Increasing Accuracy of Hand Gesture Recognition using Convolutional Neural Network

1st Gyutae Park
Electronic engineering
Gyeongsang National University
Jinju, Republic of Korea
erbxo4321@gnu.ac.kr

2nd V.K. Chandrasegar
Electronic engineering
Gyeongsang National University
Jinju, Republic of Korea
san5432155@gmail.com

3rd JoongGun Park
JD Co., Ltd
Jinju, Republic of Korea
jddesign0@gmail.com

4th Jinhwan Koh
dept. of Electronics Engineering/Engineering Research Institute (ERI)
Gyeongsang National University
Jinju, Republic of Korea
jikoh@gnu.ac.kr

***Abstract*—Human gestures play important roles in the interaction between humans and machines. These human gestures are becoming more important, yet complex gesture input and noise induced by external elements are important problems to solve in order to improve the accuracy of hand gesture recognition methods. Convolutional Neural Networks (CNN) are offered as a technology that can solve this problem in this research. CNN has the advantage of being able to learn image data, and this technology will greatly improve human-machine interaction accuracy. Data was extracted using Vivaldi antennas with a frequency bandwidth of 7.4-9.0 GHz and gain characteristics of 8 dB in five sign language operations, and data that went through the preprocessing process was learned through CNN. The classification results of the proposed CNN showed about 90% accuracy.**

Keywords—IR-UWB Radar, 2D-FFT, Hand Gesture, CNN, Machine Learning

.

I. INTRODUCTION

Human gestures play a very important role in the interaction between humans and machines. A representative example is a technology that replaces a switch or remote control that requires existing physical contact with only a gesture [1]. However, while the importance of hand gesture recognition technology increases, the accuracy of hand gesture recognition technology is still insufficient. The impulse radio ultra-wideband radar (IR-UWB RADAR) technology is effective in addressing these issues.

The radar we used to increase recognition accuracy is the Impulse Radar-Ultra Wide Bandwidth (IR-UWB), which uses a wide frequency band with low power, and has characteristics such as an occupancy bandwidth of 25% or more of the FCC's central frequency and an occupancy bandwidth of 500 MHz or more. And because it instantaneously transmits a very narrow pulse, there is a very low spectral power density over a very wide frequency band. These characteristics can improve the accuracy of hand gesture recognition as they provide high security and high data transmission characteristics, high resolution as accurate distance and location measurements are possible [2-4]. Recognizing human gestures using radar requires the extraction of meaningful information from the received signal, which is difficult to do in big datasets containing a variety of human gestures [5]. To overcome this issue, 2D-FFT was utilized to convert the data into 2D data with important properties, and a convolutional neural network (CNN) was used to classify the results. Recently, several neural network technologies have been studied, and results have been derived that CNN is easy to learn image data. Therefore, CNN was judged to be useful for classifying image data output by radar, so it was used for hand gesture identification [6]. The composition of this paper is as follows. This paper is structured as follows. It is divided into three parts, each of which has the following contents. Section 2 describes the theories related to radar and the Fourier transform and Section 3 describes the experimental process and results, and finally Section 4 describes the conclusions.

II. THEORY

A. IR-UWB Radar

Radars are largely divided into continuous wave radars and pulsed wave radars depending on the radio waves used. Continuous-wave radar refers to a radar that continuously radiates radio waves with a constant frequency, and pulsed wave radar refers to a radar that radiates radio waves that instantaneously increase and returns along a specific cycle. Here, there is a problem in that it is impossible to measure when the reflected wave returns because the same radio wave is continuously received in the receiving unit of the continuous wave radar. Therefore, it is difficult for the continuous wave radar to measure the distance between objects. Therefore, it is

978-1-6654-5818-4/22/$31.00 ©2022 IEEE

the frequency modulation continuous wave radar FMCW radar that has improved the distance measurement capability by modulating the frequency modulation continuous wave radar. However, since it is impossible to distinguish reflected waves (a kind of noise) generated by the speed, angle, and surrounding terrain of the target, the FMCW Doppler radar method solved the problem using the Doppler effect. In other words, when continuous waves are used, additional functions such as frequency modulation must be added to measure the most basic distance, and a high-spec signal processing system is needed to process a large amount of information that continues to flood, making the system larger and more complex. On the other hand, radars using pulse waves radiate waves with different amplitudes for a moment, so they can easily measure the distance to the counterpart by knowing when the reflected wave returned. Here, by using the Doppler effect, 3D information such as target speed and altitude and noise caused by topographic features can be removed. And the pulse-Doppler radar may constitute a small system. Therefore, since a large amount of information does not need to be processed compared to the continuous wave radar, the size of the device becomes smaller, the configuration becomes simpler, and the energy efficiency is increased.

III. EXPERIMENT AND RESULTS

An experimental environment was created as shown in Figure 1 to measure hand gestures using radar. The experiment used NVA-R661 radar from Novelda, which includes two Vivaldi antennas and has a bandwidth of 6.0-8.5 GHz and a gain characteristic of 8dB. The sampling rate of the radar is 39GS/s, and objects up to 1.5m away can be detected. The radar's Tx antenna transmits a signal toward an object, reaches the object, and the reflected signal is transmitted to the Rx antenna to receive the signal.

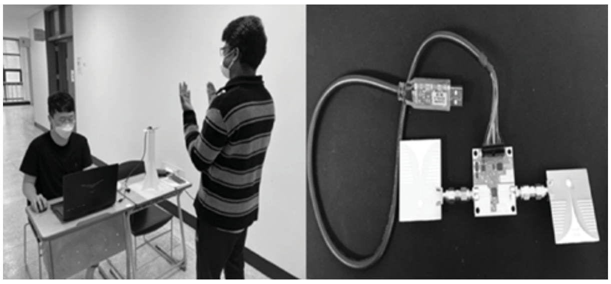

Fig. 1. Experiment environment and NVA-R661

The experiment was conducted in a long corridor where no surrounding objects existed, as shown in Figure 1, to minimize the noise component measured on the radar. In the experiment, three experimenters performed sign language operations, and measurements were performed at a distance of 50 cm from the radar. The hand movements used are five American sign language movements, the first being all done sign, the second being Eat sign, the third being More sign, and the fourth being Thank you sign. Lastly, Sorry sign.

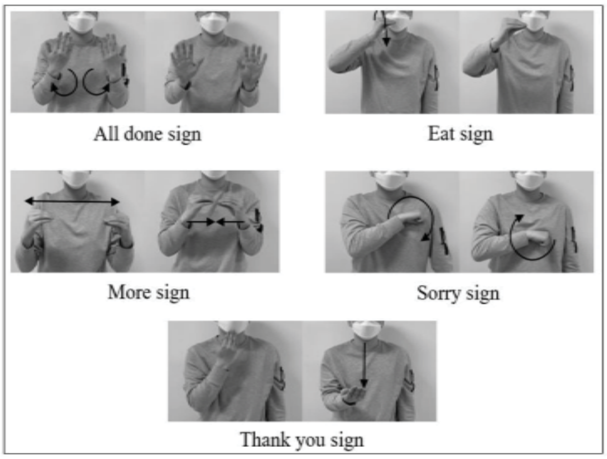

Fig. 2. Selected five sign language

In this paper, five sign language operations were measured 600 times for each operation. Of the 600 data, 500 were measured in general and 100 were measured in a form of behaviour that may be somewhat difficult to recognize (recognition distance, speed of sign language motion, height change of hand).

Figure 3 shows the measurement results of general motion, and Figure 4 shows the measurement results of motion that have changed various elements. Each result is an image obtained by adding all data and then averaging it.

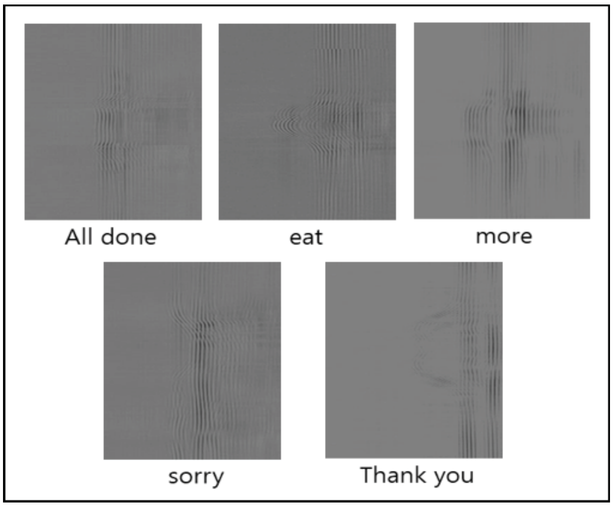

Fig. 3. The measurement results of general motion

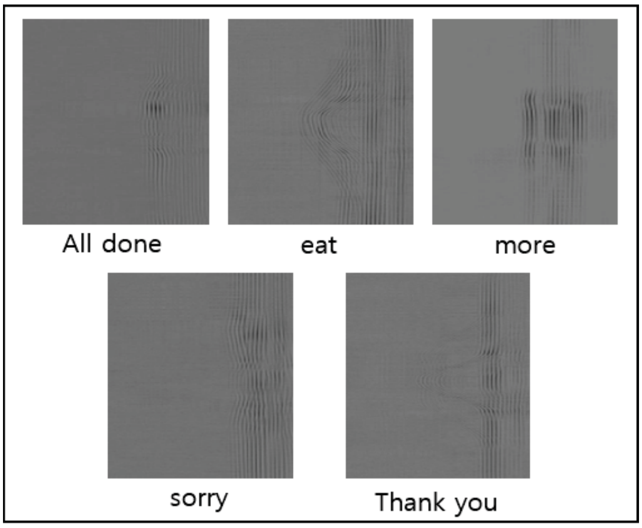

Fig. 4. The measurement results of motion that has changed various elements

In addition, a two-dimensional Fourier transform was performed to extract features from the normal hand gesture data. Image data of the five sign language operations thus obtained were converted into frequency domains through 2D-FFT using MATLAB, and frequency components at each corner, that is, zero frequency values, were collected in the middle to facilitate analysis. Figure 5 is the result of the conversion of All done sign among the five sign language operations.

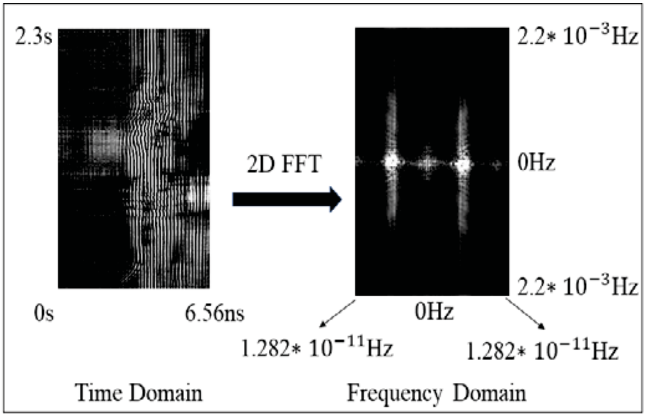

Fig. 5. Converted result (All done sign)

2D-FFT is maintained as double-type data, the same data type as the initial measurement data, but when learning, if data that is not standardized in the gradient descent algorithm of the artificial neural network is input, there is a difference in the learning process and results. To solve this problem, an 8-bit standardization process was additionally performed using Matlab.

Finally, in the normalized 2D-FFT image file, the main feature is the zero-frequency component present in the centre of the image, so only the centre portion of the image was extracted to increase the learning speed of CNN and improve recognition accuracy, and finally, data of 191 by 191 by 1 was obtained.

In this paper, learning was conducted using Matlab's Deep Learning Tool, and the CNN model proposes two CNN models. The first proposes a two-stage serial CNN model that learns by connecting two CNN terminals, and the second proposes a double parallel CNN model that connects two CNN terminals in parallel. Figure 6 shows the structure of the two models.

The filter size of the Convolution Layer used was 3 by 3, the number of filters was 32 strides 1 by 1, the padding was the same, and the weights initializer was the glorot. The pool size of the Maxpool layer was selected as 5 by 5, the stride was 1 by 1, and the padding was the same. Glorot was used as the Weights Initializer of the Fully Connected Layer.

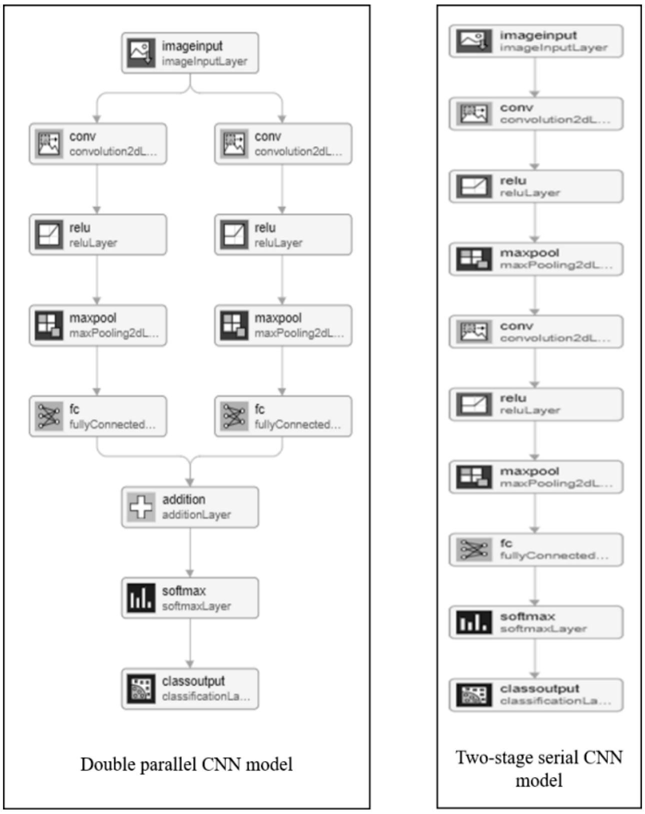

Fig. 6. Block diagram of the proposed model

The main specifications of the system in which learning was conducted are as follows.

CPU: 11th Gen Intel(R) Core(TM) i7-11700 @ 2.50GHz

RAM: 32.0GB

GPU: Geforce RTX3090

A total of four data were used for learning. Raw data(Range-Time map), data performed only 2D-FFT, data performed with 2D-FFT and normalization, and final data performed with 2D-FFT, normalization, and feature partial extraction were divided.

The evaluation method was divided into two methods.

First, the process of learning with 350 general gestures and evaluating with 150 general gestures.

The second is the process of learning with 500 general gestures and evaluating with 100 change gestures.

The results are shown in Tables I and II below.

TABLE I. TWO-STAGE SERIAL CNN MODEL

Data Type	Result of Recognition	
	The results of the general hand gesture evaluation.	*The result of the change hand gesture evaluation.*
Raw data	81.60%	51.00%
Only 2D-FFT	92.53%	64.40%
2D-FFT and normalization	91.47%	62.00%
Final data	92.53%	83.40%

TABLE II. DOUBLE PARALLEL CNN MODEL RESULT

Data Type	Result of Recognition	
	The results of the general hand gesture evaluation.	*The result of the change hand gesture evaluation*
Raw data	90.53%	35.60%
Only 2D-FFT	95.73	75.80%
2D-FFT and normalization	96.13%	80.60%
Final data	96.50%	87.20%

And for comparison with the proposed model, learning was additionally conducted using Google Net, Resnet-50, VGG-19, and AlexNet. Learning and evaluation were conducted using four types of data as mentioned above, and learning and evaluation were conducted three times to prevent fragmentary results of each model, and in the case of three, the maximum result was selected and displayed in a table. Each result is shown in Tables III, IV, V, and VI.

TABLE III. GOOGLENET MODEL RESULT

Objects Type	Result of Recognition	
	The results of the general hand gesture evaluation.	*The result of the change hand gesture evaluation*
Raw data	99.20%	71.20%
Only 2D-FFT	98.27%	86.80%
2D-FFT and normalization	95.20%	87.20%
Final data	92.13%	84.00%

TABLE IV. RESNET-50 MODEL RESULT

Data Type	Result of Recognition	
	The results of the general hand gesture evaluation.	*The result of the change hand gesture evaluation*
Raw data	99.20%	88.40%
Only 2D-FFT	98.93%	82.20%
2D-FFT and normalization	96.40%	82.40%
Final data	96.67%	84.40%

TABLE V. VGG-19 MODEL RESULT

Data Type	Result of Recognition	
	The results of the general hand gesture evaluation.	*The result of the change hand gesture evaluation*
Raw data	91.73%	47.20%
Only 2D-FFT	93.73%	86.60%
2D-FFT and normalization	92.27%	78.60%
Final data	87.33%	78.60%

TABLE VI. ALEXNET MODEL RESULT

Data Type	Result of Recognition	
	The results of the general hand gesture evaluation.	*The result of the change hand gesture evaluation*
Raw data	86.67%	59.40%
Only 2D-FFT	93.60%	83.60%
2D-FFT and normalization	92.53%	79.00%
Final data	88.93%	83.20%

Even though using Alexnet, Googlenet, Resnet-50, and VGG-19 provides better hand gesture recognition accuracy than shown in the above tables, Googlenet and Resnet-50 confirm that the accuracy is higher than the other two models. However, VGG-19 and ALEXNET have higher accuracy than serial models, but lower accuracy than parallel CNN models. The classification accuracy is very high in the case of the prominent CNN model, however, leading Googlenet and Resnet in the longer learning rate. Googlenet took 7 minutes and Resnet 10 minutes, and Googlenet and Resnet-50 took about three to five times longer than parallel models. Whereas the proposed model took only 2 minutes with comparatively higher accuracy. Moreover, a model with a shorter learning time is judged to be a more competitive model, and the model proposes a parallel model with less time execution and higher acuuracy.

IV. Conclusion

In this paper, a radar and deep learning model are proposed as a way to improve the accuracy of hand gesture recognition technology effective in interaction with machines. Five different sign language movements were directly measured by 600 for each operation using IR-UWB radar, and then made into learning data through preprocessing such as 2D-FFT, normalization, and feature extraction with MATLAB. The deep learning model used CNN Layer and Pooling Layer as the main layers, and a two-stage CNN model and a double parallel CNN model were proposed, and learning and evaluation were also conducted through several prominent models to compare the accuracy of classification results.

Compared with the classification results with existing prominent CNN models, the accuracy of the model proposed in this paper was judged to be a model with sufficient competitiveness.

In the future, it plans to conduct research to further increase the accuracy of recognition of gestures that are difficult to recognize with research plans.

References

[1] X. Guo, W. Xu, W. Q. Tang and C. Wen, "Research on Optimization of Static Gesture Recognition Based on Convolution Neural Network," 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Hohhot, China, pp. 398-3982, 2019.

[2] F. Wang, M. Tang, Y. Chiu and T. Horng, "Gesture Sensing Using Retransmitted Wireless Communication Signals Based on Doppler Radar Technology," in IEEE Transactions on Microwave Theory and Techniques, vol. 63, pp. 4592-4602, Dec. 2015

[3] U.S FCC, “Amendment of Part 97 of the Commission's Amateur Service Rules”, 2003

[4] X. Wang, A. Dinh and D. Teng, "Reliability modeling for wireless Ultra Wideband biomedical radar sensing network," 2010 International Conference on Bioinformatics and Biomedical Technology, Chengdu, China, pp. 69-73, 2010.

[5] J. Park and S. H. Cho, "IR-UWB Radar Sensor for Human Gesture Recognition by Using Machine Learning," 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Sydney, NSW, Australia, pp. 1246-1249, 2016.

[6] K. Nakada, A. Ito, H. Hatano and H. Aratame, "New Switchless and Free Positioning Gesture Recognition System Using RNN and CTC Loss Function," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, pp. 450-453, 2018.