JOURNAL OF ELECTROMAGNETIC ENGINEERING AND SCIENCE, VOL. 22, NO. 3, 335~343, MAY. 2022

https://doi.org/10.26866/jees.2022.3.r.95
ISSN 2671-7263 (Online) · ISSN 2671-7255 (Print)

Dop-DenseNet: Densely Convolutional Neural Network-Based Gesture Recognition Using a Micro-Doppler Radar

Hai Le[1] · Van-Phuc Hoang[1,*] · Van Sang Doan[2] · Dai Phong Le[1]

Abstract

Hand gesture recognition is an efficient and practical solution for the non-contact human–machine interaction in smart devices. To date, vision-based methods are widely used in this research area, but they are susceptible to light conditions. To address this issue, radar-based gesture recognition using micro-Doppler signatures can be applied as an alternative. Accordingly, the use of a novel densely convolutional neural network model, Dop-DenseNet, is proposed in this paper for improving hand gesture recognition in terms of classification accuracy and latency. The model was designed with cross or skip connections in a dense architecture so that the former features, which can be lost in the forward-propagation process, can be reused. We evaluated our model with different numbers of filter channels and experimented with it using the Dop-Net dataset, with different time lengths of input data. As a result, it was found that the model with 64 3 × 3 filters and 200 time bins of micro-Doppler spectrogram data could achieve the best performance trade-off, with 99.87% classification accuracy and 3.1 ms latency. In comparison, our model remarkably outperformed the selected state-of-the-art neural networks (GoogLeNet, ResNet-50, NasNet-Mobile, and MobileNet-V2) using the same Dop-Net dataset.

Key Words: Convolutional Neural Network, Hand Gesture Recognition, Micro-Doppler Radar.

I. Introduction

Gesture recognition is a hot topic in the field of computer vision and language technology, with the goal of interpreting human gestures using mathematical algorithms [1]. By understanding human actions, a device equipped with sensors can react as expected, such as by controlling a smart TV, a gaming box, a robot, or a computer. Nowadays, gesture recognition is popularly used for emotion recognition using vision-based methods [2, 3]. Despite achieving remarkable performance in terms of recognition accuracy, vision-based approaches are susceptible to failure under weak light conditions. To address this issue, Tan and Triggs [4] combined the strengths of robust illumination normalization, local texture-based face representations, distance–transform-based matching, kernel-based feature extraction, and feature fusion. As a result, the method achieved an 88.1% face verification rate and a 0.1% false accept rate on the challenging FRGC-204 dataset. However, this accuracy level is still low compared to those of other approaches.

An alternative to vision-based methods is the use of the fre-

Manuscript received July 14, 2021 ; Revised September 6, 2021 ; Accepted September 22, 2021. (ID No. 20210714-079J)
[1]Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam.
[2]Vietnam Naval Academy, Nha Trang, Vietnam.
[*]Corresponding Author: Van-Phuc Hoang (e-mail: phuchv@lqdtu.edu.vn)

quency-modulated continuous-wave (FMCW) radar, which can provide micro-Doppler features for gesture recognition [5]. Accordingly, Malysa et al. [6] used the hidden Markov model (HMM) for gesture recognition in their study, based on a 77-GHz FMCW radar system. By using micro-Doppler spectrogram images, HMM can classify four hand gestures with up to 82.3% overall accuracy within 30 frames. However, the study was performed with a single target and small gesture classes, which are insufficient challenges for gesture recognition tasks. In a scenario involving multiple moving subjects, Peng et al. [7] investigated the effectiveness of a 5.8-GHz FMCW radar with range-Doppler processing in recognizing human gestures. They demonstrated that portable FMCW radars could recognize human gestures in the presence of multiple moving people. Nevertheless, no gesture identification model was presented in this study. With a similar processing method of range Doppler maps from raw signals of an FMCW radar, Ryu et al. [8] introduced a feature-based gesture recognition method using a quantum-inspired evolutionary algorithm (QEA) to enhance gesture recognition accuracy. On a dataset of seven hand motion classes, gesture recognition was performed using the features selected by QEA, and 85.81% classification accuracy was obtained, higher than those of the k-nearest neighbors (81.43%) and random forest (RF, 83.33%) classifiers. As expected, the handcrafted feature extraction approaches revealed a disadvantage: their classifiers work only with small datasets and cannot cover all realistic conditions. In contrast, deep neural networks (DNNs) can automatically extract useful features to improve classification accuracy. In particular, Vandersmissen et al. [9] proposed a robust feature learning approach based on deep convolutional neural networks (DCNNs) for use in identifying persons on the basis of their gait characteristics through the micro-Doppler signatures of a low-power FMCW radar device. In the study, the DCNN model significantly outperformed the support vector machine (SVM) and RF approaches by a margin of 17%. Moreover, in the experiment with larger time windows, the DCNN model was able to further lower the error rate to 0% for above 25 seconds windows. In another study, a multidimensional parameter dataset was used for hand gesture recognition [10]. Specifically, Range-Frame-Map, Doppler-Frame-Map, and Angle-Frame-Map were included in the datasets for a meta-learning-based multi-branch network. As a result, the network obtained 97.3% recognition accuracy with seven gestures. However, the multidimensional parameters lead to high computational complexity in the preprocessing stage and lengthen the model execution time. To improve the efficiency of hand gesture recognition systems based on the 60-GHz FMCW radar, Lee et al. [11] proposed a three-dimensional (3D) CNN with an Inception structure to process the range–Doppler matrix sequence, which yielded 98.8% average accuracy. Although the 3D CNN model provided high precision, it suffered from a computational burden. The human gesture recognition task was also built on an edge-computing platform combined with an FMCW radar sensor [12]. The system, including the radar and NVIDIA Jetson Nano, was embedded. The CNN model achieved 98.47% and 93.11% average accuracy using gestures from taught and untaught subjects, respectively. In addition, a combination of long short-term memory and CNN is also employed for facilitating gesture recognition [13].

Despite remarkable achievements in gesture recognition, the aforementioned approaches still have one or more of the following drawbacks: simple experiment context, high computational cost, and slow execution time. Thus, there is still room for researching and developing a novel DCNN model that can improve human gesture recognition performance in terms of recognition accuracy and computational complexity. Motivated by the Dop-Net dataset used in the study by Ritchie et al. [14], we propose the use of a dense CNN model, the so-called Dop-DenseNet, to improve hand gesture recognition accuracy and reduce time-execution latency. Dop-DenseNet was designed with cross and skip connections to leverage the combination of the extracted features with the former ones for achieving high accuracy during the training process. The proposed model was evaluated using different numbers of filter channels and different time lengths of input data. Afterward, the Dop-DenseNet model with 64 3×3 filters was chosen to compete with several state-of-the-art networks: GoogLeNet [15], ResNet-50 [16], NasNet-Mobile [17], and MobileNet-V2 [18]. Consequently, our model remarkably outperforms other considered models in terms of accuracy and execution time.

Thus, the following are the main contributions of our study:

- We propose the Dop-DenseNet model and evaluated it using different configurations, such as changing the number of filter channels and the time lengths of the input data.
- We compared our model with the other considered ones in terms of recognition accuracy, execution time, and memory cost on the Dop-Net dataset.

The remainder of this paper is organized as follows. Section II describes our basis for using the FMCW radar for hand gesture recognition. Section III presents the proposed Dop-DenseNet and its evaluation results. Section IV compares the performances of several state-of-the-art CNN models with that of our Dop-DenseNet model for hand gesture recognition. Finally, Section VI concludes the paper.

II. Frequency-Modulated Continuous-Wave Radar for Hand Gesture Recognition

In this section, we present an operation principle for the FMCW radar that can be used for hand gesture recognition

based on micro-Doppler signatures.

1. Principles of the FMCW Radar

FMCW radars are widely used for tasks involving human gesture recognition. They transmit electromagnetic FMCW power through antennas into a space for measuring objects. By processing the reflected signal and comparing it with the duplicate of the transmitted signal, the radar can determine the target position. An example of an FMCW radar scheme for this application is shown in Fig. 1.

The signal transmitted by the FMCW radar can be expressed as follows:

$$s_t(t) = A_t \cos\left(2\pi f_c t + 2\pi \int_0^t f_T(\tau)d\tau\right), \quad (1)$$

where $f_T(\tau) = (B/T)\tau$ is the transmitter sweep frequency, f_c is the center carrier frequency, B is the bandwidth, T is the signal period, and A_t is the transmitted signal amplitude.

Due to the motion of the hand, the frequency of the receiving signal is modulated by the Doppler shift, $f_d = 2f_c v/c$. Thus, the receiving frequency is defined as follows:

$$f_R(t) = \frac{B}{T}(t - \Delta\tau) + f_d, \quad (2)$$

where $\Delta\tau = 2(R_0 + vt)/c$ is the time delay of the received signal, R_0 is the range of the hand from the radar antenna, v is the speed of hand motion, and c is the speed of light. Thus, the signal received by the radar antenna is presented as follows:

$$s_r(t) = A_r \cos\left(2\pi f_c(t - \Delta\tau) + 2\pi \int_0^t f_R(\tau)d\tau\right) + n(t), \quad (3)$$

where A_r is the received signal amplitude and $n(t)$ is the Gaussian noise. The output of the mixer is an intermediate frequency (IF) signal generated by mixing the received signal with the duplicate of the transmitted one. Afterward, the IF signal goes to a low-pass filter and an IF amplifier. If there is no reflected signal (meaning there is no detected target), the IF signal is only noise. Then, the IF signal is converted to digital form by an analog-to-digital converter for further processing in the digital signal processor. As a result, the IF signal can be written as follows:

$$s_{IF}(t) = \begin{cases} \frac{1}{2}\cos\left(\pi f_c \frac{R_0}{c} + 2\pi\left(\frac{2R_0}{c}\frac{B}{T} + \frac{2f_c v}{c}\right)t\right) + n(t) \text{ for target,} \\ n(t) \text{ for no target} \end{cases} \quad (4)$$

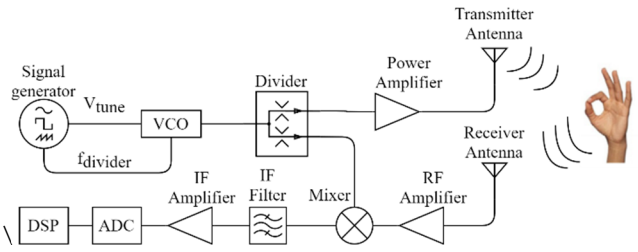

Fig. 1. Schematic of the FMCW radar for hand gesture recognition.

where R_0 is the range of the target at $t = 0$, and $n(t)$ is the Gaussian noise. It can be seen in Eq. (4) that the existence of the target is detected by the strength of the IF signal, which is represented by the signal-to-noise ratio. The moving target is determined by analyzing the micro-Doppler spectrogram of the IF signal.

2. Micro-Doppler Effect

The micro-Doppler effect is a phenomenon of secondary modulation in a reflected signal caused by the movements of parts versus a target center. The micro-Doppler effect facilitates determining the kinetic characteristics of the target; therefore, it can be used to identify hand gestures. In almost all cases, the Doppler frequency is produced by a combination of many complex and different micromovements, such as translation, rotation, and vibration. The oscillations of these micromotions produce frequency modulation phenomena on the reflected signal. They cause additional changes in the Doppler frequency shift constant of the entire translation motion of the target. For a target with only a translation motion with a constant velocity, the Doppler shift produced by this translation motion is a time-invariant function. In contrast, if a target has a vibrating or rotating motion, the Doppler frequency produced by this rotation or vibration is a function of time, which expresses a time-varying modulation of the carrier frequency of the reflected signal. The general technique for micro-Doppler analysis is to represent the time-frequency spectrum. Nowadays, conventional spectrograms defined by the short-time Fourier transform (STFT) are widely used for micro-Doppler analysis, given as follows:

$$X(\tau, \omega) = STFT\{x(t)\} = \int_{-\infty}^{+\infty} x(t)w(t - \tau)e^{-j\omega t}dt, \quad (5)$$

where $x(t)$ is the input signal of transformation, and $w(t-\tau)$ is the kernel (so-called window) function. The resolution of the STFT spectrogram was identified via the window function and the overlapping rate. It is obvious that the micro-Doppler spectrum has two dimensions: time and frequency. The value of X is a complex number; therefore, X is designated as the neural network model input, which has a size of $M \times N \times 2$, where M and N are the sizes of the frequency and time points, respectively, and 2 represents the real and imagined parts of complex values X.

III. Convolution Neural Network-based Hand Gesture Recognition

In this section, we first introduce a dataset, Dop-Net, which we used to assess CNN models. We then propose a CNN model whose performance was evaluated using the Dop-Net dataset by changing the number of filter channels of the model and the time length of input data.

1. Dop-Net Dataset

Dop-Net is an FMCW radar dataset in which each *.mat file represents the data of a person divided into different gestures recorded from that person. The FMCW radar constantly transmits a chirp signal, which increases (up-chirp) and decreases (down-chirp) its frequency linearly over time. The signal reflected from a hand is then mixed with the duplicate copy of the transmitted signal to downconvert the received signal to an intermediate frequency. The data for this classification issue are generated using a 24-GHz FMCW radar with a 750-MHz bandwidth. The other parameters of the radar are presented in Table 1.

Each gesture is made directly in front of the radar at a ≈ 30 cm distance at the same height as the antenna. The FMCW radar captures 30 seconds of the reflected signal for each gesture. The dataset provides four separate hand gestures that can be applied in human–machine interface applications. By using the STFT-based micro-Doppler analysis method and automatic segmentation with 75% overlapping, we generated 9,410 total time-frequency spectral images with a size of 200 × 200. The distribution of gestures is presented in Fig. 2, with the numbers of images for the click, pinch, swipe, and wave actions shown as 2,063, 2,026, 2,821, and 2,500, respectively.

Evidently, different hand gestures pose different micro-Doppler signatures in STFT images (Fig. 3), which facilitate machine learning, and deep learning algorithms discriminate individual hand actions. Accordingly, Ritchie et al. [14] compared different machine-learning techniques in terms of classification accuracy. As a result, a quadratic SVM classifier obtained the best accuracy (74.2%).

Table 1. Main parameters of the FMCW radar for dataset generation

Parameter	Value
Frequency (GHz)	24
Bandwidth (MHz)	750
ADC resolution (bits)	12
Transmit power (dBm)	13

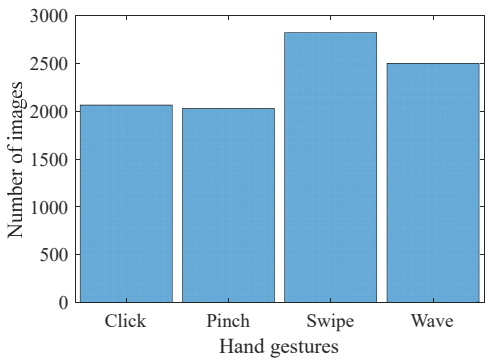

Fig. 2. Distribution of hand gestures in the Dop-Net dataset.

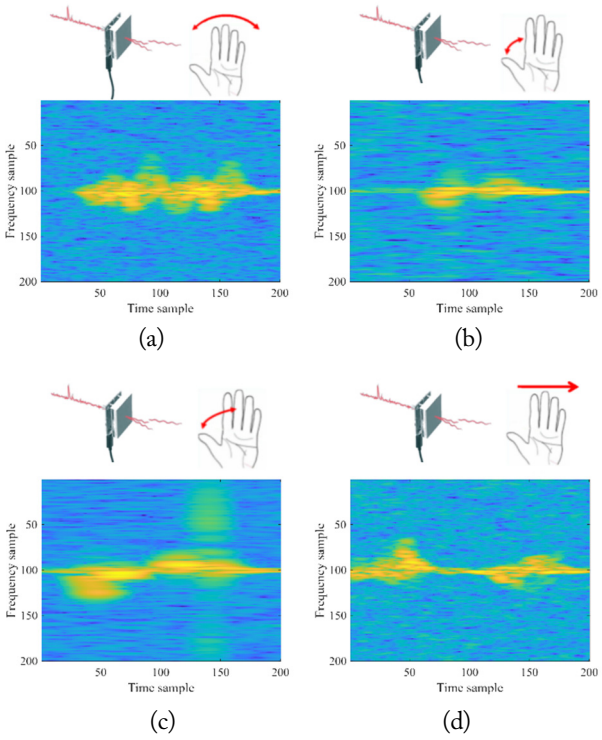

Fig. 3. Micro-Doppler spectra of different hand gestures: (a) wave, (b) pinch, (c) click, and (d) swipe.

2. Dop-DenseNet Model

For use in recognizing hand gestures, we propose a lightweight CNN model based on a densely connected architecture: Dop-DenseNet. The overall scheme of our network is detailed in Fig. 4, where cross–skip connections and depth-wise concatenations are employed for leveraging the useful former feature maps, which can be lost when going through the model's extraction backbone flow. As shown in Fig. 4, our design reuses the former feature maps twice. Indeed, the output features from the first normalization (Norm) layer are concatenated with the backbone features at the first depth-concatenation (Concat) layer, and then they are reused once more at the second Concat layer. This process is continuously repeated to the last Concat layer. The reuse of former features helps enhance the representative information of each output class and prevents gradient vanishing and overfitting problems during the training process. Therefore, this design can improve the classification accuracy of the model. In addition, maxpool layers with (2,2) strides were employed to reduce the number of learnable parameters in the model. All the convolutional (Conv) layers in the backbone flow were designed with 64 filters with a 3×3 size and a (1,1) stride. The Conv layers in skip connection flows have 64 filters with a 1×1 size and a (2,2) stride. Maxpool layers were allocated for selecting the strong features and downsampling the feature

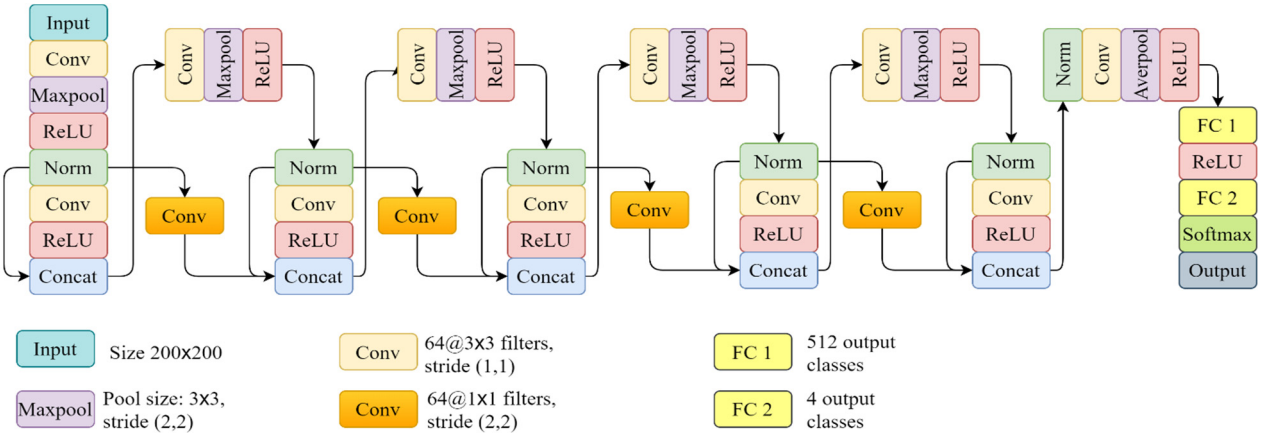

Fig. 4. Overall architecture of Dop-DenseNet.

maps. The last Conv layer is followed by an average pool layer instead of a maxpool layer. Then, two fully connected (FC) layers were deployed as classifiers. The FC 1 layer was set with 512 neurons whereas the FC 2 layer was assigned four output neurons corresponding to four classes of hand gestures (click, pinch swipe, and wave).

3. Experimental Results

The dataset of micro-Doppler images was randomly divided into the training set (80%) and the validation/test set (20%). The Dop-DenseNet model was experimented with on the dataset with the same training and test configuration. Specifically, training options were set up, as presented in Table 2. The training and testing processes were executed on a laptop with a Core-i5 9300H, RAM 16 GB CPU and a GTX 1660ti 6 GB GPU.

The learning/convergence progress was plotted in Fig. 5, where it can be observed that the cross-validation accuracies are appropriate for the training ones and approach ≈ 100% after 2,350 iterations.

Table 2. Main parameters of the FMCW radar for dataset generation

Parameter	Value
MiniBatchSize	64
MaxEpochs	30
InitialLearnRate	0.01
LearnRateDropFactor	0.5
LearnRateDropPeriod	2
Solving method	SGDM

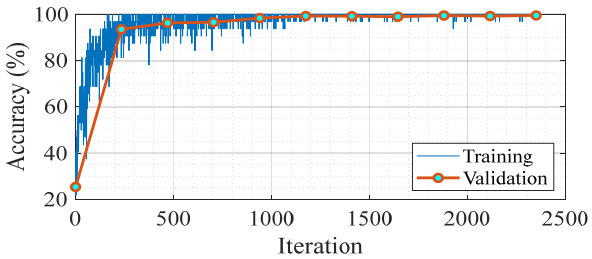

Fig. 5. Training and validation accuracies of Dop-DenseNet.

The confusion matrix that presents the classification accuracy of our proposed model for four hand gestures is shown in Fig. 6, where it can be seen that the Dop-DenseNet model gained a very high rate of correct hand gesture classification. Specifically, the model predicted the swipe and wave gestures with 100% accuracy. Two other methods also achieved a high correct classification rate, with 99.7% and 99.9% for the click and pinch gestures, respectively.

Next, we analyzed the accuracy and prediction time of the Dop-DenseNet model by changing the number of filter channels in the Conv layers to 8, 16, 32, and 64 channels. The results comparison of the different numbers of filter channels in Fig. 7 shows that the Dop-DenseNet model can improve classification accuracy by increasing the number of filters, but this will make the model bigger and slower. Specifically, the model's performance improved by about 1.1% when the number of filter channels increased from 8 to 16. Nevertheless, it enlarged the model capacity to about 50,300 learnable parameters and raised

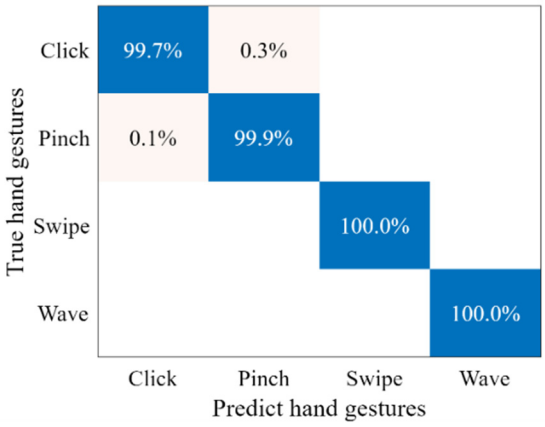

Fig. 6. Confusion matrix of hand gesture recognition using the Dop-DenseNet model.

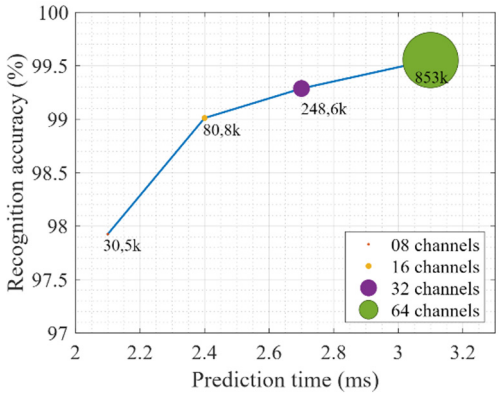

Fig. 7. Performance comparison of Dop-DenseNet when configured with different numbers of filter channels.

the prediction time delay from 2.1 ms to 2.4 ms. When we changed the number of channels in the model from 16 to 64, the classification accuracy improved by only about 0.5%. However, the model structure significantly increased by about 772,000 learnable parameters, and its processing rate was reduced to 3.1 ms.

To determine the impact of time length on the gesture recognition performance of Dop-DenseNet, we assessed the model with different time lengths (100, 200, 300, and 400 time bins). The results in Fig. 8 show that the longer time length makes the model processing slower despite improving the gesture recognition accuracy. Specifically, it can be seen in Fig. 8 that the model's accuracy was significantly improved when the time length was changed from 100 bins to 200 bins. However, when the time length was changed from 200 bins to 300 and 400 bins, the model obtained only a small improvement in accuracy. In addition, the execution time of the model was slower with time-length increments.

IV. Comparison of Implementation Results

In this section, we compare the performance of the Dop-DenseNet model with those of several state-of-the-art CNN models (GoogLeNet, DarkNet, ResNet-50, NasNet-Mobile, and MobileNet-V2) on the Dop-Net dataset. In the following subsections, we briefly describe the structures of the CNN models. Afterwards, we present the results of each model's training with 80% of the dataset, and of the model's validation and testing with the remaining 20% of the dataset. Finally, we discuss the results of the performance comparison.

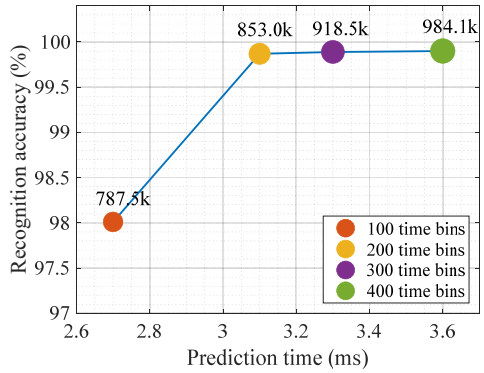

Fig. 8. Performance comparison of Dop-DenseNet with different time lengths of input data.

1. Brief Description of CNN Models

1.1 GoogLeNet

GoogLeNet [15] is a CNN with 22 layers. It was built to classify images into 1,000 object categories, such as keyboard, mouse, pencil, and many different animals. The GoogLeNet model is based on the Inception architecture. It uses Inception modules, which allow the network to choose between multiple convolutional filter sizes in each block. An Inception network stacks these modules on top of each other, with occasional max-pool layers with a stride of 2 to halve the resolution of the grid. The network has learned different feature representations for a wide range of images with a 224 × 224 size. However, on the Dop-Net dataset, the GoogleNet input size was reassigned to the 200 × 200 size, with "zero center" normalization. The output classes of the fully connected layer were changed from 1,000 to 4, which is appropriate for the number of gestures in the Dop-Net dataset.

1.2 ResNet-50

ResNet-50 [16] is a variant of the ResNet model, which has 48 convolution layers, along with 1 maxpool layer and 1 average pool layer. Therefore, it is 50 layers deep. As a result, it has 3.8 × 109 floating point operations. This architecture is used for computer vision tasks such as image classification, object detection, and localization. It can also be applied to other tasks to give them the benefit of depth and to reduce the computational expenses. The original ResNet-50 was trained on more than a million images from the ImageNet database to classify images into 1,000 object categories, similar to GoogLeNet. As a result, the network has learned rich feature representations for a wide range of images. The network has an input size of 224 × 224; however, in this work, we redesigned the 200 × 200 input size and the output size of four classes.

1.3 NasNet-Mobile [17]

NasNet stands for Neural Architecture Search Network, which was developed by the Google Brain team. The NasNet model employs the two main functionalities of normal and reduction cells. The idea of NasNet is to search for the best combination of parameters of the given search space: filter sizes, output channels, strides, number of layers, and others. As a result, NasNet

achieved state-of-the-art results in the ImageNet competition. However, it requires computation power. In this work, we applied NasNet-Mobile, a small version of NasNet, to classify hand gestures on the basis of the micro-Doppler feature maps analyzed by the FMCW radar. Similar to other networks, we redesigned the input and output sizes of NasNet-Mobile to make it appropriate for the Dop-Net dataset.

1.4 MobileNet-V2

MobileNet-V2 [18] is the second version of the MobileNet-V1 model developed by Google. The MobileNet-V2 model is a better module as an inverted residual structure, which plays the role of a backbone for feature extraction, was introduced to it. Hence, MobileNetV2 achieves state-of-the-art performance in object detection and semantic segmentation. In MobileNet-V2, two types of blocks are used: a residual block with a stride of 1 and another one with a stride of 2 for downsizing. Each block has three layers: the first layer is a 1×1 convolution with a ReLU activation function; the second layer is a depth-wise convolution; and the third layer is another 1×1 convolution but without any nonlinearity.

2. Comparison of Results

The performances of the Dop-DenseNet model (with a configuration of 64 3×3 filters and an input data time length of 200 time bins) and of four other CNN models are compared in Table 3, where it can be seen that our proposed model has the least number of learnable parameters (≈0.85 millions). It remarkably outperforms the other models in terms of accuracy (overall, 99.87%) and execution time (≈3.1 ms). MobileNet-V2 achieved the second-highest performance, with 99.38% classification accuracy, 4.0 ms prediction time, and ≈2.22 million learnable parameters. The NasNet-Mobile model with ≈4.3 million learnable parameters showed the worst average accuracy (98.77%) and prediction time (5.4 ms). Interestingly, ResNet-50 has a large architecture of ≈23.68 million parameters, but its average classification accuracy (99.18%) is lower than that of MobileNet-V2. The comparison of all the five CNN models with machine learning techniques in [14] showed that the CNN models remarkably outperform the machine learning techniques.

V. Conclusion

This paper briefly presents the principles of the FMCW radar with the use of micro-Doppler signatures for hand gesture recognition. For use in recognizing hand gestures, we propose Dop-DenseNet, a lightweight CNN model with cross-dense connections and a skip connection. Our model achieved the highest overall accuracy (99.87%) and the fastest execution time (3.1 ms) when trained and tested on the Dop-Net dataset with a time length of 200 bins. We then compared the performance of the Dop-DenseNet model with 64 3×3 filters with those of four well-known CNN models: GoogLeNet, ResNet-50, NasNet-Mobile, and MobileNet-V2 in terms of classification rate, prediction time, and number of learnable parameters. All the five models were trained and tested on the Dop-Net dataset, with the same training and testing option configuration. The Dop-DenseNet model remarkably outperformed the four other models in terms of classification accuracy, prediction time, and structural size.

References

[1] J. Kobylarz, J. J. Bird, D. R. Faria, E. P. Ribeiro, and A. Ekart, "Thumbs up, thumbs down: non-verbal human-robot interaction through real-time EMG classification via inductive and supervised transductive transfer learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 12, pp. 6021-6031, 2020.

[2] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193-4203, 2017.

[3] Q. Gao, Y. Chen, Z. Ju, and Y. Liang, "Dynamic hand gesture recognition based on 3d hand pose estimation for human-robot interaction," *IEEE Sensors Journal*, 2018. https://10.1109/JSEN.2021.3059685

[4] X. Tan and B. Triggs, "Enhanced local texture feature sets

Table 3. Performance comparison of Dop-DenseNet with different state-of-the-art CNN models

Model	Classification accuracy (%)					Time (ms)	Number of learnable parameters
	Average	Click	Pinch	Swipe	Wave		
GoogleNet	98.97	99.37	99.28	97.92	98.92	3.7	5,971,380
ResNet-50	99.18	98.73	98.56	100	100	6.6	23,684,100
NasNet-Moblie	98.77	98.10	99.28	97.92	100	5.4	4,271,232
MobileNet-V2	99.38	99.37	99.28	98.96	100	4.0	2,215,044
Dop-DenseNet	99.87	99.7	99.90	100	100	3.1	852,996

for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635-1650, 2010.

[5] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human-computer-interaction: a review," *Remote Sensing*, vol. 13, no. 3, article no. 527, 2021. https://doi.org/10.3390/rs13030527

[6] G. Malysa, D. Wang, L. Netsch, and M. Ali, "Hidden Markov model-based gesture recognition with FMCW radar," in *Proceedings of 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Washington, DC, 2016, pp. 1017-1021.

[7] Z. Peng, C. Li, J. Munoz-Ferreras, and R. Gomez-Garcia, "An FMCW radar sensor for human gesture recognition in the presence of multiple targets," in *Proceedings of 2017 First IEEE MTT-S International Microwave Bio Conference (IMBIOC)*, Gothenburg, Sweden, 2017, pp. 1-3.

[8] S. J. Ryu, J. S. Suh, S. H. Baek, S. Hong, and J. H. Kim, "Feature-based hand gesture recognition using an FMCW radar and its temporal feature analysis," *IEEE Sensors Journal*, vol. 18, no. 18, pp. 7593-7602, 2018.

[9] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3941-3952, 2018.

[10] Z. Fan, H. Zheng, and X. Feng, "A meta-learning-based approach for hand gesture recognition using FMCW radar," in *Proceedings of 2020 International Conference on Wireless Communications and Signal Processing (WCSP),* Nanjing, China, 2020, pp. 522-527.

[11] H. R. Lee, J. Park, and Y. J. Suh, "Improving classification accuracy of hand gesture recognition based on 60 GHz FMCW radar with deep learning domain adaptation," *Electronics*, vol. 9, no. 12, article no. 2140, 2020. https://doi.org/10.3390/electronics9122140

[12] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz, and N. Pohl, "Real-time radar-based gesture detection and recognition built in an edge-computing platform," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10706-10716, 2020.

[13] M. Chmurski and M. Zubert, "Novel radar-based gesture recognition system using optimized CNN-LSTM deep neural network for low-power microcomputer platform," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART)*, Virtual Event, 2021, pp. 882-890.

[14] M. Ritchie, R. Capraru, and F. Fioranelli, "Dop-NET: a micro-Doppler radar data challenge," *Electronics Letters*, vol. 56, no. 11, pp. 568-570, 2020.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1409.4842.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, 2016, pp. 770-778.

[17] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 8697-8710.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 4510-4520.

Hai Le

received his M.Sc. degree in electronic engineering from Le Quy Don Technical University in 2012. He is currently pursuing his Ph.D. in electronic engineering at the Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam. His research interests include radar systems, signal processing, and deep learning techniques.

Van-Phuc Hoang

received his Ph.D. in electronic engineering from the University of Electro-Communications (UEC), Tokyo, Japan, in 2012. He was a postdoc researcher and visiting scholar at UEC, Telecom Paris, France, and the University of Strathclyde, Glasgow, UK in 2012–2018. He is currently an associate professor and director at the Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam. His research interests include digital circuits and systems, hardware security, embedded systems for the Internet of Things, and intelligent system integration with deep learning techniques.

Van Sang Doan

received his M.Sc. and Ph.D. degrees in electronic systems and devices from the Faculty of Military Technology, University of Defence, Brno, Czech Republic, in 2013 and 2016, respectively. He was awarded an Honors degree by the Faculty of Military Technology of the University of Defence three times, in 2011, 2013, and 2016. From 2019 to 2020, he was a postdoctoral research fellow at the ICT Convergence Research Center, Kumoh National Institute of Technology, South Korea. He is currently a lecturer at the Faculty of Communication and Radar, Naval Academy in Nha Trang City, Vietnam. His current research interests include radar and sonar systems, signal processing, and deep learning.

Dai Phong Le

received his Ph.D. in radar and navigation from Saint Petersburg Electrotechnical University, Russia, in 2012. He is currently the head of the Research Group on Radar and Microelectronics at the Institute of System Integration, Le Quy Don Technical University, Hanoi, Vietnam. His research interests include radar and sonar systems, signal processing, and RF integrated-circuit design.