AN ABSTRACT OF THE THESIS OF

Cheng-Che Shih for the degree of Master of Science in Electrical and Computer Engineering presented on December 5, 2022.

Title: People Counting System Using MmWave MIMO Radar with 3D Convolutional Neural Network

Abstract approved: _____

Thinh Nguyen

Recent advances in computing, communication, and artificial intelligence (AI) technologies have made our world more interconnected and data-rich than ever with the proliferation of smart devices and sensors. As a result, we are increasingly dependent on electronic devices and sensors to automate away life's mundane parts.

For example, in business settings, people counting systems can be used to automate the data collection for advertisement, and revenue projections as well as reduce energy costs using adaptive HVAC operations. However, a naive implementation of people counting systems may result in revealing some unintended information about the users/customers and higher power consumption from operating the systems continuously. In this thesis, we study a mmWave Multiple Input Multiple Output (MIMO) radar sensor system for detecting the number of people in a confined space with the aims of low power consumption and minimal leakage of user information. In particular, we showed that a 3D convolutional neural network can accurately determine up to 4 people in a typical size room using a surprisingly minimal number of mmWave signatures (less than 10) as its inputs.

People Counting System Using MmWave Radar with 3D Convolutional Neural Network

by
Cheng-Che Shih

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented December 05, 2022
Commencement June 2023

Master of Science thesis of Cheng-Che Shih presented on December 05, 2022

APPROVED:

Major Professor, representing Electrical and Computer Engineering

Head of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Cheng-Che Shih, Author

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

PEOPLE COUNTING SYSTEM USING MMWAVE MIMO RADAR WITH 3D CONVOLUTIONAL NEURAL NETWORK

Cheng-Che Shih, Xinrui Zhou, Thinh Nguyen

School of Electrical Engineering and Computer Science, Oregon State University

ABSTRACT

Recent advances in computing, communication, and artificial intelligence (AI) technologies have made our world more interconnected and data-rich than ever with the proliferation of smart devices and sensors. As a result, we are increasingly dependent on electronic devices and sensors to automate away life's mundane parts. For example, in business settings, people counting systems can be used to automate the data collection for advertisement, and revenue projections as well as reduce energy costs using adaptive HVAC operations. However, a naive implementation of people counting systems may result in revealing some unintended information about the users/customers and higher power consumption from operating the systems continuously. In this thesis, we study a mmWave Multiple Input Multiple Output (MIMO) radar sensor system for detecting the number of people in a confined space with the aims of low power consumption and minimal leakage of user information. In particular, we showed that a 3D convolutional neural network can accurately determine up to 4 people in a typical size room using a surprisingly minimal number of mmWave signatures (less than 10) as its inputs.

Index Terms— People Counting, CNN, mmWave Radar Sensor

1. INTRODUCTION

Thanks to the recent advances in computing, communication, and AI technologies, today's world is more interconnected and data-rich than ever. We are increasingly dependent on electronic devices to aid us in many facets of our lives from work to play. In particular, sensing and automation technologies play an important role in providing us with both convenience and time efficiency. However, if not carefully designed, these sensing and automation technologies may reveal unintended private information about the users. Furthermore, as the world is moving toward a sensor-based economy with hundreds of billions of sensors deployed in all business sectors from medicine and engineering to agriculture and entertainment, there is a critical need to design low-power sensors. Therefore, the key component to the successful sensor-based economy is the proliferation of sensors and automation technologies that overcome privacy and power consumption concerns. In this thesis, we study the people counting systems as an example of a potentially low-power sensor system integrated with AI technology to automatically count the number of people present in a confined space.

People counting technologies have been around for a long time. A widely used people counting technology is the door counters that have been used everywhere from retail stores and hotels to movie theaters and train stations. Recently, more sophisticated people counting systems have been developed and they play an important part in business management for retailers. Enterprises use these advanced people counting systems to track and determine the number of people, their movement patterns, and habits in a retail store or office space. This information help businesses make decisions on how to utilize the space and HVAC systems efficiently and improve customer experience. This information also ensures the safety of building occupants in emergencies by providing safe escape routes in real-time during a fire or earthquake. Furthermore, based on the recent pandemics, many argue that knowing the accurate number of people and their movement in a confined area in real-time can be a valuable piece of information in designing effective strategies for future pandemics.

Many advanced people counting systems use high resolution sensor data such as videos and images. High-resolution data provide more detailed information and therefore make it easier for the people counting algorithms to determine and track the number of people. For exam-

ple, video surveillance systems using computer vision technologies not only can automatically count and track people but also provide automatic identification of people and their activities. In many situations, people's identities and their activities are considered private information. In such cases, it is crucial to protect this information. In addition, video surveillance systems consume more power than necessary for detecting and tracking people. As a result, other RF-based technologies use less power and do not reveal private information.

In this thesis, we study a mmWave Multiple Input Multiple Output (MIMO) radar sensor system for detecting the number of people in a confined space with the aim of low power consumption and minimal leakage of user's information. In particular, we showed that a 3D convolutional neural network can accurately determine up to four 4 people in a typical size room using a surprisingly minimal number of mmWave signatures ($<$10) as its inputs. The outline of the thesis is as follows. In Section 2, we will describe some recent advances in people counting systems. In Section 3, we describe the proposed system, the data collection methodology, the feature extraction procedure, and the proposed 3D convolutional neural network for counting the number of people. In Section 4, we show the experimental results to validate our approach and offer a few concluding remarks in Section 5.

2. RELATED WORK

Due to recent advances in AI and the reduction in sensor costs, there has been a rise in work related to people counting, identification, and tracking systems. A large portion of these works is focused on tracking people using videos and images. In particular, identifying and tracking people and their activities are accomplished using the certain image and video pixels associated with heads, faces, and movements. Non-real-time systems use existing pictures of video images [1, 2, 3] while others employ specialized real-time face detectors to identify and track people [4, 5, 6, 7] through surveillance cameras. In [4], the authors use multiple cameras to count the number of moving and standing people. They use a combination of multi-view fusion detection methods and particle tracking to achieve high accuracy. In [5], the authors use a camera to detect the face and determine whether the face is real or not and use it to count the number of people. This approach has 80% accuracy. In [6], the authors use a face detector and

combine a new scale invariant Kalman filter with a kernel-based tracking algorithm to track faces. The accuracy of discriminating real face trajectories is 93%. In [7], the authors use a single overhead-mounted camera and analyze an image area consisting of a set of virtual counting lines to result in an accuracy of 93%. In [1], the authors use surveillance videos and the R-CNN TensorFlow model to analyze. The accuracy is 94%. In [2], the authors propose a scale-driven convolutional neural network model to analyze head features to solve people counting and localization in low-density and high-density. In [3], the authors use convolutional neural networks (CNN) for counting and positioning people given aerial shots of visible and infrared images.

While video and image-based systems are predominant due to their rich data and the large market penetration of video cameras, their uses are limited to settings where privacy or power consumption is less of a concern. Videos and images provide detailed information for accurate counting and tracking of people, but they also reveal unintended information such as people's identities and their activities. Furthermore, video and camera-based systems consume more power than necessary to accurately count the people. As a result, the people counting systems using radio frequency (RF) have been proposed in the past few years [8, 9, 10, 11, 12, 13, 14, 15]. These systems have the advantages of (a) not revealing unnecessary information about the people and (b) using less power.

In [8], the authors use the Ultra-Wide-Band (UWB) radar and a CNN model for counting people. The system can count up to 10 people randomly walking in an area of 55 m^2 with 97% accuracy. These UWB systems however consume high bandwidth and UWB systems may interfere with other existing wireless communication systems. In [9], the authors use Frequency-modulated continuous-wave radar (FMCW), but it requires computation to eliminate mirror targets and takes a long time to compute for multiple chirps. In [10], the authors use low-cost FMCW MIMO (Multiple-Input-Multiple-Output) radar and propose the combination of the Range-Azimuth map and spectro-gram/cadence velocity diagram (CVD). In [11], the authors track and identify multiple subjects in real-time using the sparse point-cloud sequences obtained from a low-cost mm-wave radar. The accuracy is 91.62% operating at 15 frames per second. In [12], the authors use millimeter waves to recognize gait and propose a

novel deep-learning driven wave gait recognition method called mmGaitNet. the accuracy is 90% for a single person, and 88% for five persons. In [13], the authors use only passive and sensorless transponders to classify gestures of arms and legs. The accuracy of this system is 80- 90%. In [14], the authors all use Infra-Red (IR) sensors, but the infrared frequency is easily affected by the surrounding, and it can only control one gadget at a time in on-screen and control applications. In addition, it is difficult to operate in non-LOS (non-Line of Sight) settings. In [15], the authors use WiFi signals and infrared sensors to determine the number of people in a predefined area.

3. SYSTEM DESCRIPTION

In this section, we describe the proposed sensor hardware, the data set creation procedure, and the architecture of a 3D convolutional neural network for counting the number of people.

3.1. Sensor

Fig. 1 shows the set-up of our experiment in a 4m x 2m area. We use the IWR6843ISK, a mmWave active radar made by Texas Instruments to scan the room. It is a low-power device with a 3.3V/5V and 2.5A power supply. It operates in a frequency range of 60 to 64GHz with a temperature range from -20°C to 60°C. There are 4 receive and 3 transmit onboard antennas. Each antenna has a 120-degree azimuth field of view and a 30-degree elevation field of view. Fig. 1 shows where the radar is mounted relative to where the people are walking. To control the experiment parameters, the floor is marked with tapes running along the vertical and horizontal directions. People are instructed to walk in these directions, in addition, to walking in random directions. We also have video cameras to record the video data of the experiments. Videos are not used in determining the number of people, rather they are used to verify the correctness of data collection and post-analysis. The IWR6843ISK is installed on a TI evaluation board MMWAVEICBOOST. The evaluation board is connected to a PC and provides tools for data collection, processing, and visualization.

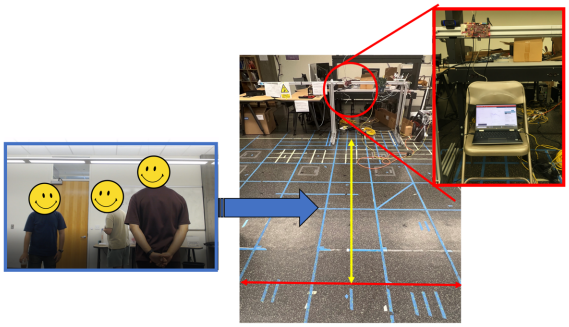

Fig. 1. Experimental setup for counting the number of people present in the room. People of different heights are walking in an area of 4m x 2m.

3.2. Dataset

To collect the training data, we employed a total of 12 people of different heights ranging from 5'2" to 5'10 in experiments. They include 7 males and 5 females. The collected data is divided into 4 groups: Group 0, Group 1, Group 2, Group 3, and Group 4 with Group n containing n people walking in the room simultaneously. For control analysis, the people are instructed to walk in random directions as well as in certain well-defined directions during each 15-minute episode. Fig. 2 The total amount collected in terms of hours for each group is shown in Table 1.

Table 1. Amount of training data in terms of hours for each group. The data is sampled at 10 frames per second

Group	0P	1P	2P	3P	4P
Time	2hr	8.5hr	8.5hr	4.5hr	1.5hr

90% of data are used for training and the remaining 10% are used for testing. Furthermore, to ensure generalization, the testing data do not contain the same people in the training data.

3.3. Feature Extraction

The steps for data acquisition and feature extraction are as follows. The IWR6843ISK scans the room by sending out RF chirps, measures the reflected RF signals, processes them into features (objects), and transferred them to the PC as shown in Fig. 3. Specifically, feature extraction (object detection) is accomplished in several steps. First, 1D range FFT is used on the data from receive antennas for every chirp that is connected with the chirping pattern on the transmit antennas during the chirp processing period. Next, 2D FFT, CFAR detection, peak grouping, and angle of arrival processing are performed. 2D FFT is meant to remove clutters. CFAR is a classical adaptive constant-false-alarm rate to remove noise and interference. Peak grouping is used to detect the objects based on the relative SNR. The estimation of azimuth and elevation are used to determine the x, y, z positions of each object. Note that the x, y, z positions are estimated using the frequency shifts rather than the time of flight. Doppler velocity of an object, i.e., the relative velocity with the direction of the radar, is also estimated. These identified objects/features are sent to the PC in real time for storage. The process repeats for the next groups of chirps. Fig. 4 shows an example of detected features/objects with their x, y, z positions and the relative SNR vs the distance. In the position, graphs there are 3 detected objects. in the relative SNR graph, there are four detected objects/features. The number of objects/features can be controlled by the users by setting appropriate thresholds. As will be shown later, only a few objects are needed ($<$10) to be able to count the number of people accurately. Data are collected and processed in frames at the rate of 10 frames per second.

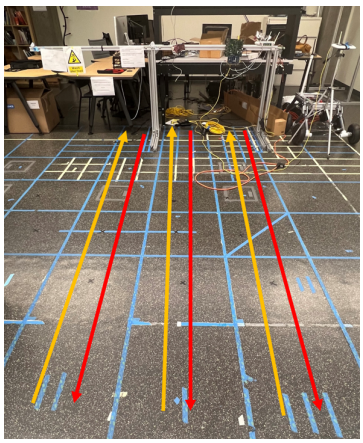

Fig. 2. Example of subjects walking on the fixed routes during the data collection process

3.4. 3D Convolutional Neural Network

The proposed 3D convolutional neural network for counting the number of people in the room is shown in Fig. 5. Each training example to the 3D CNN consists of 5 seconds worth of continuous 3D frames that consist of x, y, z positions, the Doppler velocities of the detected objects together with their labels (0,1, 2, 3, or 4 people). For example, a 5-minute RF data segment is shaped into 60 continuous 3D frames. The size of each frame is 50x20x4, where 50 represents 5x10 seconds as a result of sampling data at 10 frames per second, 20 represents detected objects within each frame, and 4 represents coordinates and Doppler velocity information (x, y, z, v). As shown in Fig. 5, the 3D CNN consists of three convolutional layers, with each layer followed by a max pooling layer and batch normalization. The parameters for each convolutional layer such as kernel size is shown in Fig. 5. LeakyReLU is also used after each convolutional layer. Connected to the last convolutional layer are two fully connected layers with 1024 and 256 nodes. The last fully connected layer is flattened and softmax is used to produce the probability distribution of the number of people. We use entropy as a loss function for the proposed 3D CNN.

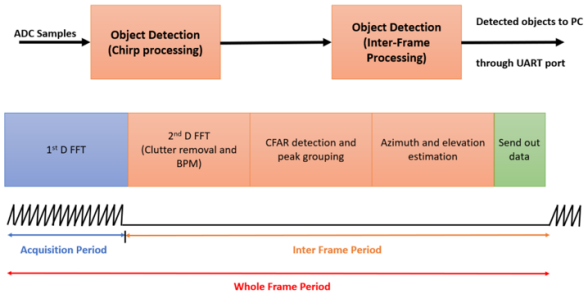

Fig. 3. Top level data path for processing receiving signals.

4. RESULTS

In this section, we show the results of counting the different numbers of people. In this first set of results, we only train the proposed 3D CNN on the dataset that contains only 0 or 1 person. Thus, the CNN will predict whether is one or no person in the room. Fig. 6 shows the loss and accuracy as a function of epochs for both training and testing. Low loss and high accuracy are better. As seen, the 3D CNN learns and converges quickly after 30 epochs and achieves an

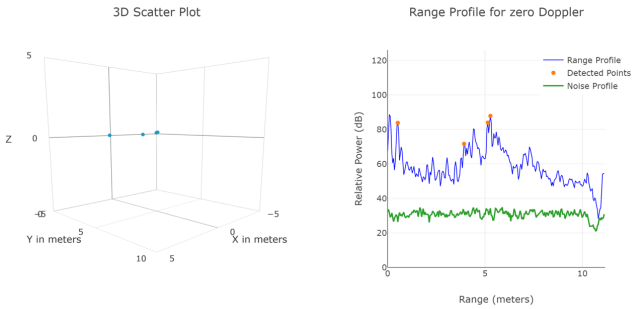

Fig. 4. Examples of x, y, z positions of the detected object (left) and the relative SNR of the detected objects vs. distance

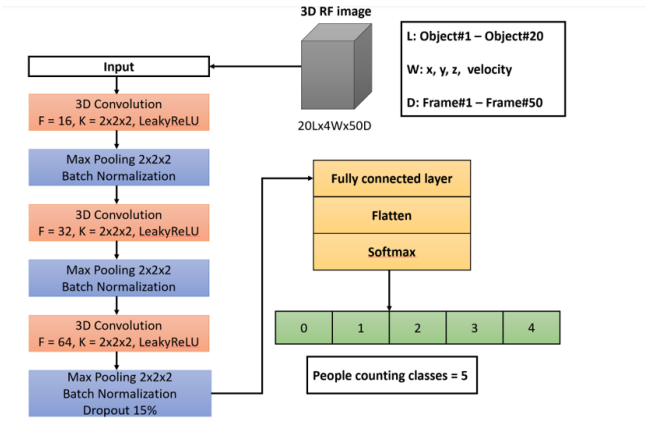

Fig. 5. Architecture of the proposed 3D convolutional neural network. The inputs are a batch of 50 frames. Each frame consists of a number of objects with their positions and Doppler velocities.

accuracy of 100%. Fig. 7 shows the joint distribution between the true and predicted labels for the case of counting 0 to 1 person.

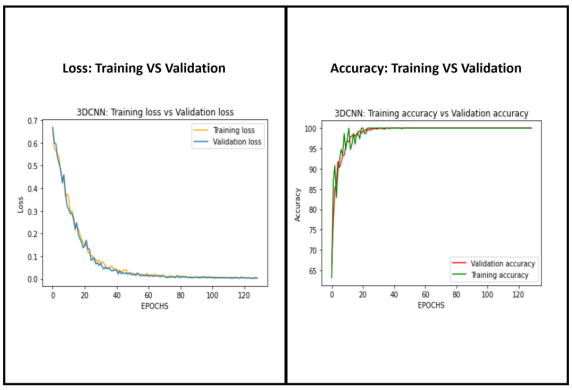

Fig. 6. Loss and accuracy as functions of epochs for the case of detecting 0 or 1 person

Next, Fig. 8 shows the loss and accuracy as a function of epochs for both training and testing for the case of counting the number of people from 0 to 2. In this case, the CNN is

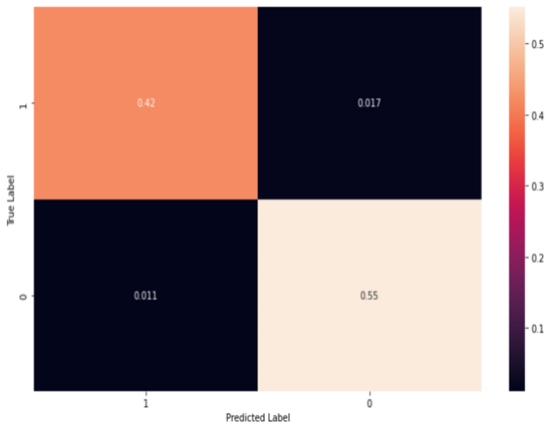

Fig. 7. The joint distribution between the true and predicted labels for the case of counting 0 to 1 persons

trained using the dataset that has a maximum of 2 people in the room. As seen, the accuracy of 100% is obtained after training for 40 epochs. Fig. 9 shows the joint distribution between the true and predicted labels for the case of counting 0 to 2 persons.

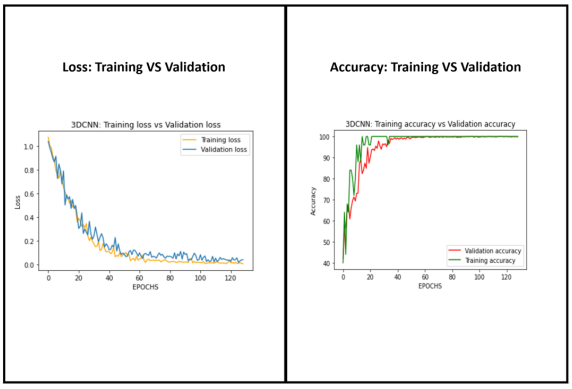

Fig. 8. Loss and accuracy as functions of epochs for the case of counting from 0 to 2 persons

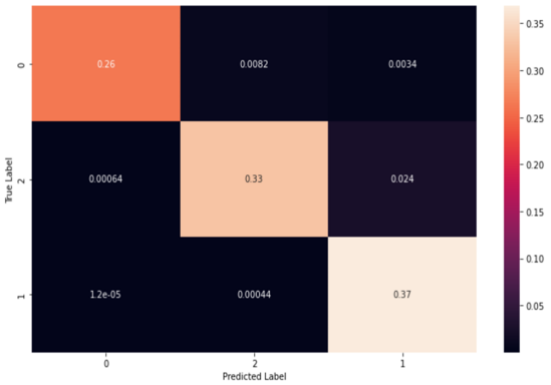

Fig. 9. The joint distribution between the true and predicted labels for the case of counting 0 to 2 persons

Next, Fig. 10 shows the loss and accuracy as a function of epochs for both training and testing for the case of counting the number of people from 0 to 3. In this case, the CNN is trained using the dataset that has a maximum of 3 people in the room. As seen, an accuracy of more than 95% is obtained after 125 epochs. Fig. 11 shows the joint distribution between the true and predicted labels for the case of counting 0 to 3 persons.

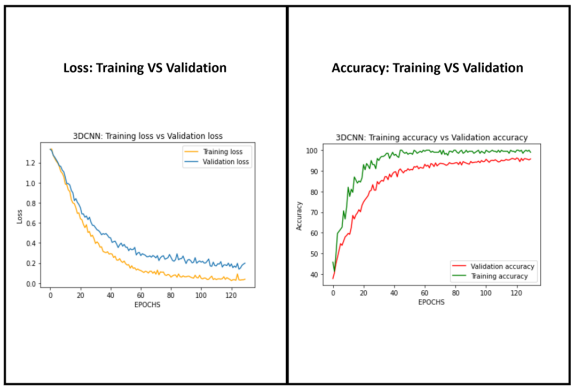

Fig. 10. Loss and accuracy as functions of epochs for the case of counting from 0 to 3 persons

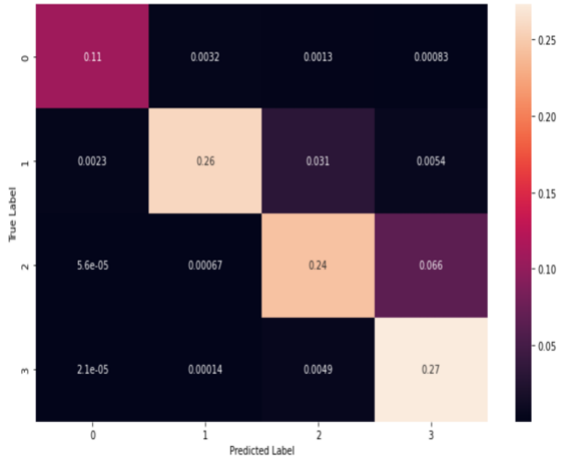

Fig. 11. The joint distribution between the true and predicted labels for the case of counting 0 to 3 persons

Finally, Fig. 12 shows the loss and accuracy as a function of epochs for both training and testing for the case of counting the number of people from 0 to 4. In this case, the CNN is trained using the dataset that has a maximum of 4 people in the room. similar to the case of counting 0 to 3 persons, an accuracy of more than 95% is obtained after 125 epochs. Fig. 13 shows the joint distribution between the true and predicted labels for the case of counting 0 to 4 persons.

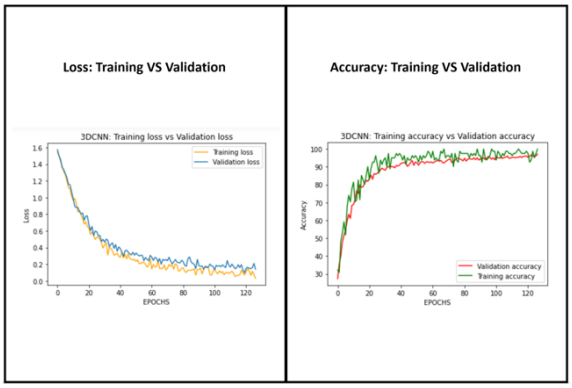

Fig. 12. Loss and accuracy as functions of epochs for the case of counting from 0 to 4 persons

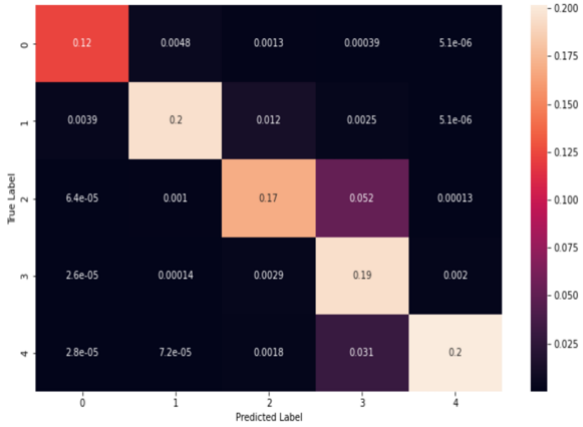

Fig. 13. The joint distribution between the true and predicted labels for the case of counting 0 to 4 persons

5. CONCLUSION

We described a people counting system that is developed by using a low-cost RF sensor (IWR6843ISK + MMWAVEICBOOST) together with a 3D-CNN classifier. The system can accurately count from 0 to 4 people in a 4m x 2m small indoor space. We show that only a small number of features are needed for accurate counting. Specifically, with less than 10 detected objects, each object is specified by a small amount of information, namely its position, Doppler velocity, and relative SNR. The results demonstrate a viable application of a low-power MIMO sensor for a high-accuracy people-counting system that does not reveal much information about the subjects. On the other hand, there are uncertain on whether the results can be generalized in general settings such as larger environments, and different types of people with various heights and gaits. For future work, we plan to improve the system by adjusting the current 3D-CNN model and adding the 3D point cloud but keeping the low-cost setup, but still aiming to achieve

good performance for scenarios involving more than four persons.

6. REFERENCES

[1] Shaik Abdul Nabi, Yadavelly Sahithi, Nikhat Sheereen, Thakur Aakanksha, and Korla Rajkumar Reddy, "Real-time people counting for surveillance videos," *JOURNAL OF ALGEBRAIC STATISTICS*, vol. 13, no. 3, pp. 523–530, 2022.

[2] Saleh Basalamah, Sultan Daud Khan, and Habib Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71576–71584, 2019.

[3] Joaquín Filipic, Martín Biagini, Ignacio Mas, Claudio D Pose, Juan I Giribet, and Daniel R Parisi, "People counting using visible and infrared images," *Neurocomputing*, vol. 450, pp. 25–32, 2021.

[4] Huadong Ma, Chengbin Zeng, and Charles X Ling, "A reliable people counting system via multiple cameras," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, pp. 1–22, 2012.

[5] Tsong-Yi Chen, Chao-Ho Chen, Da-Jinn Wang, and Yi-Li Kuo, "A people counting system based on face-detection," in *2010 Fourth International Conference on Genetic and Evolutionary Computing*, 2010, pp. 699–702.

[6] Xi Zhao, Emmanuel Delleandrea, and Liming Chen, "A people counting system based on face detection and tracking in a video," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 67–72.

[7] Javier Barandiaran, Berta Murguia, and Fernando Boto, "Real-time people counting using multiple lines," in *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, 2008, pp. 159–162.

[8] C-T Pham, VS Luong, D-K Nguyen, HHT Vu, and M Le, "Convolutional neural network for people counting using uwb impulse radar," *Journal of Instrumentation*, vol. 16, no. 08, pp. P08031, 2021.

[9] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham, "Cubelearn: End-to-end learning for human motion recognition from raw mmwave radar signals," *arXiv preprint arXiv:2111.03976*, 2021.

[10] Liyuan Ren, "People counting using low-cost fmcw mimo radar: Achieving tracking for counting and classification of groups of people using fmcw radar," 2022.

[11] Jacopo Pegoraro and Michele Rossi, "Real-time people tracking and identification from sparse mm-wave radar point-clouds," *IEEE Access*, vol. 9, pp. 78504–78520, 2021.

[12] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang, "Gait recognition for co-existing multiple people using millimeter wave sensing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 849–856.

[13] Sara Amendola, Luigi Bianchi, and Gaetano Marrocco, "Movement detection of human body segments: Passive radio-frequency identification and machine-learning technologies.," *IEEE Antennas and Propagation Magazine*, vol. 57, no. 3, pp. 23–37, 2015.

[14] Hessam Mohammadmoradi, Sirajum Munir, Omprakash Gnawali, and Charles Shelton, "Measuring people-flow through doorways using easy-to-install ir array sensors," in *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2017, pp. 35–43.

[15] Radosveta Sokullu, "People/animal counting – integrated sensor based and wifi/machine learning based system," in *2022 8th International Conference on Energy Efficiency and Agricultural Engineering (EEAE)*, 2022, pp. 1–4.