IEICE Communications Express, Vol.13, No.8, 295–298

ComEX

Special Cluster in Emerging Communication Technologies in Conjunction with Main Topics of ICETC 2023

LETTER

Fingerprint localization scheme with correction for missing values in training data and data augmentation

Togo Shinomiya[1], **Satoru Aikawa**[1, a)], **and Shinichiro Yamamoto**[1]

Abstract This study discusses an indoor localization method using the radio signal strength indicator (RSSI) of a wireless local area network (LAN). An indoor localization method that adapts a convolutional neural network (CNN) to the fingerprint method is used. In this method, the CNN learns the access point (AP) information for each coordinate using the RSSI and media access control (MAC) addresses obtained from the wireless LAN APs and compares them with the AP information received from the user to estimate the user location. However, data collection for learning is costly when using CNNs. In addition, there is a problem of missing data owing to various factors when collecting AP information. Therefore, data augmentation is proposed as a method to reduce the cost of data collection while maintaining accuracy and is performed after correcting for missing values. However, data augmentation can produce unrealistic data. This paper proposes a method for correcting missing values in measurement data as a solution to this problem.

Keywords: localization, Wi-Fi, received signal strength indicator, fingerprint method, convolutional, neural network

Classification: Navigation, guidance and control systems

1. Introduction

In recent years, Global Navigation Satellite System (GNSS)-based localization techniques have been used for localization services. However, the accuracy of indoor localization is significantly reduced because of the attenuation of satellite signals. Therefore, this study uses an indoor localization method that adapts a convolutional neural network (CNN) to the fingerprint method, which involves estimating the user location from the radio signal strength indicator (RSSI) emitted by a wireless local area network (LAN) access point (AP).

The fingerprint method consists of three steps. First, the AP information measured at the coordinates set on the map is recorded in a database (DB). Second, the AP information received from the user is converted into user data (UD). Finally, DB and UD are compared with CNN to estimate the user location [1].

However, training a CNN requires a large amount of measurement data, increasing the required time. Reducing the cost by reducing the data reduces the localization accuracy; thus, a trade-off exists between the two. Therefore, data augmentation is proposed as a method to reduce the time cost while maintaining localization accuracy [2]. Data augmentation is a technique to increase the training data used for machine learning. It is often used in the fields of image recognition and natural language processing [3, 4]. In this study, it is a method of rearranging the RSSI of measured data to increase the amount of data.

This study focused on preaugmentation measurement data to improve the accuracy of data augmentation. One of the current problems is that measurement data are missing because of radio wave attenuation and the effects of the AP frequency band. A method is proposed to compensate for these missing values and increase the information content of the preaugmentation data.

2. Localization using the fingerprint method

2.1 Fingerprint method

The fingerprint method estimated the user location by comparing DB and UD. Figure 1 shows the fingerprint method for localization. DB coordinates or reference points (RPs) were set on the map for localization, and AP information, such as RSSI and MAC address, was received from all APs that can be confirmed at each RP and recorded in the DB. The RP with the most similar data was then used as the estimation result by comparing the UD, which was the AP information received by the user, with the DB. A CNN is used for comparison.

2.2 Fingerprint method using CNN

A CNN is a deep learning model that can retain and learn the positional relationships of multidimensional arrays, such as images. In localization technology, methods have been investigated to adapt CNNs, which can retain location infor-

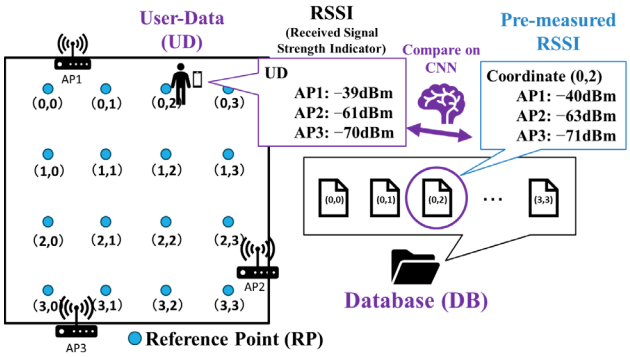

Fig. 1 Localization using the fingerprint method

[1] Graduate School of Engineer, Department of Information Electronics, University of Hyogo, Himeji-shi, Hyogo, 671-2280, Japan

[a)] aikawa@eng.u-hyogo.ac.jp

DOI: 10.23919/comex.2024TCL0010
Received February 7, 2024
Accepted February 19, 2024
Publicized March 12, 2024
Copyedited August 1, 2024

mation, to fingerprint method [5, 6].

When adapting the CNN to the fingerprint method, the localization area is divided into a grid pattern. Moreover, the RSSI measured from the APs was used for locations where APs were present, and −100 dBm which in effect represents the receiving load was set for locations where no APs were present. The number of nodes in the input layer represents the number of APs observed, and the RSSI observed for each AP is used as the input. The number of nodes in the output layer represents the number of DB coordinates, and the user-presence probability is the output for each coordinate using a softmax function. The coordinates with the highest presence probability are used as the estimated location of the user.

3. Proposed method

3.1 Fingerprint method using data augmentation

The measurement results of collecting AP information using the fingerprint method were recorded by stopping and scanning the RSSIs of all RPs multiple times. Therefore, AP information was recorded at different times on the same RP for a number of scans. In this study, data augmentation created a pseudoDB by combining the RSSIs of each AP at different times [2]. The measurement data on which the data augmentation was based was used as source data.

Figure 2 shows the selection of RSSIs for data augmentation. The RSSIs were randomly combined, as the number of RSSI combinations was enormous, based on the number of APs and scans. As the RSSIs in this augmented data were randomly selected with the same probability, the environmental changes over time and the positional relationships between APs were not considered. In contrast, source data contain information based on real environmental changes and positional relationships between APs. As this is the information that must be added to the data to be learned by the CNN, the source data was combined with randomly combined augmented data. This data was called hybrid data [2].

Figure 3 shows an example of the hybrid data. Hybrid data was created by combining augmented data with multiple data sources. It is assumed that the amount of augmented data and the number of times the source data are combined will vary depending on the measurement environment; therefore, it is necessary to determine the optimum parameters for each measurement environment. Here, the number of AP scans of an RP is expressed as "scan/RP," and 10 scans/RP is defined as "one unit" [2].

Time Stamp	RP	RSSI [dBm]	
		AP1	AP2
09:15:10	(0,2)	−48	−33
09:15:29	(0,2)	−45	−27
09:15:44	(0,2)	−46	−34
⋮	⋮	⋮	⋮

Data Augmentaion		
(0,2)	−45	−34
(0,2)	−46	−34
(0,2)	−46	−33
(0,2)	−48	−33
(0,2)	−45	−34
(0,2)	−48	−34
(0,2)	−45	−27
(0,2)	−46	−27
(0,2)	−45	−34
⋮	⋮	⋮

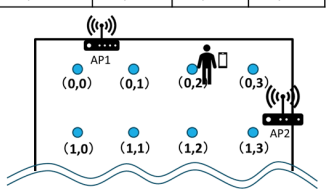

Fig. 2 Random selection of RSSIs for data augmentation

3.2 Correction of missing-values

In the fingerprint method, data may be missing when AP information is measured. This may be because the data can't be received the data because the RSSI of the AP is too low or because nearby APs are transmitting on the same channel in the same frequency band, causing radio interference that prevents reception. In the first case, either a missing value or a very low RSSI (Less than −80 dBm) is obtained continuously. In this case, the missing value has a good effect as it provides information that "this AP has a low overall RSSI". In the second case, there are sudden missing values between stable RSSIs. These missing values have a negative effect on the information provided by the AP. Especially in the second case, if the augmentation is performed on measurement data with missing values, unrealistic data will be created. Each time the missing value is selected, it has a worse effect on the augmented data, making it more unrealistic.

The missing-value correction is set at $R_{low} = -100$ dBm for the first case, which is caused by a low RSSI. In this study, the missing-value correction is intended for cases where data are missing, although the previous RSSI is stable and a strong level is detected. A threshold R_{th} was set to distinguish this abnormal missing value owing to radio attenuation. If the RSSI to be corrected is less than R_{th}, then R_{low} is set.

The missing-value correction is handled separately in DB and UD. In DB, an average value was set from the RSSIs before and after the missing value. R_{low} is set if the average value before and after the missing value is less than R_{th} or if there are missing values either before or after. For UD, only the RSSI immediately before the missing value is referenced and set, because there is no future data. R_{low} is set if the immediately preceding RSSI is lower than R_{th} used in the DB, or if the immediately preceding value is a missing value.

Figure 4 shows the missing-value correction. The threshold is set to $R_{th} = -70$ dBm, so all the missing values for AP1 are set to R_{low}; the missing values for AP2 are set to −32 dBm because the average value before and after RSSIs is at

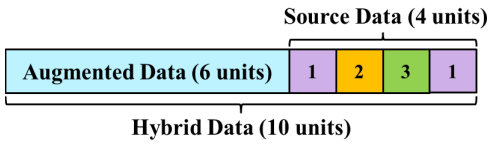

Fig. 3 Construction of the hybrid data

Source data

Time Stamp	RP	RSSI [dBm]	
		AP1	AP2
09:15:10	(0,2)	−84	−33
09:15:29	(0,2)	nan	−29
09:15:44	(0,2)	nan	−34
09:15:52	(0,2)	−81	nan
09:16:04	(0,2)	nan	−30
09:16:13	(0,2)	−83	−28
⋮	⋮	⋮	⋮

Filled nan data

RSSI [dBm]	
AP1	AP2
−84	−33
−100	−29
−100	−34
−81	−32
−100	−30
−83	−28
⋮	⋮

average

$R_{th} = -70$dBm

Fig. 4 Missing-value correction

R_{th}. Since the optimum threshold R_{th} for the missing-value correction is likely to vary depending on the environment, the parameters for each measurement environment must be investigated.

3.3 Data augmentation after correcting for missing-values

If the measured data are augmented by the data as they are, unrealistic data may be produced. Factors that lead to unrealistic data include the failure to reflect environmental changes over time and the random inclusion of sudden missing values. In this paper, the use of data corrected for missing values allows for more meaningful data augmentation. Therefore, the data augmentation in 3.1 is performed using the measurement data with missing-value correction in 3.2.

4. Validation

4.1 Determining thresholds for missing-value correction

The effectiveness of the method was confirmed by simulations to determine the optimal thresholds for correcting missing values in DB and UD.

Figure 5 shows the measurement environment. A total of 23 RPs were measured at 3-m RP spacing on the sixth floor of Building B at the University of Hyogo Himeji Campus for Engineering. The source data were measured at 200 scans/RP and split into 10 and 50 scans/RP. The data were also measured at 50 scans/RP on another day and split into 30 and 20 scans/RP, which were used as validation and test data, respectively.

The accuracy when the missing values are corrected is checked by varying the threshold from −80 to −60 dBm in steps of −5 dBm. Figure 6 shows the validation results for the corrected source data of 10 and 50 scans/RP.

As shown in Fig. 6, when the source data were 10 scans/RP, the highest accuracy was achieved at the −75-dBm threshold, whereas the −70-dBm threshold almost had the same accuracy level. Compared with the accuracy of the uncorrected data, the accuracy improved by approximately 4.6%. When the source data is 50 scans/RP, the −80-dBm and −70-dBm thresholds show the highest and the second-highest accuracies, respectively.

These results indicated that the optimum threshold varies depending on the measurement data. In this study, the threshold for the subsequent missing-value correction was set at −70 dBm.

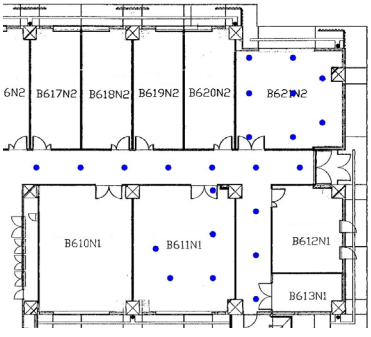

Fig. 5 Validation environment: building (University of Hyogo Himeji Campus for Engineering); 3-m RP spacing (23 RPs)

4.2 Data augmentation after correction of missing-values

From the results obtained in 4.1, missing-value correction was performed when the threshold was set to $R_{th} = -70$ dBm, and the corrected data were used for data augmentation. The amount of hybrid data is 10 units from the study presented at the International Conference on Education Technology and Computers (ICETC) [2]. The augmented data was created so that the number of units is between one and 10, and the source data was divided into one unit and combined for the remaining number of units. As shown in Fig. 7(a), the augmented data is six units, so the source data were combined for four units. At this point, the source data were three units; therefore, four units are combined for the first unit, which is in the order of the earliest measurement time. As shown in Fig. 7(b), only the source data for 10 units were combined for comparison with the proposed method.

The validation environment was the same as in 4.1, i.e., 200 scans/RP of data were split into 10 scans/RP, the threshold was corrected for the missing values with −70 dBm from 4.1, and data augmentation was performed.

Figure 8 shows the measurement results. The horizontal axis in Fig. 8 represents the number of units of augmented data, where six represents the position estimation accuracy with hybrid data, as shown in Fig. 7(b). Meanwhile, "0" represents the localization accuracy for data combined from only 10 units of source data, as shown in Fig. 7(a). As shown in Fig. 8, the highest accuracy was achieved when 10 units of augmented data were created. This is thought to

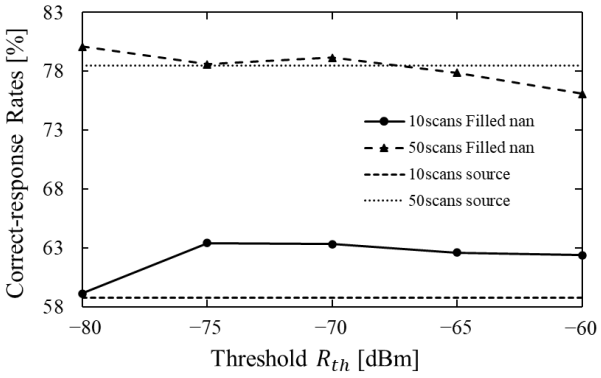

Fig. 6 Variation in position augmentation accuracy with the threshold for missing-value correction

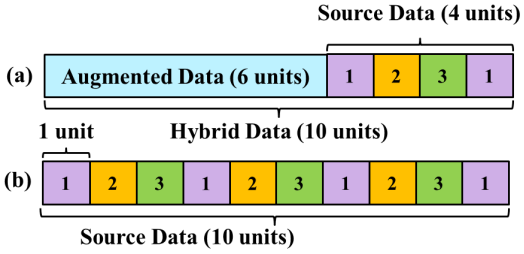

Fig. 7 Splitting of source data to create hybrid data

be because when the number of RSSIs in augmented data is small, the remaining units are repetitions of the source data, resulting in overlearning and lower accuracy. Therefore, it was assumed that as the number of augmented data units increased, the amount of information in the hybrid data increased and the accuracy improved. This suggests that accuracy will improve as the number of units increases.

4.3 Comparison with the proposed method

The localization accuracy between the proposed data augmentation method and data augmentation after correction for missing values was compared. The measurement environment was the same as in 4.1, i.e., the source data (200 scans/RP measurement data divided into 10 scans/RP) and the hybrid data (a combination of five units of augmented data and five units of source data, which is the most accurate from the results of the study presented at ICETC) [2] were used in the data augmentation method. The corrected data with the missing-value correction threshold $R_{th} = -70$ dBm, which has the highest accuracy from Fig. 6, was used as the correction of missing-value method. The data augmented by 10 units only with the missing-value-corrected data shall be used in the correction of missing-value method.

Figure 9 shows the validation results of these methods. It is shown that the proposed method improves the localization accuracy by 12.2% and 3.8% compared with the source data and conventional method, respectively.

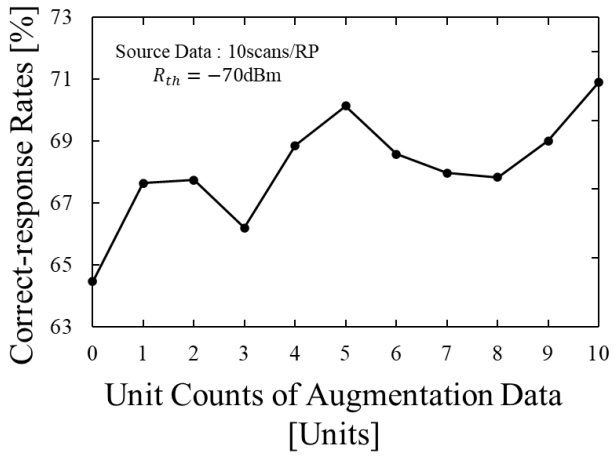

Fig. 8 Correct response rates of the proposed method versus the number of units of augmentation data for different numbers of source data scans

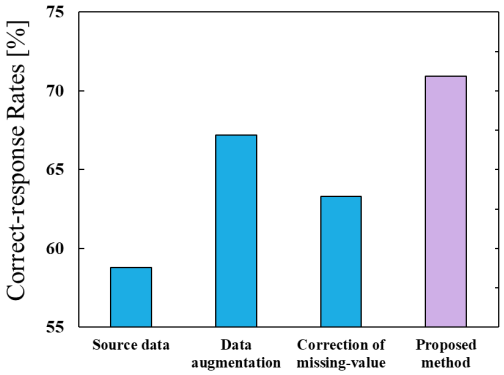

Fig. 9 Comparison of localization accuracy between the proposed and conventional methods

5. Conclusion

Indoor localization by adapting CNNs to the fingerprint method has the problem that the cost of collecting data for training is high. To solve this problem, there are data augmentation methods that increase the amount of data from a small amount of data. However, data augmentation has the potential to produce unrealistic data, so a method to compensate for missing values in the measured data has been proposed and its effectiveness has been experimentally evaluated. The results show that the proposed method can improve the accuracy.

Future work needs to address environmental change over time in data augmentation. Environmental change includes long-term and short-term time changes, and the method proposed in this paper corresponds to the short-term time course. Long-term environmental changes include changes in installations and changes in AP locations. These changes need to be considered.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number 21K04065.

References

[1] T. Muramatsu, S. Aikawa, and S. Yamamoto, "CNN localization using AP inverse position estimation," IEEE Conference Antenna Measurements and Applications (CAMA2019), pp. 1–3, Oct. 2019. DOI: 10.1109/CAMA47423.2019.8959663

[2] T. Shinomiya, S. Aikawa, and S. Yamamoto, "A fingerprint localization scheme using data augmentation," Proc. ICETC2023, no. P1-32, Nov. 2023. DOI: 10.34385/proc.79.P1-32

[3] C. Shorten and T.M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, July 2019. DOI: 10.1186/s40537-019-0197-0

[4] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: a survey," *AI Open*, vol. 3, pp. 71–90, June 2022. DOI: 10.1016/j.aiopen.2022.03.001

[5] X. Song, X. Fan, C. Xiang, Q. Ye, L. Liu, Z. Wang, X. He, N. Yang, and G. Fang, "A novel convolutional neural network based indoor localization framework with WiFi fingerprinting," *IEEE Access*, vol. 7, Aug. 2019. DOI: 10.1109/ACCESS.2019.2933921

[6] M.Z. Karakusak, H. Kivrak, H.F. Ates, and M.K. Ozdemir, "RSS-based wireless LAN indoor localization and tracking using deep architectures," *Big Data and Cognitive Computing*, vol. 6, no. 3, p. 84 Jan. 2022. DOI: 10.3390/bdcc6030084