Simultaneous Crowd Estimation in Counting and Localization Using WiFi CSI

Hyuckjin Choi, Tomokazu Matsui, Shinya Misaki, Atsushi Miyaji, Manato Fujimoto, and Keiichi Yasumoto
Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan
E-mail: {choi.hyuckjin.cc0, matsui.tomokazu.mo4, misaki.shinya.mq9, miyaji.atsushi.ly2, manato, yasumoto}@is.naist.jp

2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN) | 978-1-6654-0402-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IPIN51156.2021.9662572

***Abstract*—In the field of crowd estimation, most non-visual approaches confine their objective to only crowd counting, whereas there are a number of vision-based researches which can estimate both the number and location of people. By observation, we figured out that the WiFi channel state information (CSI) also contains the potential characteristics for both estimations. In this paper, we propose a user-device-free simultaneous crowd estimation system that enables both crowd counting and localization simultaneously, by WiFi CSI and Machine Learning. The originality of this study is that we leverage the CSI bundles as the source for extracting features that contain characteristics depending on the dynamic state (counting) and static state (localization). By experiments during three different-time sessions, we confirm that we could achieve up to 94% counting accuracy and 95% localization accuracy by k-fold cross-validation.**

***Index Terms*—crowd estimation, counting, localization, channel state information, WiFi, device-free, machine learning.**

I. Introduction

People in modern society are living in an enormous amount of floating information. Visible light, various kinds of sensor signals, radio signals connecting our mobile devices to base stations might be including useful information which has a great potential to be utilized in many applications. In recent years, researchers have been trying to convert those ubiquitous resources into practical and useful information in their fields, such as crowd estimation. Crowd estimation is a technique to count the number of people or estimate the crowd density within a certain area. It is considered a critical issue in terms of occupancy sensing for energy saving, route guidance for shoppers or travelers, and crowd control during huge gatherings or emergency evacuations.

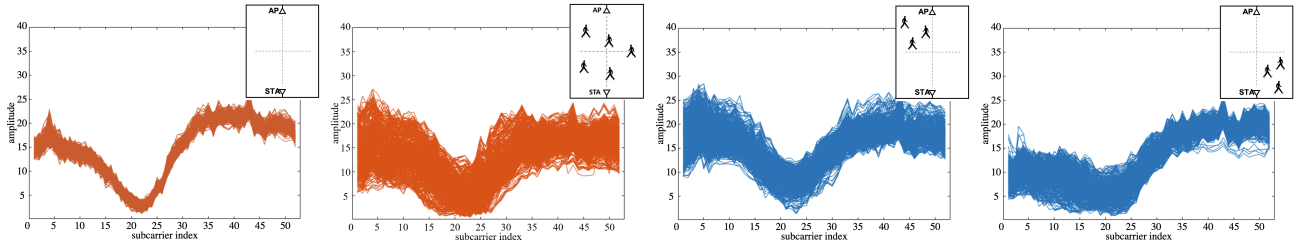
In a perspective of universality, the most common ways to estimate crowd information are vision-based approaches [1], [2]. Since these camera-based methods can predict the number of people in a crowd on both a small scale and large scale, it is naturally considered a universal solution for crowd estimation. It can simply count people by head detection using images or videos from the surveillance cameras. However, the major constraint of vision-based methods is that the occlusion of people could lead to underestimation. Moreover, it might not be possible to install a camera in every area or room, therefore the difficulty of deployment is also a non-negligible issue because of the high installation cost, in addition to privacy concerns. Meanwhile, the WiFi channel state information (hereafter, CSI) is spotlighted and utilized in many research works such as [3], [4], as a promising solution for crowd estimation because of the ubiquity of WiFi signals.

From another significant perspective of estimation diversity, most non-visual approaches only focus on crowd counting or density estimation, even though there are plenty types of solutions for crowd estimation based on non-visual sensors or radio signals such as wireless sensor network [5], PIR sensor [6], BLE [7] or WiFi [8]. As a stream of indoor localization, there is a study enabling multi-target tracking by using indoor radio tomography [9], but they have assessed their system with only two targets, which would be difficult to be considered that it is fully corresponding to crowd estimation. Vision-based approaches, meanwhile, have an advantage in that they can intuitively estimate not only the number of people but also their locations, as shown in [10], [11]. Through this work, we examine the potential of WiFi CSI as a base for simultaneous crowd estimation systems as with vision-based researches. With this simultaneous crowd estimation, we can expect advanced services such as targeted air-conditioning, evacuation route guidance, or social distance alert, based on estimated crowding level.

In this paper, we propose a new approach for crowd estimation that satisfies crowd counting and crowd localization simultaneously. The key emphasis of this work is the feasibility of precise prediction for the location of the crowd as well as the number of people using a user-device-free WiFi CSI-based system and Machine Learning (ML). The major contributions of this paper are summarized as follows:

- Proposal of simultaneous crowd estimation system for both counting and localization, unlike other existing non-visual approaches focusing on only crowd counting.
- Novel attempt to utilize a new system platform for crowd estimation using WiFi IoT devices (ESP32) and its state-of-the-art CSI toolkit, as compared to conventional methods with poor accessibility.
- Designing an experiment that enables to train two classifiers for counting and localization by appropriately regrouping data gathered from one-time data collection.
- Achieving convincing estimation accuracies by devising meaningful features and using ML, in spite of less number of links compared to other MIMO (Multiple-Input Multiple-Output) antenna WiFi systems.

The rest of this paper is organized as follows. In Section II, we first briefly address the WiFi CSI and its solutions. We



(a) Empty Room (b) Five People (c) One Section (d) Another Section

Fig. 1. Tendency of CSI Bundle Depending on Different Situations.

then show our observation in terms of CSI characteristics in Section III. The proposed system for simultaneous crowd estimation is described in Section IV. We present our performance evaluation and its result in Section V and discuss the results in Section VI. Finally, we conclude the paper in Section VII.

II. Channel State Information

In this section, we briefly present the basics of WiFi CSI, currently usable platforms, and a new promising CSI IoT solution.

A. Preliminaries

Many research works, such as crowd estimation, localization and activity recognition, are leveraging a WiFi sensing technique thanks to some solutions for access to WiFi CSI open to the public. CSI represents an estimate of the impulse response of the propagation channel between a transmitter and a receiver in the orthogonal frequency-division multiplexing (OFDM) transmission system. When we denote the OFDM system in the frequency domain, it is modeled as:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the transmitted and received complex vectors, and $\mathbf{n}$ and $\mathbf{H}$ are noise vector and channel information matrix, respectively. Since CSI is an estimate of $\mathbf{H}$, it can be denoted as $\hat{\mathbf{H}}$ which is obtained from a transmitter. $\hat{\mathbf{H}}$ contains the information of amplitude attenuation and phase shift of each subcarrier in the form of complex numbers, therefore, these measurements can be denoted as:

$$CSI = \hat{\mathbf{H}}^l = ||\hat{\mathbf{H}}^l||e^{j\angle\hat{\mathbf{H}}^l} \tag{2}$$

where $||\hat{\mathbf{H}}^l||$ and $\angle\hat{\mathbf{H}}^l$ are the CSI amplitude and phase measurement of l^{th} link (or channel). In this study, we use only the amplitude measurements for our system.

B. Conventional CSI & WiFi IoT Solution (ESP32)

There are two representative WiFi CSI-enabled solutions, Intel 5300 NIC-based CSI tool [12] and Qualcomm Atheros CSI tool [13]. Those have been widely utilized as CSI-enabled platforms in various publications so far. However, both Intel and Atheros solutions require a PC or WiFi router that is equipped with particular WiFi modules. This fundamentally restricts the accessibility to CSI data and it also may cause inconvenience in device deployment. Moreover, the Intel CSI tool has a constraint that it can provide CSI measurements of only 30 subcarriers out of 64 subcarriers. Therefore, some researchers modify those tools to fit them into their own systems.

In early 2020, an ESP32 CSI toolkit has been presented as a CSI solution, emphasizing its convenience and accessibility [14]. In this study, we use the Espressif ESP32 WiFi/Bluetooth modules with this toolkit, as a CSI measuring device for WiFi sensing. The ESP32 module operates on the mode of 2.4 GHz WiFi (bandwidth: 20 MHz), and CSI data of 52 subcarriers are obtainable. Since the ESP32 module has a single antenna, it can only exploit signals from fewer channels than other two-by-two or three-by-three MIMO WiFi architectures. As a result, we could obtain a relatively small amount of CSI data. Nevertheless, this low-cost, low-power, compact WiFi module has a great advantage in terms of easy and flexible deployment. We suppose that these compact devices have the potential to become a promising CSI IoT solution. Therefore, it makes sense to realize and evaluate a crowd estimation system using ESP32 which is a cheap and widespread WiFi platform.

III. Observations

WiFi CSI provides measurements of the signal amplitude and phase information at the subcarrier level. To investigate the CSI amplitude data, we look into a subcarrier-amplitude plot that shows the signal magnitudes of each subcarrier within a certain time interval. In our system, for example, the time-series CSI data is segmented into time-windows to convert it into overlapped CSI curves (as we will describe in Section IV). In one time-window, we call the overlapped CSI curves a CSI bundle. Figure 1 shows the CSI bundles in several different situations. CSI bundle shows a specific tendency in terms of the width and shape, therefore, it reveals a couple of characteristics in accordance with the propagation condition between WiFi Access Point (AP) and Station (STA), which is changed by moving objects or channel circumstance. Those properties can be represented in dynamic and static state-dependent characteristics, which are described in the following subsections.

A. Dynamic State-dependent Characteristic

For crowd counting, we verify the bundle-width variation depending on the number of people. If there is no person

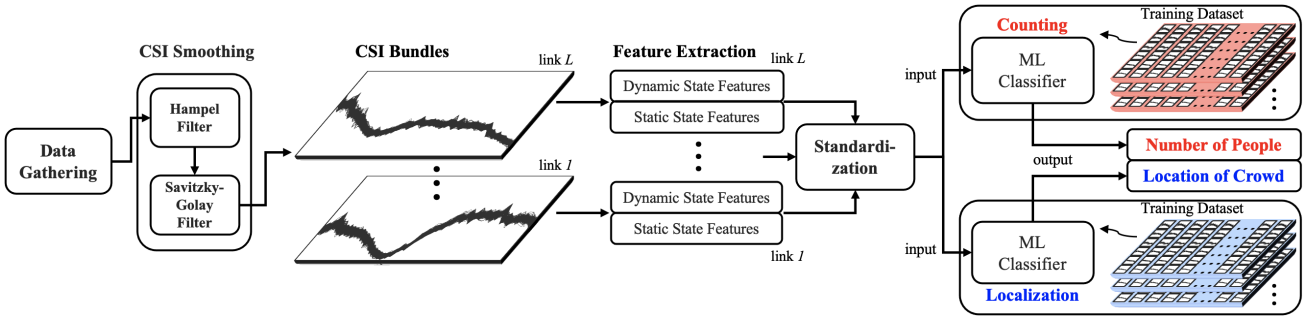


Fig. 2. Proposed System Flow.

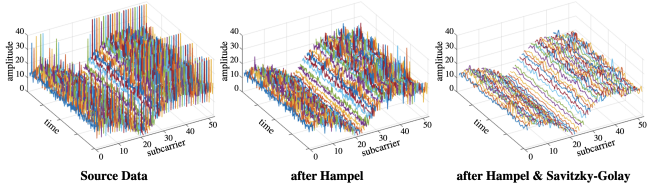


Fig. 3. CSI Smoothing Results.

between a WiFi link, the signal multipath or scattering effect is nearly constant and signal variation only comes from thermal noise or interference. Therefore, the CSI amplitudes across all the subcarriers are relatively stable. On the other hand, as the number of people in the area increases, the multipath environment becomes more and more complex due to increased obstacles. As a result, the amplitudes fluctuate widely and the CSI bundle naturally has thicker width. Figure 1 (a) and (b) represent the CSI bundles of the cases when an area is empty and five people are freely walking in the area, respectively.

B. Static State-dependent Characteristic

In a CSI bundle, we can also recognize a particular shape depending on the structural difference. The shapes of CSI curves are basically formulated by the inner circumstance of a target area. However, a cluster formed by a number of people consistently moving around within a limited area constantly affects the multipath environment of the WiFi signal. Consequently, this continuous influence affects the formation of shape tendency of the CSI bundle as well. Figure 1(c) and (d) show the difference of CSI-bundle-shape formation between two different situations when three people are freely walking within one section and another section of an area.

IV. Simultaneous Crowd Estimation System

In this section, we propose a simultaneous crowd estimation system that enables both crowd counting and localization.

A. Outline

The ultimate objective of this study is to verify the system can estimate not only how many people are in a particular area, but also which specific location of that area people are gathering at. We achieve both crowd counting and location estimation using identically processed CSI datasets, i.e., our proposed system does not require a dedicated dataset for each prediction. Therefore, we devise effective features for dynamic and static state-dependent characteristics as well as using common statistical features, in order to train the ML classifiers. Figure 2 shows the comprehensive flow of our system. We describe the system flow in the following sections, including the scheme and method of data processing and feature extraction in detail.

B. CSI Pre-processing

After gathering the WiFi CSI data which is obtained as a form of complex vector, the system first calculates amplitude values throughout the entire subcarriers. Since the CSI data is considerably noisy, it is necessary to remove the redundant components from the calculated amplitude values. For this smoothing process, we apply two kinds of filters, one is Hampel filter for eliminating spike noises, the other is Savitzky-Golay filter for removing overall white noise without distorting the tendency of the signal. Figure 3 shows the amplitudes of the time-series CSI before applying filters, after applying Hampel filter, and after applying both Hampel and Savitzky-Golay filters, respectively.

After that, the time-series amplitude values are accumulated and converted into numbers of CSI curves, then segmented into given-size time windows. The time window of CSI bundle $\mathbf{A}$ can be denoted as:

$$\mathbf{A}_k = \begin{bmatrix} a_{1,1}^{(k)} & a_{1,2}^{(k)} & \cdots & a_{1,j}^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i,1}^{(k)} & a_{i,2}^{(k)} & \cdots & a_{i,j}^{(k)} \end{bmatrix} \tag{3}$$

where $a_{i',j'}^{(k)} (1 \leq i' \leq i, 1 \leq j' \leq j)$ is an amplitude value of $j'^{\,th}$ subcarrier at $i'^{\,th}$ packet in k^{th} time window. If there are multiple links or channels in the system, $\mathbf{A}$ can also be denoted as $\mathbf{A}_k^1$, $\mathbf{A}_k^2$, $\cdots$, $\mathbf{A}_k^l$. In our work, $l = 2$ because we install two WiFi links, and we empirically set each time window to contain six seconds of CSI data with three seconds overlapping. This CSI bundle $\mathbf{A}_k$, which is consisting of CSI curves in a 6 s time-window, becomes a base unit for feature extraction.

C. Feature Extraction

As we describe in Section III, there are two kinds of characteristics that we can recognize from each CSI bundle,

TABLE I
EXTRACTED FEATURES.

Category	Feature List
Common Statistical Features	Average, Median, Variance, Standard deviation, Max, Min, Lower quartile, Upper quartile
CSI bundle-based Features	Interquartile range, Adjacent difference, Lower extreme, Upper extreme, Interextreme range, Median of Euclidean distance, PEM, Fitted curve, $1st$ Derivative
RSS-based Features	RSS variance and stardard deviation

those are dynamic state-dependent (width) and static state-dependent (shape) characteristics. Therefore, we need to figure out the features that effectively reflect the state of the target area in terms of both dynamic and static characteristics, for Machine Learning. Table I shows all the features we used in this paper.

First, we calculate several common statistical features from each subcarrier in time domain. These features basically reflect the information of width increase of a CSI bundle as the number of people increases, or the shape change of the CSI bundle as the location transition of the crowd.

It is necessary to figure out a way to enhance our system's estimation performance with some more effective features as well as statistical ones. Therefore, we now address the features which can be extracted from the CSI bundles. The *Interquartile range* is the width between the lower quartile and upper quartile. This *Interquartile range* can also reflect the difference of CSI bundle width following the number of people. The *Adjacent difference* means the sum of the numerical differences between one subcarrier and adjacent subcarriers on both sides. This is to reflect the relationship between adjacent subcarriers to the classifier, in terms of less-varying or heavily-varying subcarrier according to the state of measuring space. This *Adjacent difference* adj is denoted as:

$$adj_{i,j} = \sum_{s=1}^{S} (|a_{i,j} - a_{i,j-s}| + |a_{i,j} - a_{i,j+s}|) \tag{4}$$

where S is the number of adjacent subcarriers on both sides which will be included in adj calculation. In this paper, we decide to use $S = 2$ through empirical test.

We also devise the following feature components for avoiding the impact of time-varying change of the CSI measurement. Firstly, the packet difference $\mathbf{D}$ is calculated across all subcarriers, which is denoted as:

$$\mathbf{D} = \begin{bmatrix} a_{2,1} - a_{1,1} & a_{2,2} - a_{1,2} & \cdots & a_{2,j} - a_{1,j} \\ \vdots & \vdots & \vdots & \vdots \\ a_{i,1} - a_{i-1,1} & a_{i,2} - a_{i-1,2} & \cdots & a_{i,j} - a_{i-1,j} \end{bmatrix} \tag{5}$$

From each column of $\mathbf{D}$, we calculate q_j and Q_j, which are the lower and upper quartile values. Then, we can get values of ext_j (*Lower extreme*) and EXT_j (*Upper extreme*) by the following equation:

$$\begin{aligned} ext_j &= q_j - \alpha(Q_j - q_j) \\ EXT_j &= Q_j + \alpha(Q_j - q_j) \end{aligned} \tag{6}$$

where α is a weight of extreme value. The *Interextreme range* can be obtained by $EXT_j - ext_j$. Similarly, we can calculate the Euclidean distance between the one CSI vector and its previous CSI vector. The Euclidean distance $eucd$ at i^{th} packet can be denoted as:

$$eucd_i = \sqrt{(a_{i,1} - a_{i-1,1})^2 + \cdots + (a_{i,j} - a_{i-1,j})^2} \tag{7}$$

then, the *Median of Euclidean distance* in one time-window can be obtained by $\textbf{median}(eucd_2, eucd_3, \cdots, eucd_i)$.

We also adopt a metric *PEM* as one of our features, which is proposed in [15]. *PEM* means the percentage of non-zero elements in the dilated CSI matrix. This metric basically follows the tendency of CSI variation, i.e., the more CSI amplitude values fluctuate intensively by many moving people, the more *PEM* metric gets increased. Briefly saying, this method firstly forms the two-dimensional initial matrix for each subcarrier, which contains "1" elements along with the shape of CSI amplitude signal (single "1" in each column), the other elements are all "0"s. Then, a dilation mask goes through the matrix and makes all the adjacent elements of "1"s also become "1"s. The authors called this process matrix dilation. In this dilation process, there would be plenty of overlapped "1"s if the original CSI signal is stable, this consequentially results in a less percentage of non-zero elements. Otherwise, the *PEM* will show a relatively high percentage. In this way, the *PEM* vector (consisting of every subcarrier's *PEM*) can represent the number of people in the area.

To be specific, we assume that there is a CSI amplitude matrix $\mathbf{A}$ with the size of $i \times j$. Here, i is the number of subcarriers and j is the number of packets. We firstly generate the all "0" initial matrix $\mathbf{M}_0$ with the size of $R \times j$. Next, we calculate r by the following equation:

$$r = \left\lfloor \frac{\mathbf{A}_{ij} - \mathbf{A}_{min}}{\mathbf{A}_{max} - \mathbf{A}_{min}} (R - 1) \right\rfloor + 1 \tag{8}$$

where R is the number of rows, i.e., the matrix resolution of $\mathbf{M}_0$. When r is decided, we set the elements on r^{th} row and j^{th} column in $\mathbf{M}_0$ to "1". These "1"s will be expanded by $D \times D$ dilation mask. As a result, we get the dilated matrix M and the *PEM* vector. In this study, we empirically define the parameters for matrix resolution R and dilation coefficient D as $R = 100$ and $D = 15$.

Meanwhile, we apply polynomial curve fitting along the average line of the CSI bundle. The *Fitted curve* shows the shape of the CSI bundle itself. We also use the *1st Derivative* of the *Fitted curve*, to clarify at which point of the fitted curve has peaks, valleys, or sharp slopes. These features particularly contribute to the localization part of crowd estimation by the shape difference of each CSI bundle. Finally, we utilize variance and standard deviation calculated from

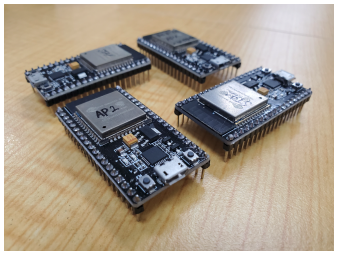


Fig. 4. ESP32 modules.

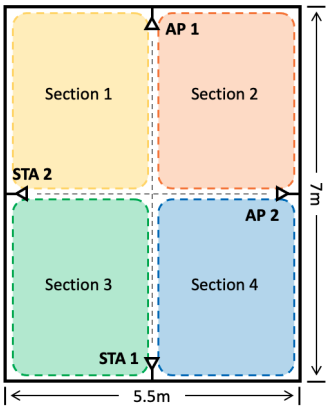


Fig. 5. Experiment plane.

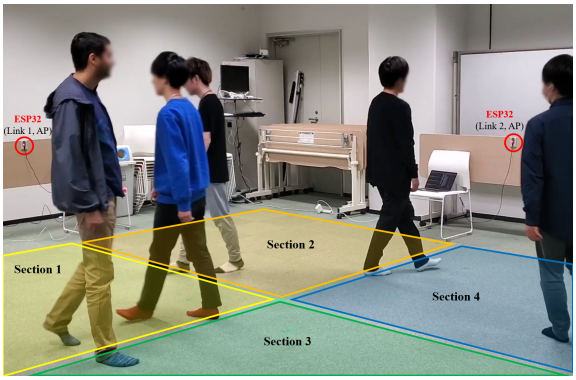


Fig. 6. Experiment Scene.

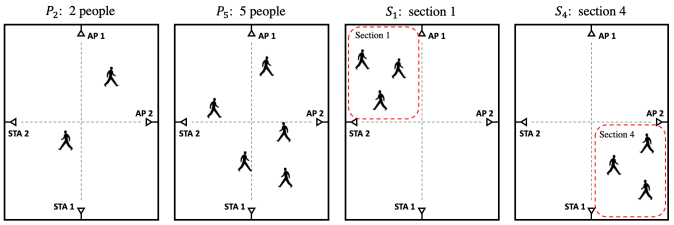


Fig. 7. P_2 and P_5 Cases in S_{oth}. Fig. 8. S_1 and S_4 Scenarios with P_3.

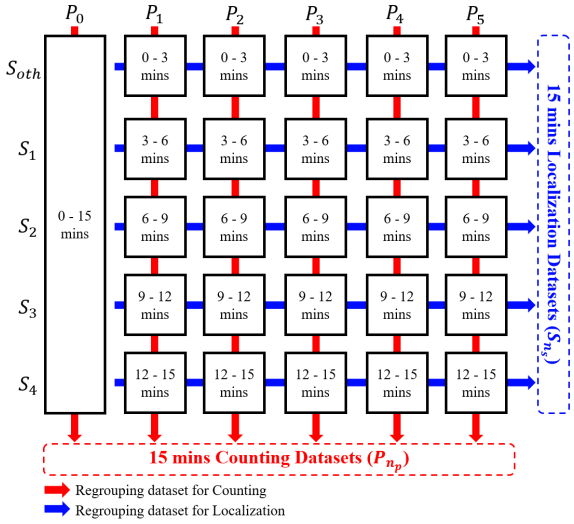


Fig. 9. Data Collecting and Regrouping Scheme.

received signal strength (RSS). These RSS-based features also show an upward trend as the number of people increases.

D. Standardization & Classification Model

After feature extraction, all features are standardized according to standard normal distribution $N(0,1)$ to fit the scales between different features, before training. Then, we train four basic ML classifiers to evaluate the performance of crowd counting and localization. In this paper, the classifiers that we used for system evaluation are Random Forest (RF), Logistic Regression (LR), Support Vector Classification (SVC), and Light Gradient Boosted Machine (LGBM).

V. Performance Evaluation

In this section, we verify the system performance of simultaneous crowd estimation through several experiments.

A. Experimental Setup

Experimental Environment: We collect the CSI data through a multi-scenario experiment with up to five participants. As mentioned in Section II, we use four ESP32 devices, which are shown in Fig. 4, with deploying them at all the edges of the experiment area. Unlike one pair of WiFi routers are usually installed in other conventional researches using Intel or Atheros CSI solutions, we place the two pairs of ESP32 modules to make two WiFi links vertically and horizontally crossing over the target area. This enables the system to faithfully observe the change of CSI measurements with regard to the movement of walking people covering the whole target area. For our experiment, we gather the 52-subcarrier-CSI data at an approximately 185 Hz packet rate. We perform the experiments in a normal size room (7 m by 5.5 m) which is divided into four equal sections for crowd localization, as shown in Fig. 5. Figure 6 shows the actual scene of our experiment.

Data Collection Scheme: To verify our insight of simultaneous crowd estimation, we design and conduct the experiments which contain five scenarios on the cases of the number of people walking in the experiment area. Here, five scenarios mean the situations that the cluster of people is walking at different sections of the area. The cases of the number of people 0-5 are denoted as P_{n_p} ($n_p = 0, 1, ..., 5$), and the scenarios related to the section number 1-4 correspond to S_{n_s} ($n_s = 1, ..., 4$, and oth indicates full-free walk). Excepting S_{oth}, the scenarios S_{n_s} are corresponding to the situation that the participants walk freely within a particular section n_s. In the scenario S_{oth}, on the contrary, the participants perform free walking all over the experiment area. The examples of P_{n_p} cases and S_{n_s} scenarios are presented in Figs. 7 and 8. In each and every case and scenario, all the participants walk randomly within the given space, without any guidance/limitation about how to walk.

As we can see in Fig. 9, all the scenario in each case of P_{n_p}

contain three minutes-long data. After collecting CSI data for every case and scenario, we combine the data into two kinds of datasets, which are for counting and localization. The datasets of P_{n_p} are constructed with all the scenarios data collected with n_p people. These 15 minutes-long datasets are used for crowd counting. On the other hand, the datasets of S_{n_s} include all the data of when the 1-5 participants are walking in section n_s. These 15 minutes-long datasets become the sources for crowd localization.

We carry out identical experiments three times in the same room, but during different times, in the morning, afternoon and evening. This is to check the difference in system performance originating from circumstance changes, such as temperature, humidity, or signal interference. The experiments in each different time are distinguished as Session 1, 2, and 3.

B. Evaluation Method

We assess the system performance using CSI amplitude data collected in three different-time sessions. The experiment results are evaluated by four basic classification models, which are mentioned in Section IV. We sequentially describe the in-session training and its test results (i.e., k-fold cross-validation), and the result of leave-one-session-out cross-validation. We apply the hyperparameter adjustment method for each ML classifier using the Optuna framework [16].

In our system, we aim to simultaneously estimate the number of people (i.e., crowd counting) and the section in which those people are walking (i.e., crowd localization). Therefore, we construct two classification models for each crowd estimation, and then those models are assessed by both k-fold cross-validation and leave-one-session-out cross-validation. The reason why we describe the results of both validation methods is that it is necessary to confirm how the system performance is affected by changed CSI measurements as time passes. Since CSI is sensitive to the transition of measuring environment, the system performance can be degraded due to minor changes occurred by some factors such as temperature, humidity, or structural transition.

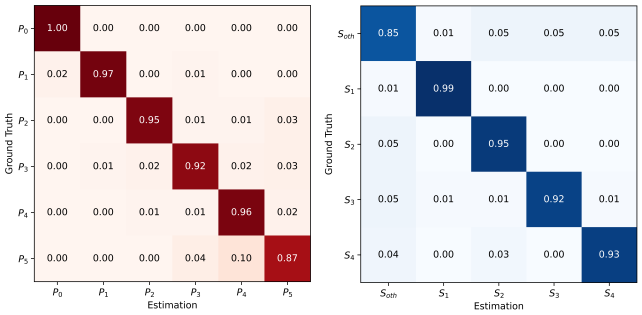
C. Results

*1) **k-fold Cross-validation**:* The k-fold cross-validation is a machine learning evaluation method to assess a trained classifier for a single session dataset. The whole dataset is split into k folds of datasets from the first. When one fold is selected as test data, the other $k-1$ folds become training data. After repeating this process k times, the system performance is derived by averaging all results from k trials. Specifically, we adopt the stratified k-fold method which splits the folds by criteria ensuring that each fold contains the same ratio of target classes data. In this study, we empirically set the number of folds as $k = 7$.To evaluate all three sessions, two classification models are constructed for each session, one for crowd counting by the features extracted from datasets P_{n_p}, the other for crowd localization by the features extracted from datasets S_{n_s}. Each session is evaluated by four classifiers mentioned in Section IV.

TABLE II
OVERALL ACCURACY OF K-FOLD CROSS-VALIDATION.

	Session 1		Session 2		Session 3	
Classifier	*c.*	*l.*	*c.*	*l.*	*c.*	*l.*
RF	.89	.91	.87	.91	.86	.93
LR	.94	.93	.90	.94	.88	.93
SVC	.93	.93	.91	.94	.89	.92
LGBM	.92	.93	.89	.92	.90	.95

c.: counting, *l.*: localization.



(a) Counting by LR (b) Localization by LR

Fig. 10. Confusion Matrices (Session 1, k-fold Cross-validation).

Table II shows the overall classification results. We can see the results of counting and localization of each session. In crowd counting, the highest accuracies among all classifiers are 94%, 91%, and 90% in Session 1, 2, and 3, respectively. Meanwhile, in crowd localization, the highest accuracies are 93%, 94%, and 95% in Session 1, 2, and 3, respectively.

Figure 10 shows the confusion matrices classified by LR in Session 1. The numbers in matrix show the prediction accuracy compared to ground truths P_{n_p} or S_{n_s}. Although the classifier would be naturally confused for accurate prediction when the number of people increases, we confirm that our system still shows over 87% of counting accuracy even in the case of five people. In crowd localization, only S_{oth} which is the scenario of the entire area walking shows under 90% prediction, but the other results of section predictions all show more than 90% in scenario S_1-S_4.

*2) **Leave-one-session-out Cross-validation**:* We have separate datasets of three sessions which are collected in the same room, on the same scenarios, but at different times. We suppose that the tendency of CSI data changes based on the time of day. In that case, a classifier trained by only a certain session's data might not be efficient for the others. However, there are not many conventional works that address the time-variant influence in CSI measurements. Hence, to confirm this variation between different sessions, we conduct leave-one-session-out cross-validation. Here, leave-one-session-out means, one whole session is selected as test data to verify a classifier trained by the other sessions. This process continues until every session becomes a test session at least once. Finally, the system accuracy is calculated by averaging all the session results.

Figure 11 presents the overall accuracy of cross-validation.

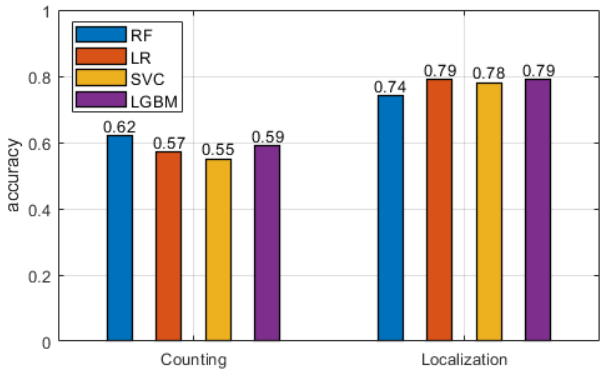
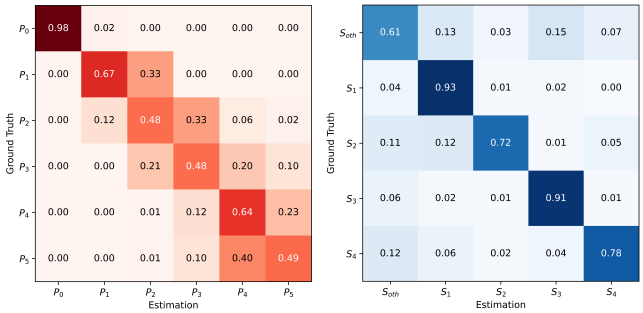


Fig. 11. Accuracy of Leave-one-session-out Cross-validation.



(a) Counting by RF (b) Localization by LR

Fig. 12. Confusion Matrices (Leave-one-session-out Cross-validation).

RF achieves the best performance with 62% for crowd counting, on the other hand, LR and LGBM show the highest accuracy with 79% in crowd localization. As with our assumption, overall accuracy decreased compared to the results by k-fold cross-validation. We suppose that CSI data is influenced by time-dependent factors. The detailed accuracy according to all the cases and scenarios can be found in Fig. 12 as confusion matrices.

VI. Discussions

In the result of leave-one-session-out cross-validation, if we look into the confusion matrix in Fig. 12(a) in a little more detail, we can see that the errors are mostly distributed right next to the correct estimation. That is, if we extend the margin of error to ± 1 person, then the average accuracy reaches nearly 95%. But still, it is necessary to consider that the performance of the crowd estimation system has declined compared to in-session training due to time-varying environmental differences, especially in the case of counting. For future work, we will address how to overcome critical issues caused by time-variant environmental parameters such as temperature, fine-structure-transition, etc. Furthermore, we are also considering the further evaluations for system robustness and adaptability, by comparison of various cases such as increasing the number of people, testing in several different-size rooms, and situation of multiple clusters on different sections. Most of all, we believe that installing two more diagonal links will considerably help the system to detect and estimate the crowd information without blind spots in the target area. Lastly, we have presented our evaluation results by classification accuracy in this work, however in case of counting, it could be more suitable to evaluate the system and show the results in a quantity of counting error. This will be fulfilled in our future work by the regression model-based analysis.

VII. Conclusions

In this paper, we proposed a simultaneous crowd estimation system that simultaneously performs both crowd counting and localization, by WiFi CSI and Machine Learning. We leveraged the CSI bundle (overlapped subcarrier-amplitude curves during a given time interval) as the source for extracting features that contain characteristics depending on the dynamic state (counting) and static state (localization). We carried out experiments at three different-time sessions for evaluation. We confirmed that we could achieve up to 94% counting accuracy and 95% localization accuracy in k-fold cross-validation. In leave-one-session-out cross-validation, we achieved up to 62% counting accuracy and 79% localization accuracy. By this study, we took the first step of using ESP32 WiFi CSI in the field of wireless sensing and have confirmed the feasibility of the ESP32-based crowd estimation system in both terms of counting and localization.

Acknowledgment

This work was supported in part by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research number JP19H05665 and JP20H04177.

References

[1] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.

[3] H. Zou, Y. Zhou, J. Yang, and C. J. Spanos, "Device-free occupancy detection and crowd counting in smart buildings with wifi-enabled iot," *Energy and Buildings*, vol. 174, pp. 309–322, 2018.

[4] J. Li, P. Tu, H. Wang, K. Wang, and L. Yu, "A novel device-free counting method based on channel status information," *Sensors*, vol. 18, no. 11, p. 3981, 2018.

[5] Y. Yuan, C. Qiu, W. Xi, and J. Zhao, "Crowd density estimation using wireless sensor networks," in *2011 seventh international conference on mobile Ad-hoc and sensor networks*. IEEE, 2011, pp. 138–145.

[6] K. Weekly, M. Jin, H. Zou, C. Hsu, C. Soyza, A. Bayen, and C. Spanos, "Building-in-briefcase: A rapidly-deployable environmental sensor suite for the smart building," *Sensors*, vol. 18, no. 5, p. 1381, 2018.

[7] A. Filippoupolitis, W. Oliff, and G. Loukas, "Bluetooth low energy based occupancy detection for emergency management," in *2016 15th International Conference on Ubiquitous Computing and Communications and 2016 International Symposium on Cyberspace and Security (IUCC-CSS)*. IEEE, 2016, pp. 31–38.

[8] S. Depatla and Y. Mostofi, "Crowd counting through walls using wifi," in *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2018, pp. 1–10.

[9] S. Nannuru, Y. Li, Y. Zeng, M. Coates, and B. Yang, "Radio-frequency tomography for passive indoor multitarget tracking," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12, pp. 2322–2333, 2012.

[10] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 1217–1226.

[12] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM CCR*, vol. 41, no. 1, p. 53, Jan. 2011.

[13] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity wifi," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: ACM, 2015, p. 53–64. [Online]. Available: http://doi.acm.org/10.1145/2789168.2790124

[14] S. M. Hernandez and E. Bulut, "Performing WiFi sensing with off-the-shelf smartphones," in *PerCom Demos 2020: 18th Annual IEEE International Conference on Pervasive Computing and Communications Demonstrations (PerCom Demos 2020)*, Austin, USA, Mar. 2020.

[15] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu, and Z. Jiang, "Electronic frog eye: Counting crowd using wifi," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 361–369.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.