Hand Gesture Recognition with Ensemble Time-Frequency Signatures Using Enhanced Deep Convolutional Neural Network

Xiang Feng*, Qun Song†, Qingfang Guo*, Duo Liu†, Zhanfeng Zhao†, Yinan Zhao†
* Weifang Medical College, Weifang, China
E-mail: fengxiang230316@163.com Tel/Fax: +86-18266634558
† Harbin Institute of Technology, Harbin, China
E-mail: Kristoffer.Song@outlook.com Tel/Fax: +86-17862703207

Abstract—Hand gesture recognition using radar has been widely applied to control electronic appliances, military appliances and so on. In this paper, we investigate the feasibility of recognizing hand gestures using fused multiple time-frequency signatures, which ensembles micro-Doppler signatures, range-time signatures and angle-time signatures on spectrograms, with an Enhanced Deep Convolutional Neural Network (EDCNN). Several typical gestures included Tick, Double pushing, Rotating clockwise, and Rotating counterclockwise, were measured using Mm-wave radar and their spectrograms investigated. Therein EDCNN was employed to classify the spectrograms, with 80% of the data utilized for training and the remaining 20% for validation. Simulation said that the classification accuracy of the proposed method was found to be 96.2%.

I. Introduction

Hand gesture recognition has been widely used in computer gaming, electronic device control, and military defense. Typically, optical sensors are the common tools to achieve gestures recognition with a high resolution [1]–[3]. However, this kind of sensors is also limited to light condition, and once sheltered from other things. Unlike optical applications, radar is not restricted by lighting condition and shelters, and also with through-object capability [4].

Recently, hand gestures recognition using radar has been a hot-topic [5]–[7]. In [5], authors utilize pulse radar and the frequency-modulated continuous-wave (FMCW) radar to measure the range of fingers, so as to track their motions. However, only considering range information and ignoring velocity would incur malfunction for continuous motions. As we known, Doppler sensors would result in a simple, cost effective, and easy approach to capturing radial velocity response [6]. In [7], authors apply the Doppler radar for hand gesture recognition, and only extract the micro-Doppler signatures and borrow convolutional neural network to make recognition. Using only micro-Doppler signatures based on FrFT, i.e., no range information, to investigate the feasibility of gestures recognizing, might lose its way for variable-motions of fingers and wrist, compared with non-rigid motion. Detailed to say, micro-Doppler signatures are represented as overlapped signatures in the joint time-frequency domain when several scatters exist, such as fingers [7] [8]. As a result, they have to carefully investigate and distinguish signatures associated with gestures. For multiple-gestures recognition, i.e., several gestures with different range and same velocity, and only using micro-Doppler cannot work well. DCNN, which is inspired by the human visual cortex, is one of the most successful deep learning algorithms [9] [10]. Therein, by training the convolutional filter and fully connected multi-layer perceptrons, DCNN simultaneously extracts and classifies some typical features. And also DCNN does not require an extraction process of handcrafted feature, while classical DCNN might not exploit the effective signatures, for its poor generalization when convolutional layer and pooling layers increasing. Novel structure will be in need for this case.

In this paper, we investigate the feasibility of recognizing hand gestures using fused multiple time-frequency signatures. Four hand gestures were investigated. As the spectrograms of these gestures only have subtle differences. Instead of the conventional supervised learning approach, the enhanced deep convolutional neural network (DCNN) is employed. Firstly, the multi-layer perceptron is used to replace the traditional linear convolution kernel to extract the special features. Then, the inception model is cascaded behind the convolution layer. At the same time, the pyramid sampling mechanism is introduced into the pooling layer to replace the conventional random sampling and maximum sampling. The pyramid multi-scale fusion strategy is used to splice features of different dimensions, and then transmitted to the fully connected layer. DCNN simultaneously extracts and classifies important features. DCNN does not require a handcrafted feature extraction process. Simulation said that the classification accuracy of the proposed method was found to be 96.2%.

II. EXPERIMENTAL SETUP AND MEASUREMENT PREPROCESSING

In this paper, we use our mm-wave radar (TI company IWAR 1441) to obtain the receiving data, which operates at 77 GHz and makes it suitable for detecting hand ges-

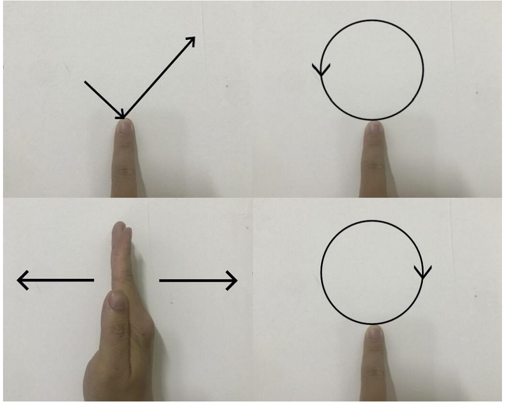

Fig. 1. Four different hand gestures.

tures. We fixed the radar to a table and executed the hand motions in the mainlobe of radar antenna. The average distance from the radar to the hands was approximately 20 cm. Four hand gestures employed in this study were (a) Tick, (b) Double pushing, (c) Rotating clockwise, and (d) Rotating counterclockwise. The employed gestures are depicted in Fig. 1.

Meanwhile, we mainly use the empirical mode decomposition (EMD) algorithm to refine the time-frequency signature [11] [12], where the spectrograms of finger motions were observed by using short-time fast Fourier transform (FFT). We ensemble the range-variation vs. time, velocity-variation vs. time, angle-variation vs. time and micro-Doppler signatures together to formulate the fused ensemble signature, as shown in Fig. 2. And each gesture was measured 50 times. As spectrum gestures of Tick, Double pushing, Rotating clockwise, and Rotating counterclockwise are almost similar because their radial velocities are analogous, even with some diverse motion directions. However, the small variation in hand movements has some peculiar features that can be observed and distinguished.

III. Novel Enhanced Deep Convolutional Neural Network Formulation

To investigate micro-Doppler signatures, spectrograms of finger motions were observed through short-time fast Fourier transform (FFT) with size of 256. We mainly use Empirical Mode Decomposition (EMD) to prepossess the ensemble signatures, and introduce our novel EDCNN neural networks to extract some typical features and make classification.

Our EDCNN primarily consists of several convolutional filters and pooling layers. The combination of convolution filters, activation function, and pooling operations constitutes the multiple perception layer and inception module, named as Net in Net model (NIN), shown in Fig. 3 and Fig. 4. Detailed to say, the convolutional filter extracts the features of a spectrogram through its convolution process, the coefficients of the convolutional filter are trained by a given dataset. In sub-model embedded-net Conv1, there are 2 multiple perception layers, with kernel size/stride 3×3/2 and 1×1/2, respectively. Moreover, there are four sub-Inception models with kernel size/stride 1×1/1, kernel size/stride 1×1/1-3×3/1, kernel size/stride 1×1/1-5×5/1, and kernel size/stride 3×3/1-1×1/1, respectively. The activation function in each convolutional filter is highly nonlinear such that it can describe the nonlinear relationship between inputs and outputs.

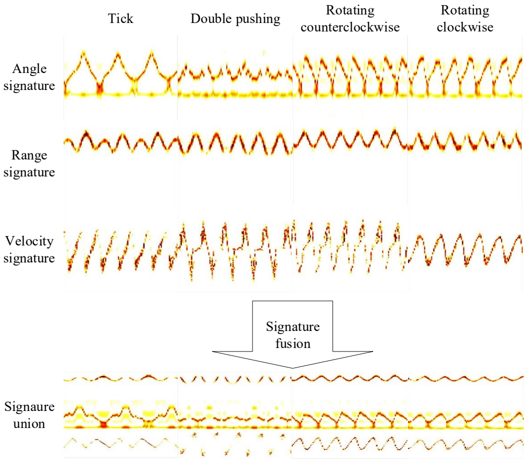

Fig. 2. Fused ensemble signature formulation.

Typically, Spatial Pyramid Pooling (SPP) is used for data dimension reduction, and also signature fusion in different dimension, as shown in Fig. 5. These three pooling layers enable the final output to be more robust to noise, by selecting a maximum value or a mean value with kernel size/stride 2×2/2, 3×3/3, 5×5/4, respectively. Note that, the dropout operation of these nodes is used to prevent overfitting in a regularization scheme. This enables EDCNN to prevent co-adaption among its nodes. The coefficients of convolution filters and weights of the final fully connected layers are trained. We also use the back propagation algorithm with a stochastic gradient descent (SGD) as a training algorithm. In other words, the trained convolutional filters work as a feature extractor, and the fully connected perceptron functions as a classifier. Finally, Fig. 4 shows the architecture of EDCNN.

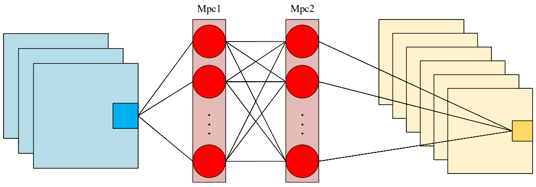

Fig. 3. The diagram of embedded-net convolution layer.

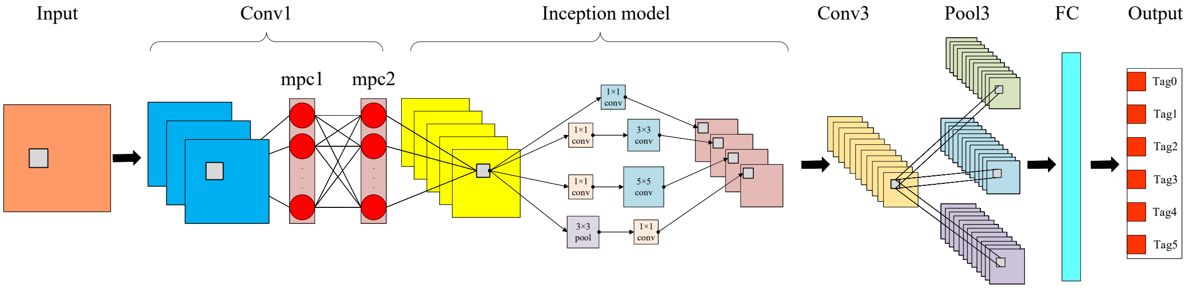

Fig. 4. The architecture of EDCNN.

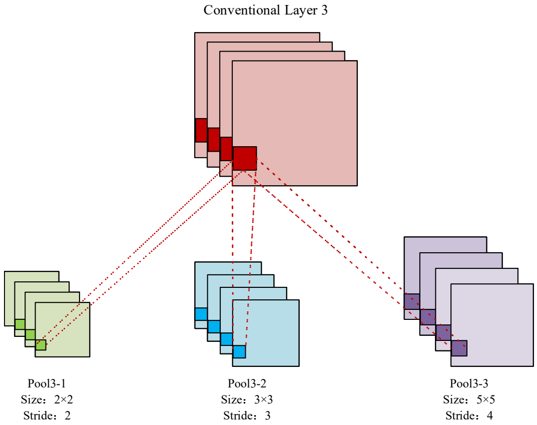

Fig. 5. The diagram of Spatial Pyramid Pooling.

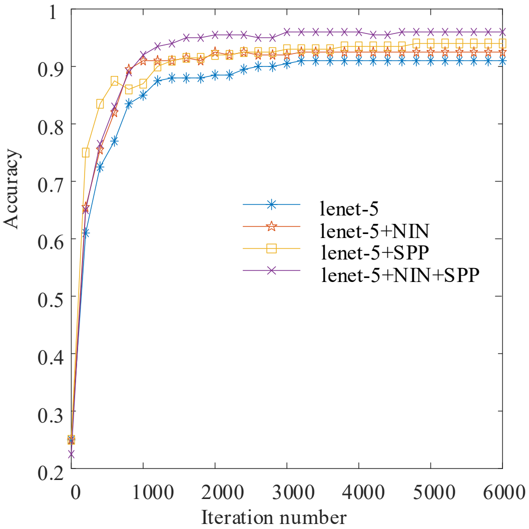

Fig. 6. Accuracy comparison of different models for united signatures.

IV. SIMULATION AND ANALYSIS

As we known, four signatures, i.e., velocity-time signature, range-time signature, angle-time signature and united signature, have been preprocessed using EMD. In this section, we use LeNet-5, LeNet-5+SPP, LeNet-5+NIN and LeNet-5+SPP+NIN to extract typical signatures of different gesture, and then make comparison. Among the 400 pieces of data measured from the single participant, 80% was used as training data and 20% as testing data. Each spectrogram was resized to 60-by-60 and the values normalized from zero to one. We used 5-fold validation to obtain valid accuracy by dividing the measured data into different training datasets and test datasets. Each DCNN structure with 6000 iterations and learning rate 0.0005 have been set. Our computer environment is i7-4510U CPU@2.00GHz 2.60GHz Caffe+Matlab, The final accuracy comparison of different models vs. signatures has been listed in Table I and Fig. 6.

In Table I, we can find that, the novel LeNet-5+SPP+NIN has achieved the best accuracy no matter in angle signature, range signature, velocity signature and united signature. In Fig. 6, as the number of iterations increases, the recognition capability of each network gradually increases, but LeNet-5+SPP+NIN network also has always been in a leading position, while the traditional LeNet-5 may lose it way. This is because the embedded multi-layer perceptrons and Inception structure and pyramid pooling method together make the feature mining network equip with multi-scale deep feature extraction and fusion capability, and also get rid of drawbacks of traditional LeNet-5. Typically, the Inception structure and pyramid pooling method makes the novel network reduce the over-fitting and be more robust.

To further evaluate the feature learning ability of the embedded network, different hyperparametric learning rates are set to study the recognition accuracy in 6000 iterations (Fig. 7), and the learning rate is lr=0.001, 0.0008, 0.0005 and 0.00005, 0.00001. It can be seen from Fig. 7 that when the learning rate is high, it is difficult for the recognition network to reach the global optimal solution after training for a certain number of times, but may fall into the local optimal solution and cannot continue to improve its accuracy; when the learning rate is low (lr=0.0003) The model accuracy rate rises slowly, which increases the time for identifying network training. Only when the learning rate is moderate (such as lr=0.0005), the proposed model can achieve the highest accuracy in a relatively short time.

TABLE I
The final accuracy comparison for different network models

Model	Angle Signature	Range Signature	Velocity Signature	United Signature
LeNet-5	82.5%	80.3%	80.1%	91.3%
LeNet-5+SPP	90.5%	82.5%	82.4%	94.1%
LeNet-5+NIN	85.5%	82.6%	82.5%	92.6%
LeNet-5+SPP+NIN	92.7%	90.0%	87.5%	96.2%

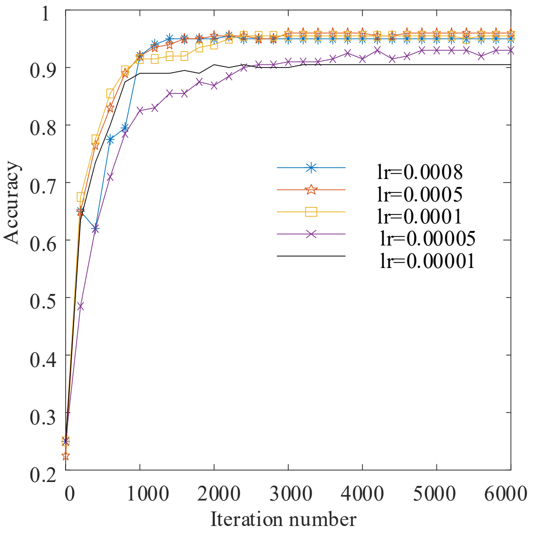

Fig. 7. Accuracy comparison of Lenet-5+NIN+SPP under different learning rates.

V. Conclusions

The conclusion goes here.In this study, we investigated EDCNN to classify human hand gestures. Four hand gestures were measured using Doppler radar and their spectrograms analyzed. The classification accuracy of proposed method was found to be 96.2%, and obvious better than other networks. However, it should be noted that micro-Doppler signatures can vary depending on aspect angle and distance to the radar, as shown in our experiments, and the novel united signatures can conquer this in some sense. In future work, we plan to measure many gestures from various human subjects to train a EDCNN for general-purpose hand gesture recognition.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61771160 and Natural Science Foundation of Shandong Province under Grants ZR2019BF037.

References

[1] G. R. S. Murthy and R. S. Jadon, “A review of vision based hand gestures recognition, ” International Journal of Information Technology and Knowledge Management, vol. 2, no.2, pp. 405–410, 2009

[2] S. Mitra and T. Acharya, “Gesture recognition: A survey,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no.3, pp. 311–324, 2007

[3] A. Boyali and N. Hashimoto, “Spectral Collaborative Representation based Classification for hand gestures recognition on electromyography signals,” Biomedical Signal Processing and Control, vol. 24, pp. 11–18, 2016.

[4] Y. Kim and B. Toomajian, “Hand Gesture Recognition Using Micro-Doppler Signatures With Convolutional Neural Network,” IEEE Access, vol. 4, pp. 7125–7130, 2016.

[5] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, “Multi-sensor system for drivers hand-gesture recognition,” 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8, 2015.

[6] P. Hugler, M. Geiger, and C. Waldschmidt, “RCS measurements of a human hand for radar-based gesture recognition at E-band,” 2016 German Microwave Conference (GeMiC), pp. 259–262, 2016.

[7] Y. Kim and H. Ling, “Human Activity Classification Based on Micro-Doppler Signatures Using a Support Vector Machine,” IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 5, pp. 1328–1337, 2009.

[8] D. P. Fairchild and R. M. Narayanan, “Classification of human motions using empirical mode decomposition of human micro-Doppler signatures,” IET Radar, Sonar & Navigation, vol. 8, no. 5, pp. 425–434, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[10] Y. LeCun, Y. Bengio, and G. Hinton,“Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.

[11] B. Dekker, S. Jacobs, A. Kossen, M. Kruithof, A. 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.

[12] Y. Wang, X. Wu, W. Li, Z. Li, Y. Zhang, and J. Zhou, “Analysis of micro-Doppler signatures of vibration targets using EMD and SPWVD,” Neurocomputing, vol. 171, pp. 48–56, 2016.