment environments and the performance of the proposed algorithms. The location error or location accuracy use the value corresponding to 67%.

A. DATA COLLECTION

Two kinds of data are collected to perform the following evaluation, which are DT data and massive MDT reported by more than 100 in-service outdoor macro BS respectively. These BSs work in the 1.9G TDD band. It is worth mentioning that the data of the indoor micro BS is not collected because the UEs accessing the micro BS are in the indoor environment by default. In offline stage, massive MDT data is applied for building IOPD in IO identification phase, and GNSS MDT data extracted from all MDT data is used to train Bayesian classifier in indoor localization phase. DT data with indoor and outdoor labels and specific location information are used to verify the accuracy of the algorithm.

The experiment area is a part of downtown area in Langfang, Shijiazhuang Province, China, which is a typical urban area. For the case where the serving cell is an indoor micro cell, since its small coverage area, the building there the UE is located can be easily determined by the accessed serving cell. Therefore, in our experiment, we only consider the macro cell to cover indoors, which is relatively difficult. We choose 4 building for experiment and then conduct fixed-point tests indoors and outdoors in every direction of the building. The test locations are recorded manually and associated with the MDT data collected on the network side to form a test data set. The experiment area is illustrated in Fig. 2, where test buildings are marked by red triangles, and the test locations are marked by black dots.

TABLE 5. Parameters used.

Characteristics	Definition
iter	100
tol	0.001
T	$\max(8.5-0.5TA, 4)$

B. IO IDENTIFICATION

In this section, all MDT data in the previous day are collected to build the IOPD, and DT data with indoor or outdoor label are used to estimate the accuracy of IO identification algorithm.

Parameters used in this section are shown in Table 5, where *iter* is the number of EM iterations to perform, *tol* is the convergence threshold and EM iterations will stop when the lower bound average gain is below this threshold. T is the threshold to determine whether the parameters obtained by IO identification algorithm are available.

In offline training stage, we group about 50 million MDT data by eNodeBID, cellID, TA and SNPCI to get more than 20 thousand sets. Then, according to EM algorithm, GMM parameter vectors of these sets are obtained. In order to verify the feasibility of GMM-based IO identification algorithm, we visualize these parameters. Figure 3 shows the relationship between indoor and outdoor RSRP distribution obtained from GMM parameters and the original data RSRP distribution, which consist of six sub-graph. The corresponding data of each sub-graph are randomly sampled from more than 20000 sets. The horizontal axis and vertical axis represent the values of RSRP and Probability Density Function (PDF) respectively. In the legend, 'original' represents the distribution of the original RSRP in the set, 'gau_0' and 'gau_1' denote the estimated indoor RSRP Gaussian distribution and outdoor RSRP Gaussian distribution respectively, gau_mix is the result of Gaussian mixture distribution after plus gau_0 and gau_1 distribution. The specific GMM algorithm parameters in each sub-graph are shown in Table 6.

It is obvious that in the regions with different distances from the BS(TA is approximately equivalent to the distance), almost all sets of original RSRP follow bimodal distribution, and the parameters obtained by GMM algorithm can well fit the distribution, which proves the feasibility of our GMM algorithm in identifying indoor and outdoor environments.

It is worth mentioning that the distribution of RSRP looks like one Gaussian distribution when TA = 6(nearly 400 meters away from the BS), but we still regard it as bimodal distribution and solve the parameters of GMM algorithm. The reason for this is that there is a logarithmic relationship between the RSRP and the distance from the BS. According to the nature of the logarithmic function, the farther away from the BS, the smaller the change rate of RSRP, which leads to the smaller difference between indoor and outdoor RSRP, resulting in the unclear bimodal characteristics.

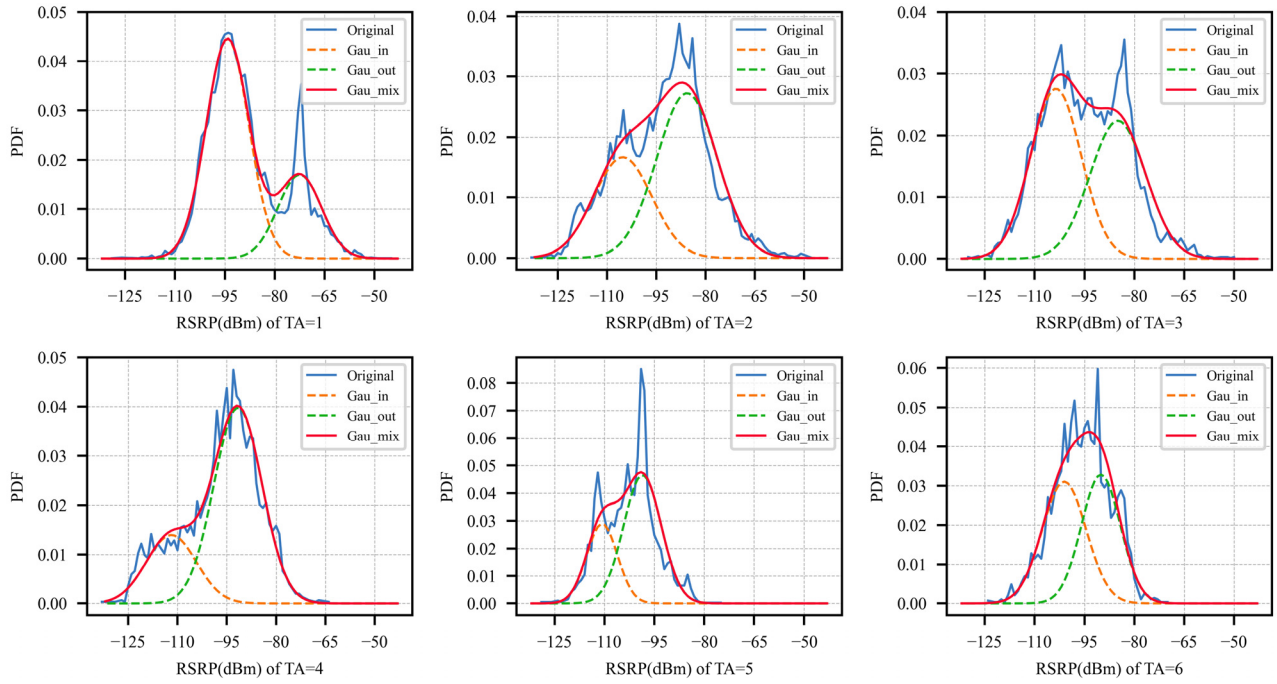

FIGURE 3. Relationship between indoor and outdoor RSRP distribution obtained from GMM parameters and original data RSRP distribution.

TABLE 6. Indoor and outdoor parameters vectors.

TA	Indoor (p_0, μ_0, σ_0)	Outdoor (p_1, μ_1, σ_1)
1	(0.72, -94.2, 6.4)	(0.23, -72.5, 6.6)
2	(0.35, -105.9, 8.5)	(0.65, -86.0, 9.3)
3	(0.53, -103.6, 7.6)	(0.47, -85.0, 8.4)
4	(0.29, -111.1, 8.0)	(0.71, -91.3, 7.2)
5	(0.34, -110.7, 4.6)	(0.66, -98.6, 5.7)
6	(0.52, -101.1, 6.6)	(0.48, -90.1, 5.9)

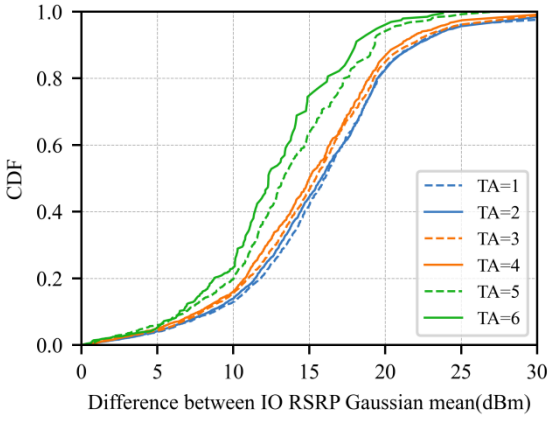

FIGURE 4. The CDF of difference between outdoor and indoor RSRP Gaussian mean.

In order to further illustrate the feasibility of the algorithm, we made statistics on the fitting results of GMM algorithm for 20000 sets, as shown in Figure 4, where x-axis is defined as the difference δ between outdoor and indoor RSRP Gaussian mean value, where $\delta = \mu_1 - \mu_0$, and y-axis is CDF value.

TABLE 7. IO identification accuracy of each dt data set.

Item	Building1	Building2	Building3	Building4
Indoor label	309	385	296	247
Indoor true predict	282	352	248	229
Outdoor label	352	331	323	284
Outdoor true predict	294	309	302	253
Accuracy	87%	92%	89%	91%

The larger the δ is, the more obvious the indoor and outdoor distribution is, and the more obvious the identifiability is. Figure 4 also shows that the distribution of δ becomes more compact with the increase of TA, which corresponds to the logarithmic function relationship between RSRP and distance. In addition, in all kinds of TA(distance) scenes, the proportion of δ greater than 10 is about 80%, which is consistent with the penetration loss results of indoor and outdoor signals in the same area in the actual environment. These results further demonstrate the feasibility of the algorithm.

Table 7 shows the IO identification accuracy of each DT set obtained through 4 buildings. The precision is defined as

$$P(y) = 1 - \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'| \tag{23}$$

where y_i and y_i' represent the label and algorithm prediction results respectively, N is the number of sample sets. It can be obtained that the accuracy of IO identification algorithm can reach about 90% in four buildings.

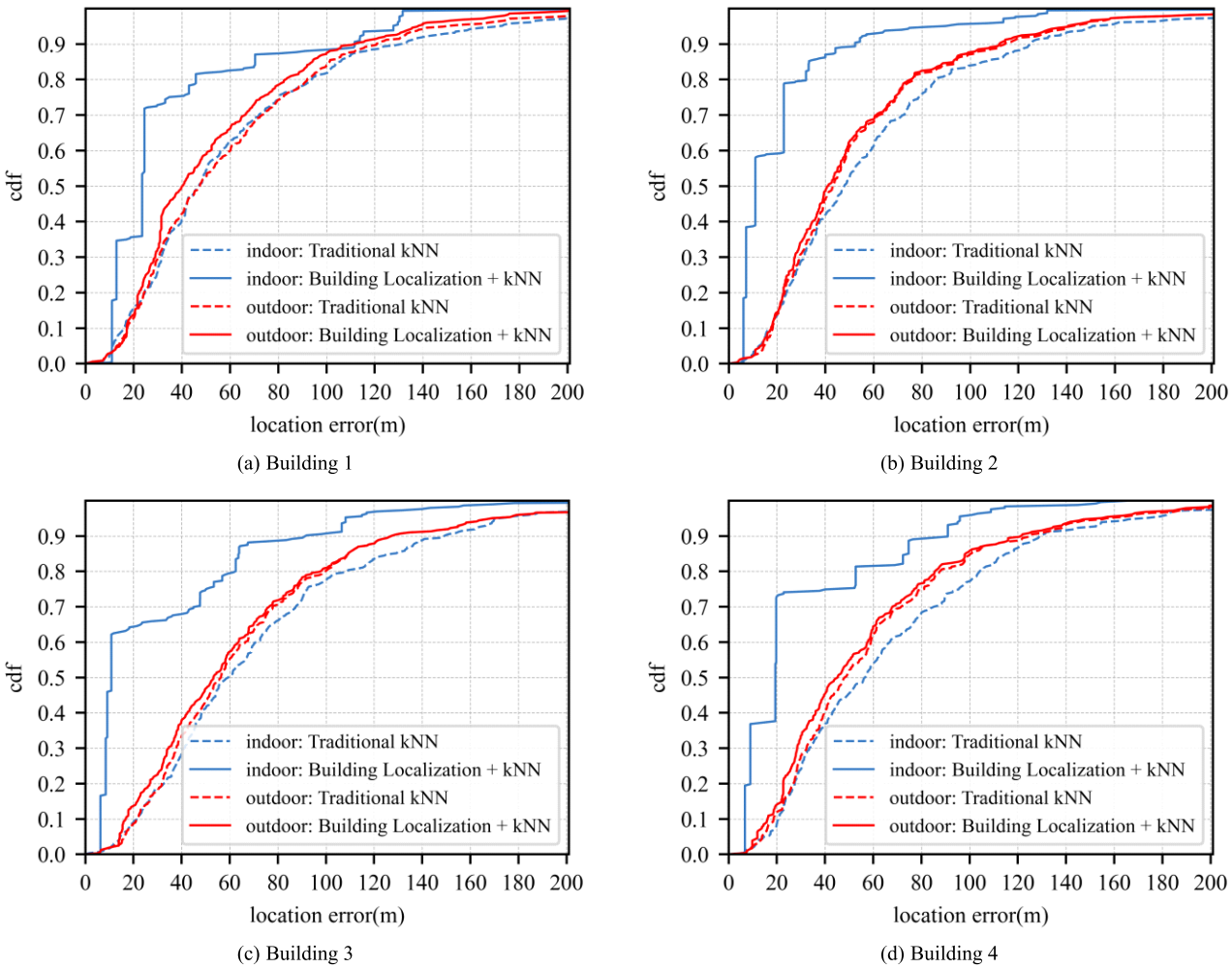

FIGURE 5. Location accuracy CDF comparison on datasets sampled inside buildings.

TABLE 8. Building predict accuracy.

Item	Building1	Building2	Building3	Building4
Indoor true predict	282	352	248	229
Building true predict	235	301	179	177
Recall ratio	73.3%	81.5%	72.2%	77.3%

C. INDOOR BUILDING LOCALIZATION

On the basis of IO identification, we continue to verify the performance of building localization algorithm. Indoor MDT data are located to a specific building, and the further building predict accuracy of indoor true predicted MDTs is shown Table 8.

The median of recall ratio of building true prediction is 75.3%.

We have treated the building localization problem as a supervised multi-classification problem. MDT classified indoor will be located to a building, but not the specific location of UE in the building. In order to evaluate the location accuracy, we use the plane center of the located building instead. The MDTs predicted outdoor are located by outdoor localization algorithm. Outdoor localization algorithm is not the focus of this article. We utilize an improved high-accuracy kNN localization method proposed in [7].

Then, compare the location accuracy CDF of the proposed full-scene localization system and the traditional outside localization system. Figure 5 shows the location accuracy CDF comparison of the two localization systems on the datasets sampled inside and near the 4 buildings separately.

Clearly, on the datasets sampled inside buildings, the proposed full-scene localization system with IO identification and building localization can improve the location accuracy significantly. Most of the test points sampled inside building can be judged as indoor and be located to buildings, which is represented as steps in the second cumulative distribution function (CDF) of each subfigure because the judged indoor points are located to the plane center of the building, rather than scattered points in the building. The remaining few points are misjudged as outdoor by IO identification algorithm and located using traditional k-NN algorithm, which is represented as smooth segments in the second CDF of each subfigure.

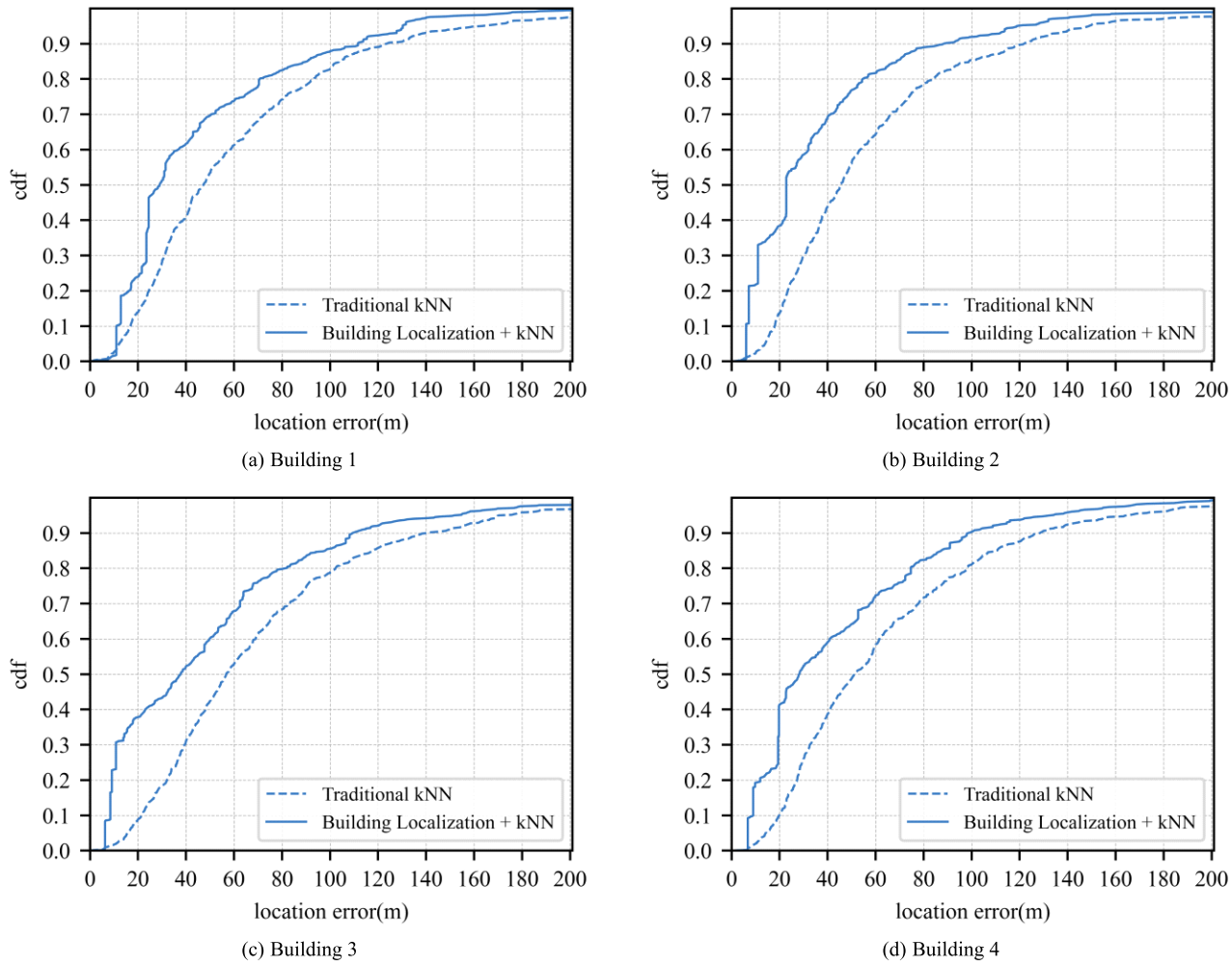

FIGURE 6. Location accuracy CDF comparison on complete datasets.

TABLE 9. Building predict accuracy.

Item	Location accuracy of proposed method(m)	Location accuracy of traditional method(m)
Building1	45.8	68.1
Building2	37.9	63.1
Building3	58.7	77.5
Building4	52.8	73.1
Mean	49.3	70.6

On the datasets sampled near buildings, the location accuracy is slightly improved because the GNSS MDT around the buildings is also used to train building classifier offline, most of the outside points misjudged as indoor by IO identification algorithm can be located to the nearest building.

Figure 6 shows the location accuracy CDF comparison on the joint datasets sampled inside and near the 4 buildings. Table 9 lists the location accuracy on each building datasets.

The mean of location accuracy is 49.3m(@67%).

VI. CONCLUSION

In this paper, a commercial high-accuracy, low-cost and real-time indoor building-level localization system is proposed, which is applicable for locating the MDT data in the LTE cellular communication network system. The system optimizes the traditional fingerprint localization algorithm, which utilizes low cost and easily collected MDT data to assist indoor localization, instead of manually collecting and maintaining high cost indoor DT data. In order to compensate for the loss of location accuracy, we innovatively divide the online process into two phases: IO identification phase and indoor localization phase. A real-time and precise GMM-based unsupervised algorithm is applied to identifying if the non-GNSS MDT data is in indoor environment in IO identification phase. Then, a multi-class classification algorithm based on Bayesian classifier is used to locate indoor MDT data to the specific building. The results of experiments conducted in an in-service LTE network using more than 100 LTE base stations demonstrate that the proposed technique yields a IO identification accuracy of 90% and an indoor location accuracy of 49.3m(@67%) respectively, which provides at least 30.2% enhancement in location accuracy compared to traditional technology without DT FD.

It is worth mentioning that since the concepts of penetration loss, TA, RSRP and MDT are applicable to LTE and NR,

our indoor localization system is also applicable to NR to a certain extent.

For our future work, many issues need to be further investigated. In order to meet the higher location accuracy requirements of different services, we will explore a more accurate indoor localization algorithm to locate UE to specific building floor without using DT FD. For the case of outdoor UEs access to indoor micro station caused by the mobility configuration, we will study the IO identification and indoor localization algorithm in the case of accessing indoor micro station. Considering that GNSS MDT naturally contains outdoor tags, we are trying to apply partially supervised classification algorithm to IO identification phase, and compare the accuracy with the algorithm in this paper. In addition, we are also working on extending our system to different scenes, such as high-speed train scene, trajectory matching algorithm, etc. We will also study the compatibility of our localization system with 5G NR.

REFERENCES

[1] *Hadoop*. [Online]. Available: https://hadoop.apache.org
[2] *Spark*. [Online]. Available: https://spark.apache.org
[3] *Storm*. [Online]. Available: https://storm.apache.org
[4] *Flink*. [Online]. Available: https://flink.apache.org
[5] *Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA);Radio measurement collection for Minimization of Drive Tests (MDT); Overall description;Radio Measurement Collection for Minimization of Drive Tests (MDT)*, document TS 37.320 V15.0.0, 3GPP, Jun. 2018.
[6] J. Turkka, T. Hiltunen, R. U. Mondal, and T. Ristaniemi, "Performance evaluation of LTE radio fingerprinting using field measurements," in *Proc. Int. Symp. Wireless Commun. Syst.*, Aug. 2015, pp. 450–466.
[7] *NG Radio Access Network (NG-RAN); Stage 2 functional specification of User Equipment (UE) positioning in NG-RAN*, document TS 38.305 V16.1.0, 3GPP, 2020.7.
[8] W. Fang and B. Ran, "An accuracy and real-time commercial localization system in LTE networks," *IEEE Access*, vol. 8, pp. 120160–120172, 2020.
[9] B. Uehrer and R. R. Michael, *Overview of Global Positioning Systems*. Hoboken, NJ, USA: Wiley, 2019, pp. 655–705.
[10] L. Shao, S. Yang, H. Liu, and J. Li, "Research on location method of pipe climbing robot based on gyroscope," in *Proc. IEEE Int. Conf. Mechatronics Automat.*, Aug. 2018, pp. 238–242.
[11] D.-D. Nguyen, A. Elouardi, S. A. Rodriguez Florez, and S. Bouaziz, "HOOFR SLAM system: An embedded vision SLAM algorithm and its hardware-software mapping-based intelligent vehicles applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4103–4118, Nov. 2019.
[12] K. Guan, L. Ma, and X. Tan, "Vision-based indoor localization approach based on SURF and landmark," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Sep. 2016, pp. 1–5.
[13] T. Koike-Akino, "Fingerprinting-based indoor localization with commercial MMwave WiFi: A deep learning approach," *IEEE Access* vol. 8, pp. 84879–84892, 2020.
[14] F. Wang, J. Feng, and Y. Zhao, "Joint activity recognition and indoor localization with WiFi fingerprints," *IEEE Access* vol. 7, pp. 80058–80068, 2019.
[15] Z. Lingyan, "DeFi: Robust training-free device-free wireless localization with WiFi," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8822–8831, Sep. 2018.
[16] X. Ye, X. Yin, X. Cai, A. Perez Yuste, and H. Xu, "Neural-Network-Assisted UE localization using radio-channel fingerprints in LTE networks," *IEEE Access*, vol. 5, pp. 12071–12087, 2017.
[17] L. Ma, N. Jin, Y. Zhang, and Y. Xu, "RSRP difference elimination and motion state classification for fingerprint-based cellular network positioning system," *Telecommun. Syst.*, vol. 71, no. 2, pp. 191–203, Jun. 2019.
[18] H. Liu, Y. Zhang, and X. Su, "Mobile localization based on received signal strength and Pearson's correlation coefficient," *Int. J. Distrib. Sensor Netw.*, vol. 15, pp. 1–10, Aug. 2015.
[19] M. Ibrahim and M. Youssef, "CellSense: An accurate energy-efficient GSM positioning system," *IEEE Trans. Veh. Technol.*, vol. 61, no. 1, pp. 286–296, Jan. 2012.
[20] *A 2016 Global Research Report On The Indoor Positioning Market*. [Online]. Available: https://www.indooratlas.com/wp-content/uploads/2016/09/A-2016-Global-Research-Report-On-The-Indoor-Positioning-Market.pdf
[21] E. Carvalho, B. S. Faical, G. P. R. Filho, P. A. Vargas, J. Ueyama, and G. Pessin, "Exploiting the use of machine learning in two different sensor network architectures for indoor localization," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2016, pp. 1–7.
[22] G. P. R. Filho, L. A. Villas, V. P. Gonçalves, G. Pessin, A. A. F. Loureiro, and J. Ueyama, "Energy-efficient smart home systems: Infrastructure and decision-making process," *Internet Things*, vol. 5, pp. 153–167, Mar. 2019.
[23] G. P. R. Filho, L. A. Villas, H. Freitas, A. Valejo, D. L. Guidoni, and J. Ueyama, "ResiDI: Towards a smarter smart home system for decision-making using wireless sensors and actuators," *Comput. Netw.*, vol. 135, pp. 54–69, Apr. 2018.
[24] F. Zhu, C. Luo, M. Yuan, Y. Zhu, Z. Zhang, T. Gu, K. Deng, W. Rao, J. Zeng, "City-scale localization with Telco big data," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 1–5.
[25] A. Shokry and M. Torki, "Youssef: DeepLoc: A ubiquitous accurate and low-overhead outdoor cellular localization system," in *Proc. SIGSPATIAL/GIS*, 2018, pp. 339–348.
[26] N. Wu, Y. Xiong, H. Wang, and J. Kuang, "A performance limit of TOA-based location-aware wireless networks with ranging outliers," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1414–1417, Aug. 2015.
[27] F. Benedetto, G. Giunta, and E. Guzzon, "Enhanced TOA-based indoor-positioning algorithm for mobile LTE cellular systems," in *Proc. Workshop Positioning Navigat. Commun.*, 2011, pp. 137–142.
[28] Z. Wei, Y. Zhao, X. Liu, and Z. Feng, "DoA-LF: A location fingerprint positioning algorithm with millimeter-wave," *IEEE Access*, vol. 5, pp. 22678–22688, 2017.
[29] Y. Zheng, M. Sheng, J. Liu, and J. Li, "Exploiting AoA estimation accuracy for indoor localization: A weighted AoA-based approach," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 65–68, Feb. 2019.
[30] Y. Li, G. Shi, X. Zhou, W. Qu, and K. Li, "Reducing the site survey using fingerprint refinement for cost-efficient indoor location," *Wireless Netw.*, vol. 25, no. 3, pp. 1201–1213, Apr. 2019.
[31] T. Hiltunen, J. Turkka, R. Mondal, and T. Ristaniemi, "Performance evaluation of LTE radio fingerprint positioning with timing advancing," in *Proc. 10th Int. Conf. Inf., Commun. Signal Process. (ICICS)*, Dec. 2015, pp. 1–8.
[32] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures," *Neural Comput.*, vol. 8, no. 1, pp. 129–151, Jan. 1996.
[33] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, nos. 2–3, pp. 131–163, 1997.
[34] D. Heckerman, "Bayesian networks for data mining," *Data Mining Knowl. Discovery* vol. 1., no. 1, pp. 79–119, 1997.
[35] *Technical Specification Group Radio Access Network; Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TS 38.901 NR; V16.1.0, 3GPP, Dec. 2019.

WEI FANG (Member, IEEE) graduated from the Wuhan University of Technology. He did Joint Ph.D. Training with the University of Wisconsin–Madison, Madison, WI, USA, for a period of two years. He is currently working as a Lecturer with the School of Automation, Wuhan University of Technology.

CHANGJUN XIE (Member, IEEE) received the Ph.D. degree in vehicle engineering from the Wuhan University of Technology (WHUT), Wuhan, Hubei, China, in 2009.

From 2012 to 2013, he was a Visiting Scholar with UC Davis, Davis, CA, USA. He is currently a Professor with the School of Automation, WHUT. He is also the Vice Dean of the School of Automation, WHUT. He has published over 50 articles and over 40 articles are indexed by SCI or EI. His research interests include the control strategy of fuel cell vehicles, intelligent and connected vehicle, and vehicle control and optimization of new energy vehicles.

BIN RAN received the Ph.D. degree from the University of Illinois, Chicago, IL, USA, in 1993. He is currently a Professor with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA, and the Director of the Research Center for Internet of Mobility, Southeast University, Nanjing, China. He is one of the co-founders of the Chinese Overseas Transportation Association, where he was the first Chairman. He has authored or coauthored over 90 articles in international journals, including the *Transportation Science*, *Transportation Research: Part B*, and IEEE Access.