IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 February 2023, accepted 25 February 2023, date of publication 9 March 2023, date of current version 21 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3254537*

RESEARCH ARTICLE

# Accuracy Enhancement of Hand Gesture Recognition Using CNN

**GYUTAE PARK, VASANTHA KUMAR CHANDRASEGAR, (Student Member, IEEE), AND JINHWAN KOH, (Member, IEEE)**

Department of Electronic Engineering, Gyeongsang National University, Jinju, Gyeongsangnam 52828, South Korea

Corresponding author: Jinhwan Koh (jikoh@gnu.ac.kr)

This research was funded by National Research Foundation of Korea (NRF): NRF-2020R1F1A1062177 and Institute for Information & communication Technology Planning & evaluation (IITP): 2022-00156409.

**ABSTRACT** Human gestures are immensely significant in human-machine interactions. Complex hand gesture input and noise caused by the external environment must be addressed in order to improve the accuracy of hand gesture recognition algorithms. To overcome this challenge, we employ a combination of 2D-FFT and convolutional neural networks (CNN) in this research. The accuracy of human-machine interactions is improved by using Ultra Wide Bandwidth (UWB) radar to acquire image data, then transforming it with 2D-FFT and bringing it into CNN for classification. The classification results of the proposed method revealed that it required less time to learn than prominent models and had similar accuracy.

**INDEX TERMS** Hand gesture, CNN, deep learning, IR-UWB radar, 2D-Fast Fourier Transform.

## I. INTRODUCTION

Human hand gestures are important as a means of human communication and play a key function in the recently developed Human-machine interfaces technology.

A typical example is a technology that uses hand gestures to replace switches or remote controls that require physical contact [1]. While the importance of hand gesture recognition technology is increasing, the accuracy of hand gesture recognition technology still needs to be improved. An effective way to solve this problem is to use Impulse Radio - Ultra Wide Bandwidth(IR-UWB) RADAR [2]. This radar provides low power and wide frequency band and has features such as an occupied bandwidth of 25% or more of the FCC-regulated center frequency (bandwidth of 500 MHz or more) [3].

Since it instantaneously transmits a very narrow pulse, there is a very low spectral power density over a wide frequency band. These characteristics can improve the accuracy of hand gesture recognition as they provide high security, high data transmission characteristics, and high resolution, as accurate distance and location measurements are possible [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

Recognizing human gestures using radar requires obtaining meaningful information from the received signal, which is challenging to do in big datasets containing a variety of human gestures [5]. To overcome this issue, 2D-Fast Fourier Transform was utilized to convert the data into 2D data with essential properties in the radar systems [6], [7], [8]. And a convolutional neural network (CNN) was used to classify the results. Several neural network methods have recently been analyzed, and the results show that CNN is effective in learning from image data [9]. As a result, CNN was shown to be efficient for classifying image data generated by radar, and it was employed for hand gesture recognition [10], [11]. Other existing papers propose methods to improve accuracy in the data measurement process (such as radar selection, radar parameter adjustment, noise removal, etc.) or a method to improve accuracy through a new model in Deep learning, but there was no case of improving accuracy through a combination of 2D-FFT and CNN. In addition, in this paper, considering that human gestures are not always similar, the evaluation is conducted by dividing them into two sets: a similar gesture data set and a non-similar gesture set in the evaluation process of the CNN model.

This paper's structure follows: Section II describes the methodology, Section III describes the experiment and

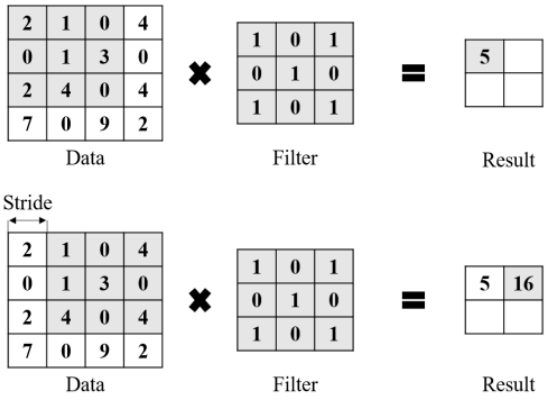


**FIGURE 1.** Convolution process.

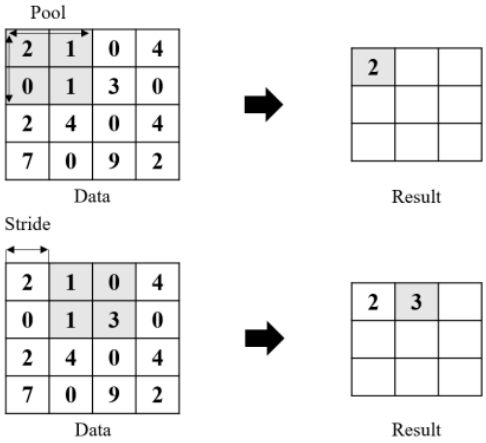


**FIGURE 2.** Maxpooling process.

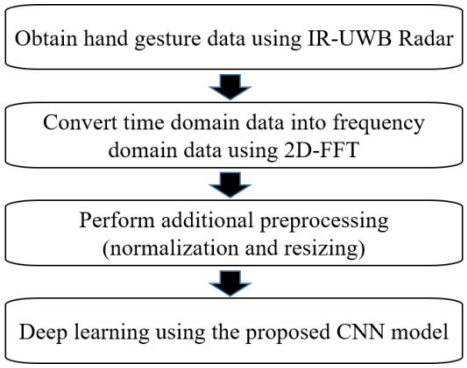


**FIGURE 3.** Block diagram of the experiment procedure.

simulation, Section IV describes the experimental results, and Section V describes the conclusion.

## II. METHODOLOGY

This section deals with IR-UWB radar, CNN, and 2D-Fast Fourier Transform.

IR-UWB Radar transmits radio frequency pulses with extremely short durations (nano-or picoseconds), resulting in a wide bandwidth. There are four technological advantages to using the IR-UWB radar. The first advantage is that it has a large channel capacity because of its short durations, which allows for lower transmission power. Second, the short time duration of the pulses reduces the effect of multipath because the arrival of the pulses may be separated and filtered at the receiver. The third advantage is that the high temporal resolution makes process timing much more precise [12], [13].

Furthermore, regarding environmental adaptability, and privacy protection, radar outperforms visible light, infrared, acoustic technologies, and cameras for non-contact detection methods [14], [15].

Due to the input of high-dimensional data, CNN is a suggested neural network to improve existing artificial neural networks with much longer training times. CNN has a convolution layer and a pooling layer as its main layers. The convolution layer performs convolution processes with filters of a specific size moving above the input data. The movement interval of the filter is called a stride. Figure 1 is an example of a convolution process.

A pooling layer is a layer that reduces the size of the output data from the convolution layer. Pooling includes maxpooling, used to obtain the maximum value from a pool of a specific size, and average pooling, to get the average value from a pool of a specific size. The data size is reduced because it does not require all the data from the output from the convolution layer. In other words, the expected effects of these layers are various positive effects, such as data size reduction, power consumption reduction, and learning time reduction. Figure 2 below shows the process of Maxpooling.

The Fourier Transform is used to convert the signal from the time domain into the frequency domain and vice versa. Extending this method to two dimensions, if the Fourier transform is applied to each axis of the image data, the frequency component of the image data can also be obtained, which contains the data's inherence features. Similarly, effectively using 2-D FFT, CNN can classify images more efficiently and thus enhance its performance.

## III. EXPERIMENT AND SIMULATION

This section deals with experimental and simulation processes. The main content is as follows: experimental environment and equipment, selected five hand gestures, obtained data, and proposed CNN models.

The following is the procedure for the experiment. First, obtain the data of hand gestures using IR-UWB radar. Second, convert data in the time domain to data in the frequency domain using 2D-FFT. Third, perform additional preprocessing to improve the performance of CNN. Finally, do deep learning using the proposed CNN model. Figure 3 is a block diagram of the experimental procedure.

The experiment was conducted in a long corridor with no surrounding objects to reduce the noise component detected on the radar. In this experiment, we obtained the hand gesture data using IR-UWB (NVA-R661) radar, which contains two Vivaldi antennas and features a bandwidth of 6.0–8.5 GHz and a gain characteristic of 8 dB, shown in figure 4.

In this experiment, five American sign language gestures were selected, such as the "All done" sign, "Eat" sign,

**FIGURE 4.** Experiment settings and NVA-R661.

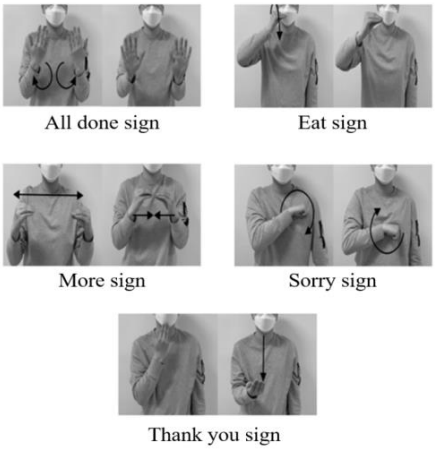


**FIGURE 5.** Selected five sign languages.

"More" sign, "Sorry" sign, and "Thank you" sign, as shown in Figure 5.

Since a person's sign language behavior is not always the same, several factors should be considered, such as the height of the person, the shape of the sign language gestures, the speed of the sign language behavior, and the measurement distance. Therefore, the experiment was conducted with two measurements, method I and II. As the method, I involve fixing the person at a distance of 50 cm from the radar and measuring the same hand gestures 500 times. Method II was performed by positioning the three people with different hand shapes, heights, and sign language behaviors at 40 cm to 60 cm from the radar and measuring hand gestures 100 times. In method II, hand gestures were generated 20 times for each increasing distance of 5 cm, from 40 cm to 60 cm (40cm, 45cm, 50cm, 55cm, 60cm).

Figure 6 shows the average data of all 500 data sets measured using the method I.

Figure 7 shows the average data of all 100 data sets measured using method II. By comparing Figures 6 and 7, it can be seen that the difference between the shape of data in Figure 6 and Figure 7.

In addition, a two-dimensional Fourier transform was performed to convert from the time domain to the frequency domain to obtain the frequency value of the hand gesture data. After converting the image data into the frequency domain, the zero frequency value of each corner was moved to the middle. Figure 8 shows the vertical-axis and horizontal-axis values for each domain of data.

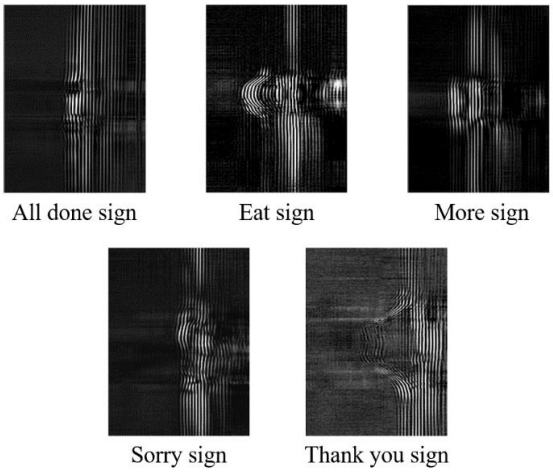


**FIGURE 6.** Acquired hand gestures data using Method I.

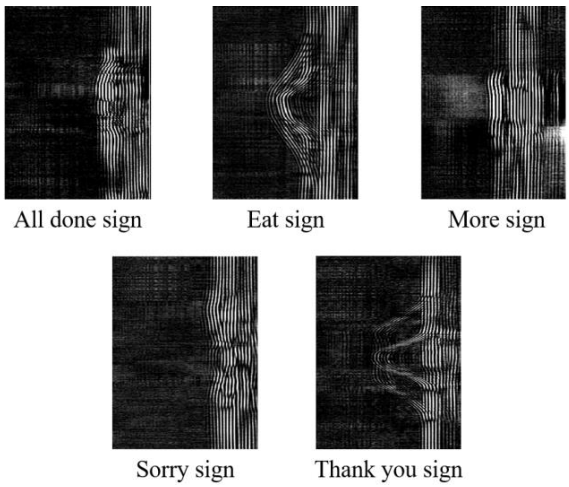


**FIGURE 7.** Acquired hand gestures data using Method II.

In this experiment, the obtained dataset's data type was double-type. The input of such unnormalized data causes a difference in the calculation result of the gradient descent algorithm of the neural network. However, an 8-bit normalization process was performed to overcome this issue by converting the type from double type to int type. And since an essential part of the normalized 2D-FFT image data is the zero-frequency component at the center of the image, only that part was cut and stored separately to minimize the processing time.

In this paper, two types of CNN models are proposed. The first model is a two-stage serial CNN model, and the second model is a double parallel CNN model. Two-stage serial CNN model is one in which two sets of layers are connected by a convolution, relu, and maxpool layer in series. A double parallel CNN model is one in which two sets of layers are connected by a convolution, relu, and maxpool layer in parallel.

While classifying data using CNN, the process is divided into two parts: a learning process and a testing process. The learning process refers to the process of learning so that the system can classify what data and the test process refers

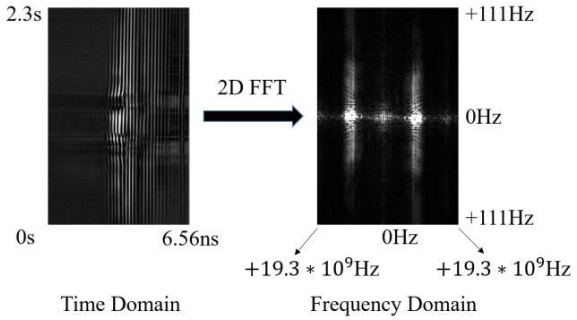


**FIGURE 8.** Converted result (All done sign).

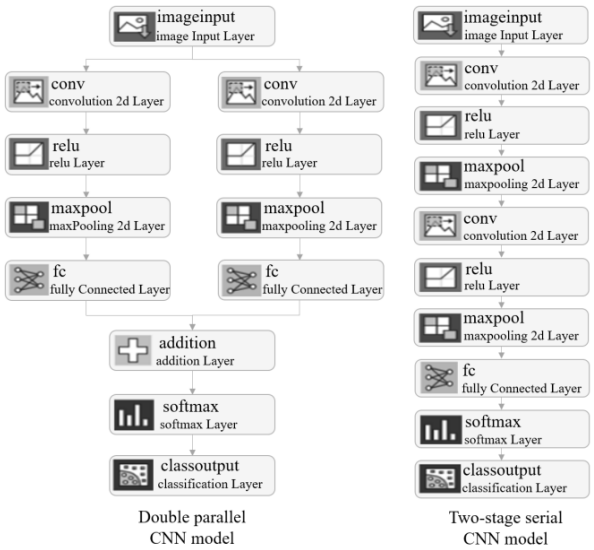


**FIGURE 9.** Block diagram of the proposed CNN models.

to checking the performance of how accurately the system can classify. Two separate inputs need to be provided for the learning and testing processes. As explained previously, the data obtained in this experiment is separated into two categories. The first is 500 datasets of method I, and the second is 100 datasets of method II. Using these data, a case 1 classification simulation was performed, with 350 (70%) of the method1 datasets put into the learning process and the remaining 150 (30%) used in the test process. And a case-to-case classification simulation was performed, in which all method I data was put into the learning process and method II data was put into the testing process.

Case 1 is the classification of gestures when the majority of people use a similar form of sign language. However, the actual situation is not as simple as in case 1. Most people use sign language in a variety of forms, and if the system is to classify hand gestures, it must also classify more complicated forms of gestures, such as in Case 2.

The convolution layer's filter size was 3 by 3, the number of filters was 32, the stride was 1 by 1, the data was padded to the same size as the Convolution layer's input, and the weightsInitializer was glorot. The Maxpool layer's Pool Size was 5 by 5, the stride was 1 by 1, and the data was padded to the same size as the Maxpool layer's input. The FullyConnected layer's weightsInitializer is glorot.

**TABLE 1.** Two-stage serial CNN model results.

| Input data | Results of Recognition | |
|---|---|---|
| | Case 1 | Case 2 |
| Raw data | 74.42% | 42.93% |
| Only 2D-FFT | 91.38% | 56.73% |
| 2D-FFT and normalization | 89.82% | 59.60% |
| Resized 2D-FFT and normalization | 91.78% | 80.27% |

**TABLE 2.** Double parallel CNN model results.

| Input data | Results of Recognition | |
|---|---|---|
| | Case 1 | Case 2 |
| Raw data | 88.40% | 32.33% |
| Only 2D-FFT | 95.60% | 71.67% |
| 2D-FFT and normalization | 95.29% | 78.07% |
| Resized 2D-FFT and normalization | 95.20% | 86.00% |

**TABLE 3.** GoogLeNet model results.

| Input data | Results of Recognition | |
|---|---|---|
| | Case 1 | Case 2 |
| Raw data | 96.71% | 70.13% |
| Only 2D-FFT | 97.97% | 84.53% |
| 2D-FFT and normalization | 94.09% | 83.07% |
| Resized 2D-FFT and normalization | 90.27% | 83.33% |

The main specifications of the system that conducted learning are CPU: 11th Gen Intel (R) Core (TM) i7-11700 @ 2.50GHz, RAM: 32.0GB, and GPU: Geforce RTX3090.

## IV. COMPUTATION RESULT

For comparison with the proposed model, learning was additionally conducted using GoogLeNet, ResNet-50, VGG-19, and AlexNet. However, VGG-19 and AlexNet have higher accuracy than the two-stage serial CNN model but lower accuracy than the double parallel CNN model. And the classification accuracy is very high in the cases of GoogleNet, ResNet-50, and the dual parallel model, but the double parallel model outperforms GoogleNet and ResNet-50 regarding

**TABLE 4.** ResNet-50 model results.

| Input data | Results of Recognition | |
|---|---|---|
| | Case 1 | Case 2 |
| Raw data | 98.98% | 88.40% |
| Only 2D-FFT | 98.04% | 80.20% |
| 2D-FFT and normalization | 95.69% | 79.80% |
| Resized 2D-FFT and normalization | 95.29% | 81.27% |

**TABLE 5.** VGG-19 model results.

| Input data | Results of Recognition | |
|---|---|---|
| | Case 1 | Case 2 |
| Raw data | 91.20 | 41.33% |
| Only 2D-FFT | 92.18% | 83.60% |
| 2D-FFT and normalization | 89.33% | 78.87% |
| Resized 2D-FFT and normalization | 84.53% | 72.53% |

**TABLE 6.** AlexNet model results.

| Input data | Results of Recognition | |
|---|---|---|
| | Case 1 | Case 2 |
| Raw data | 91.20% | 41.33% |
| Only 2D-FFT | 92.18% | 83.60% |
| 2D-FFT and normalization | 89.33% | 78.87% |
| Resized 2D-FFT and normalization | 84.53% | 72.53% |

learning time. The double-parallel model takes 2 minutes, GoogLeNet takes 7 minutes, and ResNet-50 takes 10 minutes. GoogLeNet and ResNet-50 take three and five times longer than the double parallel model.

The proposed models (two-stage serial model and double parallel model) took only 2 minutes with comparatively high accuracy. Moreover, a model with a shorter learning time is considered more competitive, and the proposed parallel model executes in less time and with the same accuracy as GoogLeNet and ResNet-50. Tables 1 to 6 below are tables that summarize the accuracy of each model. The tables show the

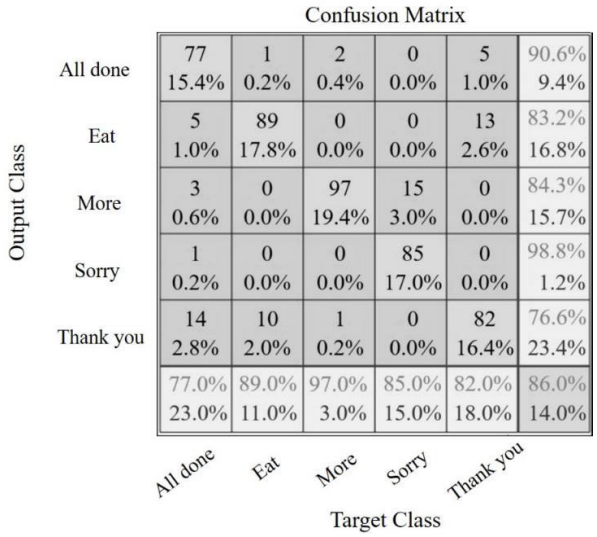


**FIGURE 10.** Double Parallel CNN Model's confusion matrix.

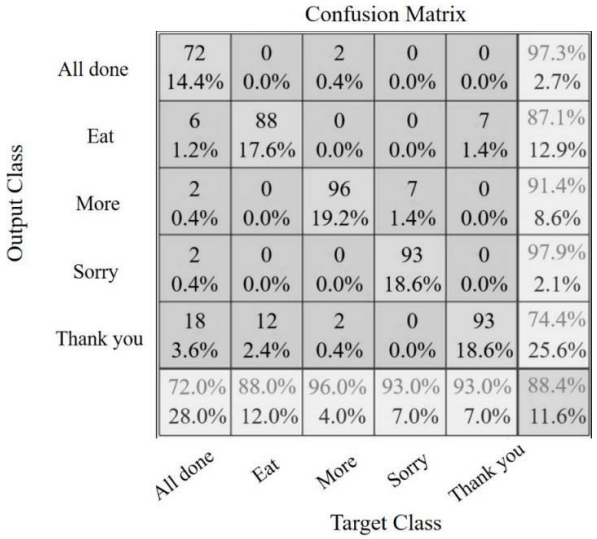


**FIGURE 11.** ResNet-50 Model's confusion matrix.

recognition results with various input data, such as raw radar data, 2D-FFT of the raw data, normalization, and finally, resizing of the data.

Tables 1 and 2 show the classification results of the two-stage serial CNN model and the double parallel CNN model proposed in this paper. Dual parallel models recorded higher accuracy than two-stage serial models, and both models achieved higher accuracy improvements using preprocessing. In particular, the effect of preprocessing may be confirmed in case 2.

Table 3 to Table 6 show the classification results of prominent CNN models. Learning is performed with a large number of layers in superior models. Therefore, their accuracy is reasonably good even without the preprocessing process in the result of cases 1 and 2 when compared to the proposed models. Even so, it is evident that when 2D-FFT is used, the accuracy improves. However, not all preprocessing enhances accuracy, and preprocessing for each structure must

be appropriately applied. Prominent models use a large number of layers for learning, as mentioned earlier, so they have a longer execution time compared to the proposed model. It took about 3 minutes for AlexNet, 7 minutes for VGG-19 and GoogLeNet, and 10 minutes for Resnet-50. The figure below is the confusion matrix of the Double Parallel CNN model and the confusion matrix of ResNet-50, recorded with the highest accuracy, as mentioned in case 2.

## V. CONCLUSION

In this research, a combination of 2D-FFT and CNN is proposed to increase the accuracy of hand gesture recognition techniques. Five American sign languages(''All done,'' ''Eat,'' ''More,'' ''Sorry,'' and ''Thank you'')were selected and measured 500 times under similar conditions and 100 times under non-similar conditions with IR-UWB radar. In addition, the obtained data were preprocessed with 2D-FFT, 8bit-Normalization, and Resizing to increase CNN classification accuracy, which was input into the Double parallel CNN model and the Two-stage serial CNN model for simulation. Several prominent models were also simulated by inputting the obtained data to validate the proposed model's performance.

The conclusion of this paper is as follows: Although it cannot always be asserted that all preprocessing increases accuracy, experiment results show that classification accuracy is increased when combining 2D-FFT preprocessing with the CNN deep learning model. Among the prominent models, namely GoogLeNet and ResNet-50, have achieved high accuracy without using 2D-FFT. The suggested double-parallel CNN model with 2D-FFT achieved higher accuracy than the conventional models. Furthermore, the proposed Double Parallel CNN Model takes about 2 minutes to finish, which is much faster than GoogLeNet and ResNet-50. Thus, this paper's proposed combination of 2D-FFT and CNN validates a highly competitive model.

## REFERENCES


[1] X. Guo, W. Xu, W. Q. Tang, and C. Wen, ''Research on optimization of static gesture recognition based on convolution neural network,'' in *Proc. ICMCCE*, Hohhot, China, Oct. 2019, p. 398.

[2] T. J. Daim and R. M. A. Lee, ''The effect of body position on IR-UWB radar sensor-based hand gesture speed recognition and classification system,'' in *Proc. I2CACIS*, Shah Alam, Malaysia, Jun. 2022, pp. 158–162.

[3] *FCC 02-48*, U.S. Federal Commun. Commission, Washington, DC, USA, 2003, pp. 15–17.

[4] F.-K. Wang, M.-C. Tang, Y.-C. Chiu, and T.-S. Horng, ''Gesture sensing using retransmitted wireless communication signals based on Doppler radar technology,'' *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 12, pp. 4592–4602, Dec. 2015, doi: 10.1109/TMTT.2015.2495298.

[5] X. Wang, A. Dinh, and D. Teng, ''Reliability modeling for wireless ultra wideband biomedical radar sensing network,'' in *Proc. ICBBT*, Chengdu, China, 2010, pp. 69–73.

[6] A. Rizik, A. Randazzo, R. Vio, A. Delucchi, H. Chible, and D. D. Caviglia, ''Feature extraction for human-vehicle classification in FMCW radar,'' in *Proc. ICECS*, Genoa, Italy, 2019, pp. 131–132.

[7] R. Wan, Y. Song, T. Mu, and Z. Wang, ''Moving target detection using the 2D-FFT algorithm for automotive FMCW radars,'' in *Proc. CISCE*, Haikou, China, Jul. 2019, pp. 239–243.

[8] M. Song, J. Lim, and D.-J. Shin, ''The velocity and range detection using the 2D-FFT scheme for automotive radars,'' in *Proc. IC-NIDC*, Beijing, China, Sep. 2014, pp. 507–510.

[9] L. Yuan, ''Remote sensing image classification methods based on CNN: Challenge and trends,'' in *Proc. CONF-SPML*, Stanford, CA, USA, Nov. 2021, pp. 213–218.

[10] J. Park and S. H. Cho, ''IR-UWB radar sensor for human gesture recognition by using machine learning,'' in *Proc. ICMCCE*, Sydney, NSW, Australia, Dec. 2016, pp. 1246–1249.

[11] K. Nakada, A. Ito, H. Hatano, and H. Aratame, ''New switchless and free positioning gesture recognition system using RNN and CTC loss function,'' in *Proc. CSCI*, Las Vegas, NV, USA, Dec. 2018, pp. 450–453.

[12] A. G. Yarovoy, X. Zhuge, T. G. Savelyev, and L. P. Ligthart, ''Comparison of UWB technologies for human being detection with radar,'' in *Proc. EURAD*, Munich, Germany, Oct. 2007, pp. 295–298.

[13] A. G. Yarovoy and L. P. Ligthart, ''UWB radars: Recent technological advances and applications,'' in *Proc. IEEE Radar Conf.*, Waltham, MA, USA, Apr. 2007, pp. 43–48.

[14] S. Skaria, A. Al-Hourani, and R. J. Evans, ''Deep-learning methods for hand-gesture recognition using ultra-wideband radar,'' *IEEE Access*, vol. 8, pp. 203580–203590, 2020, doi: 10.1109/ACCESS.2020.3037062.

[15] M. Jia, S. Li, J. L. Kernec, S. Yang, F. Fioranelli, and O. Romain, ''Human activity classification with radar signal processing and machine learning,'' in *Proc. UCET*, Glasgow, U.K., Aug. 2020, pp. 1–5.


**GYUTAE PARK** received the B.S. degree in electronic engineering from Gyeongsang National University, Jinju, South Korea, where he is currently pursuing the M.S. degree with the Electronic Engineering Department. His research interests include radar signal processing, sensors, and machine learning.

**VASANTHA KUMAR CHANDRASEGAR** (Student Member, IEEE) was born in Kolar Gold Fields, Karnataka, India, in 1983. He received the B.E. and M.Tech. degrees in electrical and electronics engineering from Visvesvaraya Technological University, Bangalore, India, and the Ph.D. degree with the Electronic Engineering Department, Gyeongsang National University, Jinju, South Korea, in 2019. From 2012 to 2018, he worked as an Assistant Professor at the electrical and electronics engineering department at (2012 to 2016) and Alliance University (2017 to July 2019), Bangalore, India. His research interests include radar signal processing sensors and machine learning.

**JINHWAN KOH** (Member, IEEE) received the B.S. degree in electronics from Inha University, Incheon, South Korea, and the M.S. and Ph.D. degrees in electrical engineering from Syracuse University, Syracuse, NY, USA. He is currently a Professor with the Department of Electronic Engineering, Engineering Research Institute, Gyeongsang National University, Jinju, South Korea. His current research interests include radar signal processing, remote sensing, and AI applications.