CNN-Based Millimeter-Wave Radar Signal Gesture Recognition

Xia Yichen
1023040903
Nanjing University of Posts and Telecommunications
School of Computer Science
Nanjing, China

***Abstract*—In the future, human-computer interaction is set to become inevitable. Although there are already various technologies for achieving human-computer interaction, such as Wi-Fi sensing, sound waves, ultrasonic sensing, and visible light sensing, they all have their respective drawbacks, significantly limiting the future development of human-computer interaction. This paper introduces a millimeter-wave gesture recognition solution that incorporates a customized neural network to describe and extract intrinsic gesture features. This allows for training with a limited dataset and the recognition of untrained gestures. The system was implemented on a personal computer, and its recognition performance was tested. The experiments demonstrate that the system exhibits a high level of recognition accuracy.**

***Index Terms*—gesture recognition, convolutional neural network.**

I. Introduction

In recent years, computer technology has permeated various aspects of our daily lives, making human-computer interaction inevitable. It is widely believed that computers and display technologies will continue to advance. The gateway allowing communication between humans and machines or computers is referred to as the human-machine interface. Keyboards, mice, and touch screen sensors represent traditional methods of human-computer interaction. However, these methods are becoming bottlenecks in developing user-friendly interfaces, and human gestures can serve as a more natural means to establish an interface between humans and computers. Short-range radar possesses the capability to detect subtle movements while maintaining high accuracy. Radar sensors have demonstrated potential in research areas such as posture estimation, vital sign detection, and gesture recognition. Gesture recognition based on millimeter-wave radar is a crucial aspect of natural human-computer interaction, complementing traditional methods such as mouse, keyboard, or touch-based interfaces. This technology promises users the ability to control household appliances, smartphones, smart speakers, or other IoT devices by making predefined gestures. To achieve this contactless gesture recognition, various techniques have been proposed using different mediums, such as sonar [1], [2] and Wi-Fi signals [3], [5] as sensors. However, these methods suffer from coarse granularity, limiting their ability to recognize subtle gestures. The latest solutions offer fine-grained perception but come with high costs and lack penetration capability. Millimeter-wave sensing presents ideal advantages, including high-range precision and the ability to penetrate walls, resulting in more convenient and intriguing use cases, such as when wearing gloves or dealing with oily fingers. Furthermore, millimeter-wave radio, as a prominent technology in emerging 5G networks, is poised to be ubiquitous, deployed on billions of mobile wireless devices. Therefore, millimeter-wave radar holds the potential to become an omnipresent sensing tool. Convolutional Neural Networks (CNNs) are a class of feedforward neural networks that include convolution operations and exhibit a deep structure, representing one of the prominent algorithms in deep learning. CNNs are specifically designed for supervised training and are oriented towards recognizing invariance in two-dimensional shapes. They function as a particular multilayer perceptron with the capability of learning representations, enabling translation-invariant classification of input information based on their hierarchical structure. Consequently, CNNs can be applied for extracting features from millimeter-wave radar signals in this paper and recognizing gestures. In this study, I devised a radar signal gesture recognition system based on Convolutional Neural Networks, allowing users to control applications through predefined gestures without physical contact. Experimental results demonstrate that this system achieves a high level of recognition accuracy.

II. Related Works

A. Millimeter-Wave Sensing

Millimeter-wave, as the primary frequency band of 5G radio, is an electromagnetic wave with wavelengths ranging from 1 to 10 millimeters. In addition to enabling multi-bandwidth ultra-high-speed wireless links, millimeter-wave can also be applied to various sensing tasks. Recently, researchers have employed millimeter-wave radio for environmental mapping [6], [7], material identification [8], vital sign monitoring [9], human activity recognition [10], user identification [11], and gait recognition [12]. Moreover, to achieve more engaging human-computer interactions, they have explored how to perceive human gestures, such as finger tracking.

B. Other Non-contact Sensing Methods

Before the advent of millimeter-wave radio, researchers utilized optical signals, Wi-Fi signals, and other methods to

achieve various human-computer interactions. For instance, visible light sensing can recognize gestures with fairly high accuracy [16]–[18]. However, it cannot perceive gestures in dark or non-line-of-sight scenarios and carries the risk of privacy leakage. There are also researchers utilizing infrared three-dimensional structured light sensing. Due to the shorter wavelength of infrared compared to millimeter waves, infrared light sensing can accurately simulate hand movements, enabling gesture recognition [19]. However, it has drawbacks such as high energy consumption, susceptibility to interference, and the risk of privacy leakage. In addition, acoustic and ultrasonic sensing methods are also prevalent. Due to the relatively lower frequency of acoustic and ultrasonic waves, utilizing them for sensing purposes entails lower power consumption. Most studies leverage the Doppler effect to detect gestures [2] and track fingers [1]. For instance, Air gesture, a new type of human-computer interaction introduced by Huawei, utilize ultrasonic waves to detect and recognize user gestures, enabling intelligent interaction and providing users with a novel experience. Currently, it supports five types of air gestures: upward wave, downward wave, leftward wave, rightward wave, and press operation. However, acoustic and ultrasonic methods are susceptible to environmental noise, resulting in lower recognition accuracy compared to millimeter-wave techniques. Before the advent of millimeter-wave radio, researchers primarily relied on Wi-Fi signals and commercial devices to implement various natural human-machine interactions. Initially, much research focused on monitoring human body postures [4], [5]. In recent years, existing studies on gesture recognition can be classified into two categories: a) recognizing unique gestures from different individuals [14], and b) recognizing the same set of gestures [3]. However, most of them cannot detect gestures with small motion amplitudes; they can only sense large and straightforward movements of the entire hand, such as pushing a hand along a straight line. Only WiFinger [15] is capable of performing this task, but it is overly sensitive to changes in the surrounding environment, which has a certain impact on practical applications.

III. Problem Statement

The principle of millimeter-wave gesture sensing is to capture radar signals reflected from human arm gestures, analyze the changes caused by the reflections, and thereby capture the variation patterns of different gestures. Millimeter-wave radar emits frequency-modulated continuous-wave signals, a mature modulation method for obtaining dynamic information from reflection points. The radar has a fine distance resolution of 4 centimeters, allowing it to treat arm gestures as a dynamic set of reflected points, referred to as a "point cloud," after signal processing. Each point contains its dynamic information, including distance, Doppler, and position. Therefore, gestures can be described by a large number of such point clouds. Given the substantial information and multiple implicit relationships in the point cloud data for each gesture, most studies prioritize sorting the scattered information by extracting multiple features. However, the change patterns of gestures remain not sufficiently distinct in the extracted features. Therefore, this paper adopts a convolutional neural network to extract features and perform gesture recognition. The paper predefines four gestures, namely, pressing a button with fingers, flipping the entire hand, moving the hand forward and backward, and moving the hand left and right.

IV. Solutions

A. Keras Network.

Keras is an advanced deep learning application programming interface that provides a simple and intuitive interface for building, training, and deploying neural networks. Developed by François Chollet in 2015, Keras is dedicated to user-friendliness, modularity, and extensibility. Keras is designed to allow users to quickly implement ideas without delving into the intricacies of low-level details. Its concise and intuitive API makes neural network construction easy. It has underlying support for various deep learning frameworks, including TensorFlow, Theano, and Microsoft Cognitive Toolkit. The Keras network has the following advantages: 1)Allows for simple and rapid prototyping, emphasizing user-friendliness, modularity, and scalability. 2)Supports both convolutional networks and recurrent networks, as well as combinations of the two. 3) Seamlessly runs on both CPU and GPU. The number of users using Keras is also very high, second only to TensorFlow.

B. Overview of neural networks.

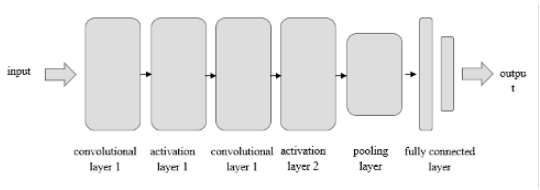

Fig. 1. Neural Network Structure

Millimeter-wave gestures contain only a few dozen elements, which is different from images with millions of pixels. Therefore, a simple convolutional neural network shown in Figure 1 was designed. Firstly, two convolutional layers are used, each followed by an activation layer to extract features from millimeter-wave gestures. Due to the sparse nature of the point cloud, a max-pooling layer is subsequently added to select features crucial for classification. Finally, two fully connected layers are employed to analyze the classification and output its computed scores. The loss function for this neural network is

$$\begin{aligned} L(Y,c) &= -\log\left(\frac{\exp(Y[c])}{\sum_i \exp(Y[i])}\right) \\ &= -Y[c] + \log\left(\sum_i \exp(Y[i])\right) \end{aligned} \tag{1}$$

Where Y is the computed score, c is the index of the accurately recognized gesture class, and i is the index of other gesture classes. The loss function commonly used for classification tasks in deep learning is cross-entropy. However,

there are two different activation functions that can be used for classification: sigmoid and softmax. Since the former can only provide the probability of whether the input belongs to a particular label, it is used for binary classification or multi-label classification. The latter, on the other hand, is used for multi-class classification to determine if the input belongs to one of multiple labels. In this case, the latter is more suitable for the multi-class classification task in this paper. Therefore, the paper adopts the softmax activation function with cross-entropy as the loss function.

C. Network Input

Unlike images with fixed shapes, millimeter-wave gesture samples consist of variable reflection points. Therefore, this paper transforms each reflection point into a consistent shape as the input to the network. Due to the instability across different frames, this study explores the variation patterns of each gesture by extracting Doppler frame contours. To avoid interference from positional information, the analysis focuses solely on the Doppler and reflection intensity of each point. Additionally, considering temporal information, this paper treats it as the horizontal axis. In other words, this input illustrates how the gesture is distributed over time across different Doppler frames, reflecting the correlations between different frames. Specifically, this paper sums the energy for points in the point cloud with the same Doppler and index.

$$\varphi = \left[\sum \epsilon_{d,j}\right]_{I_d \times I_j} \tag{2}$$

Where φ represents the input to the neural network, ϵ denotes the reflection intensity of each point in the point cloud, d is the Doppler of each point in the point cloud, j represents the time frame, I_d is the total number of points with the same Doppler in the point cloud, and I_j is the duration of frames during the gesture, with I_j set to 30 in this study.

D. Concrete Realization

The neural network in this paper takes a single-channel matrix φ as input, with a size of 90×90, and its output is Y from the second fully connected layer. The input of Convolutional Layer 1 is a 90×90 color RGB image, with an output channel of 32, a kernel size of 3×3, a stride of 1, and preserving the boundaries of the convolution results. The max-pooling layer has a kernel size of 2×2, a stride of 1, and a padding of 1. The output channel of Convolutional Layer 2 is 32, with a kernel size of 3×3, a stride of 1, and not preserving the boundaries of the convolution results. Two activation layers follow the respective convolutional layers. Therefore, the output size of the fully connected layer is 4, which corresponds to the categories of gestures to be classified. Training stops after 50 iterations. The neural network is optimized using the Adam optimizer in this paper. The implementation of this neural network is done using the Keras framework.

V. Experiment

There are a total of 160 images from web resources which are combined into a data set. Firstly, in the `dataset.py` file, the original images are converted to RGB format, then cropped to a specific size, and finally transformed into h5 files to serve as the dataset for training the neural network.

A. Confusion Matrix

The confusion matrix is a situation analysis table used in machine learning to summarize the predictions of a classification model. It consolidates the records in the dataset based on two criteria: the actual class and the predicted class by the classification model. In the matrix, rows represent the true values, and columns represent the predicted values. In this matrix, the element (i, j) indicates the number of times the i-th gesture was correctly identified as the j-th gesture, as detailed in Figures 2 and 3.

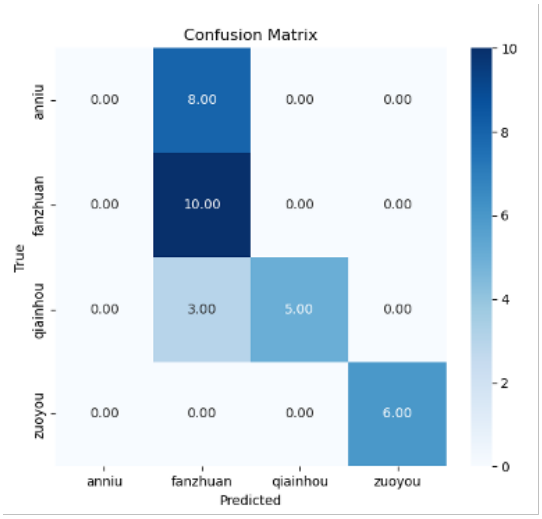

Fig. 2. Confusion Matrix with 30 Iterations

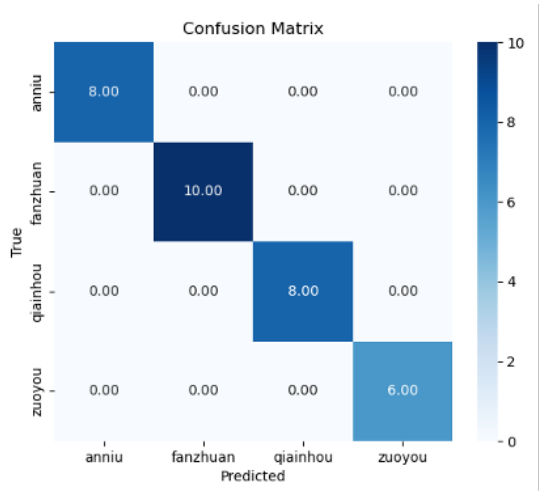

Fig. 3. Confusion Matrix with 50 Iterations

From the above two figures, we can observe the following two points:

1) After 30 iterations, the neural network model shows a relatively high error rate of 37.5% in recognizing arm movements in the test set. However, after 50 iterations,

all 32 actions in the test set are accurately identified, indicating a higher recognition accuracy of the neural network model.

2) When recognizing human gestures, the neural network model sometimes misclassifies arm movements as flipping actions and is more likely to make completely incorrect predictions for finger-button operations. This may be due to the lack of spatial positional information in the neural network input, as it flattens three-dimensional gestures into one dimension, operating only on a range scale and ignoring the spatial and physical significance of the input images.

B. The Impact of Iteration Count on Accuracy

Due to the impact of the training iteration count on the final accuracy of the neural network model, relevant experiments were conducted. The influence of training iteration count on model accuracy of the training set is shown in Figure 4.

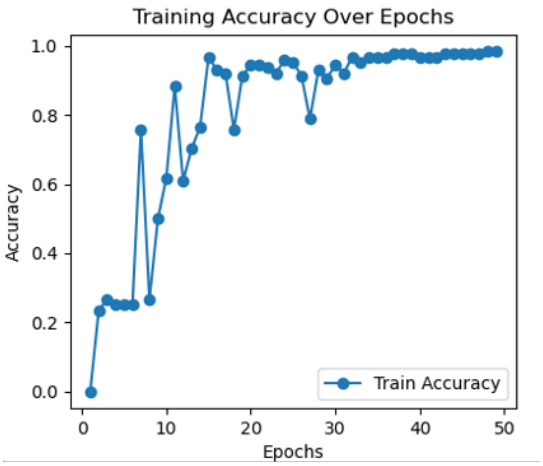

Fig. 4. The Impact of Iteration Count on Recognition Accuracy

From Figure 4, it can be observed that when the training iteration count is less than 30, the recognition accuracy generally increases with the increase in iteration count. However, as the iteration count continues to increase, the change in recognition accuracy is not significant, gradually approaching 98.4%.

VI. Summary

This paper designs, implements, and evaluates a convolutional neural network for gesture recognition based on millimeter-wave sensing. The network allows users to directly control devices through individual gestures. To achieve this capability, a customized convolutional neural network is proposed in this paper to learn specific rules of variation from gesture feature relationships.

References

[1] W. Mao, J. He, and L. Qiu, "CAT: High-precision acoustic motion tracking," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2016, pp. 69–81.

[2] H. Watanabe and T. Terada, "Improving ultrasound-based gesture recognition using a partially shielded single microphone," in *Proc. ACM Int. Symp. Wearable Comput. (ISWC)*, 2018, pp. 9–16.

[3] N. Yu, W. Wang, A. X. Liu, and L. Kong, "QGesture: Quantifying gesture distance and direction with WiFi signals," in *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol. (IWMUT)*, 2018, pp. 1–23.

[4] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "RF-based fall monitoring using convolutional neural networks," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–24, 2018.

[5] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3D tracking via body radio reflections," in *Proc. 11th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, 2014, pp. 317–329.

[6] T. Wei, A. Zhou, and X. Zhang, "Facilitating robust 60 GHz network deployment by sensing ambient reflectors," in *Proc. 14th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, 2018, pp. 213–226.

[7] A. Zhou, S. Yang, Y. Yang, Y. Fan, and H. Ma, "Autonomous environment mapping using commodity millimeter-wave network device," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2019, pp. 1126–1134.

[8] H.-S. Yeo, G. Flamich, P. Schrempf, D. Harris-Birtill, and A. Quigley, "RadarCat: Radar categorization for input & interaction," in *Proc. 29th Annu. Symp. User Interface Softw. Technol. (UIST)*, 2016, pp. 833–841.

[9] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Monitoring vital signs using millimeter wave," in *Proc. 17th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, 2016, pp. 211–220.

[10] A. D. Singh, S. S. Sandha, L. Garcia, and M. B. Srivastava, "RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proc. 3rd ACM Workshop Millimeter Wave Netw. Sens. Syst. (mmNets)*, 2019, pp. 51–56.

[11] H. Liu et al., "Accurate and robust user identification and authentication through hand-gesture sensing with mmWave radar," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, to be published.

[12] Z. Meng et al., "Gait recognition for co-existing multiple people using millimeter wave sensing," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 849–856.

[13] T. Wei and X. Zhang, "mTrack: High-precision passive tracking using millimeter wave radios," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, 2015, pp. 117–129.

[14] M. Shahzad and S. Zhang, "Augmenting user identification with WiFi based gesture recognition," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–27, 2018.

[15] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "WiFinger: talk to your smart devices with finger-grained gesture," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (Ubicomp)*, 2016, pp. 250–261.

[16] H. G. Doan, H. Vu, and T. H. Tran, "Recognition of hand gestures from cyclic hand movements using spatial-temporal features," in *Proc. 6th Int. Symp. Inf. Commun. Technol. (SoICT)*, 2015, pp. 260–267.

[17] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *Int. J. Comput. Vis.*, vol. 126, pp. 430–439, Apr. 2018.

[18] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2015, pp. 1–7.

[19] Y. Zhang, T. Gu, C. Luo, V. Kostakos, and A. Seneviratne, "FinDroidHR: Smartwatch gesture input with optical heartrate monitor," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp 1–42, 2018.