Device-Free Indoor People Counting Using Wi-Fi Channel State Information for Internet of Things

Yen-Kai Cheng and Ronald Y. Chang
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
Email: {ykc, rchang}@citi.sinica.edu.tw

***Abstract*—People/crowd counting is a critical technique in many people-centric Internet of Things (IoT) applications, e.g., security monitoring and energy management for smart homes. Device-free people counting systems can in general be categorized as image-based and non-image-based. Non-image-based methods have the advantages of being economical and nonintrusive, as only ambient wireless signals from off-the-shelf wireless devices such as Wi-Fi are used. In this paper, we propose a non-image-based people counting system based on the deep neural network (DNN) model using fine-grained physical-layer wireless signatures such as Wi-Fi channel state information (CSI). Only one Wi-Fi transmitter and one laptop receiver are required, and people are not required to wear or carry any equipment (i.e., device-free). A novel feature space expansion scheme that incorporates the dynamic information of CSI measurements is proposed for the DNN model to enhance its performance. Real testbed experiments showed that the proposed system can achieve as high as 88% average correct classification rate in estimating the exact number of the crowd of size up to nine people in the most general indoor scenario.**

I. Introduction

People/crowd counting is crucial in many people-centric Internet of Things (IoT) applications, e.g., traffic management, elderly monitoring, smart home energy management, etc. Classical solutions to this problem can be broadly categorized as image-based and non-image-based methods. With the advances of imaging technologies, an image-based counting model estimates the crowd density by analyzing the human characteristics in high-resolution images in the pixel, texture, or object level, often achieving a superb detection accuracy [1]. However, the sensitivity to the lightness of the scenario background, the high computational costs, the privacy concerns, etc. are limiting factors that could confine the applicability of image-based methods.

Non-image-based methods perform people counting based on wireless signals (ultrasound, infrared, Wi-Fi, etc.), especially the received signal strength (RSS) as an indicator of signal propagation through the space [2]. The advantages of these methods are that they are economical and practical, as off-the-shelf wireless devices such as Wi-Fi can be used and no extra equipment is required. Many studies on RSS-based people counting systems observed that the RSS profiles are sensitive to the change in the number of people in the environment and attempted to correlate the two [2]–[4]. In [3], 16 sensor nodes in a 18×18 m^2 room were used and an RSS fingerprint approach was adopted to estimate the crowd density. The approach was shown achieving 86% accuracy in classifying three categories of crowd density, i.e., 0–3 people, 4–10 people, and more than 10 people. In [2], a mathematical model was proposed to estimate the probability distribution of the number of people, based on the observation that people may obstruct the line of sight and there is a scattering effect on the transmitted signal. The counting model was tested in both indoor and outdoor environments with up to nine testers, and an error rate of less than two people was shown. In [4], a smartphone-based people counting system called Wi-Counter was proposed. A noise-reduction process on the Wi-Fi signals and a five-layer neural network as the estimation model were employed. Wi-Counter yields an error rate of $< 15\%$ as compared to previous studies (33% in [5] and 22% in [6]) in counting 50 people in the same testbed environment (a 96 m^2 classroom).

RSS measurements are subject to environmental noises and multipath fading. To improve the estimation accuracy, expanded RSS data acquisition is often performed, which inevitably incurs more deployment, time, and/or computational costs (e.g., more transmitters need to be installed). Recently, channel state information (CSI) has been introduced as an alternative signature to overcome the above challenges [7]. CSI provides fine-grained physical-layer information, such as multipath signal components (i.e., subcarriers) with amplitude/phase information for each subcarrier, as compared to the coarse-grained RSS. CSI has recently found many applications such as localization [8], gesture recognition [9], human activities recognition [10], breathing and heart rates tracking during sleep [11], customer's behavior analysis [12], etc.

In the application of indoor people counting, CSI profiles can potentially provide more discriminative features and are more sensitive to the number of people in the environment than RSS profiles, without the need to install many transmitters and receivers. CSI-based people counting was studied in [13], [14]. The relationship between the number of people and the collected CSI was theoretically examined in [13]. It was reported that 98% and 70% estimation errors are less than or equal to two persons for indoor and outdoor environments, respectively, in the testing with a maximum of 30 people. In

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grants MOST 104-2628-E-001-002-MY2 and MOST 106-2628-E-001-001-MY3.

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

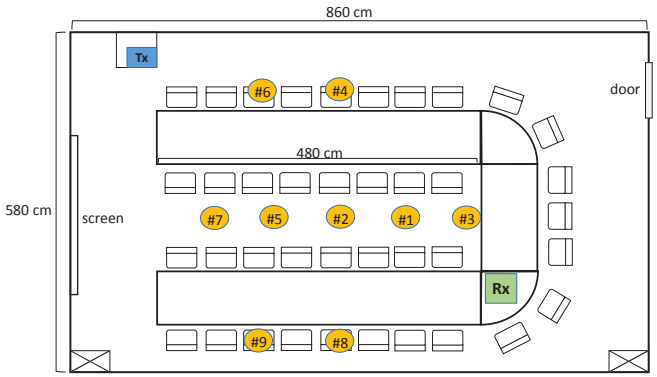


Fig. 1. Floor plans of the conference room in the basement of CITI, AS. The transmitter (a Wi-Fi AP) and receiver (an ASUS laptop computer with Ubuntu 10.04 LTS and Intel Wi-Fi Wireless Link 5300 802.11n MIMO radios) are placed at the marked locations. In the n-people condition ($n = 1, \ldots, N$), where $N = 9$ is shown, the n people assume locations from $\#1$ to $\#n$ (for offline training and online testing with fixed locations; see Table I).

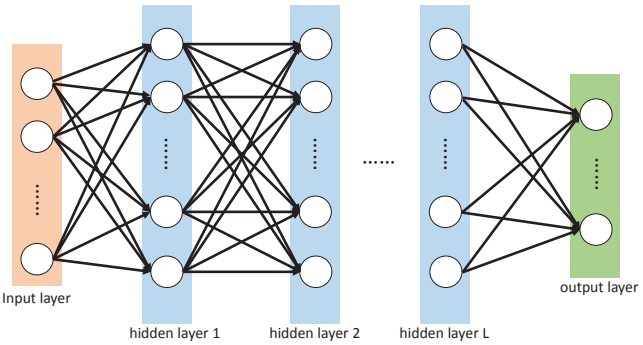


Fig. 2. Structure of a neural network model.

[14], a dimension-reduction method to reduce a CSI vector into a two-dimensional feature space was introduced and a linear classifier was used for crowd counting. Classification tests showed that the model trained in one room can also be used in the other two rooms with different sizes for testing. Specifically, in the testing with a maximum of 7 people, about 80–90% estimation errors are less than or equal to two persons. Note that the CSI-based people counting systems in [13], [14] both considered some error tolerance (e.g., two persons) in evaluating the methods, with the correct classification rates of *exact* numbers of the crowd being low (see Fig. 10(a) of [13] and Tables 1–2 of [14]).

In some applications, such as home security, estimating the exact number of the crowd is desired. In this paper, we are motivated to propose a people counting system, with the objective of estimating the exact number of people in the target environment. The proposed system is based on deep neural network (DNN) as the classifier and CSI as the features, with the smallest economical deployment of one transmitter and one receiver. Our real-world experiments showed that the proposed system can detect the exact number of the crowd of size up to 9 people (10 classes) with around 88% accuracy in the most general indoor scenario.

The remainder of the paper is organized as follows. Sec. II describes the indoor people counting problem. Sec. III presents the proposed method. Sec. IV presents the experimental results and discussion. Finally, Sec. V concludes the paper.

II. The Device-Free Indoor People Counting Problem

We consider an indoor environment where there is an unknown number of people in the environment. The people are not expected to wear or carry any particular equipment or device (i.e., device-free). A pair of transmitter (e.g., a Wi-Fi access point (AP)) and receiver (e.g., a wireless network adapter) are preinstalled at fix locations in the environment, and the CSI between the transmitter and receiver is measured at the receiver. The objective is to estimate the *exact* number of people in the indoor environment based on the CSI measurements.

More specifically, we consider the conference room in the basement of Research Center for Information Technology Innovation (CITI), Academia Sinica (AS), Taipei, Taiwan, as shown in Fig. 1. A pair of transmitter (i.e., a commercial Wi-Fi AP with two antennas) and receiver (i.e., an ASUS laptop computer with Ubuntu 10.04 LTS and Intel Wi-Fi Wireless Link 5300 802.11n MIMO radios with two antennas [7]) are placed at the marked locations. There is an unknown number of up to N people in the conference room. The dimensions of the conference room, the locations of the transmitter and receiver, and the locations of people are depicted in Fig. 1. The nearest distance between any two locations of people is at least 100 cm. Note that in our experiments, as will be detailed in Sec. IV, we consider both scenarios of fixed locations of people (as shown in Fig. 1) and arbitrary locations of people (where people assume arbitrary locations in the room) for online testing.

III. The Proposed DNN-Based Approach to Indoor People Counting Using CSI Features

A. Deep Neural Network Model

DNN has found many applications such as face detection, speech recognition and natural language processing, etc. DNN originated from shallow neural network, the early term of artificial neural network (ANN). The basic structure of a neural network model is composed of three parts: an input layer, L hidden layers, and an output layer, as illustrated in Fig. 2. In the figure, circles represent neuron-like nodes, inspired from the human's nervous system, which are basic units of the neural network. The layers are organized in a hierarchical way, and the nodes of two adjacent layers are fully connected. In many applications, an increasing number of the hidden layers improves the performance at the cost of higher computational complexity. The relationship between the $(l-1)$th and lth hidden layers can be described by

$$\mathbf{a}_l = \sigma_{\text{sig}}(\mathbf{W}_l \mathbf{a}_{l-1} + \mathbf{b}_l), \quad l = 2, \ldots, L \tag{1}$$

where $\mathbf{a}_l$ is a $a_l \times 1$ vector and $\mathbf{a}_{l-1}$ is a $a_{l-1} \times 1$ vector, with a_l and a_{l-1} being the numbers of nodes in lth and $(l-1)$th layers, respectively; $\mathbf{W}_l$ and $\mathbf{b}_l$ are the weight matrix and bias vector in the lth layer, respectively, whose values are randomly initialized before training; $\sigma_{\text{sig}}(\cdot)$ is the sigmoid function commonly used as the activation function in a neural network model, which is to produce a nonlinear decision boundary via nonlinear combinations of the weighted inputs. Note that $\mathbf{a}_0$, which is a $a_0 \times 1$ vector, represents the input layer, where a_0 is the dimension of the feature space for each input data. The sigmoid function is given by

$$\sigma_{\text{sig}}(t) = 1/(1+\exp(-t)) \tag{2}$$

where the input of the function is a linear combination of weights and inputs in each hidden layer, and the output of the function will be restricted to $(0,1)$. At the output layer, we use the softmax function $\sigma_{\text{soft}}(\cdot)$ to yield a categorical distribution of the desired classes $\hat{\mathbf{y}}$ based on the nodes from the last hidden layer $\mathbf{a}_L$, i.e.,

$$\hat{\mathbf{y}} = \sigma_{\text{soft}}(\mathbf{a}_L). \tag{3}$$

We select the class with the maximum value from $\hat{\mathbf{y}}$ as our prediction.

With the many hidden layers in the DNN model, a common way to train the DNN model is back-propagation. For the ith training sample, its label $\mathbf{y}_i$ is a $1 \times (N+1)$ vector where the $(n+1)$th element will be one and the other elements will be zero if there are n people in the environment (considering that the first element corresponds to the no-people condition). The estimation by the DNN model $\hat{\mathbf{y}}_i$ is also a $1 \times (N+1)$ vector calculated from (3), where each value in the vector indicates the probability for each condition.

Given $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$, DNN can iteratively adjust the parameters of the weight matrix with the objective of reducing the training errors from the last hidden layer to the first hidden layer. Cross-entropy is used to define the training error e_k during the kth iteration, i.e.,

$$e_k = -\frac{1}{S}\sum_{i=1}^{S}\sum_{j=1}^{N+1} y_{i,j}\log \hat{y}_{i,j}^{k} \tag{4}$$

where S is the total number of the training data, $y_{i,j}$ is the jth element of $\mathbf{y}_i$, and $\hat{y}_{i,j}^{k}$ is the jth element of $\hat{\mathbf{y}}_i$ in the kth iteration.

Once the training error e_k is calculated, the parameters of weight matrix and bias vector in the kth iteration will be updated in the $(k+1)$th iteration based on the gradient descent. We denote $\mathbf{W}_l$ and $\mathbf{b}_l$ in the kth iteration by $\mathbf{W}_l^{(k)}$ and $\mathbf{b}_l^{(k)}$, and the gradient descents of $\mathbf{W}_l^{(k)}$ and $\mathbf{b}_l^{(k)}$ by $\Delta\mathbf{W}_l^{(k)}$ and $\Delta\mathbf{b}_l^{(k)}$. Thus, we have

$$\mathbf{W}_l^{(k+1)} = \mathbf{W}_l^{(k)} + \Delta\mathbf{W}_l^{(k)} = \mathbf{W}_l^{(k)} - \eta(\partial e_k/\partial \mathbf{W}_l^{(k)}) \tag{5}$$

$$\mathbf{b}_l^{(k+1)} = \mathbf{b}_l^{(k)} + \Delta\mathbf{b}_l^{(k)} = \mathbf{b}_l^{(k)} - \eta(\partial e_k/\partial \mathbf{b}_l^{(k)}) \tag{6}$$

where η is the learning rate. The training will continue until the training error is less than the specified threshold or the training iteration exceeds a predetermined maximum number.

B. DNN-Based Indoor People Counting Model with CSI Features

A DNN-based approach to people counting generally comprises two phases: offline training phase and online testing phase. During the offline training phase, the CSI is measured for the cases of zero to N people in the environment, where N is the known maximum number of people in the application. The collected CSI that corresponds to the n-people condition $(n = 0, 1, \ldots, N)$ is denoted by $\mathbf{X}_n$, which is a $M_n \times N_{\text{sub}}$ matrix, where M_n is the total number of CSI measurements and N_{sub} is the total number of subcarrier indices for each sampling time. (For example, in our testbed environment, since the transmitter and receiver each has two antennas, and there are 30 subcarriers for each transmit-receive antenna pair in the 802.11n system, we have $N_{\text{sub}} = 2 \times 2 \times 30 = 120$.) The training data are collected as $\mathbf{X}_{\text{train}} = [\mathbf{X}_0^\top, \ldots, \mathbf{X}_N^\top]^\top$. A counting model f is trained with the training data $\mathbf{X}_{\text{train}}$ in the offline phase. Then, in the online phase, we collected another set of data, i.e., testing data $\mathbf{X}_{\text{test}}$, for testing the model f.

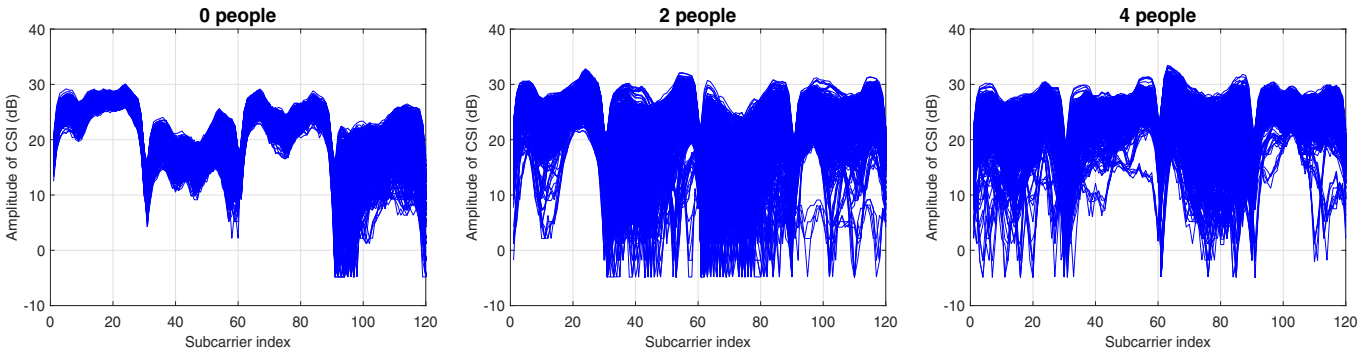
Fig. 3 exemplifies the measured CSI for the no-people, two-people, and four-people conditions in the considered indoor environment. As can be seen, the collected CSI data carry sufficiently discriminative features with respect to the number of people in the indoor environment. Thus, the CSI data are relevant features that are fed into a counting model for estimating the number of people in the indoor environment.

C. The Proposed Feature Space Expansion

To mitigate the effect of environmental noises on CSI measurements and the performance of the prediction model, preprocessing of the input data, such as noise reduction and feature selection, is often performed (see, e.g., [4], [10]). Noise reduction, or denoising, aims to remove noises from the input data by employing some kind of filtering. Feature selection aims to remove features that are noisy or irrelevant based on a statistical method (e.g., entropy) or physical meaning (e.g., max-mean). Both methods are commonly used in the model training. However, noise reduction requires additional time/computational complexity in performing filtering, and feature selection could compromise the robustness of the model due to the reduced feature space.

The nodes in the hidden layers (often called anchor nodes) in a DNN model, in theory, will be trained to present the frequently occurring patterns of the input data or of the previous hidden layer, which provides the same effect of feature selection [15]. This effective feature selection capability of the DNN model, if coupled with the expanded feature space of the input data to provide more information to the model for selection, could enhance the performance of the model. Thus, we propose a feature space expansion scheme, where the differences between the CSI measurements and the average CSI measurements in each n-people condition are calculated to provide "dynamic information" which comprises the additional features. Specifically, we calculate

$$\mathbf{X}'_n = \mathbf{X}_n - \bar{\mathbf{X}}_n, \quad n = 0, 1, \ldots, N \tag{7}$$



(a) (b) (c)

Fig. 3. Amplitude of CSI for $N_{\text{sub}} = 120$ channels (30 subcarriers for 2×2 transmit-receive antenna pairing) measured in some time duration, for (a) no-people, (b) two-people, and (c) four-people conditions (all fixed locations).

TABLE I
SUMMARY OF CSI MEASUREMENTS IN ONLINE AND OFFLINE PHASES

Phase	Scenario	# collected samples
Offline training	Fixed locations of people	$10,937$
Online testing (Scenario A)	Fixed locations of people	$10,565$
Online testing (Scenario B)	Arbitrary locations of people	$11,818$

where $\bar{\mathbf{X}}_n$ is the sample mean vector of $\mathbf{X}_n$. Then, we expand the training data as

$$\mathbf{X}'_{\text{train}} = [\mathbf{X}_{\text{train}}, \mathbf{X}'] \tag{8}$$

where $\mathbf{X}' = [\mathbf{X}'^{\top}_0, \ldots, \mathbf{X}'^{\top}_N]^{\top}$ has the same dimensions as $\mathbf{X}_{\text{train}}$.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

We conduct real-world experiments in the testbed environment shown in Fig 1. In the offline training phase, the participating people are instructed to remain in their designated locations as shown in Fig 1. There is no restriction to their poses (standing, sitting, etc.) or activities (talking, using computers, etc.) in the designated locations. This simulates, for example, a real office environment. CSI measurements are performed for the n-people condition, for $n = 0, 1, \ldots, N$, where $N = 9$. The collected data are used as the training data for the multi-classification DNN model. In the online testing phase, we consider two scenarios: A) fixed locations of people (i.e., the same scenario as in the offline training phase), and B) arbitrary locations of people, where the participating people assume arbitrary locations in the room and the locations may vary in different instances of data collection. The duration of each CSI collection (offline or online) is about 3 minutes. There are about 1,000 samples collected for each n-people condition, and thus the total number of samples is around 10,000, as summarized in Table I.

In our work, we use three hidden layers ($L = 3$) to build an encoder structure of the DNN model [16], and the three hidden layers (from first to third) contain 120, 60, and 30 nodes, respectively. The learning rate is set to 0.004 for the first layer and 0.0025 for other layers. Also, similar to [17], [18], we follow the auto-encoder method to pre-train the layers before the back-propagation training (fine-tune training) of the entire DNN model, to accelerate the training process and improve the efficiency of the model. We set the number of training iterations to pre-train the three hidden layers to 400, 200, and 150, respectively, and we set 400 iterations for the fine-tune training.

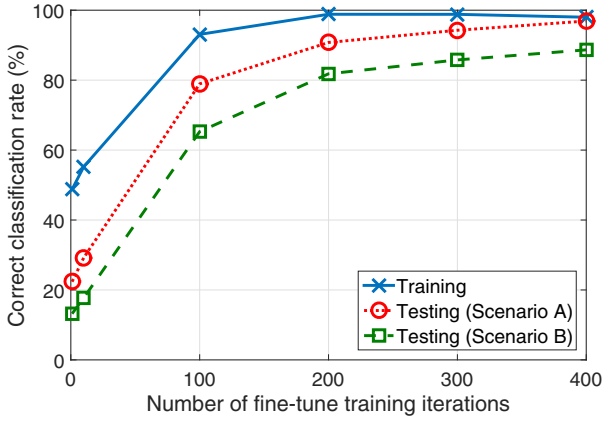


Fig. 4. The correct classification rate (CCR) vs. the number of fine-tune training iterations in the DNN model with feature space expansion, for the offline training phase and two scenarios in the online testing phase.

We compare the proposed DNN-based method with k-nearest neighbor (k-NN) [19] and Gaussian maximum likelihood estimation (MLE) [20]. We set $k = 1, 3, 5, 7$ for k-NN. The proposed DNN-based method is also compared with its no-feature-expansion counterpart, for which we set the number of training iterations to pre-train the three hidden layers to 90, 60, and 30, respectively.

B. Results and Discussion

The proposed DNN model is well-trained. We first examine whether the DNN model in our system is well-trained.

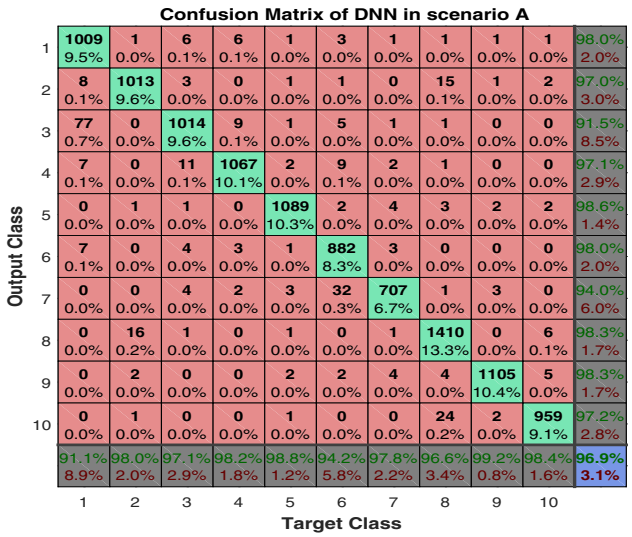
TABLE II
THE CORRECT CLASSIFICATION RATE FOR k-NN, MLE, AND THE PROPOSED DNN, WITH AND WITHOUT FEATURE SPACE EXPANSION AND IN TWO TESTING SCENARIOS

w/o feature space expansion	k-NN ($k=1$)	MLE	DNN
Scenario A	47.1%	40.99%	53.85%
Scenario B	21.95%	19.99%	48.87%
w/ feature space expansion	k-NN ($k=7$)	MLE	DNN
Scenario A	73.87%	44.34%	96.90%
Scenario B	40.5%	23.78%	88.66%

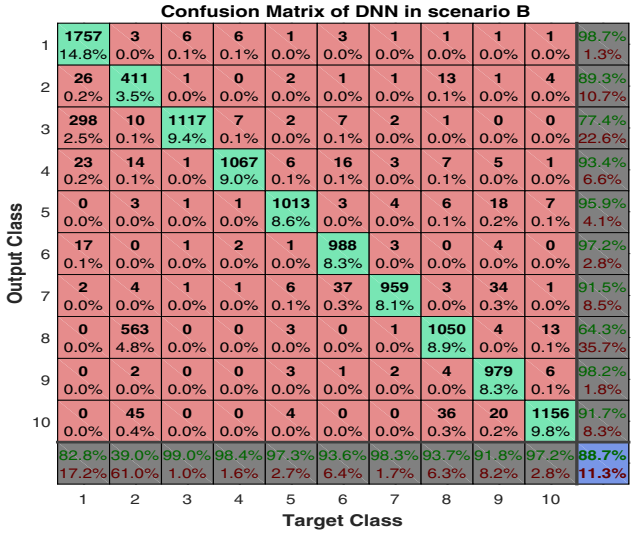
Fig. 4 plots the average (over 900 runs) correct classification rate (CCR) of the DNN model vs. different fine-tune training iterations. The CCR characterizes the percentage of correct estimation by the model (i.e., the estimated number of people is the same as the actual number of people in the environment). As can be seen, the performance of the DNN model largely improves with an increasing number of fine-tune training iterations, and 400 fine-tune training iterations, as set in our simulations, produce a well-trained DNN model.

The proposed DNN model yields higher correct classification rates. Table II compares the CCR performance of k-NN, MLE, and DNN with and without feature space expansion for two testing scenarios. The best result of k-NN among $k = 1, 3, 5, 7$ is presented for each condition. As can be seen, DNN outperforms k-NN and MLE in all conditions. This can be attributed to the fact that, as previously stated, the multiple layers with different numbers of anchor nodes in the DNN model pose a similar effect of feature selection, which mitigates the effect of environmental noises on CSI measurements. The performance of MLE could be limited by the fact that the measured CSI data are not necessarily Gaussian distributed, which is the underlying assumption in the MLE method. The table also shows that the performance of all methods improve with feature space expansion, especially DNN, suggesting that feature space expansion could be particularly beneficial for DNN with an inherent feature selection capability. DNN with feature space expansion achieves a high performance of 88.66% CCR for Scenario B, which is the most general scenario.

The proposed DNN model using CSI features is robust to people location variability and is highly suitable for the indoor people counting application. Fig. 5 shows the confusion matrices for DNN with feature space expansion in the two testing scenarios, which illustrate the classification distributions. The indices 1–10 of the target/output class represent the numbers of actual/estimated people being 0–9. As can be observed, first, there are in general more instances of misclassification for Scenario B than for Scenario A, resulting in a lower average CCR for Scenario B, as shown in Table II. Second, a higher value of an off-diagonal entry for Scenario A largely corresponds to a higher value of the same entry for Scenario B (e.g., 0-people condition misclassified as 2-people condition, 1-people condition misclassified as 7-people condition, 7-people condition misclassified as 9-people condition, etc.), showing that the dominant cases of misclassification are largely consistent for Scenarios A and B. This suggests that people locations, which are different in Scenarios A and B, are not the key factor affecting the classification in the DNN model, but the number of people is. This suggests the high applicability and practicality of the proposed DNN model using CSI features for indoor people counting. However, the relationship between the number of



(a)



(b)

Fig. 5. The confusion matrices for the DNN model with feature space expansion for (a) Scenario A and (b) Scenario B. The confusion matrix shows the numbers and percentages of correct (diagonals) and incorrect (off-diagonals) classification. The cells in gray in the bottom row show the *recall* of the model for each true class (i.e., the percentages of correct (top number) and incorrect (bottom number) classification for each true class). The cells in gray in the right-most column show the *precision* of the model for each predicted class. The bottom-right cell in blue shows the average correct (top number) and incorrect (bottom number) classification rates.

people and the measured CSI is still not fully discovered and deserves further investigation.

V. Conclusion

We have proposed an enhanced people counting system based on the DNN model as the classifier and fine-grained wireless signatures of CSI as the features. A three-layer DNN was employed, with a feature space expansion scheme for the model. The performance of the proposed system was verified in a real indoor testbed environment. Only one Wi-Fi transmitter and one laptop receiver were deployed. The results showed that the proposed feature space expansion scheme can improve the performance of various models. Furthermore, the proposed three-layer DNN model with feature space expansion demonstrates the highest correct classification rates of 96.90% and 88.66% in estimating the exact numbers of the crowd of size up to nine people, in indoor scenarios with fixed and arbitrary locations of people, respectively.

References

[1] J. C. S. J. Junior, S. R. Musse, and C. R. Jung, "Crowd analysis using computer vision techniques," *IEEE Signal Process. Mag.*, vol. 5, no. 27, pp. 66–77, Sep. 2010.

[2] S. Depatla, A. Muralidharan, and Y. Mostofi, "Occupancy estimation using only WiFi power measurements," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1381–1393, 2015.

[3] Y. Yuan, J. Zhao, C. Qiu, and W. Xi, "Estimating crowd density in an RF-based dynamic environment," *IEEE Sensors J.*, vol. 13, no. 10, pp. 3837–3845, 2013.

[4] H. Li, E. C. Chan, X. Guo, J. Xiao, K. Wu, and L. M. Ni, "Wi-Counter: smartphone-based people counter using crowdsourced Wi-Fi signal data," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 4, pp. 442–452, 2015.

[5] M. Nakatsuka, H. Iwatani, and J. Katto, "A study on passive crowd density estimation using wireless sensors," in *4th Intl. Conf. on Mobile Comput. and Ubiquitous Netw. (ICMU)*, 2008.

[6] C. Xu, B. Firner, R. S. Moore, Y. Zhang, W. Trappe, R. Howard, F. Zhang, and N. An, "SCPL: indoor device-free multi-subject counting and localization using radio signal strength," in *ACM/IEEE Int'l Conf. on Inf. Process. in Sensor Networks (IPSN)*, 2013, pp. 79–90.

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

[8] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 25, Nov. 2013.

[9] W. He, K. Wu, Y. Zou, and Z. Ming, "WiG: WiFi-based gesture recognition system," in *24th Int'l Conf. on Computer Commun. and Networks (ICCCN)*, 2015, pp. 1–7.

[10] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proc. of the 21st Annual Int'l Conf. on Mobile Comput. and Netw.*, 2015, pp. 65–76.

[11] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf WiFi," in *Proc. of the 16th ACM Int'l Symposium on Mobile Ad Hoc Networking and Comput.*, 2015, pp. 267–276.

[12] Y. Zeng, P. H. Pathak, and P. Mohapatra, "Analyzing shopper's behavior through WiFi signals," in *Proc. of the 2nd Workshop on Physical Analytics*, 2015, pp. 13–18.

[13] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu, and Z. Jiang, "Electronic frog eye: Counting crowd using WiFi," in *Proc. of IEEE INFOCOM*, 2014, pp. 361–369.

[14] S. D. Domenico, M. D. Sanctis, E. Cianca, and G. Bianchi, "A trained-once crowd counting method using differential WiFi channel state information," in *Proc. of the 3rd Int'l Workshop on Physical Analytics*, 2016, pp. 37–42.

[15] C.-C. J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. of Visual Commun. and Image Representation*, vol. 41, no. 36, pp. 406–413, Nov. 2016.

[16] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015.

[17] M. Mimura, S. Sakai, and T. Kawahara, "Deep autoencoders augmented with phone-class feature for reverberant speech recognition," in *IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2015, pp. 4365–4369.

[18] O. Plchot, L. Burget, H. Aronowitz, P. Mat *et al.*, "Audio enhancing with DNN autoencoder for speaker recognition," in *IEEE Int'l Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2016, pp. 5090–5094.

[19] S. Kumari and S. K. Mitra, "Human action recognition using DFT," in *Third National Conf. on Computer Vision, Pattern Recognition, Image Process. and Graphics (NCVPRIPG)*, 2011, pp. 239–242.

[20] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Trans. on Image Process.*, vol. 21, no. 4, pp. 2160–2177, Apr. 2012.