Bachelor Thesis

Anti-Aging Calibration Methodology with User Log-Oriented Anomaly Detection for Wi-Fi Fingerprinting Localization

Ritsumeikan University
College of Information Science and Engineering

Kohei Yamamoto

January 2018

Abstract

Lately, Wi-Fi localization based on received signal strength indicator (RSSl) has been used as the main approach to estimate positions within doors. However, aged-deterioration caused by dynamic environmental changes causes a decline in the accuracy. Therefore, model calibration is inseparable during certain periodic cycles. To dates, transfer learning is utilized to retrain a model using a small set of additional and supervised datasets. However, this overhead is still heavy because the method determines the reference points to obtain additional datasets either randomly or comprehensively, and this point results in the destabilization of the recovery. In this study, a system named No-Sweat Detective is proposed to detect the anomalous reference points employing co-occurrence of Wi-Fi source derived from unsupervised datasets, namely, a user log uploaded from the users of location-based services. The experiment in our test-bed depicted that No-Sweat Detective could detect anomalies by reproducing the environmental changes. An extensive five-month-long experiment demonstrated the redundancy against a dynamic real situation. Further, we confirmed that the No-Sweat Detective could materialize the anti-aged-deterioration, which can stabilize the recovery of the model by 61.9% and can suppress the aged-deterioration by 10.9% at most in comparison with that of the existing methods.

Contents

List of Figures

List of Tables

Chapter.1

INTRODUCTION

Location-based computing has been significantly growing in the market of Internet of Things (IoT) [1]. While the most common technology for geolocation measurement is Global Positioning System (GPS), there are some places where the signal from positioning satellites cannot be obtained: indoor space and underground districts. Instead of such technology, several alternative methods have been investigated to implement indoor localization: Bluetooth Low Energy (BLE) [2], Pedestrian Dead Reckoning (PDR) [3], Indoor MEssaging System (IMES) [4], and so on.

Among them, especially **Wi-Fi localization** based on RSSI (Received Signal Strength Indicator) has become an increasingly heated modality due to the ubiquity of Wi-Fi accessibility [5]. Dissimilar to other indoor localization methods using BLE or IMES, Wi-Fi localization does not require any additional installation for Wi-Fi environments. Although there are some types of algorithm to perform Wi-Fi localization, fingerprinting algorithm [6] is mostly used and it has been proving industry practice. Typically, such Wi-Fi localization consists of two phases: training and operating phases. In the training phase, a **localization model** is created from the initial **labeled observations** (collection of Wi-Fi observations labeled as reference points, which collected manually). During the operating phase, the target location is estimated by comparing the model with the current observation.

However, there can be **aged-deterioration** of the model over time because environmental changes, such as furniture movements, access point (AP) movements, and obstacle constructions may occur, that can affect the accuracy of localization. In addition, automatic power adjustment module implemented in modern AP, which optimizes the transmitting power, then it also brings fluctuation in Wi-Fi environments. This types of fluctuation and of aged-deterioration are not only limited to fingerprinting algorithm but also all other types of Wi-Fi localization algorithm, which utilizes a model which is created from observation results such as localization employing a kalman filter, a particle filter, and so on[7]. Thus, model calibration is indispensable during certain periodic cycles. Nevertheless, the manual recollection of labeled observations requires considerable effort and involves time-consuming overhead [8]. This point has been one of the main drawbacks of such Wi-Fi localization before such localization system comes into use.

To date, numerous researchers have tried to solve this disadvantage. Some research have been trying to calibrate or reconstruct the model with less or no effort [9, 10, 11, 12, 13, 14, 15, 16, 17]. In addition, there has been some kinds of research trying to fix deceived effects in Wi-Fi environments by detecting environmental changes [18, 19, 20]. However, these methods force localization environments and humans to install additional infrastructure such as motion sensors and rich digital devices. Moreover, they require map information including AP location information with paying much effort. Thus, their methods is not fitting practical use. As of late, employing transfer learning has been a major approach to reduce the effort involved in calibration [21, 22, 23, 24, 25]. In these methods, an attempt has been made to reflect the contemporary localization environments in the current model through retraining the model by adding a small number of labeled observations at every calibration [26]. This approach can considerably reduce the number of labeled observations required to retrain the model; however, labeled observations are either randomly or comprehensively picked up from the reference points in the localization site, and this point results in the **destabilization** of the recovery at calibration. As shown above, the existing methods thus far for anomaly detection require much manual effort and rich additional infrastructure, and methods for transfer learning do not take the recovery stability into their consideration.

In this study, we propose a system named the **No-Sweat Detective**, that materializes **no effort** anomaly detection specialized in any types of Wi-Fi localization with **stable and higher recovery** at every calibration of the model. The No-Sweat Detective can detect reference points, that are close to the environmental changes with no effort by utilizing a user log, which is automatically uploaded from location-based services. Therefore, the No-Sweat Detective does not require any manual work other than recollecting labeled observations at detected reference points to retrain the model as same as the existing methods. The No-Sweat Detective is run on the user log, namely, the **unlabeled observations** (collection of Wi-Fi observation without label of reference point). Apparently, the user log can be obtained with no effort as long as such location-based services continue. Employing the co-occurrence derived from the unlabeled observations, the No-Sweat Detective vectorizes each AP as a vector model, which describes relative positions with other APs. After that, the similarity, in other words, normality of the vector model is calculated chronologically to detect anomalous areas that are likely to be environmentally changed. This detection is done by performing density-based clustering and nullifying untrustworthy Wi-Fis coming from APs that are close to where environmental changes occur. Finally, the No-Sweat Detective determines the reference point at which it can convergently recollect the labeled observations. Then this approach consequently can be expected to realize retraining of the model with stable and higher recovery.

The main contributions of this work can be summarized as follows.

- We proposed a system named No-Sweat Detective, namely, no effort anomaly detection specialized in any types of Wi-Fi localization with stable and higher recovery at calibration.
- We observed that the No-Sweat Detective worked even in the dynamic real situation such as huge underground districts, based on a five-month-long study and datasets.
- We verified that the No-Sweat Detective could materialize the anti-aged-deterioration, which can stabilize the recovery of the model by 61.9% and can suppress the aged-deterioration by 10.9% at most in comparison with that of the existing methods.

Chapter.2

RELATED WORK

In this section, we give a brief overview of existing localization algorithm based on Wi-Fi RSSI and of research trying to be resistant to the aged-deterioration of the model.

2.1 Types of Wi-Fi Localization

Lately, Wi-Fi localization has emerged as a popular way to position somebody or something within doors. The methods of most localization system based on Wi-Fi RSSI are classified into two major categories [27]: trilateration algorithm [28] and fingerprinting algorithm [29]. Trilateration algorithm can get the target location by calculating the distance between the target object and three reference points at minimum, however, this algorithm heavily relies on AP location information. Thus, much effort have to be paid before it comes into use in industry practice. On the other hand, fingerprinting algorithm can estimate the target position by matching observations, namely, some characteristics of RSSI which is location unique and this algorithm does not require any prior information unlike the former algorithm. Moreover, this localization algorithm has been known to provide higher accuracy on average than any other methods [30]. For this reason, the latter algorithm is widely used and designed for indoor localization and it has been proving the real world practice. However, as explained in the previous section, the aged-deterioration of the Wi-localization model is significant drawbacks in fingerprinting localization, which caused by dynamic environmental changes. To minimize these kinds of aged-deterioration and fluctuation in the Wi-Fi environments, some filters such as a kalman filter and particle filter are applied, however, the aged-deterioration cannot be completely removed from the model[7]. For these reasons, Wi-Fi localization system requires calibration of the model during certain period cycles.

2.2 Countermeasure against Aged-Deterioration

Naturally, the model can be wholly calibrated by obtaining new labeled observations from all the reference points in the localization site, however, this makes effort for deployment and maintenance of the localization system rather high, therefore, there has been many research conducted to counteract the aged-deterioration or reduce calibration effort.

A system named Calibree [15] and other research [16, 17] utilize radio propagation model for creating the initial model. They showed unnecessity of secondary calibration in their systems. These methods uses gaussian distributions to stochastically estimate location. Although it is remarkable that they achieved no need of secondary calibration of the model, their methods require high-resolution calibration for making the initial model. Besides, they demand some structural information of buildings to create such sophisticated model. This point gets in the way of industry use and much effort has to be devoted beforehand.

Some systems named TuRF [13], QRFC [14], UnLoc [12], and other work [9, 10, 11] have proven that they can obtain labeled observations in real time with no effort. TuRF, QRFC, and UnLoc can label observed unlabeled observations as reference points by estimating the user's location using step counter sensor and by annotating its location to each observation. However, to force the user to put on some sensors or to have additional installment to environments [10, 11] are too assumed from the view of real use, thus, these methods are hardly thought practical.

There has been some related research have tried to counteract the aged-deterioration issue by detecting signal changes in wireless networks. Song et al. [18] and have shown that they could detect node redeployment in network environment by focusing on the change of node neighborship and the change of measured distances between nodes. However, their approach originally requires the custom hardware in addition to exact location information of nodes. Moreover, they do not take detection of radio wave variation into their account, which caused by environmental changes. Meng et al. [19] have proposed a probabilistic localization method to detect signal effects which are severely distorted by unexpected environment changes. However, they have only proven the redundancy in a experimental test-bed but in a real world situation.

Furthermore, all the process in their approach is based on how to reduct the effects of such distorted signal in the initial model, namely, the initial model is used the whole time. Thus, they do not target calibration in their case. Ohara et al. [20] have shown that they could detect environmental changes observing CSI (Wi-Fi channel state information), however, not only that they require the pair of transmitter and receiver to get such information, CSI is scarcely used in real situation from the view of security. Furthermore, their target of environmental changes is limited inside a room, thus, it is unknown that their approach can work in outer areas where more intricate changes expected to occur. Incidentally, as for anomaly detection unique to the No-Sweat Detective, the method regards the high recall rate most importantly more than the precision rate since even uncertain signals have the possibility that eventually repercuss the accuracy of the model. This is the point dissimilar to most other research of anomaly detection in the field of security [31] as they secure the combination of the high recall rate and the high precision rate.

As of late, employing transfer learning has been a major approach to reduce the effort involved in calibration [21, 22, 23, 24, 25]. In these methods, an attempt has been made to reflect the contemporary localization environments in the current model through retraining the model by adding a small amount of labeled observations at every calibration [26]. Yang et al. [22] and other research [23] have proven the higher accuracy with much less calibration effort with a small amount of labeled observations, however, they require comprehensively installed anchor nodes to obtain it. Yin et al. [25] have proposed a localization system which can adapt to dynamic environmental changes, which materialized by regression based algorithm and model-tree based algorithm. Another research [24] tried to cope with performance degradation of the model by using support vector machine (SVM) regression analysis. Albeit they have only evaluated the improvement in their experimental test-bed, their method has provided much good results in practical approach for the degrades over time with adding labeled fingerprints in small quantity. This approach can considerably reduce the number of labeled fingerprints required to retrain the model; however, they are randomly picked up from the reference points in the localization site, and this results in a distabilization of recovery factor at calibration. As shown above, the existing methods thus far for anomaly detection require much manual effort and rich additional infrastructure, and methods for transfer learning do not take the recovery stability into their consideration.

Therefore, labeled fingerprints must be convergently recollected at the specific reference points where environmentally changed.

With the No-Sweat Detective, we can identify the reference points where supposed to be calibrated intensively, besides, the detection process can be performed with no effort using the user log uploaded from the user of location-based services.

Chapter.3

LOCALIZATION PROTOCOL WITH NO-SWEAT DETECTIVE

In this section, we describe the system overview to explain our methodology with respect to the existing methods.

3.1 Two Types of Observations

In our system, we utilize two types of datasets of Wi-Fi observations, namely, labeled observations and unlabeled observations. As examples shown in Table 3.1 and Table 3.2, both types of observations have the following common data structure: Time, BSSID, ESSID, and RSSI (a unit is dBm), each describes, absolute time that the observation is obtained or scanned, Basic Service Set Identifier (BSSID) that is a unique ID to identify each AP, Extended Service Set Identification (ESSID) that is also an ID to distinguish each AP, and RSSI, respectively. Generally, Wi-Fi localization uses BSSID to individuate each AP since ESSID are sometimes overlapped among multiple frequency bands coming from the different AP.

Obviously, the difference between the two types of observations is whether reference point is labeled to each observation or not. As mentioned in the previous sections, labeled observations are datasets of observations used for creating the initial signal model and for retraining the model over time. Labeled observations are obtained by collectors (e.g. a provider of location-based services) at each reference point in the site where they want to implement Wi-Fi localization. Thus, “Labeled” itself literally means that where the observation is collected is labeled to each observation. Typically, these labeled observations are gathered during certain period cycles (e.g.

Table.3.1 Data structure of one labeled observation.

AP	Time	BSSID	ESSID	RSSI	Reference Point
A	1509885150	00:09:b4:70:1d:c7	00PASSPORT	-55	123
B	1509885150	12:09:b4:70:15:f6	FREE_Wi-Fi	-80	123
C	1509885150	b4:c7:99:16:07:34	Secure_Wi-Fi	-40	123

Table.3.2 Data structure of one unlabeled observation.

AP	Time	BSSID	ESSID	RSSI	Reference Point
A	1509885153	00:09:b4:70:1d:45	10PASSPORT	-70	Unknown
B	1509885153	12:09:b4:70:15:56	LINK_Wi-Fi	-45	Unknown
C	1509885153	b4:c7:99:16:07:a4	Public_Wi-Fi	-40	Unknown

every week, every two weeks). In our experiments, labeled observations are also used as grand truth to verify our methodology.

On the other hand, unlabeled observations are datasets of observations used for detecting anomalous AP and specific reference points where supposed to be performed site survey at calibration. Unlabeled observations are obtained automatically from the user of location-based services (e.g. indoor navigation applications on mobile devices) in the site where any types of Wi-Fi localization services are running . By contrast to labeled observations, “Unlabeled” itself literally means that where each observation is obtained is never known since each observation are conducted unconsciously by the users and are not supervised by anyone. Naturally, these unlabeled observations can be automatically accumulated on a server as long as such services continue, more importantly, the amount is in large quantities.

3.2 System Overview

In Figure 3.1, an overview of the No-Sweat Detective is depicted. The No-Sweat Detective consists of several modules: the Access Points Anomaly Detector (APAD), the Reference Points Anomaly Detector (RPAD), and the TransFer Learning (TFL). First, the APAD obtains unlabeled observations as inputs to detect APs that are close to where the environmental changes occurred. After that, the RPAD receives detected AP information such that the RPAD can detect the reference points which should be selected for the next calibration. Then the TFL retrains the model using

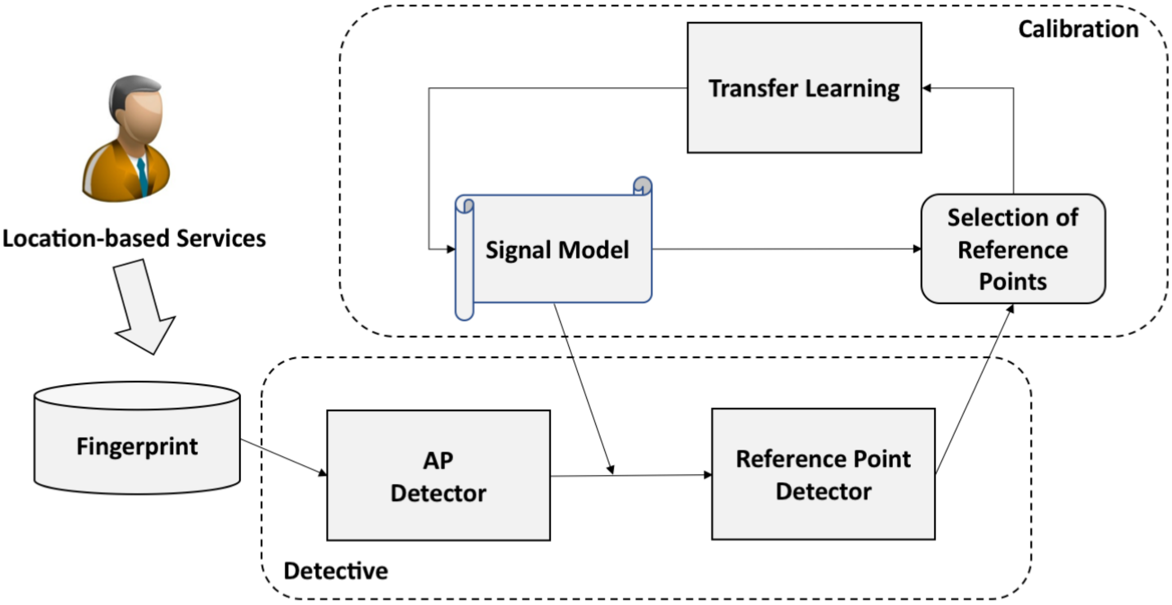

Figure.3.1 No-Sweat Detective architecture.

the labeled observations recollected at the detected reference points, then this process continuously goes live as long as location-based services are under operation.

3.3 Access Points Anomaly Detector (APAD)

As described previously, we utilize unlabeled observations which automatically uploaded from the user of location-based services and mobile applications to detect anomalies. Then a module called APAD of the No-Sweat Detective handles the detection of anomalous APs which are close to the place where environmentally changed. However, unlabeled observations are those for which the point of observation is unknown. Thereby, APAD utilizes co-occurrence analyzes relative position with other APs to detect changes in the Wi-Fi environments.

The APAD does not use all the unlabeled observations for the analysis. In fact, the APAD only picks up the unlabeled observation whose maximum RSSI has a value that is greater than the upper threshold $vecFilt$, because the APAD primarily works on the premise that the unlabeled observation was most likely to be observed close to a certain AP. In this context, the qualified unlabeled observations are illustrated as follows: each R indicates the observed BSSID and each lower letter r indicates the observed RSSI.

$$(R_{1_{r1}}, R_{2_{r2}}, R_{3_{r3}}, \cdots, R_{x_{rx}}) \qquad \{vecFilt < max(r1, rx)\} \tag{3.1}$$

Next, the APAD vectorizes the unlabeled observation. Here is an illustration that how a vector model of an AP A is created. When the unlabeled observation above indicates that the RSSI coming from an AP is the strongest, it can be vectorized as a vector model A with a lower threshold $vecWidth$ which determines the size of the vector model. This vector model is described as follows:

$$\vec{A} = (R_{1_{r1}}, R_{2_{r2}}, R_{3_{r3}}, \cdots, R_{x_{rx}})\,\{vecFilt < max(r1, rx), vecWidth < rx\} \tag{3.2}$$

This vector model expresses the relative position of the AP A with respect to the other APs around A. Then, the next time the vector model A' of AP A is created, the cosine similarity calculated as follows:

$$\cos(\vec{A}, \vec{A'}) = \frac{\vec{A} \cdot \vec{A'}}{|\vec{A}||\vec{A'}|} = \frac{\vec{A}}{|\vec{A}|} \cdot \frac{\vec{A'}}{|\vec{A'}|} = \frac{\sum_{i=1}^{|V|} A_i A'_i}{\sqrt{\sum_{i=1}^{|V|} A_i^2} \cdot \sqrt{\sum_{i=1}^{|V|} {A'}_i^2}} \tag{3.3}$$

The comparison of vector models is calculated based on all the dimension of vector model A and of vector model A'. For instance, when the dimension of A does not contain some dimension of A', 0 is designed to be assigned. This process is chronologically applied to other BSSIDs as well. When the environmental changes occur, the similarity of AP that is close to the distorted environments will cause a drop in the value; thus, the APAD also installs the threshold of cosine similarity $CosSim$ as it shows the normality of the AP and of Wi-Fi environments around the AP.

3.4 Reference Points Anomaly Detector (RPAD)

After the execution of APAD, the next step is to detect environmentally changed area and its contained reference points to recollect labeled observations at the next calibration of the model. Originally, the localization at the reference points which are close to environmental changes heavily gets effects by Wi-Fi distortion, thus, such reference points need to be detected as anomaly. The RPAD detects environmentally changed areas that correspond to the reference points that it contains. However, we must address two cases: the first case in which the AP is singularly detected by the APAD, and the second case in which the AP is plurally detected by the APAD.

In the first case, the RPAD detects the reference points as shown in Figure 3.2. Here is illustration that AP A, B, C, and D are thought to exist and many unlabeled

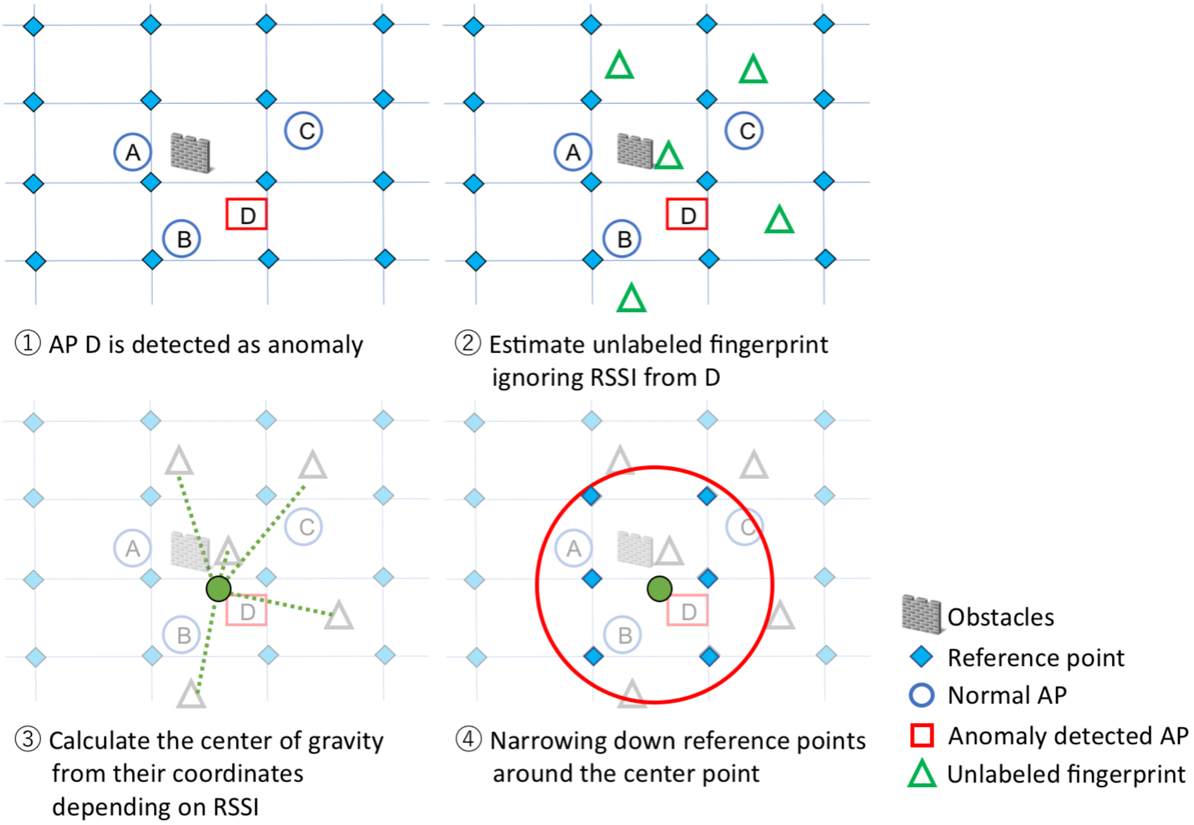
A
B
C
D
① AP D is detected as anomaly
② Estimate unlabeled fingerprint ignoring RSSI from D
③ Calculate the center of gravity from their coordinates depending on RSSI
④ Narrowing down reference points around the center point
Obstacles
Reference point
Normal AP
Anomaly detected AP
Unlabeled fingerprint

Figure.3.2 Flow to specify reference points when APAD shows singular detection.

observations are observed before environmental changes occurred. After the APAD detects D because it is close to the environmental changes, the RPAD roughly estimates the location where each unlabeled observation was scanned while ignoring the Wi-Fi coming from D as an anomaly. Next, the RPAD gets the weighted average center coordinates of the estimated locations. At this point, unlabeled observations which received RSSI more than or equal to -45 from the anomalous AP D are targeted since unlabeled observations which received weak RSSI may not return the trustworthy location estimate. Then, RPAD calculates the weighted average center coordinates by roughly categorized weights: unlabeled observations which received from the anomalous AP D more than or equal to -46 and less than -40 as 1.0, more than or equal to -40 and less than -35 as 1.5, and more than or equal to -35 as 2.0. Finally, the RPAD detects the reference points within a radius ϵm from the center, then, at the detected reference points, a provider of location-based services goes manually recollect labeled observations to use it at the next calibration of the model.

In the second case, it is likely that huge changes occurred in the Wi-Fi environments;

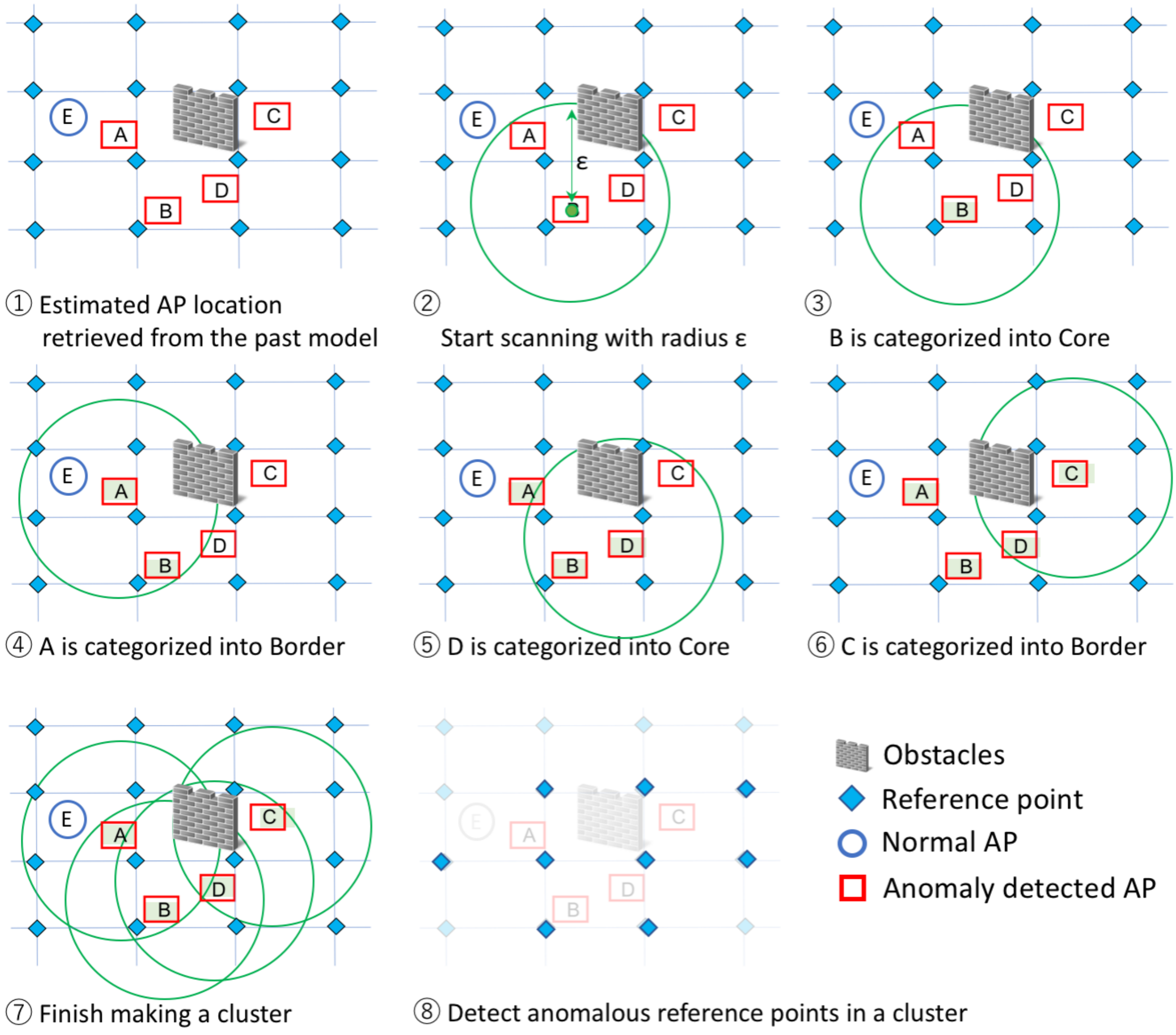
E
A
B
C
D
ε
① Estimated AP location retrieved from the past model
② Start scanning with radius ε
③ B is categorized into Core
④ A is categorized into Border
⑤ D is categorized into Core
⑥ C is categorized into Border
⑦ Finish making a cluster
⑧ Detect anomalous reference points in a cluster
Obstacles
Reference point
Normal AP
Anomaly detected AP

Figure.3.3 Flow to specify reference points when APAD shows plural detections.

therefore the aforementioned algorithm cannot be applied because the RPAD has to ignore too many key Wi-Fis to estimate the center. This problem brings lower reliability into RPAD to search the area where anomalous APs or environmental changes are most likely located. Thus, we apply Density-based spatial clustering of applications with noise (DBSCAN) [32] as shown in Figure 3.3. DBSCAN is the most well known method to conduct density-based clustering in the field of data mining introduced by Ester et al.. DBSCAN does not require the number of clusters that result in arbitrarily shaped clusters; therefore, this algorithm is suitable for our approach because the shape and the number of the clusters is never unknown. Originally, DBSCAN has two parameters: a radius ϵ and $minPts$. Besides, all the data points are categorized into three different concepts: a data point p becomes a core point if ϵ-neighborhood of p contains $minPts$ data points at least, becomes a border point if ϵ-neighborhood of p contains less than $minPts$ data points, and becomes outlier other than above. We regard the estimated locations of APs which are detected by the APAD as points since it is assumed that the collective displacement of plural AP hardly to occur and huge environmental changes have repercussion in a wide range, thus, if plurally anomaly detected, that is most likely because of environmental changes such as new obstacle construction. Here is an illustration that AP A, B, C, D, and E are thought to exist and AP A, B, C, and D were detected as anomalies. After that, the RPAD retrieves each anomalous AP locations from the past model. Then RPAD starts performing DBSCAN with radius ϵ, in this illustration, 2 is assigned to $minPts$ considering the high collection rate of reference points. After the scanning, B is categorized into core point since it has A and D within ϵ. Then the scan is shift to A, then A is categorized into a border point since it only has one within ϵ. Following the same way on D and C, they are categorized into core point, border point, respectively. Finally, after the completion of the scanning, we can see the arbitrary shape cluster as described in ⑦; then, we can assume that huge environmental changes most likely occurred in the clusters; thus, the RPAD detects reference points contained in the clusters and the labeled observations are recollected for the next calibration of the model.

On the two cases above, the RPAD sequentially executes the first case, searching only if the anomalous AP is not contained in the clusters created by DBSCAN. Provisionally, 20 is assigned to ϵ in our real scenario experiment described in the next section, considering the high collection rate of the reference points.

3.5 Transfer Learning (TFL)

As explained in the past sections, in the process of Wi-Fi localization, the initial model is initially created from the initial dataset of the labeled observations collected from all the reference points. After a while passed from making the initial model, a small amount of additional datasets (labeled observations) gathered at corresponding reference points are used to retrain the model. Thus, we need two types of labeled datasets to implement calibration. At this point, TFL receives a small amount of labeled observations and the TFL should be replaceable implying that the No-Sweat Detective should be a framework which can be applied to any transfer learning methods. Thereby, we design to prove the performance of the No-Sweat Detective combined with two different transfer learning methods[26] in this study: the MixTrain and the Lasso methods. The MixTrain method is closer to the basis of transfer learning than the Lasso method. The MixTrain is the method which learns the parameters θ itself from all the datasets such that the algorithm retrains the model by adding the regularization term of L1 norm, written as $(\sum_{i=1}^{|\theta|} |\theta_i|)$; this is simply for keeping the weights given to features from being overfitted by the general L1 norm. On the other hand, the Lasso is the method which learns parameters from the variation of parameters using L1 norms, written as $(\sum_{i=1}^{|\theta|} |\theta_i^{(k-1)} - \theta_i^{(k)}|)$; this regularization term is written as it performs regularization by minimizing the variation between the parameter θ at period $k-1$ and θ at period k.

Chapter.4

VALIDATION AND RESULTS

In this section, we explain environmental settings including the datasets used in the experiments that we conducted and the results.

4.1 Environmental Settings

We conducted two different experiments to prove the function of the anomaly detection module of the No-Sweat Detective and the effectiveness in the real world. We show details of both environmental settings in following 4.1.1 and 4.1.2 respectively.

4.1.1 Laboratory Datasets

We first validated whether the No-Sweat Detective can detect anomalies or not by reproducing environmental changes in our test-bed on the fourth floor of a 10-story building in the Tokyo Institute of Technology[*1], as shown in Figure 4.1. The floor has the dimension of 799m^2 (17m × 47m) contains three labs, one meeting room, four private rooms, one hallway. As shown in Figure 4.1, we especially used the meeting room and hallway as the localization site, besides, we set 105 reference points within 1m^2 of the floor, covering 348m^2 (12m × 29m) dimension to obtain labeled observations. We gathered labeled observations from one scanning per point and conducted it repeatedly ten times a day with Nexus5. Then, we collected 2,100 labeled observations in total within two days as 1,050 per day. In addition, we deployed six pairs of APs in the hallway, as shown Figure 4.2.

As an evaluation of the No-Sweat Detective, it was run on the measurements over

[*1] https://www.titech.ac.jp/english/

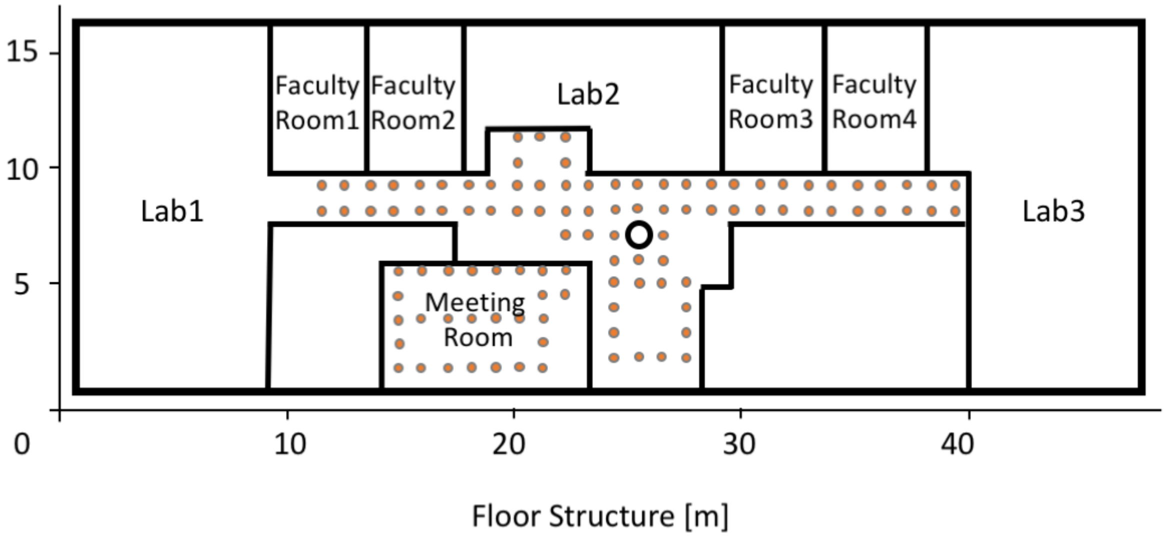

Figure.4.1 Laboratory datasets setting with reference points.

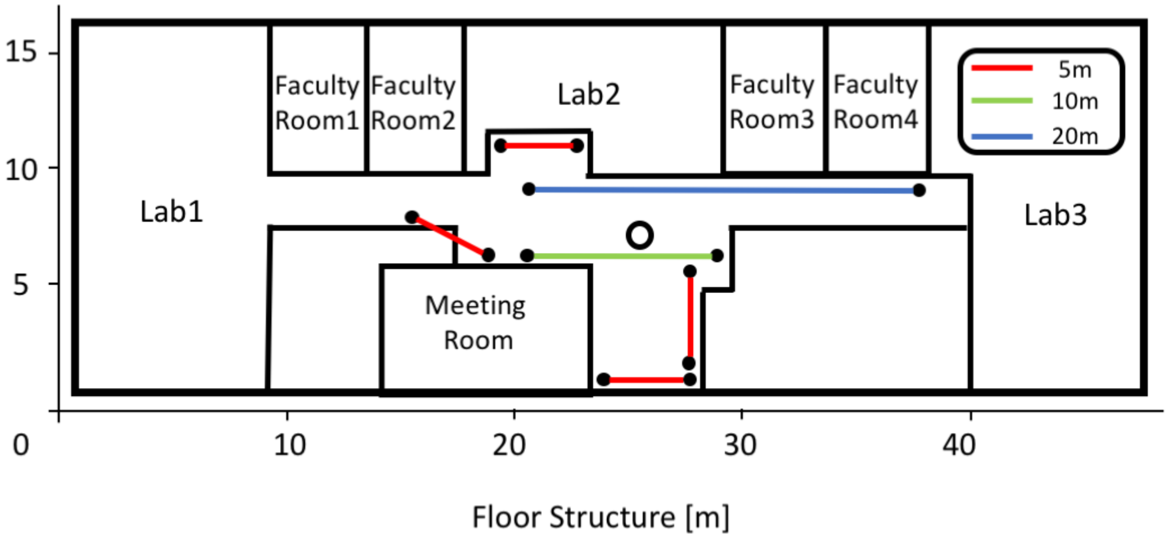

Figure.4.2 Pairs of displacement of AP.

the first day and the second day. We assume that the radio wave along with RSSI received from APs are distorted due to the environmental changes such as appearance of new obstacles and APs displacements. Therefore, it is reasonable enough to reproduce the environmental distortions in the Wi-Fi environments by displacements of APs. Following that, over the two days, we displaced APs to another location to reproduce environmental changes; six displacements consisting of four 5m shifts, one 10m shift, and one 20m shift were made. Finally we observed if the No-Sweat Detective can detect the anomalies.

4.1.2 Underground District Datasets

We conducted further experiments to validate the performance in the real underground district[*2] in Osaka city where is the one of the most complicated area in Japan, as shown in Figure 4.3 The reasons why we chose this place as our next experiment location are strongly based on the following statements.

- This area is notorious for the most spatially complex structure like a maze especially in Osaka city, therefore, there are several location-based services and applications developed to support and to allow navigations for the user. In other words, we can obtain unlabeled observations in large quantities with relatively easy access.
- This area is also infamous for its murderous congestion of people, this fatal condition for Wi-Fi localization can let us simultaneously prove the redundancy of the No-Sweat Detective against fluctuation of Wi-Fi caused by human bodies, multi-path fading, and other unexpected environmental changes.

We set 39 reference points at around 8m intervals, which follows industry practice since we collect labeled observations at the reference points actually used in the existing location-based service utilizing Wi-Fi localization. The localization site covers the dimension of 4615m^2 (71m × 65m) where has clothing store, accessories store, cafe, and so on. We gathered labeled observations, around six scannings per point with Nexus5, every two weeks repetitively for five months. We collected 2,693 labeled observations in total. As for unlabeled observations, we obtained it through the mobile navigation application named "Umechika-Navi"[*3]. We obtained 764 unlabeled observations in total over the five months at this localization site, note that these unlabeled observations were originally filtered, thus, the amount was much enough to run anomaly detection.

The No-Sweat Detective was run on the measurements of labeled observations and unlabeled observations over the five months to detect anomalous reference points, then we conducted calibration every two weeks to retrain the model. We also tried to investigate how the amount of reference points used for each calibration will affect the aged-deterioration, therefore, we compared the accuracy of the final model with the

[*2] http://whity.osaka-chikagai.jp/en/

[*3] http://www.umechikanavi.jp/

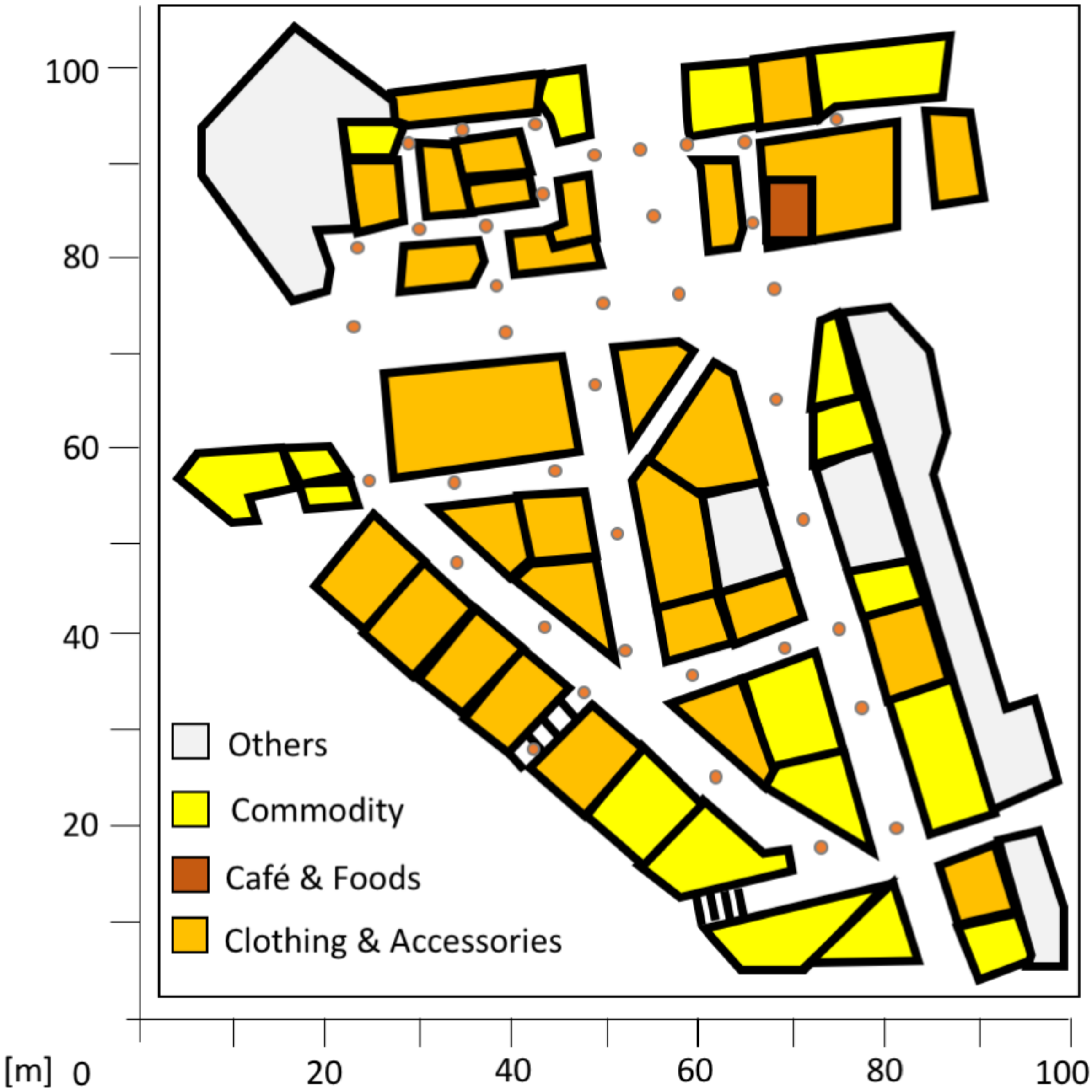

Figure.4.3 The part of huge underground districts where Wi-Fi localization and performance validation implemented.

existing methods, by varying the amount of reference points used for each calibration: 10%, 20%, 30%, 60%, and 100% of all the reference points. We recollected labeled observations at the reference points detected by the No-Sweat Detective to retrain the model, the rest of labeled observations were randomly picked up from the other reference points when there is the number remaining. Finally, we observed the accuracy of the model retrained with the same amount of labeled observations after 10 iterative trials, in comparison with the existing transfer learning methods and with the model

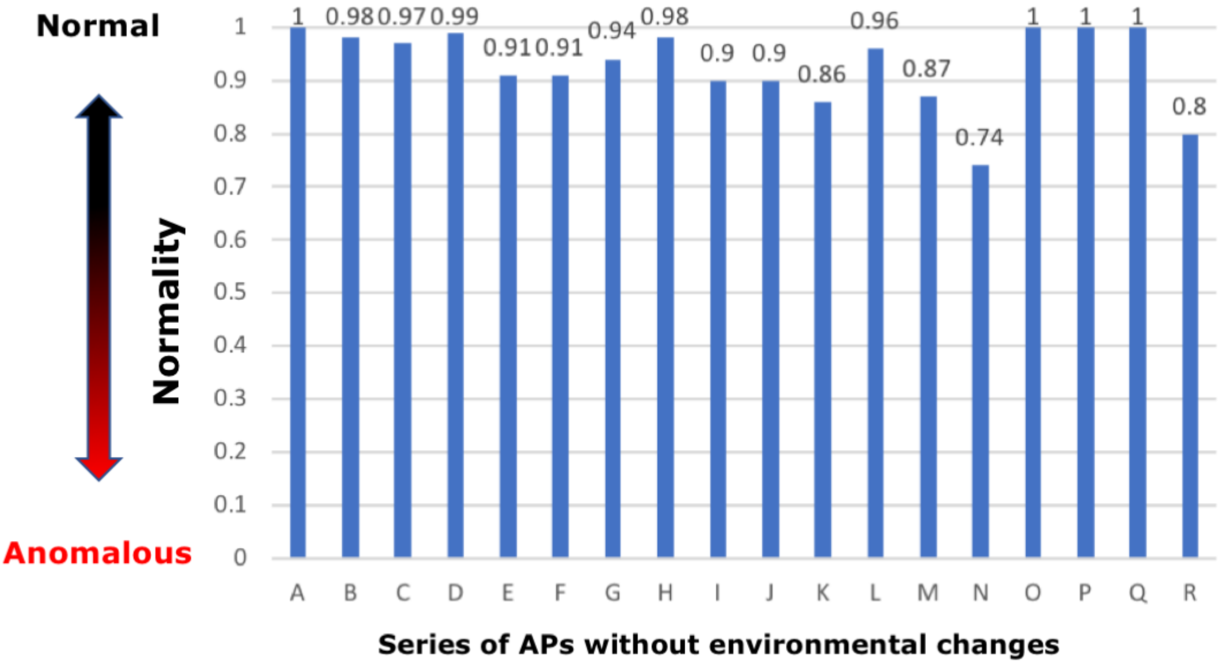

Figure.4.4 Normality of the AP and of the environments around the AP without environmental changes over two days.

which never retrained over five months (NonTrain) as a baseline of the evaluation.

As explained in 3.5, we had the MixTrain method and the Lasso method as the most well known transfer learning methods. Therefore, we made comparisons with them by applying the No-Sweat Detective.

4.2 Results in Laboratory Datasets

After running the No-Sweat Detective on the measurements with the provisional parameters $\{vecFilt, vecWidth\}$ = {-35, -35}, it have shown that 18 vector models were created which all covered six pairs of AP. In Figure 4.4, the vertical axis represents the normality of each AP and of the environments around the AP, and the horizontal axis represents the series of APs without environmental changes. In Figure 4.5, the vertical axis represents the normality of each AP and of the environments around the AP, and the horizontal axis represents the series of APs with environmental changes: numbers in the horizontal axis depicts the distance of displacement of APs. As shown above, APs have overall high normality with 1 as maximum and 0.745 as minimum in the non-distorted APs in Figure 4.4. Conversely, APs have overall low

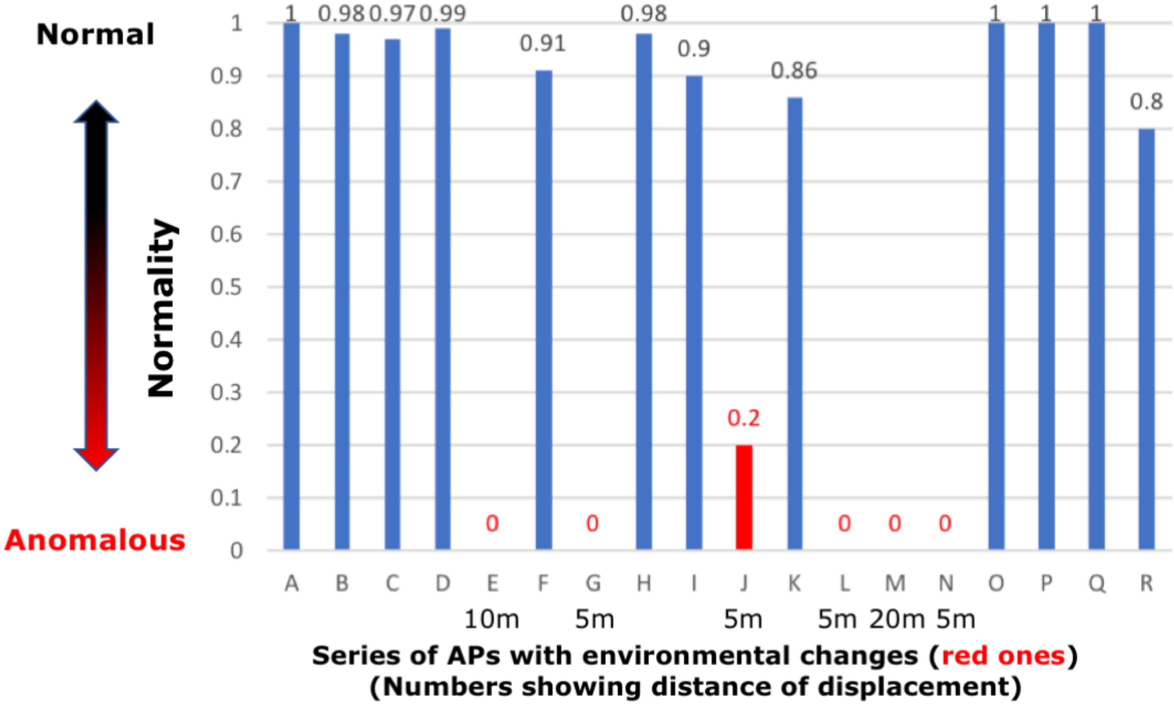

Figure.4.5 Normality of the AP and of the environments around the AP with environmental changes (red ones) over two days.

normmality with 0.208 as maximum and 0 as minimum in the distorted APs in Figure 4.5. Therefore, we confirmed that the No-Sweat Detective could detect anomalies by reproducing environmental changes. The thing should be focused on in both graphs is that five pair of six pair of APs which we prepared for the experiment are judged as completely different vector models by the No-Sweat Detective. This is because both APs of each pair have perfectly unique family in itself one another. Therefore, as shown above, we verified that the anomaly detection module of the No-Sweat Detective could perform as we expected to detect anomaly close to the environmental changes.

4.3 Results in Underground District Datasets

In the view of our evaluation, there are three aspects to evaluate the performance of the No-Sweat Detective applied to the existing methods. First, we investigate how much the No-Sweat Detective can suppress the aged-deterioration to the lower levels after 10 iterative trials, by measuring the average error of localization model with sliding the amount of reference points used in each calibration: 10%, 20%, 30%, 60%,

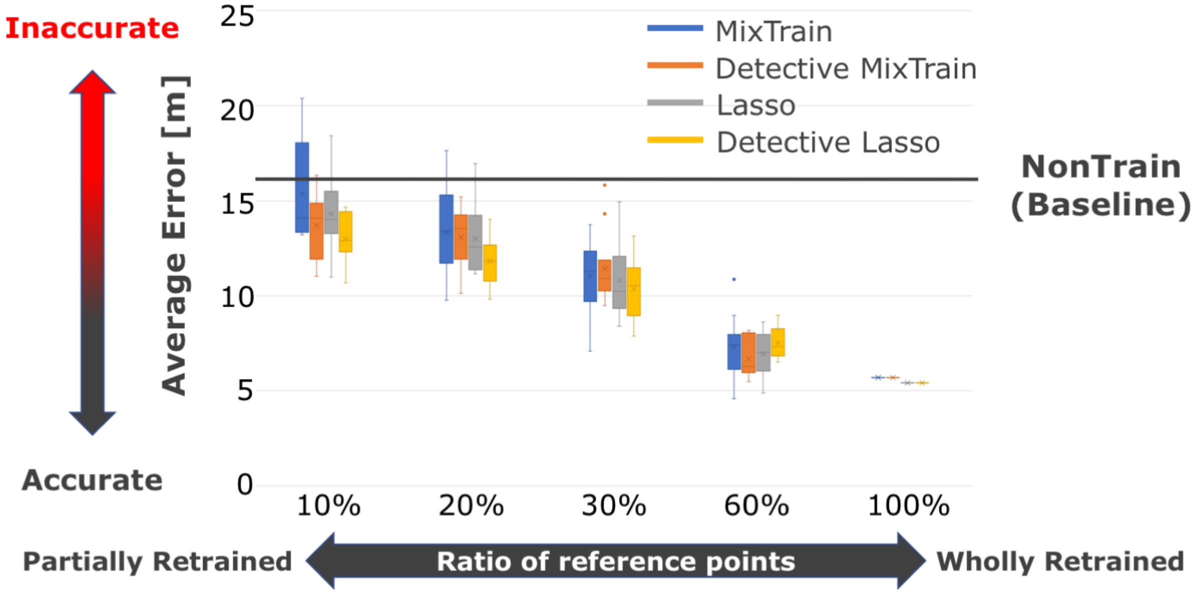

Figure.4.6 Average error of the final model compared to the existing methods after 10 trials, varying the number of reference points used for each calibration: 10%, 20%, 30%, 60%, and 100% of all the reference points.

and 100% out of all the reference points. Second, we evaluate how the cumulative distribution of estimate error will look changed by the performance of the No-Sweat Detective. Last, we evaluate the mean of and the variance of average error of the final model after 10 iterative trials.

After 10 iterative trials running on the measurements of labeled observations with provisional parameters $\{vecFilt, vecWidth, CosSim\}$ = {-43, -59, 0.65}, 36 BSSIDs were observed; four different BSSIDs were detected as anomalies and they were all covered by 10 anomaly detected BSSIDs in the unlabeled observations. After creating the initial model using the initial dataset of labeled observations, we confirmed that the average error of the initial model was 2.32m, and for the model which was never retrained (the NonTrain), the error was 15.96m. Thereby, we also could confirm that the aged-deterioration of the model occurred in the localization site through the observation period.

Evaluation of Average Error

Figure 4.6 shows the accuracy of the MixTrain, the MixTrain with the No-Sweat Detective (Detective MixTrain), the Lasso, the Lasso with the No-Sweat Detective (Detective Lasso) after 10 trials on five-month-long observation, as the vertical axis describes the average error and the horizontal axis represents the amount of reference points used at calibration. According to the results, we observed that there was no such huge difference between the existing methods and the methods combined with the No-Sweat Detective in the 60% selection and in the 100% selection. Focusing on the 10% selection, we observed that the Detective MixTrain drastically shrunk the average error and significantly stabilized the recovery of the model at calibration compared to the MixTrain. The same is true of the Lasso, the Detective Lasso shrunk the average error and stabilized the calibration. Focusing on the 20% selection, the Detective MixTrain still shrunk the average error, and the Detective Lasso shrunk the average error more remarkable than that of the Detective MixTrain. Even focusing on the 30% selection, the Detective Lasso still could suppresses the average error.

Most surprisingly, in terms of the NonTrain (baseline), the box of the MixTrain exceeded the baseline, however, the Detective MixTrain could prevent from being over the baseline focusing on the 10% selection. This outcome implies that the No-Sweat Detective could prevent the existing transfer learning methods from being overfitted to the biased additional datasets that are selected randomly. As shown above, the No-Sweat Detective have shown the powerful performance when we designed to select the small amount of additional datasets, in comparison with the existing methods.

Evaluation of Cumulative distribution

Next, in Figure 4.7 the cumulative distribution of estimate error focusing on the median of 10% selection in Figure 4.6, including the NonTrain method, is shown. The vertical axis shows percentile and the horizontal axis represents the estimate error. Obviously, the graph indicates that the line representing the NonTrain became saturated because of the long and round path and the line got flocked to the right side of the graph. From examining the comparison between the Lasso and the Detective Lasso, we found that the line of the Detective Lasso significantly had shifted to the left side more than that of the Lasso although the difference of the median between the Detective Lasso and the Lasso was small. This is implying that the model could be prevented from getting aged-deteriorated by being retrained with labeled observations recollected at felicitous reference points. We also verified that the line representing the performance of the Detective MixTrain overall got flocked to the left side more than that of the MixTrain by 80%; the part from 0 up to 0.8 of the percentile, although there was no difference of the median between the Detective MixTrain and the MixTrain. As for the part from 0.8 up to 1 of the percentile, we assume that there were some

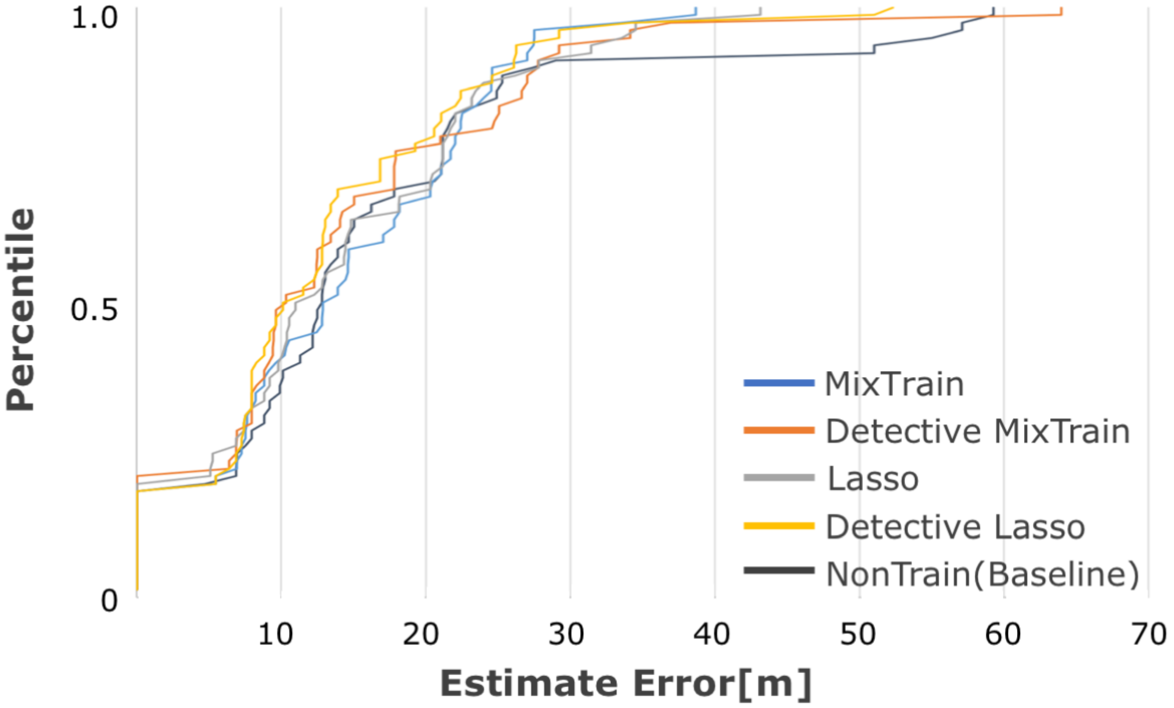

Figure.4.7 Cumulative distribution of estimate error extracted of median from 10% selection

potential holes in the localization site where the unlabeled observations could not be obtained much enough; the amount was enough but the coverage; anyway, the result have shown that Detective MixTrain generally won the MixTrain as it wins up to 0.8 of the percentile. Therefore, these results also support the effectiveness of the No-Sweat Detective.

In light of these outcomes, we confirmed that the No-Sweat Detective can be applied to the existing transfer learning methods to maintain the overall higher accuracy and to suppress the aged-deterioration over long-term operations.

Evaluation of Mean of and Variance of Average Errors

Figure 4.8 shows the mean of and the variance of the average error after 10 trials on five-month-long observation as the left vertical axis describes the mean of the average error, the right vertical axis describes the variance, and the horizontal axis represents the ratio of reference points used at calibration. According to the result, we observed that the mean converges to around 5.5 in all the methods as the ratio of reference points that are used for calibration increases. Besides, the variance overall converges to lower values in all methods as the ratio of reference points that are used

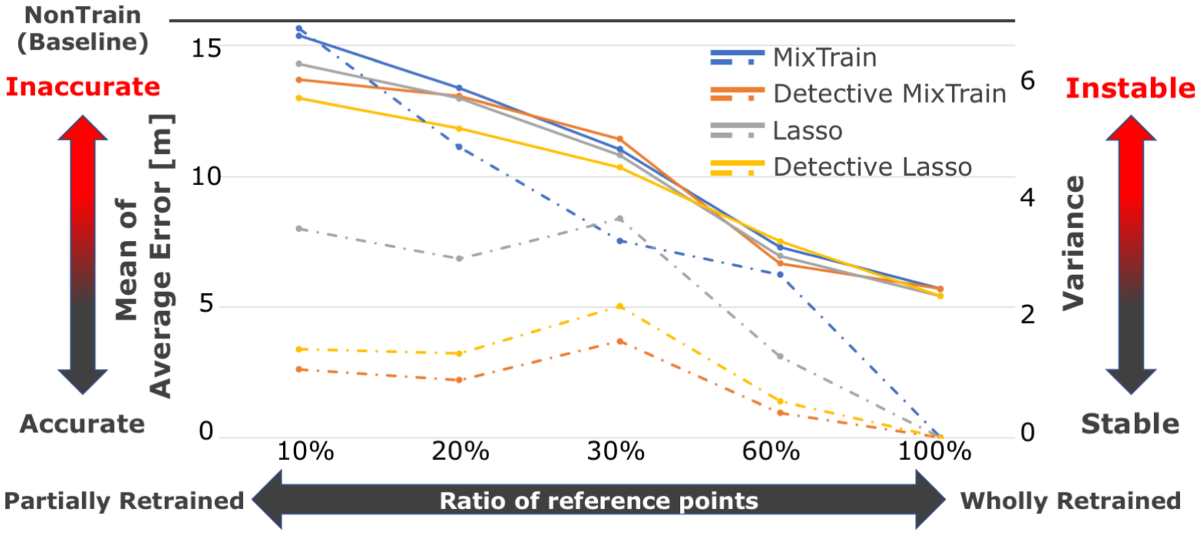

Figure.4.8 Mean of and variance of the average error after 10 trials over five-month-long observation.

for calibration increases.

Casting a spotlight on the mean graph, we validated that both of the Detective MixTrain and the Detective Lasso overall reducted the mean compared to the original methods: the MixTrain and the Lasso respectively. Focusing on the 10% selection, we observed that the Detective MixTrain could achieve the largest suppression of the aged-deterioration from the mean 15.38m down to 13.70m in comparison with the MixTrain. On the other hand, focusing on the 20 % selection, we observed that the Detective Lasso could achieve the largest suppression of the aged-deterioration from the mean 14.31m down to 13.00m in comparison with the Lasso.

Casting a spotlight on the variance graph, we also validated that both of the Detective MixTrain and the Detective Lasso overall hugely reducted the variance compared to the original methods: the MixTrain and the Lasso respectively. Focusing on the 10% selection, we observed that the Detective MixTrain could achieve the most remarkable suppression of the variance from 6.85 down to 2.61 in comparison with the MixTrain, and the Detective Lasso could achieve the most remarkable suppression too from 3.50 down to 1.48.

As shown above, we confirmed that the No-Sweat Detective could suppress the aged-deterioration by 10.92% against the MixTrain and by 9.15% against the Lasso at most. The same is true of the variance, we confirmed that the No-Sweat Detective could stabilize the calibration by 61.90% against the MixTrain and by 57.71% against the Lasso at most. In light of these outcomes, we verified that the No-Sweat Detective could stabilize the calibration and could suppress the aged-deterioration with higher accuracy in comparison with the existing transfer learning methods.

Chapter.5

CONCLUSION

In this study, we proposed a system named the No-Sweat Detective to detect environmental changes with no effort, and to stabilize the model calibration with higher recovery. The No-Sweat Detective can detect reference points which are close to the environmental changes by employing co-occurrence derived from unsupervised observations, namely, a user log which is automatically uploaded from the user of location-based services. We conducted two different types of experiments to prove the function of the anomaly detection module of the No-Sweat Detective and of the performance in the real world.

The former experiment in our test-bed have shown that the No-Sweat Detective could detect anomalies by reproducing the environmental changes. The latter experiment with five-long-month observation have demonstrated that the redundancy against a dynamic real situation, and we verified that the No-Sweat Detective could materialize the anti-aged-deterioration, which can stabilize the recovery of the model by 61.9% and can suppress the aged-deterioration by 10.9% at most in comparison with that of the existing methods.

As for our forthcoming challenges, we are willing to find a way to implement an autonomous calibration system. Specifically, we will define an algorithm that can substitute unlabeled observations for labeled observations and conceive how comparable the localization accuracy depicted by the autonomous calibration system is with that of the current manual calibration system, combined with our anomaly detection method, a probability theory, and further profound data mining theory.

Acknowledgement

For my thesis supervisor Prof. Nobuhiko Nishio, Ritsumeikan University, I would like to thank you for your helpful discussions and dependable assistance on computational theories for this thesis.

For my research supervisor Dr. Kota Tsubouchi, Yahoo Japan Research, the author is grateful to you that you offered some modules and numerical wisdom related to the topic in this thesis.

For Masato Sugasaki, Tokyo Institute of Technology, the author would like to thank you for your on-demand technical assistance and your offer of some modules and datasets used in this thesis.

For Kohei Kanagu, Ritsumeikan University, I am grateful to you for your technical assistance with the experiments to gather datasets used in this thesis.

For Other Yahoo Japan, my Lab, and my family Members, I also thank you all that you offered such a favorable environments to tackle this research.

Bibliography

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 1082–1090, New York, NY, USA, 2011. ACM.

[2] A. Thaljaoui, T. Val, N. Nasri, and D. Brulin. Ble localization using rssi measurements and iringla. In *2015 IEEE International Conference on Industrial Technology (ICIT)*, pp. 2178–2183, March 2015.

[3] W. Kang and Y. Han. Smartpdr: Smartphone-based pedestrian dead reckoning for indoor localization. *IEEE Sensors Journal*, Vol. 15, No. 5, pp. 2906–2916, May 2015.

[4] N. Kohtake, S. Morimoto, S. Kogure, and D. Manandhar. Indoor and outdoor seamless positioning using indoor messaging system and gps. In *International Conference on Indoor Positioning and Indoor Navigation (IPIN'11)*, 2011.

[5] P. Bahl and V. N. Padmanabhan. Radar: an in-building rf-based user location and tracking system. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, Vol. 2, pp. 775–784 vol.2, 2000.

[6] B. Abdelghani and Q. Gao. Improved fingerprinting localization with connected component labeling based on received signal strength. In *2016 International Conference on Progress in Informatics and Computing (PIC)*, pp. 198–204, Dec 2016.

[7] F. Evennou and F. Marx. Advanced integration of wifi and inertial navigation systems for indoor mobile positioning. *EURASIP J. Appl. Signal Process.*, Vol. 2006, pp. 164–164, 2006.

[8] Y. Li, Z. He, J. Nielsen, and G. Lachapelle. Using wi-fi/magnetometers for indoor

location and personal navigation. In *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–7, 2015.

[9] Y. Kim, H. Shin, and Hojung Cha. Smartphone-based wi-fi pedestrian-tracking system tolerating the rss variance problem. In *IEEE International Conference on Pervasive Computing and Communications (PerCom2012)*, pp. 11–19, 2012.

[10] Y. C. Chen, J. R. Chiang, H. H. Chu, P. Huang, and A. W. Tsui. Sensor-assisted wi-fi indoor location system for adapting to environmental dynamics. In *Proceedings of the 8th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '05, pp. 118–125, New York, NY, USA, 2005. ACM.

[11] S. Chen, Y. Chen, and W. Trappe. Exploiting environmental properties for wireless localization and location aware applications. In *2008 Sixth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 90–99, March 2008.

[12] 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, pp. 197–210, 2012.

[13] C. Li, Q. Xu, Z. Gong, and R. Zheng. Turf: Fast data collection for fingerprint-based indoor localization. In *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–8, Sept 2017.

[14] H. H. Liu. The quick radio fingerprint collection method for a wifi-based indoor positioning system. *Mobile Networks and Applications*, Vol. 22, No. 1, pp. 61–71, Feb 2017.

[15] A. Varshavsky, D. Pankratov, J. Krumm, and E. Lara. Calibree: Calibration-free localization using relative distance estimations. In *Proceedings of the 6th International Conference on Pervasive Computing*, Pervasive '08, pp. 146–161, 2008.

[16] M. M. Atia, M. Korenberg, and A. Noureldin. A consistent zero-configuration gps-like indoor positioning system based on signal strength in ieee 802.11 networks. In *Proceedings of the 2012 IEEE/ION Position, Location and Navigation Symposium*, pp. 1068–1073, April 2012.

[17] P. Barsocchi, S. Lenzi, S. Chessa, and F. Furfari. Automatic virtual calibration of range-based indoor localization systems. *Wirel. Commun. Mob. Comput.*, Vol. 12, No. 17, pp. 1546–1557, December 2012.

[18] H. Song, L. Xie, S. Zhu, and G. Cao. Sensor node compromise detection: The location perspective. In *Proceedings of the 2007 International Conference on Wireless Communications and Mobile Computing*, IWCMC '07, pp. 242–247, 2007.

[19] W. Meng, W. Xiao, W. Ni, and L. Xie. Secure and robust wi-fi fingerprinting indoor localization. In *2011 International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–7, Sept 2011.

[20] K. Ohara, T. Maekawa, and Y. Matsushita. Detecting state changes of indoor everyday objects using wi-fi channel state information. In *PACM Interact. Mob. Wearable Ubiquitous Technol.*, 2017.

[21] C. Luo, H. Hong, and M. C. Chan. Piloc: A self-calibrating participatory indoor localization system. In *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, pp. 143–153, April 2014.

[22] B. Yang, J. Xu, J. Yang, and M. Li. Localization algorithm in wireless sensor networks based on semi-supervised manifold learning and its application. *Cluster Computing*, Vol. 13, No. 4, pp. 435–446, Dec 2010.

[23] A. M. Bernardos, J. R. Casar, and P. Tarrío. Real time calibration for rss indoor positioning systems. In *2010 International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–7, Sept 2010.

[24] Y. Tian, B. Denby, I. Ahriz, P. Roussel, R. Dubois, and G. Dreyfus. Practical indoor localization using ambient rf. In *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1125–1129, May 2013.

[25] J. Yin, Q. Yang, and L. Ni. Adaptive temporal radio maps for indoor location estimation. In *Third IEEE International Conference on Pervasive Computing and Communications*, pp. 85–94, March 2005.

[26] P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pp. 110–, 2004.

[27] N. S. Kodippili and D. Dias. Integration of fingerprinting and trilateration techniques for improved indoor localization. In *2010 Seventh International Conference on Wireless and Optical Communications Networks - (WOCN)*, pp. 1–6, Sept 2010.

[28] D. Ruiz, J. Ureña, J. C. García, C. Pérez, J. M. Villadangos, and E. García. Efficient trilateration algorithm using time differences of arrival. *Sensors and Actuators A: Physical*, Vol. 193, No. Supplement C, pp. 220 – 232, 2013.

[29] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg. Compass: A probabilistic indoor positioning system based on 802.11 and digital compasses. In *WINTECH*, 2006.

[30] Y. Gu, A. Lo, and I. Niemegeers. A survey of indoor positioning systems for wireless personal networks. *Commun. Surveys Tuts.*, Vol. 11, No. 1, pp. 13–32, January 2009.

[31] N. Sayegh, I. H. Elhajj, A. Kayssi, and A. Chehab. Scada intrusion detection system based on temporal behavior of frequent patterns. In *MELECON 2014 - 2014 17th IEEE Mediterranean Electrotechnical Conference*, pp. 432–438, April 2014.

[32] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.