UJIIndoorLoc: A New Multi-building and Multi-floor Database for WLAN Fingerprint-based Indoor Localization Problems

Joaquín Torres-Sospedra*, Raúl Montoliu*, Adolfo Martínez-Usó†, Joan P. Avariento*,
Tomás J. Arnau*, Mauri Benedito-Bordonau*, and Joaquín Huerta*
*Institute of New Imaging Technologies, Universitat Jaume I, Avda. Vicente Sos Baynat S/N, 12071, Castellón, Spain.
†Dpto. de Sistemas Informáticos y Computación, Universitat Politècnica de València, Valencia, Spain.

***Abstract*—Although indoor localization is a key topic for mobile computing, it is still very difficult for the mobile sensing community to compare state-of-art localization algorithms due to the scarcity of databases. Thus, a multi-building and multi-floor localization database based on WLAN fingerprinting is presented in this work, being its public access granted for the research community. The here proposed database not only is the biggest database in the literature but it is also the first publicly available database. Among other comprehensively described features, full raw information taken by more than 20 users and by means of 25 devices is provided.**

I. Introduction

Many real world applications need to know the localization of a user in the world to provide their services. Therefore, automatic user localization has been a hot research topic in the last years. Automatic user localization consists of estimating the position of the user (latitude, longitude and altitude) by using an electronic device, usually a mobile phone. Outdoor localization problem can be solved very accurately thanks to the inclusion of GPS sensors into the mobile devices. However, indoor localization is still an open problem mainly due to the loss of GPS signal in indoor environments.

A spectacular growth of indoor localization studies has been witnessed during the last decade (see Figure 1), and the WLAN-based ones is the basis for many indoor localization approaches. This is mainly due to the proliferation of both wireless local area networks (WLANs) and mobile devices. Nowadays WLANs can be found anywhere, and mobile phones have increasingly become an indispensable part of our daily lives and, therefore, we can safely expect that the user is at the same location than the mobile device. The last generation of these devices (also known as smartphones) not only provides programmable abilities but they carry embedded sensors [1] like GPS, accelerometer, gyroscope, microphone, camera, bluetooth, etc. which have even been used to study social interactions [2] or predict human behavior [3] among many other studies.

WLAN Fingerprint-based positioning systems are based on the Received Signal Strength Indicator (RSSI) value. Commonly, two phases are needed: calibration and operation [5]. In the calibration phase, a radio map of the area where the users should be detected is constructed. Later, during the operational phase, a user obtains the signal strength of all *visible* access points of the WLAN that can be detected from his/her position and creates a *test* sample. This sample is sent to the server to be compared with the *training* samples of the radio map. Basically, the user's location corresponds to the position associated with the most similar sample in the radio map.

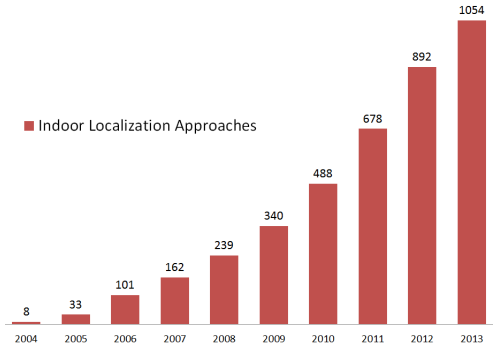

Fig. 1. Timeline and number of research works records on indoor localization from 2004 to 2013 (records collected in September 2013 [4]).

One of the major advantages of the WLAN fingerprint-based methods is that they do not require the installation of any additional hardware since they use the existing WLAN infrastructure. Therefore, the location of the user can be obtained without additional infrastructures and costs. However, WLANs were not natively designed to support a positioning function. Taking into account the existing obstacles introduced by the indoor environment (including reflections and multi path interference) the spread of radio signal in indoor environments is very hard to predict [5]. In addition, in WLAN-based positioning systems, the user typically carries the mobile device with him/her, being his/her motion or how the device is carried an important factor that affects the measured RSSI values [6].

Although there are many papers in the literature trying to solve the indoor localization problem using a WLAN fingerprint-based method, there still exists one important drawback in this field which is the lack of a common database for comparison purposes. Each approach presents its estimated results using its own database and describes how the experiment was carried out. Under these conditions, it is not possible to compare different methods since the particularities of each experiment are hardly reproducible. In the Pattern Recognition and Machine Learning research fields, the common practice is

to test the results of each proposal either using a well-known dataset or providing the dataset used. In this way, researchers are able to fairly compare different methodologies in the literature. For instance, the *UCI Machine Learning Repository*[1] is a well-known example [7] in this sense. However, in the WLAN fingerprint-based indoor localization field does not exist such kind of database.

In this paper, the *UJIIndoorLoc* database is presented to overcome this gap. We expect that the proposed database will become the reference database to compare different indoor localization methodologies. As far as we know, the proposed database is the first public accessible database in this field and researchers can access to the database following this url[2].

The main contribution of this work is the creation and the presentation of the *UJIIndoorLoc* database which is the biggest database in the literature as it was previously mentioned. It would also be the first publicly available database that could be used to make comparisons among different methods in this field. The main characteristics of the database[3] are:

- It covers a surface of $108703m^2$ including 3 buildings with 4 or 5 floors depending on the building.
- The number of different places (reference points) appearing in the database is 933.
- 21049 sampled points have been captured: 19938 for training/learning and 1111 for validation/testing.
- Dataset independence has been assured by taking Validation (or testing) samples 4 months after Training ones.
- The number of different wireless access points (WAPs) appearing in the database is 520.
- Data were collected by more than 20 users using 25 different models of mobile devices (some users used more than one model).

Two Android applications have been used to create the database *CaptureLoc* and *ValidationLoc*. Both applications use as a reference map services that are published in ArcGIS server. These services contain the geographic information of the building interiors as well as the training reference points localization. Using these services, the applications created show the maps to improve the user localization for training and validation. Data were collected at three multi-floor buildings of the Jaume I University [4] (UJI).

The rest of the paper has been organized as follows: Section II presents the related work. Section III explains in detail the proposed database. How this database has been created is commented in Section IV. Section VI describes the most important challenges we contemplate using the proposed database. Finally, in Section VII some conclusions are given.

[1] http://archive.ics.uci.edu/ml/

[2] http://www.geotec.uji.es/ujiindoorloc-database/

[3] The first version of the database only covers 3 buildings, our idea is to cover all the university facilities (30 buildings), so the final version of the database will be approximately 10 times higher.

[4] http://smart.uji.es/

II. Related work

Indoor positioning and localization literature is vast. In [8], authors categorised approaches according to the technique used for localization into several paradigms, including calibration-free localization [9], WLAN based techniques [10], Dead-reckoning [11], simultaneous localization and mapping (SLAM) [12] and multi-modal sensing [13]. Another classification can be found in [14], where fingerprint-based indoor localization has been particularly classified into two categories: *infrastructure-based* and *infrastructure-less* approaches. Infrastructure-based approaches rely on the deployment of customized Radio-Frequency beacons (RFID, infrared, ultrasound, bluetooth, led lights, etc.) that can be carefully optimized for a particular purpose. The main drawback of these approaches is that they need their own customized hardware. However, infrastructure-less approaches use the already available wireless signals to profile a location, taking advantage of the powerful mobile phones sensors. Our work is an infrastructure-less approach since we use the already available WLAN access points (WAPs) to construct the database by using mobile phones.

There are also many works dealing with the indoor localization problem by using WLAN-based techniques. Table I shows the main characteristics of the data used in some of the most important papers in this field. In [5] a new fingerprint-based method was proposed which uses a previously stored map of the signal strength at several positions and determines the position using similarity functions and majority rules. According to the authors, their proposed method is able to obtain high rates determining the building, the floor and the place, with an average error around 3 meters. The database used in [5] has (see Table I) 9358 sample points taken from 2 buildings of 3 floors each one, with 101 different WAPs in the database. However authors do not provide information about the number of users or the number of different devices used to capture the samples. Information about the covered surface was also not provided. The database used in [5] was the biggest one in the previous literature but it is not public and it has 55% less samples than the one we propose here, 80% less number of WAPs and 60% less places, among other data.

TABLE I. Main characteristics of the data used in some of the most important papers in the indoor localization field: *Number of buildings* (N_B), *Surface* (Surf.), *Number of floors* (N_F), *Number of places* (N_P), *Number of Samples* (N_S), *Number of WAPs* (N_W), and *Number of Devices* (N_D). N/A stands for information not available or not provided by authors.

Work	N_B	$Surf.$	N_F	N_P	N_S	N_W	N_D
[5]	2	N/A	3	392	9358	101	N/A
[15]	1	N/A	1	96	2880	206	2
[16]	1	N/A	N/A	N/A	N/A	9	N/A
[17]	1	$980m^2$	1	50	N/A	N/A	N/A
[14]a	3	$2730m^2$	3	120	N/A	434	1
[14]b	3	$69000m^2$	1	13	N/A	379	1
Our	3	$108703m^2$	4/5	933	21049	520	25

In [15] a different approach that uses only the rankings of the RSSI values is used. Authors argue that their method is better to avoid the well known problem of having hardware and software differences between user devices, that produces

that the RSSI reported by the current mobile device may differ from the RSSI in the database, and therefore this can degrade the positioning accuracy. According to the information provided by authors, the database used is also significantly smaller than the proposed in [5]. For instance, the number of WAPs, which is the only criterion in which work is superior to [5], is still 60% less than ours. In [16] two improvements are proposed to the common way used to solve the WLAN-based fingerprint problem. The first issue is for differing antenna attenuation among different devices. The second is for dealing with environments where not every beacon is visible everywhere. Although the methods proposed in [15] and in [16] are both promising, since they try to solve some of the common problems in this field, the databases used (see Table I) are very small and are different.

Other WLAN-based works as [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33] have been not included in Table I since they do not provide information about the characteristics of the database or the information provided shows that their databases are very small with respect to the one proposed in this paper or even the one used in [5]. The last three rows of the Table I show information about the databases used in two papers [17], [14] which are not WLAN-based. [17] uses a RF-based method, and not all information about the database is available. [14] proposed a FM signal-based method. They performed experiments in two scenarios named as [14]a and [14]b in Table I. In both cases, the authors tested their proposed FM-based method together with a WLAN-based methodology. The database of the second scenario has similar size to the one proposed in this paper, but it is still smaller than ours. For instance, it covers 37% less surface, and the number of WAPs is 28% smaller. In addition, just one device were used in their experiments.

In general, each indoor localization work used their own data to publish the results, and therefore it is quite difficult to make comparison among the proposed methods. A common database is needed to perform this task. As it has been commented before, in the Pattern Recognition and Machine Learning research fields, it is a common practice to share databases to allow other researchers to provide comparable results when using their proposals. Up to date, WLAN-based indoor localization methods have relatively small databases which in some cases, come just from one building being often captured using a small number of devices or by a small number of users. The database presented in this paper overcomes all these lacks.

III. *UJIIndoorLoc* DATABASE DESCRIPTION

In this section, the proposed database is completely described. First, Section III-A provides details about the information published for each capture (i.e. for each sampled point or record). Second, Section III-B shows how the whole database has been split into sets for training and validation purposes.

A. Description of elements stored

As it was previously introduced, the whole proposed database contains 21049 records. Each record is directly related to a single capture and it contains the following 529 numeric elements:

001–520	RSSI levels
521–523	Real world coordinates of the sample points
524	BuildingID
525	SpaceID
526	Relative position with respect to SpaceID
527	UserID
528	PhoneID
529	Timestamp

As an example, Table II shows an extract of one record. Due to the length of this record, only two RSSI levels are shown and being the values of the spatial coordinates truncated to one decimal. The example corresponds to the 7754th record from the training set.

1) RSSI Levels: The most important information for WLAN fingerprinting comparison purposes are the WAPs detected and their RSSI level values. In the proposed database, this information represents the 98% of the data given in each record (520 vector positions out of 529) as a 520-element vector of integer values. These values represent the RSSI levels whereas the WAP identifiers (MAC addresses) are linked to the vector positions. Table III expands the example record given in Table II showing its RSSI levels. This representation has been adopted due to the number of different WAPs detected on the three buildings (that is, 520) and the fact that Android provides integer RSSI levels.

The method *getScanResults()* included in WifiManager[5] Android class has been used to obtain the list of detected WAPs from each localization in each capture. This list, shown in Table IV, contains the MAC address and the corresponding intensity level for any detected WAP. The MAC addresses are coded as strings, and the RSSI levels correspond to negative integer values[6] measured in dBm, where $-100dBm$ is equivalent to a very weak signal, whereas $0dBM$ means that the detected WAP has an extremely good signal. Although the list shown in the example is sorted according to the intensity value, this ordering depends mostly on the device (model and Android version).

The database includes 520 WAPs identified by the MAC address. These addresses have been alphabetically sorted and sequentially renamed to WAP_{nnnn}. We use these new identifiers instead of the MAC addresses due to privacy reasons.

A total number of 520 WAPs appear in the database and the 520-element vector from each record contains the raw intensity levels of the detected WAPs from a single WiFi scan. Obviously, not all the WAPs are detected in each scan. For instance, only 14 WAP identifiers were detected in the scan example shown in Table IV. The RSSI levels for these WAPs remain unaltered, using the artificial value $+100dBm$ by default in those WAPs that have not been detected by the device.

It is important to mention how the RSSI values are distributed in the proposed database. Figure 2 introduces the

[5]http://developer.android.com/reference/android/net/wifi/WifiManager.html
[6]http://developer.android.com/reference/android/net/wifi/ScanResult.html

TABLE II. EXAMPLE OF ONE DATABASE ENTRY (7754-TH RECORD). IT WAS CAPTURED ON JUNE, 4TH 2013 (12:02:22PM GMT+02) BY USER11 WITH A *HTC Wildfire S A510e* (ANDROID VERSION 2.3.5). THE DEVICE DETECTED 14 WAP (NEGATIVE RSSI VALUES) ON THE REFERENCE POINT LOCATED OUTSIDE OFFICE 111 ON THE THIRD FLOOR OF THE *TI* BUILDING.

[1]	...	[520]	[521]	[522]	[523]	[524]	[525]	[526]	[527]	[528]	[529]
WAP_{001}	...	WAP_{520}	Longitude	Latitude	Floor	BuildingID	SpaceID	Rel.Pos.	UserID	PhoneID	Time
-97	...	+100	-7594.7...	4864983.9...	3	0	111	2	11	13	1370340142

TABLE III. EXTRACT OF THE VECTOR THAT REPRESENTS THE RSSI VALUES. MAC ADDRESSES HAVE BEEN ANONYMIZED DUE TO PRIVACY REASONS.

WAP_{001}	...	WAP_{031}	WAP_{032}	WAP_{033}	WAP_{034}	WAP_{035}	WAP_{036}	...	WAP_{520}
-97	...	+100	-97	+100	+100	-65	-65	...	+100

TABLE IV. EXTRACT OF THE WAPS LIST WITH 14 RSSI VALUES PROVIDED BY *getScanResults()*.

pos. in list	WAP Identifier	RSSI level
1^{st}	WAP_{032}	$-97dBm$
2^{nd}	WAP_{001}	$-97dBm$
3^{rd}	WAP_{268}	$-97dBm$
4^{th}	WAP_{150}	$-94dBm$
	...	
11^{th}	WAP_{036}	$-65dBm$
12^{th}	WAP_{035}	$-65dBm$
13^{th}	WAP_{142}	$-48dBm$
14^{th}	WAP_{143}	$-46dBm$

frequency distribution of all the individual intensity levels recorded in our database (374234 intensity values). Although these levels range from $-104dBm$ to $0dBm$, the number of individual measures inside range $[-45dBm...0dBm]$ is insignificant and cover only 1.7% of total RSSI levels recorded in our database. Similarly, the number of single measures reporting a value lower than $-95dBm$ is also insignificant (2.3% of total RSSI measures).

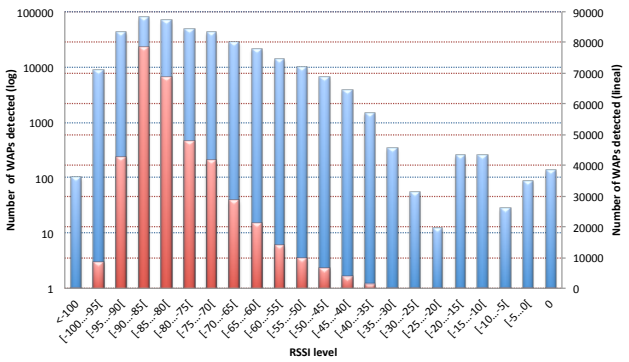

Fig. 2. Frequency distribution of the number of times that a RSSI value appears in the proposed database. Red bars stand for the values in linear scale (right scale) and blue bars stand for the values in logarithmic scale (left scale).

Figure 3 shows the number of WAPs detected in a single capture. This number ranges from 0 (where there is not any WiFi coverage) to 51. So, localizations with no coverage have not been removed from the database. Finally, the average number of WAPs scanned in each capture is 17.92, therefore, approximately 500 elements of the previously described vector contain out of range values (represented as $+100dBm$). It is worth saying that, according to our experiments, the main factors that affect to the number of WAPs reported by a WiFi scan are the location, the phone model (Android version and hardware) and how the device is held, although we are aware that there exist other factors as is demonstrated in [6], [5].

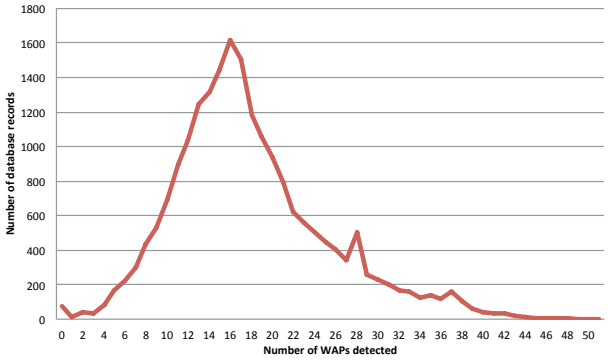

Fig. 3. Frequency distribution of the Number of WAPs that are detected on a single capture.

2) Real-world coordinates: Real-world coordinates are represented in each sample/capture by means of three values in each record (vector positions from 521 to 523), the longitude and latitude coordinates (in meters with UTM from WGS84) and the floor of the building. An example of these values is shown in Table V.

TABLE V. DETAILED COORDINATES AND FLOOR OF THE PLACE WHERE THE STORED CAPTURE ON THE 7754TH RECORD WAS TAKEN.

Longitude	Latitude	Floor
-7594.736999999732	4864983.902400002	3

3) Space identifiers: BuildingID, vector position 524, is an integer value (from 0 to 2) that corresponds to the building in which the capture was taken. Figure VI shows: the UJI University campus (left); the three buildings of the School of Technology and Experimental Sciences (center image), hereafter ESTCE; and a zoom inside the third floor of the *TI* building (right). This figure has been introduced to show where the example point (Table V) is exactly located. Table VI shows the *BuildingID* for each building of the ESTCE.

TABLE VI. RELATION BETWEEN *BuildingID* AND THE REAL BUILDING.

Building ID	Real Building
0	ESTCE - TI
1	ESTCE - TD
2	ESTCE - TC

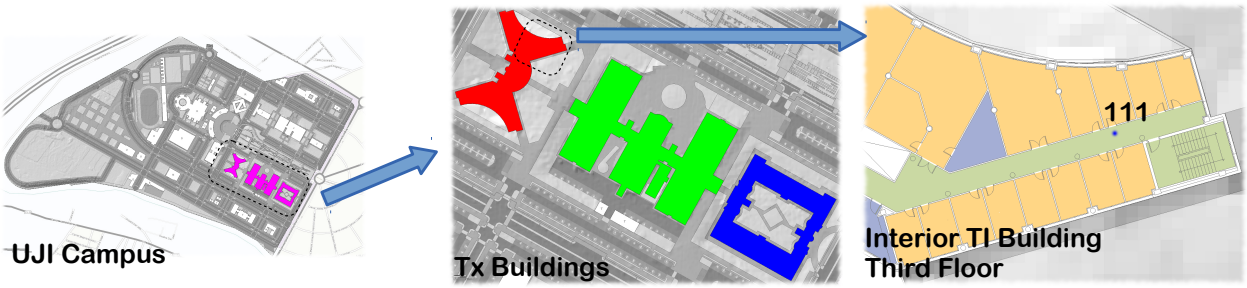

Fig. 4. Map of the UJI Riu Sec Campus and zoom on the Tx Buildings. Pink refers to the *ESTCE - Tx* building on the UJI Campus map (left). On the *Tx* building zoom (right): red refers to *TI* building, green corresponds to *TD* building and blue stands for *TC* building. On the interior of *TI* building, the blue point is the reference point.

The position 525 of the vector that represents each sample is called *SpaceID* and contains a single integer value that, in this case, is used to identify the particular space (offices, labs, etc.) where the capture was taken. The *relative position with respect to the space* is also provided in the 526 position and it denotes if the capture was taken inside (value 1) or outside (value 2) the space at the corridor. Outside means in front of the door of the space. Following with our previous example, the values of these fields on the 7754-th record are shown in Table VII. According to these values, the capture was taken on the reference point located outside office 111 at *ESTCE - TI* building (*TI*).

TABLE VII. REFERENCE POINT IN WHICH THE 7754TH RECORD WAS TAKEN.

Floor	BuildingID	SpaceID	Rel.Pos.
3	0	111	2

As Section III-B describes, the database is split into two subsets: the training subset and the validation subset. In the training subset, the reference points are well-specified, being these points captured by, at least, two users. However, in the validation subset, the measures were taken at arbitrary points as would happen in a real localization system and, for this reason, these reference points (identified by *SpaceID* and *Relative position*) are not stored in the validation records. This fact is denoted by assigning the default value 0 to both fields.

4) User Identifier: *UserID*, position 527, contains an integer value that ranges from 1 to 18. This value is used to represent the 18 different users who participated in the procedure to generate the training samples. This field has not been recorded in the validation phase, so the *default value* 0 is used to denote it. The height of each user is also provided. This information could be useful because the concrete spatial position of the device has a direct impact on the measured RSSI values [6]. Table VIII shows the correspondence among *UserID*, anonymized user name and the user's height.

5) Phone Identifier: Similarly, *PhoneID* (position 528) contains an integer value to represent the Android device used in each capture. Table IX shows the correspondence between each *PhoneID* and its associated device (model and version). Moreover, the device used by each user is also shown. As mentioned before, users from USER_{0001} to USER_{0018} are those who participated in the procedure of generating the training set, whereas user USER_{0000} is used to denote that the device was used to generate the validation set.

TABLE VIII. CORRESPONDENCE AMONG *UserID*, THE ANONYMIZED USER WHO TOOK THE CAPTURE AND ITS HEIGHT.

UserID	Anonymized user	Height
0	USER_{0000} (Validation User)	-
1	USER_{0001}	170
2	USER_{0002}	176
3	USER_{0003}	172
4	USER_{0004}	174
5	USER_{0005}	184
6	USER_{0006}	180
7	USER_{0007}	160
8	USER_{0008}	176
9	USER_{0009}	177
10	USER_{0010}	186
11	USER_{0011}	176
12	USER_{0012}	158
13	USER_{0013}	174
14	USER_{0014}	173
15	USER_{0015}	174
16	USER_{0016}	171
17	USER_{0017}	166
18	USER_{0018}	162

As it can be noticed from Table IX, the number of different devices used is 20 (25 if the Android version is considered). There have been a few cases in which some users have the same device model. Concretely, three different users own a Nexus 4 and they participated in the procedure to create the validation set. Moreover, a *LT22i* phone has been shared by two users to generate the training samples.

6) Timestamp: Finally, we have also included a *Timestamp* register in the 529 position of the vector representing the time (in *Unix time* format) in which the capture was taken. This time was set by a centralized server to avoid outliers. The timestamps provided by each device is not recorded because the device's timing settings could be different and we could not trust on the time provided by them. In fact, wrong timing settings could add noise to our records or invalidate them. E.g, a capture taken at morning could be incorrectly stored as taken at evening. Another severe case occurs if the time has not been setup, so captures taken at 00:00 Jan 1st 2001 should not be recorded because the proposed database was done in 2013.

TABLE IX. CORRESPONDENCE BETWEEN *PhoneID* AND REAL DEVICE. REAL DEVICE'S INFORMATION INCLUDES THE MODEL DESCRIPTION AND ANDROID VERSION. USERS WHO EMPLOYED THE DEVICE ARE ALSO SHOWN.

PhoneID	Android Device	Android Ver.	UserID
0	Celkon A27	4.0.4(6577)	0
1	GT-I8160	2.3.6	8
2	GT-I8160	4.1.2	0
3	GT-I9100	4.0.4	5
4	GT-I9300	4.1.2	0
5	GT-I9505	4.2.2	0
6	GT-S5360	2.3.6	7
7	GT-S6500	2.3.6	14
8	Galaxy Nexus	4.2.2	10
9	Galaxy Nexus	4.3	0
10	HTC Desire HD	2.3.5	18
11	HTC One	4.1.2	15
12	HTC One	4.2.2	0
13	HTC Wildfire S	2.3.5	0,11
14	LT22i	4.0.4	0,1,9,16
15	LT22i	4.1.2	0
16	LT26i	4.0.4	3
17	M1005D	4.0.4	13
18	MT11i	2.3.4	4
19	Nexus 4	4.2.2	6
20	Nexus 4	4.3	0
21	Nexus S	4.1.2	0
22	Orange Monte Carlo	2.3.5	17
23	Transformer TF101	4.0.3	2
24	bq Curie	4.1.1	12

B. Database division

The whole database is split into two different sets: the *Training* set and the *Validation* set. On the one hand, the training set provides fully-detailed measures whose location corresponds to predefined reference points. On the other hand, the validation set provides the same information on arbitrary points. Table X shows the information about the two sets.

TABLE X. BASIC FEATURES OF BOTH DATABASE SUBSETS

	Training	Validation
Captures	19674	1111
WAPs	465	367
RSSI Range	$[-104\ldots0]dBm$	$[-102\ldots-34]dBm$
Ref. points	933	$None^{*}$
Users	18	$Unknown^{**}$
Devices	16	11

* There was not any established reference point for validation.

** The validation stage does not store the user id in order to be more realistic.

Although both the training subset and the validation subset contain the same information, the latter includes the value 0 in some fields. These fields are: *SpaceID*, *Relative Position with respect to SpaceID* and *UserID*. As it has been commented before, this information was not recorded because the validation captures were taken at arbitrary points and the users were not tracked in this phase. This fact tries to simulate a real localization system.

IV. HOW IT WAS MADE

As mentioned in Section III, the *UJIIndoorLoc* database is split into two independent subsets: the *Training* set, and the *Validation* set. This section introduces how the two sets were created.

A. Generating the training set

An Android application called *CaptureLoc* has been developed to capture records for the training set. This application collects all the required information (See Section III) and sends it to a centralized server which permanently stores it. This collect-and-send procedure is automatically repeated ten times for each captured location due to the harsh nature of the WLAN signals propagation [34]. Interaction user-device is required to select the user identifier and to select the place where the capture will be taken. Figure 5 shows an example of the user-device interaction.

Fig. 5. Screenshot of *CaptureLoc*. On the left, example where the capture is done (red circle). Button *Send Fingerprint* starts the collect-and-send procedure. On the right, the result of a capturing process that reports four errors.

To generate the training set, all the closed spaces of the three buildings (offices, laboratories, classrooms, WCs, among other spaces) have been initially considered as important places where the captures should be done. Then, one reference point inside each space and, at least, another reference point outside each space (i.e. at corridors) have been selected as reference points for all the considered closed spaces. The point inside the space is located at the centroid of the closed space, whereas the outside point is located in front of the door. If the space has multiple accesses, we have selected one reference point per entrance (door). Figure 6 shows a graphical example of how and where the reference points are located.

Then, 18 users performed the captures to generate the training set. The reference points were uniformly distributed to the users with the restriction that any reference point should be covered by, at least, two users. Any further suggestion, advice and/or direct order were not provided to the users, and they were free to capture the assigned reference points on their own way. Figure 5 (right) shows a capture process in which there were some errors (captures 5, 8, 9 and 10 were not recorded), so the user decided to repeat the process in the same reference point. The user was in charge of deciding if the capture procedure must be repeated or not. Errors on capturing are often related to low internet coverage (either 3G nor WiFi) and they have only been reported on a few places.

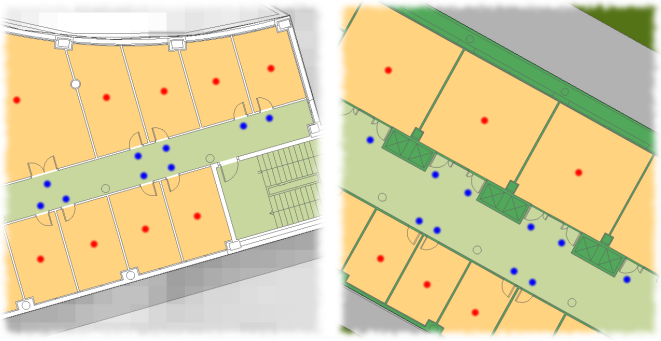

Fig. 6. Example of reference points located at the first floor of *TI* building (left) and third floor of the *TC* building (right). Red points corresponds to the reference points inside closed spaces, where blue points stands for the reference points taken in front of the door/doors (outside the spaces).

Finally, the 18 users have covered 924 reference points. In general, each reference point has been registered by, at least, two users, so more than 18400 training samples were expected to be recorded (19937 records were finally obtained). There have been a few cases in which the user has repeated the capture procedure due to connection errors on the first trial (Figure 5) and, moreover, there are some areas that have been covered by 3 users. Although all the suggested reference points located outside the spaces were captured, the users were not always allowed to capture inside some restricted spaces (chemical laboratories with biohazard labels, private offices, among other facilities).

B. Generating the validation set

Another Android application, *ValidateLoc*, has been developed to capture more points for validation purposes. This application performs the operation phase by sending the required information (only WAPs detected and RSSI levels) to a centralized server, and it gets a point inside a building (given by its longitude, latitude, floor) from the server. The localization is performed in the server side so the procedure to obtain the position from the fingerprint is totally transparent to the user. Then, the application asks the user if the provided localization is correct or not. If it is correct, the WiFi fingerprint and the successfully predicted localization are sent to the server and they are permanently stored. Otherwise, the application asks the user for the real localization and sends the information to be stored in the server side. Figures 7 and 8 show two execution examples of *ValidateLoc*.

To generate the validation samples, 14 users installed the application on their Android devices and executed it during 20 minutes (approximately) in each of the three *Tx* buildings. The users localized themselves with the application on their way, any advice/suggestion was not provided to them. Moreover, any reference point was not provided to them so they could capture WiFi signals in places which were not in the training phase. Despite last fact, users were close to reference points in most of the cases, so they were correctly located.

Finally, 1111 captures were recorded. Note that the *ValidateLoc* application only sends one validation capture after localizing the user.

V. *UJIIndoorLoc* BASELINE

In this section, the proposed database has been used with a basic indoor positioning system to provide a baseline for further comparisons. Note that the objective of this work is not to provide an accurate indoor positioning system, the objective is to provide an objective database which can be used for comparing positioning systems and other algorithms based on WLAN-fingerprinting.

We considered that the distance-based technique k-Nearest Neighbor (kNN) [35] can be used as baseline for comparison purposes. In particular, we have developed the 1NN technique ($k = 1$) in conjunction to the *Euclidean Distance* as a basic indoor localization system. The necessary steps to localize a current fingerprint are:

- The *Euclidean Distance* of the current fingerprint with respect to all the fingerprints included in the training set is calculated.
- The current fingerprint location corresponds to the Euclidean's closest training fingerprint location if there is only one candidate with the shortest distance.
- When some candidates provide the shortest distance, we apply a voting procedure to extract the "winning" building and floor. Then, the position corresponds to the average of the location provided by the Euclidean's closest training fingerprints that are on the winning building and floor. In case of tie, a localization error is raised.

So, the 1111 validation fingerprints were located using this simple positioning system and the results can be considered the baseline for this database. The results are shown in Table XI and include the *Success rate*, the *Error in positioning* and *Time*.

TABLE XI. UJIINDOORLOC RESULTS WITH 1NN IN CONJUNCTION WITH *Euclidean Distance*

Error in positioning	$7.9m$
Success rate	89.92%
Time	$495.26 \pm 0.54ms$

The *success rate* corresponds to the percentage of validation fingerprints correctly located inside the corresponding building and floor. The *error in positioning* is the average error in meters of the validation fingerprints correctly located inside the corresponding building and floor. The $Time$ field contains the average time in milliseconds required to obtain the precise location (longitude, latitude, floor) per fingerprint. $Time$ was calculated after repeating the experiments 20 times. The experiments were done by means of a non-optimized Matlab script run on a Intel's Core i7 based computer.

With the basic indoor location system the error is, in average, $7.9meters$ when the fingerprint has been located in the correct building and floor. In more than 10% of cases, the fingerprint was localized in other floor and/or building. Particularly, there were 3 out of 1111 cases (0.27%) in which the building was not correctly predicted, and 109 errors (9.81% of total cases) in detecting the correct floor. Other researchers can use the results shown here for further comparisons.

Fig. 7. Screenshots of the *ValidateLoc* application. The first image corresponds to the initial screen before localizing the user. The second one shows the localization and asks the user if the position is correct and the users presses the "yes" button. The last image warns the user that the validation fingerprint has successfully arrived to the server. Blue point stands for the predicted position and green one to the position assigned to the fingerprint.

Fig. 8. Screenshots of the *ValidateLoc* application. The first image shows again the localization and asks the user if the position is correct. In this case, the user presses the "no" button. The second image shows the screen in which the real position is introduced. The last image warns the user that the validation fingerprint has successfully arrived to the server. Blue point stands for the predicted position and the red one to the position assigned to the fingerprint.

VI. Discussion and Challenges

This database has been initially generated for indoor localization in our University Campus. Therefore, testing localization algorithms is the first use that external researchers can do with the published dataset. New indoor localization algorithms are often proposed using private datasets that are not publicly available, as it has been previously mentioned in sections I and II. Two different algorithms can be hardly compared only with the information and the results provided by the authors. This public dataset can be used to test the accuracy of any localization algorithm based on RSSI levels or for performing a comparison of localization algorithms under the same experimental framework.

However, the database is susceptible of being used on alternative problems that are discussed here. For instance:

- It could be of interest an analysis about how the internal structure of the building is related with the WLAN access points and, therefore, how the number and position of these points can be optimised without being out of WiFi range. This could obviously result in important savings in terms of hardware acquisition and installation efforts. In summary, a very desirable preprocessing step in any WiFi structure is to reduce the redundant access points keeping a complete coverage.
- WAPs location can be inferred with the databased provided. Although some WAPs are visible, the wide majority of them are hidden to human eye. Commonly, the WLAN antennas are located inside restricted areas or in the ceiling. The localization of the WAPs can strongly support the localization algorithm.
- Detection of low-coverage places such as the ones recorded in our dataset. On the one hand, adding new antennas can improve the localization algorithm because those places can not be localized by RSSI-based algorithms. On the other hand, it would improve the internet connection to those users who are on those places.
- Detection of WLAN collision places where some WAPs are emitting in the same channel, and WLAN connectivity may be degraded.
- The automatically detection of removed and new WAPs may be interesting since re-mapping could be avoided. The procedure to fully map a building requires planning, elaboration of a mapping strategy and working hours. The automatic detection may reduce the maintenance costs of fingerprint databases. In the introduced database, the validation fingerprints were taken 3 months later than the training ones, and some WAPs disappeared and new ones were introduced. From the 520 detected WAPs in the *UJIIndoorLoc* database, 312 of them were detected in training and validation phases. 153 WAPs were only detected in the training phase, and 55 new WAPs appeared in the validation phase.

Another practical applications about how devices detect the WAPs are:

- How two devices differ on obtaining the individual RSSI values on the same place. Table XII shows a summary of the RSSI values of the same WAP provided by two different devices. Although the range of possible values is similar for both devices, one of them tends to provide lower values according to the mean and median values.
- The study of anomalies in data such as the one detailed in Table XIII, where the RSSI values of the same WAP are shown for 4 different records located in the same place. One of the devices detected the highest possible value 0, but the same device detected a very low signal strength in the same place for the same WAP. This seems to be an anomaly in the data.
- According to our records, the number of RSSI scanned by a device depends on the environment and on the device itself. Table XIV shows an example of that, where the number of WAPs detected by two devices is shown. It is well known that not all devices detect the same number of WAPs. However, this database not only confirms this fact but a detailed analysis on this sense could be performed.

TABLE XII. SUMMARY OF THE RSSI VALUES OF WAP_{0034} PROVIDED BY TWO DIFFERENT DEVICES. MEASURES WERE TAKEN IN FRONT OF OFFICE 120 LOCATED ON THE FIRST FLOOR OF THE *TI* BUILDING.

PhoneID	min	max	mean	median
13	$-52dBm$	$-41dBm$	$-48.5dBm$	$-50.5dBm$
14	$-49dBm$	$-37dBm$	$-41.5dBm$	$-39dBm$

TABLE XIII. EXTRACT OF WAP_{0517} RSSI VALUES. MEASURES TAKEN IN FRONT OF OFFICE 215 (3RD FLOOR OF *TC* BUILDING).

Record	PhoneID	UserID	WAP_{0517} RSSI level
628	23	2	$-87dBm$
846	23	2	not detected
2643	19	6	$-81dBm$
2677	19	6	$0dBm$

TABLE XIV. NUMBER OF RSSI VALUES SCANNED BY TWO DIFFERENT DEVICES IN TWO SCENARIOS: 1) IN FRONT OF OFFICE 121 (1ST FLOOR OF *TI* BUILDING) AND 2) CONSIDERING THE WHOLE DATABASE.

Outside office 121

PhoneID	min	max	mean	median
13	12	17	14.9	15
14	9	22	19.5	21

General

PhoneID	min	max	mean	median
13	0	32	15.55	15
14	2	48	17.98	17

Finally, similarly to other public databases (as for instance the ones included in the *UCI Machine Learning Repository* [7]), the proposed database can be used to evaluate new *Pattern Recognition* methods, as for instance the ones related to feature selection, editing & condensation, and new classification strategies, among others.

VII. CONCLUSIONS

This paper introduces a new database for indoor localization, *UJIIndoorLoc*, on the basis of a WLAN fingerprinting (RSSI levels) environment. First, database description has been fully detailed, including the features used in the database, their meaning, and the value ranges. Second, the procedure and the applications used to generate the database have also been described. To address the problem of samples diversity and realistic approach, more than 20 users participated in generating the database. Each training reference point was initially assigned to, at least, two different users. No suggestion or advice about capturing was given to the users. In addition, the device was always held by a human user in contrast to other datasets in which the device was left on a place to take several samples. All the samples were collected by a human user because the human body partially blocks the radio wave communication [36]. Therefore, the samples taken for *UJIIndoorLoc* can be considered very realistic. Due to privacy issues, some information has been anonymized.

While the WLAN-based localization databases used in the literature tend to cover small areas [17] or one-floor buildings [14], *UJIIndoorLoc* covers three buildings with 4 or more floors and almost $110.000m^2$. Moreover, the shape and internal structure are quite different among the three buildings where the samples were collected. In addition, more than 20 people and 25 different devices have been used to generate the database, in contrast to other databases that were generated using a single device or few devices [25]. Our proposed database can also be very useful for validation and comparison purposes, since validation samples have also been provided. All these features of the proposed database make *UJIIndoorLoc* suitable for testing and benchmarking localization algorithms. Alternatively, *UJIIndoorLoc* can be used for other purposes such as analysis of device accuracy, improvement of WiFi coverage, WAPs optimization (localization and distribution), among others. The here proposed *UJIIndoorLoc* database not only is the biggest database in the literature shown but it is also the first publicly available database that could be used to make comparisons among different methods in the field. In addition, s basic positioning system has been developed using the k-Nearest Neighbor rule in order to provide a baseline for comparison purposes.

Summarizing, the scarcity of publicly available localization databases, none as far as we know, reflects the need of a common public database for research purposes such as the one here presented.

ACKNOWLEDGMENT

We would like to thank to Yasmina Andreu, Óscar Belmonte, Irene Garcia-Martí, Diego Gargallo, Nadal Francisco, Josep López, Ruben Martínez, Roberto Mediero, Javier Ortells, Nacho Piqueras, Ianisse Quizán, David Rambla, Luis E. Rodríguez, Ana Sanchís, Carlos Serra and Sergi Trilles for their help on creating this database.

We also thank Javier Fernández, Ángel Ramos, Álvaro Arranz and Guillermo Amat for their collaboration and comments, as members of *Percepción* Project (Ministerio de Industria, Energía y Comercio - Programa Avanza2 - 2012), for their useful conversations in this area.

REFERENCES

[1] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 140–150, 2010.

[2] R. Montoliu, J. Blom, and D. Gatica-Perez, "Discovering places of interest in everyday life from smartphone data," *Multimedia Tools and Applications*, vol. 62, no. 1, pp. 179–207, 2013.

[3] R. Montoliu, A. Martinez-Uso, and J. Martinez-Sotoca, "Semantic place prediction by combining smart binary classifiers," in *Proceedings of the Mobile Data Challenge by Nokia Workshop, in conjunction with International Conference on Pervasive Computing (Pervasive'12)*, 2012.

[4] Google, "Google Scholar," 2013. [Online]. Available: http://scholar.google.com/

[5] N. Marques, F. Meneses, and A. Moreira, "Combining similarity functions and majority rules for multi-building, multi-floor, wifi positioning," in *Proceedings of the 3th the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2012)*, 2012.

[6] K. Kaemarungsi and P. Krishnamurthy, "Properties of indoor received signal strength for wlan location fingerprinting." in *Proceedings of the 1th Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous'04)*, 2004, pp. 14–23.

[7] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[8] H. Wang, S. Sen, A. Elgohary, M. Farid, M. Youssef, and R. R. Choudhury, "No need to war-drive: unsupervised indoor localization," in *Proceedings of the 10th international conference on Mobile systems, applications, and services (MobiSys'12)*, 2012, pp. 197–210.

[9] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan, "Indoor localization without the pain," in *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking (MobiCom'10)*, 2010, pp. 173–184.

[10] W. G. Griswold, P. Shanahan, S. W. Brown, R. T. Boyer, M. Ratto, R. B. Shapiro, and T. M. Truong, "Activecampus: Experiments in community-oriented ubiquitous computing." *IEEE Computer*, vol. 37, no. 10, pp. 73–81, 2004.

[11] I. Constandache, R. R. Choudhury, and I. Rhee, "Towards mobile phone localization without war-driving." in *Proceedings of the 29th IEEE International Conference on Computer Communications (INFOCOM'10)*, 2010, pp. 2321–2329.

[12] B. Ferris, D. Fox, and N. Lawrence, "Wifi-slam using gaussian process latent variable models," in *In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, 2007, pp. 2480–2485.

[13] J. Lester, T. Choudhury, and G. Borriello, "A practical approach to recognizing physical activities," in *Proceedings of the 4th international conference on Pervasive Computing (Pervasive'06)*, 2006, pp. 1–16.

[14] Y. Chen, D. Lymberopoulos, J. Liu, and B. Priyantha, "Indoor localization using fm signals," *IEEE Transactions on Mobile Computing*, vol. 12, no. 8, pp. 1502–1517, 2013.

[15] J. Machaj, P. Brida, and J. Benikovsky, "Optimization of rank based fingerprinting localization algorithm," in *Proceedings of the 3rd International Conference on Indoor Positioning and Indoor Navigation (IPIN'12)*, 2012, pp. 1–7.

[16] C. Beder and M. Klepal, "Fingerprinting based localisation revisited," in *Proceedings of the 3rd International Conference on Indoor Positioning and Indoor Navigation (IPIN'12)*, 2012, pp. 1–7.

[17] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *Proceedings of the 19th IEEE International Conference on Computer Communications (INFOCOM'00)*, 2000, pp. 775–784.

[18] S. Yang, P. Dessai, M. Verma, and M. Gerla, "Freeloc: Calibration-free crowdsourced indoor localization," in *Proceedings of the 32th IEEE International Conference on Computer Communications (INFOCOM'13)*, 2013.

[19] P. Mirowski, H. Strck, P. Whiting, R. Palaniappan, M. MacDonald, and T. Kam Ho, "Kl-divergence kernel regression for non-gaussian fingerprint based localization," in *Proceedings of the 2on International Conference on Indoor Positioning and Indoor Navigation (IPIN'11)*, 2011, pp. 1–10.

[20] J. Machaj, P. Brida, and R. Pich, "Rank based fingerprinting algorithm for indoor positioning," in *Proceedings of the 2on International Conference on Indoor Positioning and Indoor Navigation (IPIN'11)*, 2011, pp. 1–6.

[21] M. Kranz, C. Fischer, and A. Schmidt, "A comparative study of DECT and WLAN signals for indoor localization," in *Proceedings of the 8th Annual IEEE International Conference on Pervasive Computing and Communications (PerCom'10)*, 2010, pp. 235–243.

[22] V. Honkavirta, T. Perl, S. Ali-Lytty, and R. Pich, "A comparative survey of wlan location fingerprinting methods." in *Proceedings of the 6th Workshop on Positioning, Navigation and Communication (WPNC'09)*, 2009, pp. 243–251.

[23] M. Gunawan, B. Li, T. J. Gallagher, A. G. Dempster, and G. Retscher, "A new method to generate and maintain a wifi fingerprinting database automatically by using rfid," in *Proceedings of the 3rd International Conference on Indoor Positioning and Indoor Navigation (IPIN'12)*, 2012, pp. 1–6.

[24] Y. Gu, A. Lo, and I. G. Niemegeers, "A survey of indoor positioning systems for wireless personal networks." *IEEE Communications Surveys and Tutorials*, vol. 11, no. 1, pp. 13–32, 2009.

[25] T. J. Gallagher, B. Li, A. G. Dempster, and C. Rizos, "A sector-based campus-wide indoor positioning system," in *Proceedings of the 1st International Conference on Indoor Positioning and Indoor Navigation (IPIN'10)*, 2010, pp. 1–8.

[26] S. Brning, J. Zapotoczky, P. Ibach, and V. Stantchev, "Cooperative positioning with magicmap," in *Proceedings of the 4th Workshop on Positioning, Navigation and Communication (WPNC'07)*, 2007, pp. 17–22.

[27] N. Swangmuang and P. Krishnamurthy, "Location fingerprint analyses toward efficient indoor positioning." in *Proceedings of the 6th IEEE International Conference on Pervasive Computing and Communications (PerCom'08)*, 2008, pp. 100–109.

[28] C. Nerguizian, C. L. Despins, and S. Affes, "Indoor geolocation with received signal strength fingerprinting technique and neural networks." in *Proceedings of the 11th International Conference on Telecommunications (ICT'04)*, vol. 3124, 2004, pp. 866–875.

[29] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy, "Precise indoor localization using smart phones." in *Proceedings of the international conference on Multimedia (MM'10)*, 2010, pp. 787–790.

[30] J. Geun Park, B. Charrow, D. Curtis, J. Battat, E. Minkov, J. Hicks, S. J. Teller, and J. Ledlie, "Growing an organic indoor location system." in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys'10)*, 2010, pp. 271–284.

[31] E. C. L. Chan, G. Baciu, and S. C. Mak, "Orientation-based wi-fi positioning on the google nexus one." in *Proceedings of the 6th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob'10)*, 2010, pp. 392–397.

[32] J. marti, J. Sales, and E. Marin, R. Jimenez-Ruiz, "Localization of mobile sensors and actuators for intervention in low-visibility conditions: The zigbee fingerprinting approach," *International Journal of Distributed Sensor Networks*, vol. 2012, p. 10, 2012.

[33] J. Marti, J. Sales, R. Marin, and P. Sanz, "Multi-sensor localization and navigation for remote manipulation in smoky areas." *International Journal of Advanced Robotic Systems*, vol. 10, p. 8, 2013.

[34] H. A. Karimi, Ed., *Advanced Location-Based Technologies and Services*. Taylor & Francis, 2013.

[35] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theor.*, vol. 13, no. 1, pp. 21–27, Sep. 1967.

[36] A. Natarajan, M. Motani, B. de Silva, K.-K. Yap, and K. C. Chua, "Investigating network architectures for body sensor networks," in *Proceedings of the 1st ACM International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments (SIGMOBILE'07)*, 2007, pp. 19–24.