

Rethinking of Feature Selection Method for Room-Level Localization using Public APs

Youngsam Kim* and Soohyung Kim*

*Authentication Research Section, ETRI, Daejeon, Korea

kimzt@etri.re.kr, lifewsky@etri.re.kr

Abstract— As WiFi infrastructures and smartphones are propagated widely various WiFi fingerprint-based indoor localization system has been proposed. One of the approaches to improve the accuracy is the feature selection method that analyse the characteristics of WiFi RSSI and construct fingerprint database. However, none of the previous studies analysed about the number of features selected despite the fact that it is important in practical system. In this paper, we propose the feature selection method with characteristics of constancy and strength and evaluate the performance according to the number of features selected.

Keywords— Indoor Localization, WiFi Fingerprint, Feature Selection, Public APs, RSSI

I. INTRODUCTION

With the rapid development of mobile communication and the pervasive computing technology, various applications are extensively trying to adopt the indoor localization technology. Recently, Google's Abacus project or a SafeNet's context-based authentication solution [1] considers the location information as an implicit authentication factor in the area of the user authentication. The location information as an authentication factor has been mainly studied using outdoor locations based on GPS in [2]-[4] while using indoor locations has not. One of the main reasons is that there are no general API of indoor localization. Since smartphones could easily access to the API for outdoor locations based on GPS, it is relatively easy to obtain many outdoor trajectory data. Indoor localization, however, needs the customized system for each target building and it makes that third parties use indoor location information for their services very limited.

Fingerprint-based indoor localization system uses the WiFi fingerprint that consists of multiple access points (APs) and its signal strength. WiFi signal is vulnerable to environmental factors such as walls, doors, furniture, and even people. In addition, it might be from a personal access point or a temporary hotspot. As a result, WiFi signal strength the smartphone received, RSSI, is not constant and stable.

To address this kind of problem, fingerprint-based indoor localization has been studied for last sixteen years since RADAR [1] system had proposed. There are two categories. One consists of localization algorithms using the supervised learning and the other includes fingerprint database construction methods using the unsupervised learning.

Supervised learning-based systems are simple and easy to understand and many researches have been conducted. [5,6,7,8,9,10,11] uses only WiFi fingerprint and [12,13,14,15,16] uses hybrid method based on both WiFi fingerprint and sensor fusion. Supervised learning-based system consists of two phases, including offline site-survey and online location estimation. The site-survey is to construct the fingerprint database (radio map) and the quality of database directly affects the accuracy of location estimation. A problem is that the cost of site-survey is proportional to the size of buildings because it is conducted by a human. Unsupervised learning-based systems [17]-[20] address the issue and propose the method for construction of the fingerprint database without site-survey. Other problem is that the distance error is several meters. This is caused by the fact that the wireless signal is vulnerable to the environmental changes. To address this issue, localization systems using both WiFi fingerprint and sensor fusion method do not completely depend on the fingerprint database but compensate the error using multiple sensors like accelerometer or orientation sensor.

In this paper, the feature selection method that can score each feature by analysing the characteristics of RSSI according to the constancy and strength [9], [10] is proposed. In addition, we show that the number of features selected affects the localization accuracy. In the user authentication domain, room-level indoor localization system could be assumed. For performance evaluation, we collect the data in the office building with eight rooms for about one month. The results of performance evaluation show the accuracy of room-level localization is about 95% with the proposed feature selection method.

The rest of this paper consists of as follows. In section 2, the related works about WiFi fingerprint-based localization is addressed. Section 3 describes the proposed methodologies, data collection and feature selection. Section 4 shows the result of performance evaluation. Section 5 draws the conclusion.

II. RELATED WORKS

WiFi fingerprint-based indoor localization has been researched for about two decades since the RADAR [5] system had proposed. WiFi fingerprint-based indoor localization system increasingly used for the practical

applications because both WiFi infrastructures and smartphones are widely deployed. However, the wireless signal becomes more vulnerable to interference or fluctuation in the complex indoor environments that the temporary access points are increasing, nowadays. It necessarily results the increasing distance error and the cost of managing fingerprint database.

RADAR system firstly proved the possibility of WiFi fingerprint-based indoor localization system. It uses pre-installed three access points and kNN (Nearest Neighbor) algorithm to estimate location. The performance of RADAR system allows about 2 meters error. Horus [6] system conducted detailed analysis of the characteristics of access points and increased the accuracy. According to the Horus system, wireless channel has temporal and spatial characteristics. In particular, it showed that the collected samples from the same access point has the autocorrelation and made the correlation modeller for it. Horus system uses Bayesian probability-based location estimation using public access points, which means system administrator has no information about the access points.

The localization system using public access points should consider the characteristics of WiFi access points. Lina et al. [7] studied about distribution of WiFi fingerprint in real world. Generally, gaussian distribution is assumed in the probabilistic localization techniques, but real distribution could be different that means real WiFi RSSI can have the double peak Gaussian distribution. Lina et al. proposed the localization system using both double peak Gaussian distribution and general Gaussian distribution.

Arsham et al. [9] studied the WiFi AP's characteristics and categorized in five characteristics, strength, stability, variance, constancy, and coverage. According to the performance evaluation in [9], constancy and strength are the main characteristics to improve the accuracy. Pei et al. [10] also studied about import access points named IAP and proposed weighted kNN algorithm based on IAP.

Meanwhile, unsupervised learning based researches that can generate the fingerprint database automatically and decrease the cost of site-survey has been conducted. Chenshu et al. proposed the crowdsourcing-based system that can construct the logical floor plan and match to the physical floor plan based on graph theory [19], [20]. The matching method consists of three phases, skeleton mapping, branch-knot-mapping, and correction. The result of the research shows that 15 logical locations are matched in total 16 locations.

In this paper, we show that the number of features selected affects the accuracy of localization system. Proposed feature selection method scores each features using strength and constancy characteristics referred to the previous studies [9], [10].

III. METHODOLOGY

In this section, data collection method for performance evaluation and proposed feature selection method are described.

A. Data Collection

WiFi fingerprint data was collected in each room in the office as depicted in figure 1 using android-based custom mobile application. Room labels for ground truth was already set to the application to reduce the human error when a user input the label of the room. The custom application used getScanResult method in WiFiManager class and the collected data consists of timestamp, mac address and RSSI of APs, and room label. Samsung Galaxy S4 is used to collect the WiFi fingerprint data and the scan rate is 0.25Hz.

The office has 14 conference rooms but only eight rooms are accessible. Each room has different size. Room 254, 255, 256 is 3x4 meters, room 271, 265, 266 is 3x5 meters, room 269 is 6x5 meters, and room 252 is 8x5 meters. The period of data collection is from august 29th to September 29th and we collected the data two or three times a week. Total number of collected data is for 11 days (0829, 0831, 0902, 0905, 0907, 0909, 0912, 0919, 0921, 0926, and 0929). The reference points in each room are various because we consider the room-level localization. To cover the multiple points in the room, data was collected in centroid of the room, along the wall, and some arbitrary points in the room.

The 0829 dataset that will be used as a training set has 90 samples per each room and others that will be used as a test set has 30~90 samples per each room.

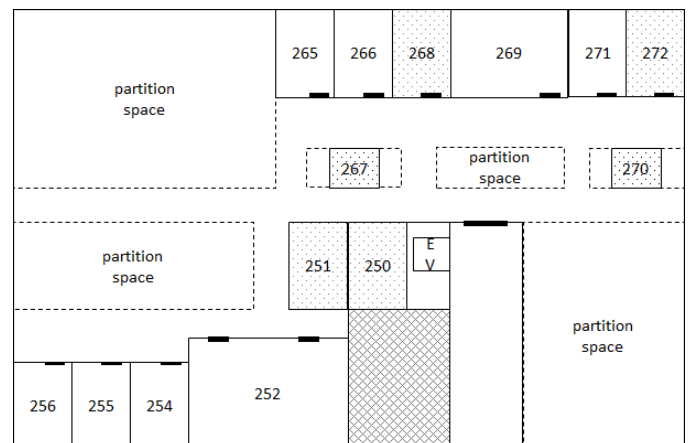


Figure 1. The floor plan where data collection took place.

B. Feature Selection

According to [9, 10], constancy and strength of RSSI are considerable to improve the localization accuracy. In this section, we describe the scoring method using both characteristics.

Let the set of labels be $L=\{l_0, \dots, l_j\}$ and the list of access points be $\mathbf{x}=\{x_0, \dots, x_{n-1}\}$, that is features with the size of n . Then, default fingerprint can be defined as $\mathbf{f}=\{f_0, \dots, f_{n-1}\}$ such that f_i is the RSSI value of an access point x_i . Fingerprint f_i is a value of RSSI ranging from 0 to -100 with normal distribution. Actually, RSSI cannot be -100. Thus, we regard it as null RSSI ($\emptyset$) that the signal from the access point is not detected

If m fingerprints are collected for each label, then fingerprint database for an arbitrary location l is defined F_l as two dimensional matrix and final fingerprint database is defined F as follows.

$$F = \bigcup_{l \in L} F_l$$

$$s. t. F_l = \{f_{ij} | 0 \leq i < m, 0 \leq j < n\}$$

To score the features, the input of the scoring method should be each column vector of F_l . Let j -th column vector of F_l be v_j . Constancy score of a feature x_j is calculated based on frequency that the each element of v_j is not null RSSI. The length of v_j is m and normalized constancy score $C_l(j)$ can be defined as follows.

$$C_l(j) = \frac{|\hat{v}_j|}{m}$$

$$s. t. \hat{v}_j = \{f_{ij} | 0 \leq i < m, f_{ij} \neq \emptyset\}$$

Strength score $S_l(j)$ is a mean value of column vector v_j excluding null RSSI and can be defined as follows.

$$S_l(j) = \frac{\sum_{s \in \hat{v}_j} s}{|\hat{v}_j|}$$

Hybrid score $H_l(j)$ is calculated from constancy score $C_l(j)$ and strength score $S_l(j)$ as follows.

$$H_l(j) = \alpha C_l(j) + (1 - \alpha) S_l(j)$$

$$s. t. T_l(j) = 1 - \frac{S_l(j)}{\tau}$$

Constant α is the weight between constancy and strength. If α is one hybrid score only considers the constancy characteristic otherwise it considers strength characteristic only. Constant τ is normalization constant for strength score.

Now, hybrid score can be calculated for all l_i in L . For each l_i , k features can be selected as x_l according to the score with descending order. Final features selected x^* is sum of each x_l for all l_i in L .

$$x^* = \bigcup_{l \in L} x_l$$

$$s. t. x_l = \{x_{l0}, \dots, x_{lk-1}\}$$

IV. PERFORMANCE EVALUATION

This section shows that the results of performance evaluation of the proposed feature selection method based on the 11 dataset. Dataset 0829 is used for training and the other 10 datasets (0831~0929) are used for test. Classification algorithm is kNN and scikit-learn library in python is used for implementation.

We evaluate the performance by comparing the accuracy of room-level classification when using all features and using

selected features. In addition, we show the accuracy is changed according to the number of features selected.

A. Hybrid constant

Proposed feature selection method has hybrid constant α that can weight constancy or strength. To determine constant α , we conducted the test varying α and the result is like figure 2. Maximum accuracy is 94.8% when α is 0.9 which is bigger when using only one characteristic. It means that considering two characteristics is meaningful even though the improvement is only 1%. Hybrid constant α should be determined based on the dataset and hybrid-scoring algorithm. In the rest of the evaluation, we set α is 0.9.

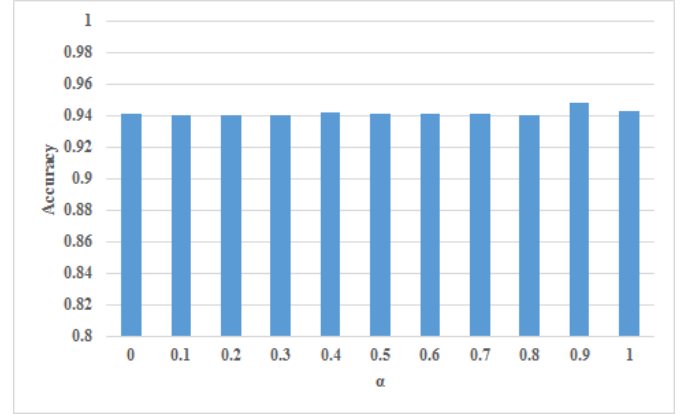


Figure 2. Localization accuracy according to hybrid constant α

B. Number of Features

Proposed feature selection method selects k features for each rooms, sums all of them, and finally selects k^* features. Figure 3 show that the accuracy of 10 each test set according to k ranging from 10 to 140. The number in parenthesis means k^* that is determined by k . Total number of features n is 232. In most test set, the accuracy is rarely changed when k is more than or equal to 100 even when using all 232 features. This result can be analysed that only about 100 features are enough to the localization without decreasing accuracy. Additionally, it can be helpful to decrease the computational cost because the kNN algorithm considers only selected features.

Maximum accuracy is achieved when k is 30, 40, and 70 depending on the test dataset. In practical system, we should determine the fixed k . In naïve approach, we could choose 40 because the average accuracy is highest. Figure 4 show the accuracy of three kinds of approaches. *NoFS* is the case that system uses all features. *AvgFS* is the naïve approach that chooses k when the average accuracy of whole test set is highest. *MaxFS* assumes that there is a specific methodology to choose optimal k . The accuracy of each case is 93.58%, 94.81%, and 95.15%, respectively.

Proposed feature selection method results in 1.6% point increase compared when no feature selection is applied. This is not remarkable but meaningful because the default accuracy (*NoFS*) is too high. It means the environment where we

collect the data was so static. If the environment is noisier, then the proposed method might result better improvement of accuracy.

Meanwhile, determining the optimal number of features selected is difficult problem. In the case of *AvgFS*, several training datasets are needed to reflect characteristics that are more general. In the case of *MaxFS*, k should be dynamically changed by an arbitrary period. Thus, *MaxFS* needs the database update and dynamic feature selection method. In the future works, we will address this problem.

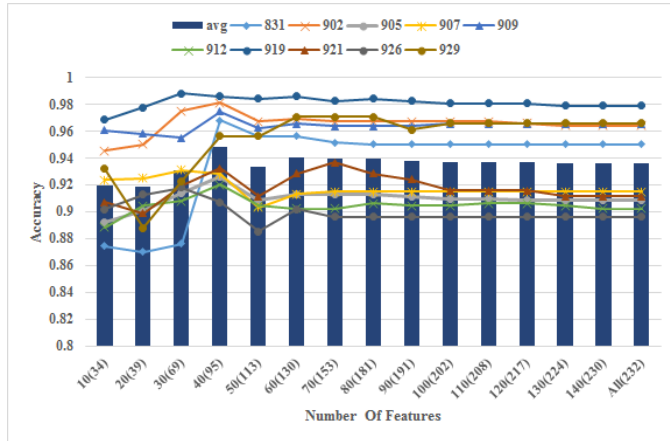


Figure 3. Localization accuracy according to the number of features selected. Labels on x axis represent $k(k^*)$

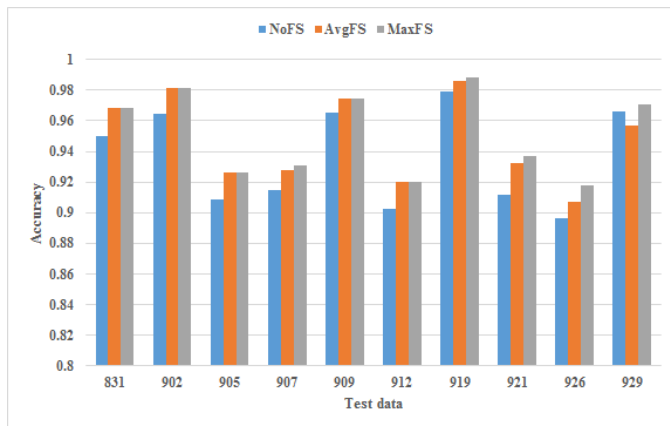


Figure 4. Localization accuracy of each test data with the number of features. *NoFS* used all features, *AvgFS* used selected features where the average accuracy is highest, and *MaxFS* used selected features where the accuracy is highest for each test data.

V. CONCLUSIONS

In this paper, we proposed hybrid-scoring algorithm using the characteristics of constancy and strength and selects the features based on that score. The number of features selected is one of the factors that can increase the accuracy in the fingerprint-based localization system. We collected the data for performance evaluation in the office with eight conference

rooms for about one month. The performance evaluation show the accuracy could be increased about 1.6% point if we can select k adequately. In future works, we plan to design and implement the localization system including database update and incremental feature selection method.

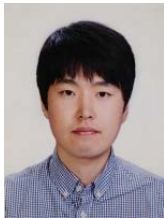
ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. B0126-16-1007, Development of Universal Authentication Platform Technology with Context-Aware Multi-Factor Authentication and Digital Signature

REFERENCES

- [1] SafeNet, "Context-Based & Step-Up Authentication Solutions," <http://www.safenet-inc.kr/multi-factor-authentication/context-based-authentication/>
- [2] Eiji Hayashi, Sauvik Das, Shahriyar Amini, Jason Hong, and Ian Oakley, "CASA: Context-Aware Scalable Authentication," *SOUPS 2013*, pp. 1-10, Newcastle, UK, July 2013.
- [3] Farid M. Naini, Jayakrishnan Unnikrishnan, Patrick Thiran, and Martin Vetterli, "Where You Are Is Who You Are: User Identification by Matching Statistics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 358-372, Feb. 2016.
- [4] Lex Fridman, Steven Weber, Rachel Greenstadt, and Moshe Kam, "Active Authentication on Mobile Devices via Stylometry, Application Usage, Web Browsing, and GPS Location," *IEEE Systems Journal*, arXiv:1503.08479v1[cs.CR], pp. 1-10, March 2015.
- [5] Paramvir Bahl and Venkata N. Padmanabhan, "RADAR: An In-Building RF-based User Location and Tracking System," *Tech. Rep. MSR-TR-00-12*, Microsoft Research, Feb. 2000.
- [6] Moustafa Youssef and Ashok Agrawala, "The Horus WLAN Location Determination System," *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, pp. 205-218, Seattle, Washington, June 2005.
- [7] Lina Chen, Binghao Li, Kai Zhao, Chris Rizos, and Zhengqi Zheng, "An Improved Algorithm to Generate a Wi-Fi Fingerprint Database for Indoor Positioning," *Sensors 2013*, vol. 13, issue 8, pp. 11085-11096, Aug. 2013.
- [8] Jungmin So, Joo-Yub Lee, Cheal-Hwan Yoon, and Hyunjae Park, "An Improved Location Estimation Method for Wifi Fingerprint-based Indoor Localization," *International Journal of Software Engineering and Its Applications*, vol. 7, no. 3, pp. 77-86, May 2013.
- [9] Arsham Farshad, Jiwei Li, Mahesh K. Marina, and Francisco J. Garcia, "A Microscopic Look at WiFi Fingerprinting for Indoor Mobile Phone Localization in Diverse Environments," *2013 International Conference on Indoor Positioning and Indoor Navigation*, Oct. 2013.
- [10] Pei Jiang, Yunzhou Zhang, Wenyan Fu, Huiyu Liu, and Xiaolin Su, "Indoor Mobile Localization Based on Wi-Fi Fingerprint's Important Access Point," *International Journal of Distributed Sensor Networks*, vol. 2015, pp. 1-8, 2015.
- [11] Jonathan Ledlie, Jun-geun Park, Dorothy Curtis, André Cavalcante, Leonardo Camara, Afonso Costa and Robson Vieira, "Molé: a Scalable, User-Generated WiFi Positioning Engine," *2011 International Conference on Indoor Positioning and Indoor Navigation*, 2011.
- [12] Anshul Rai, Krishna Kant Chintalapudi, Venkata N. Padmanabhan, and Rijurekha Sen, "Zee: Zero-Effort Crowdsourcing for Indoor Localization," *MobiCom '12*, pp. 293-304, Istanbul, Turkey, Aug. 2012.
- [13] Ahmad Abadleh, Sangyup Han, Soon J. Hyun, Ben Lee, and Myungchul Kim, "ILPS: Indoor Localization using Physical Maps and Smartphone Sensors," *15th International Symposium on "A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, June 2014.
- [14] Ahmad Abadleh, Sangyup Han, Soon J. Hyun, Ben Lee, and Myungchul Kim, "Construction of Indoor Floor Plan and Localization," *Wireless Networks*, vol. 22, issue 1, pp. 175-191, Jan. 2016.

- [15] Hongbo Liu, Yu Gan, Jie Yang, Simon Sidhom, Yan Wang, Yingying Chen, and Fan Ye, "Push the Limit of WiFi based Localization for Smartphones," *MobiCom'12*, pp. 305-316, Istanbul, Turkey, Aug. 2012.
- [16] Eladio Martin, Oriol Vinyals, Gerald Friedland, and Ruzena Bajcsy, "Precise Indoor Localization Using Smart Phones," *MM'10*, pp. 787-790, Firenze, Italy, Oct. 2010.
- [17] He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury, "No Need to War-Drive Unsupervised Indoor Localization," *MobiSys'12*, pp. 197-210, Low Wood Bay, Lake District, UK, June 2012.
- [18] Yifei Jiang, Xin Pan, Kun Li, Qin Lv, Robert P. Dick, Michael Hannigan, and Li Shang, "ARIEL: Automatic Wi-Fi based Room Fingerprinting for Indoor Localization," *UbiComp'12*, pp. 441-450, Pittsburgh, USA, Sep. 2012.
- [19] Chenshu Wu, Zheng Yang, Yunhao Liu, and Wei Xi, "WILL: Wireless Indoor Localization Without Site Survey," *2012 Proceedings IEEE INFOCOM*, pp. 64-72, 2012.
- [20] Chenshu Wu, Zheng Yang, and Yunhao Liu, "Smartphones Based Crowdsourcing for Indoor Localization," *IEEE Transaction on Mobile Computing*, vol. 14, no. 2, pp. 444-457, Feb. 2015.
- [21] Patrick Lazik and Anthony Rowe, "Indoor Pseudo-ranging of Mobile Devices using Ultrasonic Chirps," *SenSys'12*, pp. 99-112, Toronto, ON, Canada, Nov. 2012.
- [22] Patrick Lazik, Niranjini Rajagopal, Oliver Shih, Bruno Sinopoli, and Anthony Rowe, "ALPS: A Bluetooth and Ultrasound Platform for Mapping and Localization," *SenSys'15*, pp.73-84, Seoul, South Korea, Nov. 2015.



Youngsam Kim received a B.S. (2009) degree in computer engineering from Chungbuk National University and a M.S. (2011) in information security engineering from the University of Science and Technology in South Korea. In 2011, he joined future internet team in NIMS as a researcher

Currently, he is a researcher at the Electronics and Telecommunications Research Institute. His research interests include context-aware authentication, machine

learning, and security protocols.



Soohyung Kim received the B.S. and M.S. degrees in computer science from Yonsei University, Seoul, Korea, in 1996 and 1998. He received the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology(KAIST), Daejeon, Korea in 2016. He is currently a Director of Authentication Research Section in Electronics and Telecommunications Research Institute(ETRI), Daejeon, Korea. His research interests include payment

system, biometrics, identity management, network and system security.