Description of the SODIndoorLoc dataset

1. Abstract

A supplementary open dataset for WiFi indoor localization (SODIndoorLoc) was created. It can be treated as a supplement of the UJIIndoorLoc dataset. It covers three buildings and multiple floors within corridors, office rooms and meeting rooms. The total covering area is about 8000 m^2. 1802 points with different locations are arranged, the number of reference points (RPs) and testing points (TPs) are 1630 and 272, respectively. 23925 samples are recorded, 21205 for training/learning, and 2720 for testing/validation. The dataset contains 3 kinds of scenes, office room, meeting room, and corridor. Hall and corridor are seamless in these buildings, so there is no distinction between the two scenes. 105 single-band and dual-band APs were pre-installed in the 3 buildings. Locations of these pre-installed APs and CAD drawings are provided. Considering differences in the number of samples and MACs in the training data, there are 9 training sheets and 5 corresponding testing sheets in the dataset. The sampling distance between two adjacent points is about 1.2 meters in two buildings, while the distance is about 0.5 meters in a three-story building. The proposed dataset can be used for clustering, classification, and regression to compare the performance of different indoor positioning applications based on WiFi fingerprint, e.g. high-precision positioning, building, floor or fine-grained scene identification, range model simulation, rapid construction of fingerprint datasets.

2. Contact

Jingxue Bi
bijingxue19@sdjzu.edu.cn
School of Surveying and Geo-Informatics
Shandong Jianzhu University
Fengming Road No.1000, Licheng District, Jinan City, 250101, Shandong Province, China

3. Cite Request

Supplementary Open Dataset for WiFi Fingerprint-based Indoor Localization

4. Dataset Information

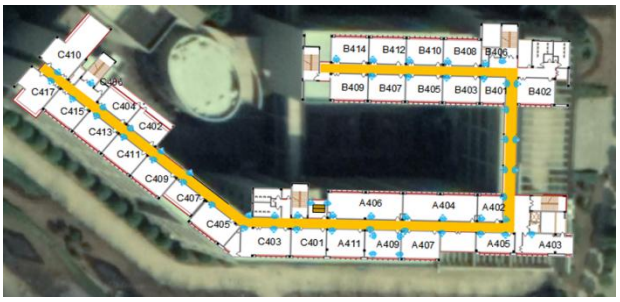
A. Description of the testing area

The total indoor area is about 8000 m^2, covering 3 buildings in different cities. As shown in Fig.1, floor plan with simple layout is overlaid on satellite image of each building. And all the WiFi collecting areas are rendered in orange color.

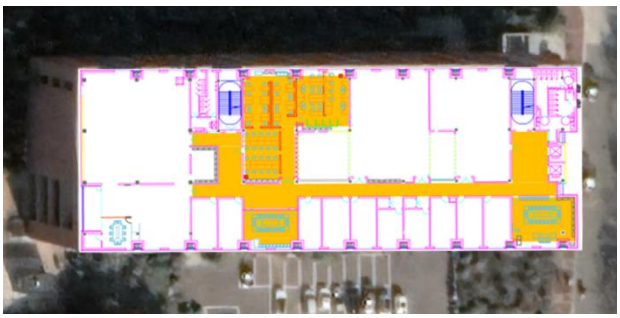
Clear floor plans are provided in 3 CAD files, which are named depending on buildings. Independent coordinate system is utilized in a CAD drawing. Coordinates of training and testing points are simply processed. Everyone could render their own floor plan by themselves.

The detailed information of 3 buildings can be found in the provided manuscript.

(a)



(b)



(c)

Fig. 1. Testing area covering 3 buildings, (a) CETC331 building with the 2nd floor, (b) HCXY building and (c) SYL building with the 4th floor

B. Description of the testing area

As the same as the UJIIndoorLoc dataset, the proposed SODIndoorLoc dataset also adopts sheets to store WiFi fingerprint records and other supplementary information. They are saved as comma separated value (CSV) files. The biggest differences from the other existing datasets are the addition of several attributes and a sheet file containing APs' location and transmission frequency.

1) The division of data sheets

There are nine training sheets and five testing sheets, as summarized in Table I. Only a pair of training and testing sheets about the the CETC331 building are provided, while there are 4 training sheets and 2 testing sheets for both the HCXY building and the SYL building.

The letter "All" indicates all detected APs are adopted to build a whole RSS vector, and the dimensions are shown in Table II. Number of samples in each building are summarized in Table III. 1630 RPs and 272 TPs with different locations are

arranged in each building, the summary of points' numbers in each building are shown in Table IV.

The letter "AP" denotes only pre-installed APs are utilized for filtering RSS vector. These simplified sheets contains less dimensions, shown in Table V.

TABLE I. SUMMARY OF TRAINING AND TESTING SHEETS

Buildings	**Training Sheets**	**Testing Sheets**
CETC331	Training_CETC331	Testing_CETC331
HCXY	Training_HCXY_All_30	Testing_HCXY_All
	Training_HCXY_All_Avg	
	Training_HCXY_AP_30	Testing_HCXY_AP
	Training_HCXY_AP_Avg	
SYL	Training_SYL_All_30	Testing_SYL_All
	Training_SYL_All_Avg	
	Training_SYL_AP_30	Testing_SYL_AP
	Training_SYL_AP_Avg	

The number "30" means that 30 samples at each RP are stored in the sheet, the letter "Avg" points out that the average of 30 samples at each RP is in the sheet.

TABLE II. DIMENSIONS OF RSS VECTOR IN DIFFERENT SHEETS

Statistic	**Buildings**			
	CETC331	***HXCY***	***SYL***	***ESTCE-TI***
dimension	52	347	363	520

TABLE III. NUMBER OF SAMPLES IN DIFFERENT SHEETS

Sheets	**Buildings**		
	CETC331	***HCXY***	***SYL***
training	955	11370	8880
testing	840	860	1020

TABLE IV. NUMBER OF POINTS IN DIFFERENT SHEETS

Sheets	**Buildings**		
	CETC331	***HCXY***	***SYL***
training	955	379	296
testing	84	86	102

TABLE V. DIMENSIONS OF RSS VECTOR IN SIMPLIFIED SHEETS

Statistic	**Buildings**		
	CETC331	***HCXY***	***SYL***
dimension	52	56	46

2) The format of training and testing sheets

Each WiFi fingerprint is characterized by the detected MACs and the

corresponding RSS. And most of wireless access points (WAPs) are with multiple bands. It is not appropriate to assign WAP as an attribute, as the UJIIndoorLoc dataset does. The detected MACs are utilized for identifying RSS values in the proposed dataset. Taking privacy reasons into account, all detected MACs in the whole building are sorted in the detected order and sequentially renamed to the combination of the letter "MAC" and the index in the order. For example, MAC*n* indicates the *n*th detected MAC and the *n*th attribute. Because numbers of detected MACs in 3 buildings are different, *n* means different values in training and testing sheets for three buildings.

If the number of all detected MACs is *n*, the range from the 1st attribute to the *n*th attribute can be expressed as MAC1, ... , MAC*n*. The $(n+1)^{th}$ and the $(n+2)^{th}$ attributes indicate coordinates in the east and north directions, named as ECoord and NCoord. Identifiers of floor level, building, scene, user, and phone are respectively indicated by from the $(n+3)^{th}$ attribute to the $(n+7)^{th}$ attribute, named as FloorID, BuildingID, SceneID, UserID, and PhoneID in sequence. The $(n+8)^{th}$ attribute indicates sample times, named as SampleTimes. The header of a sheet is shown as the first row of the Table VI. And the 2nd row of the Table VI is an example, i.e., the 12th sample of the Training_CETC331.

TABLE VI. THE HEADER OF A SHEET AND AN EXAMPLE, THE 12TH SAMPLE OF THE TRAINING_CETC331

1	**...**	***n***	***n*+1**	***n*+2**	***n*+3**	***n*+4**	***n*+5**	***n*+6**	***n*+7**	***n*+8**
MAC1		**MAC*n***	**ECoord**	**NCoord**	**FloorID**	**BuildingID**	**SceneID**	**UserID**	**PhoneID**	**SampleTimes**
– 44	...	100	47.4	18	1	1	1	4	3	1

3) RSS vector

RSS vector is a set of RSS values in the order corresponding to the first *n* attributes. MACs and corresponding RSS values can be obtained from the nearby APs in each scan. As long as the program aligns the scanned MACs to the first *n* attributes, and assigns the corresponding RSS values to the attribute values, then sets the attribute values of the undetected MACs to a certain value, e.g., 100 dBm, which is suggested by the UJIIndoorLoc dataset, a whole RSS vector can be obtained, as shown in Table VII. The scanned RSS value is a negative integer one, the unit is dBm, where – 100 dBm is equivalent to a very weak signal, whereas 0 dBm indicates an extremely good signal. The minimum RSS value in 3 buildings is – 104 dBm, a same value in the UJIIndoorLoc dataset. In subsequent operations, 100 dBm can be replaced with a very small negative number, e.g., – 105 dBm.

4) Local coordinates

The adopted coordinate system is local independent coordinate system. All coordinates in the proposed dataset are not consistent with those in CAD drawings. They have been transformed for privacy reasons. The unit of local coordinates is meter.

5) Space identifiers

FloorID, BuildingID, and SceneID are referred to as space identifiers. They are set as positive integer values from 1 to 4. FloorID ranges from 1 to 3 in CETC331 sheets, FloorID is 4 for HCXY and SYL sheets. BuildingID ranges from 1 to 3, the

CETC331, HCXY and SYL buildings are respectively set as 1, 2 and 3. There are 3 kinds of scenes in WiFi collecting area, corridor, office room and meeting room. And they are set as 1, 2 and 3 in sequence. Hall and corridor are seamless in these buildings, so there is no distinction between the two scenes. Space identifiers in the Table VII mean that the WiFi collection is in a corridor of the 1st floor of the CETC331 building.

6) User identifier

10 students participated in the WiFi RSS collection in three buildings. They are marked with numbers from 1 to 10 instead of names. The height of each user is provided. In addition, the coarse height of holding smartphone in user's hand is also supplied, as shown in Table VII. Because we think these information might be important in range model simulation and range-based localization. The unit of height is centimeter.

TABLE VII. INFORMATION OF USERS

UserID	Height	Height of phone	UserID	Height	Height of phone
1	165	109	6	176	125
2	179	132	7	178	127
3	174	123	8	177	125
4	171	116	9	182	134
5	158	101	10	181	131

7) Phone identifier

9 Android smartphones were utilized to collect WiFi data, where 2 phones were in same model. 3 bands were Xiaomi, Huawei and Samsung. The detail can be found in Table VIII.

TABLE VIII. PHONE IDENTIFIER

Model	Xiaomi 8	Xiaomi 11	Xiaomi 4	Huawei Mate8	Xiaomi 6	Samsung S7	Xiaomi 6	Redmi 4	Xiaomi 5X
ID	1	2	3	4	5	6	7	8	9

8) Sample times

Timestamp register was introduced in the UJIIndoorLoc dataset in Unix time format to represent the time of WiFi collection. But sample times is adopted in the proposed dataset to record the times of WiFi samples, ranging from 1 to 30. In the Training_CETC331 sheet and training sheets labeled by "Avg", the sample times is recorded as 1 at each sample. The sample times ranges from 1 to 30 in the training sheets labeled by "30". The sample times ranges from 1 to 10 for all the testing sheets.

9) Information of pre-installed APs

Table IX is an example of information of a pre-installed AP in the SYL building, it mainly contains space locations, MAC addresses and channel frequencies. A sheet file is provided about the information of pre-installed APs for each building. These information are vital for range model simulation and range-based localization. Space locations are recorded as the format of RPs and TPs by using ECoord, NCoord and FloorID. MAC addresses are replaced by corresponding attributes. Channel frequency

is the central frequency of WiFi channel, which can reflect whether the signal belongs to the 2.4 GHz band or 5 GHz band. The unit of channel frequency is MHz. It is note that the sheet of the HCXY building doesn't contain the last two columns, due to the pre-installed APs are single-band. The height of an AP is not provided, and the height can be customized.

TABLE IX. EXAMPLE OF INFORMATION OF PRE-INSTALLED APS

ID	ECoord	NCoord	FloorID	Attribute_2.4	Frequency_2.4	Attribute_5	Frequency_5
1	50.6	12.6	4	MAC125	2437	MAC340	5220