mmWave Radar-based Hand Gesture Recognition using Range-Angle Image

Jih-Tsun Yu, Li Yen, and Po-Hsuan Tseng
Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan
Andy0010@hotmail.com.tw, das73541@gmail.com, phtseng@ntut.edu.tw

***Abstract*—The radar sensing on fine human-motion/hand-gesture provides further human-computer interaction (HCI) experience. Most of the studies about gesture recognition with mmWave frequency modulated continuous wave (FMCW) radar adopts the range and the velocity estimated from the raw data, such as the time-frequency spectrogram, micro-doppler spectrogram, or range-Doppler image (RDI). Besides, the angle estimated using multiple receive antennas also contains rich information of gesture, especially in the discrimination among the horizontal movement of the gesture. Thus, we propose to use the range-angle image (RAI) as the input and train a model consisting of the convolutional neural network and long short term memory that is capable of recognizing hand gestures. We validate the proposed scheme based on the collection of hand gestures by several subjects in different classrooms using 77 - 81GHz mmWave radar of Texas Instrument. Based on the configuration of one transmit antenna and four receive antennas, we show that the hand gesture recognition using RAI outperforms that using RDI. Also, we adopt the fusion strategy to consider both RDI and RAI to further improve accuracy.**

***Index Terms*—FMCW Radar, Hand-Gesture Recognition, Machine Learning, Human-Computer Interaction**

I. Introduction

In the field of human-computer interaction, static/dynamic object recognition based on sensors has always been an important part. Gesture recognition to get rids of physical control interfaces such as buttons and touch screens has become a hot research topic. In the various in-car scenario, e.g., to avoid an accident due to any distraction, a gesture control assisted driving system allows users to implement a variety of complex functions through the contact-less operation.

There are many gesture recognition studies based on various sensors, such as vision-based solution using OpenCV, wearable device with five hand-held LEDs to track the gesture, and so on. Wireless gesture recognition serves as a contactless technology in which the wireless signal can penetrate through materials, such as plastics, walls, or clothes. Wireless gesture recognition still works at night and its performance is not susceptible to visible light. The privacy is also reserved in wireless-based solution compared to the vision-based. Among the wireless sensing technologies, Wi-Fi and millimeter wave (mmWave) radars draw a lot of attention nowadays.

Chanel measurements in Wi-Fi devices, such as channel state information (CSI) or received signal strength indication (RSSI) [1]–[3], can be utilized to distinguish the characteristics of different gestures. For example, [1] uses the Wi-Fi routers as an experimental platform to compare the differences in gestures by obtaining changes in CSI. [2] recognizes gestures based on changes in RSSI when the user makes a gesture. [3] extracts the feature from the CSI amplitude and classifies the gesture using the support vector machine (SVM).

Since the ranges, velocities, and angles of the objects can be estimated based on the frequency modulated continuous wave (FMCW) radar with multiple receive antennas, the gesture recognition methods using mmWave-based FMCW radar [4]–[8] further make use of these estimations to provide a better accuracy, compared to the Wi-Fi-based solutions. All of them are based on machine learning to learn the features of the gestures from different types of inputs. The time-frequency spectrogram in [6] considered the range estimation of each chirp, the micro-doppler spectrogram in [5], [7] considered the velocity estimation of each frame, and the range-doppler image (RDI) in [4], [8] considered both the range and the velocity estimation of each frame. Long short term memory (LSTM), which has the ability to analyze the causality of time series in a sequence of data, is used to recognize the features from the time sequences of hand gesture. Notice that the angle estimation has only be treated as a higher dimension of input in [8], in which the RDIs of different receive antennas were inputted to the neural network in a separate channel. The correlations among the RDIs of different antennas have not been considered thoroughly. It is worth mentioning that Soli [8] uses RDI and recognizes gestures based on an end-to-end model by connected a convolutional neural network (CNN) with an LSTM. This prototype has been implemented in Google Pixel 4.

Since a highly accurate position can be estimated from the mmWave-based radar, we explore how the angle information can assist the gesture recognition. We define up to 12 kinds of gestures. Through the pre-processing of the signal, the gesture classification model is established and trained based on a cascade of CNN with LSTM. The novelties of this paper are as follows:

- In addition to the RDI used in [4], [8], we first design to use the range-angle image (RAI) [9] as a measurement input. We show that its performance is better than the model trained with RDI.
- Moreover, we consider a fusion architecture to input

[1] This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 107-2221-E-027-040-MY2 and by the Technological University Paradigms.

978-1-7281-5207-3/20/$31.00 ©2020 IEEE

both RAI and RDI. The fusion scheme enhances the gesture recognition through the range, velocity, and angle information.

- We verify the performance using the Texas Instruments (TI) platform with the frequency ranges from 77-81GHz. Comparing to our implementation of Soli [8] adopting only RDIs with 1 transmit antenna (1T) and 4 receive antennas (4R) in the TI platform, the classification accuracy of our proposed fusion method reaches 92.74%.

This paper first introduces the principle of FMCW radar and how RDI and RAI graphics are generated in Sec. II. Sec. III-B introduces the gesture recognition system proposed in this paper, including the data collection, data pre-processing, and design of deep neural networks. Sec. IV describes the experimental results and the related discussion. Finally, Sec. V draws the conclusion.

II. Frequency Modulated Continuous Wave (FMCW) Radar

A. Range-Doppler Image (RDI)

The FMCW radar transmits a chirp signal, which is modulated as a linear change in frequency with its slope S over time. This signal propagates to hit the targeted object, bounces back, and then is received at the receive antenna. Assuming that the distance between the object and the radar is d, and the speed of light is c, then the round-trip time between the radar and this object is $\tau = \frac{2d}{c}$. Through the design of radar, the received chirp and the transmitted chirp can be directly mixed to obtain their frequency difference. That is, the output of the mixer is an intermediate frequency (IF) signal. The frequency of the IF signal f is caused by the round-trip propagation time from the radar to the object as

$$f = S \cdot \tau = \frac{S \cdot 2d}{c} \Rightarrow d = \frac{fc}{2S} \tag{1}$$

The Fourier transform of the IF signal observes the frequency components of this signal to infer the range of the object.

Suppose an object moves at a speed of v. In order to recognize the moving object, the radar continuously sends two chirp signals with a time interval of T_c. Then the phase difference between the two chirp signals arriving at the receiving end is related to the moving speed as

$$\omega = \frac{4\pi v T_c}{\lambda} \Rightarrow v = \frac{\lambda \omega}{4\pi T_c} \tag{2}$$

The FMCW radar transmits a signal sequence composed of N_c chirps, termed as a frame.

The procedure to obtain a RDI by using 2D-FFT is summarized as follows.

1) After the processing of the mixer and the ADC conversion, the samples of a chirp signal are stored in a column vector. All the samples of the chirps in a frame are cascaded column-by-column to obtain a 2D matrix.
2) The first FFT (range FFT) is performed in each column to obtain the range information of the object.
3) The second FFT (doppler FFT) is performed in each row to obtain the velocity information of the object.
4) The absolute value of the range-doppler matrix is taken to obtain the RDI.

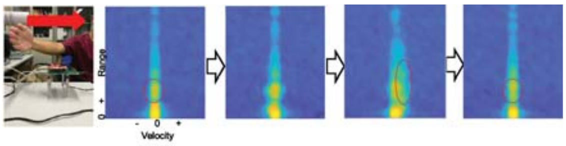


Fig. 1. RDI Example – Gesture #3: A palm swept from right to left (y-axis: range; x-axis: velocity)

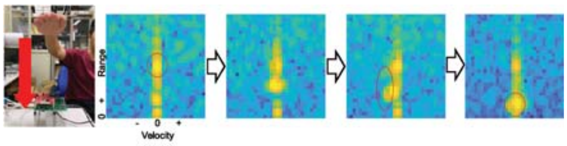


Fig. 2. RDI Example – Gesture #4: A palm swept from up to down (y-axis: range; x-axis: velocity)

B. Range-Angle Image (RAI)

Suppose that the distance between the two receive antennas is d. If the angle of arrival (AoA) of the radar signal is θ, the phase difference between the two receive antennas is $\omega = \frac{2\pi d \sin\theta}{\lambda}$. Thus, the phases change between two receive antennas can be used to estimate the angle of the object $\theta = \arcsin\frac{\omega\lambda}{2\pi d}$.

Since each receive antenna records the different phases reflected by the same object, we cascade the range-doppler matrices of M receive antennas obtained from the RDIs as a 3D signal. We first increase the dimension of the receive antenna by zero-padding from M to N, to increase the angular resolution. The third FFT is then performed along the antenna dimension to obtain the range-doppler-angle matrix. By summing this 3D matrix along the dimension of the doppler into a 2D matrix, the RAI is obtained. The positions of the objects (ranges and angles) in the environment can be observed in the RAI.

C. Observations in RDI and RAI

Figs. 1 and 2 show a couple of RDIs in a consecutive time sequence measured by a mmWave radar when a palm swept from the horizontal and the vertical directions. Figs. 3 and 4 illustrate the RAIs of the same gestures. In Fig. 2, the brightest spot in the picture moves slowly from the larger range (at the upper part of figure) to the closer range (at the lower part of figure). The velocity is negative (the left part of the figure) when the hand swipes down. The feature of the vertical movement can be observed in RDI. However, we found that the horizontal movement, e.g., a palm swept from right to left, can not be observed in the RDI thoroughly. In Fig. 1, although the palm swipe still causes the velocity changes, its range to the radar at the starting point and that at the ending point are the same. Therefore, the positions of the starting

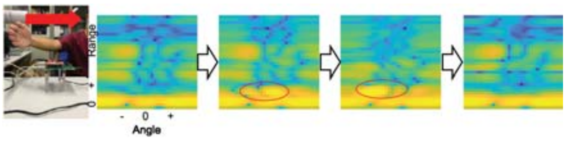


Fig. 3. RAI Example – Gesture #3: A palm swept from right to left (y-axis: range; x-axis: angle)

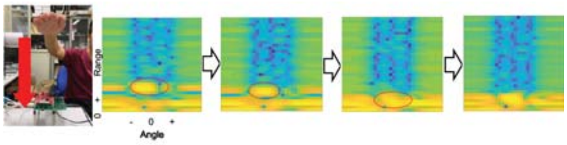


Fig. 4. RAI Example – Gesture #4: A palm swept from up to down (y-axis: range; x-axis: angle)

point and the ending point, which are usually important in defining a gesture, can not be distinguished in RDI. On the other hand, the horizontal movement of the gesture causes a change in the angle, and the vertical movement causes an increment/decrement in the range, as observed in the RAIs in Figs. 3 and 4. The adoption of the angle provides further information which can not be observed by the range and doppler.

III. Gesture Recognition System Design

A. *Gesture Design*

Fig. 5 shows the 12 gestures used in this work, which are divided into three groups according to the usage or the purpose of the comparison.

B. *Network Model*

The overall process of the proposed gesture recognition system is shown in Fig. 6. The measured RDI/RAI will be pre-processed to lower the noise and normalize its power. Based on the collected data, the proposed network structure refers to the end-to-end network composed of CNN and LSTM. The collected data is used for training and the trained model will be used to classify the gesture. Besides the comparison of using RDI or RAI as the measurement input to the same network structure, through the concept of data fusion, RDI and RAI are used as two sets of inputs to an end-to-end network.

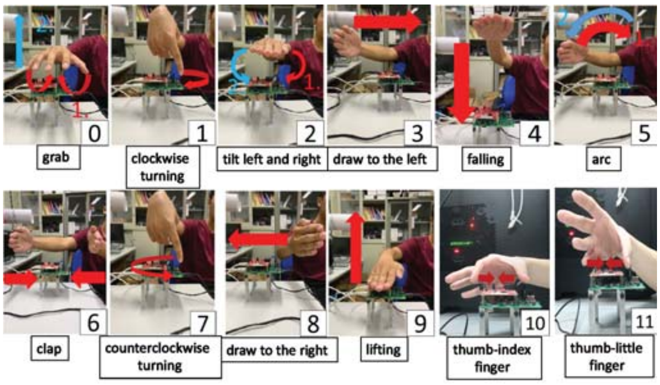


Fig. 5. 12 Different Gestures

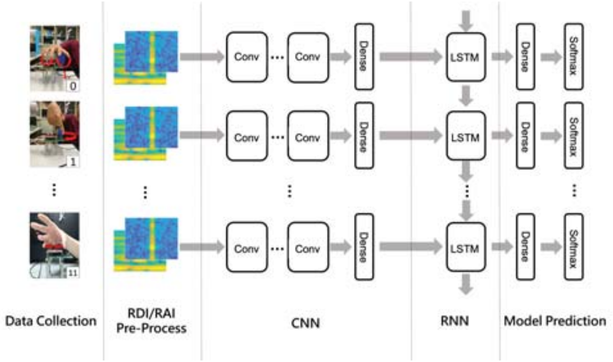


Fig. 6. Gesture Recognition System Network Model

1) Convolutional Neural Network (CNN): The input RDI/RAI of size $N \times N$ pixels converts the gesture recognition problem into an image classification problem. We use CNN and set it at the front end of the deep neural network to capture features in RDI and RAI. Batch normalization is used and a rectified linear unit (ReLU) linear activation function is used in each of the CNN. After the feature extraction of the three CNNs, the data will be re-adjusted into a flat vector of length 64×1 by the flatten layer and inputted to the next dense layer.

2) Long Short Term Memory (LSTM): Based on the RDI/RAI data from a sequence with a fixed length of N_c frames, the LSTM is adopted with the advantages in the sequence data processing. The LSTM network model is able to extract the associated features between each gesture frame-to-frame. The input and output dimensions of the LSTM are all set to 512. The time step is set to N_c frames based on the gesture data. The activation function is $\tanh$.

3) Detailed Network Settings: In addition, dropout is utilized at the second and third layers of CNN, and behind each layer of the dense network layer and the LSTM. The dropout rates are 0.4 in CNNs and 0.5 afterward to prevent over-fitting. The final part maps the output to a probability distribution over the predicted gesture using the softmax. The categorical cross-entropy is adopted as the loss function.

4) Data Fusion: Fusion is to make use of several data sets according to their different characteristics, to set several inputs using these sets, and to combine their feature after several network layers to make the final classification. Based on the above-mentioned end-to-end model, we design a deep neural networks to fuse the input of RDI and RAI. We fuse the results of the above-discussed RAI and RDI from their end-to-end models as in Fig. 7. The details of the network architecture is illustrated in this figure.

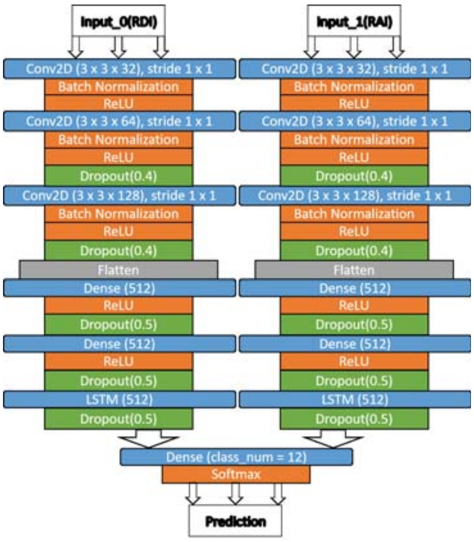


Fig. 7. Network Structure of Fusion

IV. Performance Evaluation

A. Experimental Setup

Since sufficient variation of gesture is required to train and evaluate the deep neural network-based solutions, we follow the setup in the Soli paper [8], such as using a similar number of total sequences for training, test subjects, sessions, and the percentage to split the training and the evaluation set. We asked 12 subjects to perform the 12 gestures, receiving only minimal instruction on how-to perform them. We recorded raw RDI/RAI at 25Hz and captured each gesture 20 times from all 12 subjects, over 12 sessions resulting in $12 \times 20 \times 12 = 2880$ sequences. Considering the average time spent on each gesture, we set the length of each gesture measurement to 64 frames, which takes about 2.6 seconds. Sequences of each gesture are annotated with a class label. The dataset is split in 50%-50% for training and evaluation and shuffled randomly.

The IWR1443BOOST mmWave radar development board and the DCA1000EVM real-time data acquisition board produced by TI are used to evaluate the gesture recognition performance. We set the transmission power into its minimum and the transmission range is less than 3 meters in our testing environment. There are $N_c = 32$ chirps per frame and each chirp has 64 sampling points. Therefore, the size of RDI and RAI as 64 pixels in the y-axis and 32 pixels in the x-axis is obtained from the raw data. In order to eliminate the extra background and highlight the characteristics of RDI, we first zero-padding the raw data to generate an RDI with a size of 128×64. We remove the upper part of the RDI with the range farther away and use zero-velocity as the center to cut the RDI into 32×32. On the other hand, we do not perform further post-processing of the RAI and simply discard the upper part, leaving the lower half of the RAI into size 32×32. Since each receive antenna can measure an RDI, 4 different RDIs can be obtained under 1T4R setting, i.e., the input using RDI is $32 \times 32 \times 4$. On the other hand, since the RAI is obtained from the FFT of the raw data from 4 receive antennas, the RAI is measured in 1 channel by the sum of the FFT results, i.e., the input using RAI is $32 \times 32 \times 1$.

B. Comparison of RDI, RAI and Fusion

We measure the classification accuracy per frame and list the average accuracy. The RDI and RAI databases are established from the same gesture measurements. We compare them using the same network structure. Each frame of RDI and that of RAI have the same size, i.e., 32×32, but the channel number of RDI measurement is four times large than that of RAI measurement under 1T4R. We summarize the confusion matrices using RDI and using RAI in Tables. I and II. The average accuracy is listed in Table III. The confusion matrix can be used to observe the relationship of gesture recognition error. Notice that the element of the $(n+1)$-th column of the matrix represents the probability that the gesture #n is predicted to the gestures #0 to #11. The sum of the probabilities on each column is 100%. For example, in the RDI case as shown in Table I, by looking into the case of gesture #0 in the first column, the recognition accuracy of #0:*grab* is 71.41%, and the probability of 12.18% is misjudged as #2:*tilt left and right*.

We observe some misclassification of RDI cases as follows

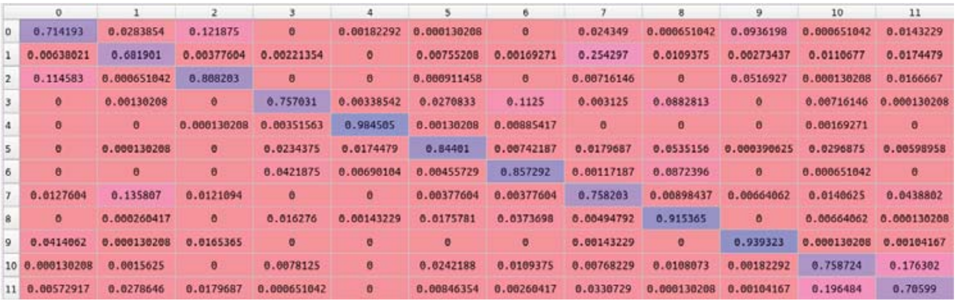
- #0:*grab* and #2:*tilt left and right* have 12.18% and 11.45% chances of being misjudged into each other.
- #1:*clockwise turning* has a probability of 25.42% to be misjudged as #7:*counterclockwise turning*, which is the largest misjudgment in the case study.
- #3:*draw to the left* is mistaken into #6:*clap* with a probability of 11.2%.
- #10:*thumb-index finger* and #11:*thumb-little finger* have 17.6% and 19.6% chances of being misjudged into each other.

By looking into the motion characteristics of the gesture itself, we can find out the similarity of these highly-misclassified gestures. For example, #0 and #2 have displacement changes of the palms up and down, and #3 and #6 have a quick swing.

In general, we observe that using RDI may have problems of weak recognition in the gestures with the horizontal displacement and the angle change, such as #1:*clockwise turning* and #7:*counterclockwise turning*. On the other hand, using RDI performs strongly in the gestures which have vertical displacement change, such as #4:*falling* or #9:*lifting*. The adoption of the angle characteristic in the gestures, which has the horizontal movement, helps them make up for the shortcomings of the lack of discrimination in the RDI. When the RAI is used, the above-mentioned gestures #0, #1, #2, #3 and #6, #10, and #11 all have significant accuracy improvements as shown in Table II. Overall, the performance of using RAI in the experiment is better than using RDI.

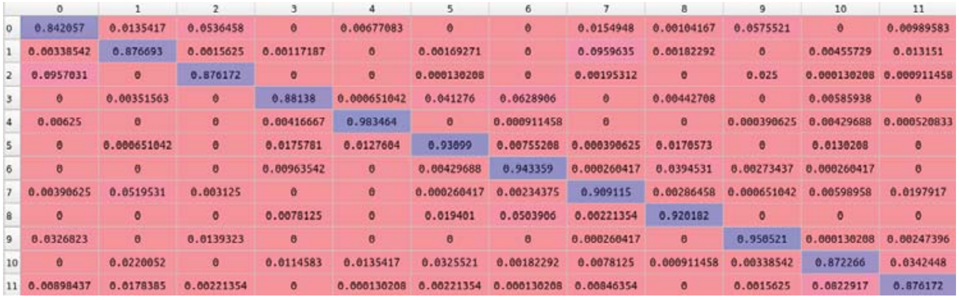
We compare the performance of fusion in Table III with the case of using only RAI or RDI in Table III. Using the fusion further improves the performance compared to the RAI case. For example, #1:*clockwise turning* in the fusion has an

TABLE I
CONFUSION MATRIX OF END-TO-END NETWORK USING RDI



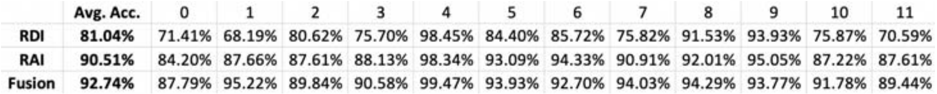
	0	1	2	3	4	5	6	7	8	9	10	11
0	0.714193	0.0283854	0.121875	0	0.00182292	0.000130208	0	0.024349	0.000651042	0.0936198	0.000651042	0.0143229
1	0.00638021	0.681901	0.00377604	0.00221354	0	0.00755208	0.00169271	0.254297	0.0109375	0.00273437	0.0110677	0.0174479
2	0.114583	0.000651042	0.808203	0	0	0.000911458	0	0.00716146	0	0.0516927	0.000130208	0.0166667
3	0	0.00130208	0	0.757031	0.00338542	0.0270833	0.1125	0.003125	0.0882813	0	0.00716146	0.000130208
4	0	0	0.000130208	0.00351563	0.984505	0.00130208	0.00885417	0	0	0	0.00169271	0
5	0	0.000130208	0	0.0234375	0.0174479	0.84401	0.00742187	0.0179687	0.0535156	0.000390625	0.0296875	0.00598958
6	0	0	0	0.0421875	0.00690104	0.00455729	0.857292	0.00117187	0.0872396	0	0.000651042	0
7	0.0127604	0.135807	0.0121094	0	0	0.00377604	0.00377604	0.758203	0.00898437	0.00664062	0.0140625	0.0438802
8	0	0.000260417	0	0.016276	0.00143229	0.0175781	0.0373698	0.00494792	0.915365	0	0.00664062	0.000130208
9	0.0414062	0.000130208	0.0165365	0	0	0	0	0.00143229	0	0.939323	0.000130208	0.00104167
10	0.000130208	0.0015625	0	0.0078125	0	0.0242188	0.0109375	0.00768229	0.0108073	0.00182292	0.758724	0.176302
11	0.00572917	0.0278646	0.0179687	0.000651042	0	0.00846354	0.00260417	0.0330729	0.000130208	0.00104167	0.196484	0.70599

TABLE II
CONFUSION MATRIX OF END-TO-END NETWORK USING RAI



	0	1	2	3	4	5	6	7	8	9	10	11
0	0.842057	0.0135417	0.0536458	0	0.00677083	0	0	0.0154948	0.00104167	0.0575521	0	0.00989583
1	0.00338542	0.876693	0.0015625	0.00117187	0	0.00169271	0	0.0959635	0.00182292	0	0.00455729	0.013151
2	0.0957031	0	0.876172	0	0	0.000130208	0	0.00195312	0	0.025	0.000130208	0.000911458
3	0	0.00351563	0	0.88138	0.000651042	0.041276	0.0628906	0	0.00442708	0	0.00585938	0
4	0.00625	0	0	0.00416667	0.983464	0	0.000911458	0	0	0.000390625	0.00429688	0.000520833
5	0	0.000651042	0	0.0175781	0.0127604	0.93099	0.00755208	0.000390625	0.0170573	0	0.0130208	0
6	0	0	0	0.00963542	0	0.00429688	0.943359	0.000260417	0.0394531	0.00273437	0.000260417	0
7	0.00390625	0.0519531	0.003125	0	0	0.000260417	0.00234375	0.909115	0.00286458	0.000651042	0.00598958	0.0197917
8	0	0	0	0.0078125	0	0.019401	0.0503906	0.00221354	0.920182	0	0	0
9	0.0326823	0	0.0139323	0	0	0	0	0.000260417	0	0.950521	0.000130208	0.00247396
10	0	0.0220052	0	0.0114583	0.0135417	0.0325521	0.00182292	0.0078125	0.000911458	0.00338542	0.872266	0.0342448
11	0.00898437	0.0178385	0.00221354	0	0.000130208	0.00221354	0.000130208	0.00846354	0	0.0015625	0.0822917	0.876172

TABLE III
AVERAGE CLASSIFICATION ACCURACY OF END-TO-END NETWORK USING DIFFERENT INPUTS



	Avg. Acc.	0	1	2	3	4	5	6	7	8	9	10	11
RDI	**81.04%**	71.41%	68.19%	80.62%	75.70%	98.45%	84.40%	85.72%	75.82%	91.53%	93.93%	75.87%	70.59%
RAI	**90.51%**	84.20%	87.66%	87.61%	88.13%	98.34%	93.09%	94.33%	90.91%	92.01%	95.05%	87.22%	87.61%
Fusion	**92.74%**	87.79%	95.22%	89.84%	90.58%	99.47%	93.93%	92.70%	94.03%	94.29%	93.77%	91.78%	89.44%

8% improvement over that in the RAI. The fusion allows the feature extracted from the range, velocity, and angle of the gesture.

V. CONCLUSION

Based on the 77-81GHz mmWave frequency modulated continuous wave (FMCW) radar, we propose a machine learning-based gesture recognition system. We first collect a large amount of gesture data with Texas Instrument (TI) mmWave-based FMCW radar and convert the measurements into a range-doppler image (RDI) or a range-angle image (RAI). The considered end-to-end network model is composed of a convolutional neural network (CNN) and a long short term memory (LSTM). It is observed that using RAI has better performance than using RDI under the setting of 1 transmit antenna and 4 receive antenna. The gesture including the horizontal movement improves significantly when using RAI. Moreover, we fuse the RDI and RAI to improve the performance further.

REFERENCES

[1] H. Abdelnasser, K. A. Harras, and Y. Moustafa, "A Ubiquitous WiFi-based Fine-Grained Gesture Recognition System," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2474–2487, Nov 2019.

[2] R. L. Lao and A. K. Wong, "Detecting hand gestures with wi-fi technology: Applications for received-signal-strength indicators in interactive interface design," *IEEE Consumer Electronics Magazine*, vol. 7, no. 2, pp. 73–82, 2018.

[3] G. Zhou, T. Jiang, Y. Liu, and W. Liu, "Dynamic gesture recognition with Wi-Fi based on signal processing and machine learning," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015, pp. 717–721.

[4] J. S. Suh, S. Ryu, B. Han, J. Choi, J. Kim, and S. Hong, "24 GHz FMCW Radar System for Real-Time Hand Gesture Recognition Using LSTM," in *2018 Asia-Pacific Microwave Conference (APMC)*, 2018, pp. 860–862.

[5] B. Dekker, S. Jacobs, A. S. Kossen, M. C. Kruithof, A. G. Huizing, and M. Geurts, "Gesture recognition with a low power fmcw radar and a deep convolutional neural network," *2017 European Radar Conference (EURAD)*, pp. 163–166, 2017.

[6] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278–3289, 2018.

[7] M. Ritchie, A. Jones, J. Brown, and H. D. Griffiths, "Hand Gesture Classification using 24 GHz FMCW Dual Polarised Radar," in *International Conference on Radar Systems (Radar 2017)*, 2017, pp. 1–6.

[8] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 851–860, 2016.

[9] J.-T. Yu, L. Yen, and P.-H. Tseng, "mmWave Radar-based Gesture Recognition using Machine Learning," in *Taiwan Telecommunications Annual Symposium*, 2020, pp. 1–4.