Human Activity Recognition with Millimeter-wave Radar Point Clouds Using Feature-fusion based Neural Networks

Zhou, Houpu

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
18
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2023-03-24
(URL)
https://doi.org/10.15002/00026287

Human Activity Recognition with Millimeter-wave Radar Point Clouds Using Feature-fusion based Neural Networks

Houpu Zhou
Graduate School of Computer and Information Sciences
Hosei University
houpu.zhou.5i@stu.hosei.ac.jp

Abstract*—This study presents a human activity recognition method based on millimeter-wave radar point clouds. A feature-fuse neural network was proposed to combine different features extracted from two subnet. First, millimeter-wave radar point cloud data is preprocessed into different formats, such as 3D voxels, 2D dual-views picture. Second, the voxels and views are fed into 3D and 2D neural networks to extract spatial and temporal information, respectively. The features extracted from the two networks are fused to enhance recognition accuracy. The research introduces innovative preprocessing techniques for both voxels and bi-views data, specifically dynamic denoising and Doppler information embedding, and demonstrated their practicality. The dataset used for the experiments contains 10 different human activity categories. Unlike other studies, it focuses on recognizing both abnormal and normal activities, particularly those that are similar to abnormal ones. This research aims to explore potential scenarios for home health monitoring and achieved a recognition accuracy of 95.30%, showing an improvement of 2% to 7.8% compared to the respective single networks. The feature fusion network model showed an accuracy rate of 95.40% during the testing for detecting abnormal activities. For challenging activity categories, the feature fusion network is able to distinguish between them more easily compared to a single network.

Keywords*—Millimeter wave radar, Point Cloud, Voxel, Bi-views, Feature Fusion, LSTM, CNN.

I. Introduction

The field of Human Activity Recognition (HAR) has gained importance in recent years and is widely applied in elderly and vulnerable population monitoring[1][2][3], virtual reality[4] and human-computer intelligent interaction [5][6]. Different types of devices, including wearables, cell phones [7], smartwatches [8], environmental sensors [9], and infrared cameras [10], LiDAR imaging devices [11] and Kister-wave radar [12][13][14] have been studied for their effectiveness in human activity recognition.

However, wearable devices have limitations for long-term home monitoring, while they are not suitable for long-term home monitoring because each user needs to use a specific device and be required worn at any time. Environmental sensors do not depend on users to own specific equipment, which has great potential in home monitoring based on activity recognition. Among these environmental sensors, visible cameras can only work in a lighting environment. There are blind areas in night monitoring and has privacy issues. Although infrared images can enable night monitoring and avoid user privacy issues, the accuracy of HAR system in complex indoor environments remains to be studied.

In this work, a new activity recognition model was proposed that uses both voxel and doppler information embedded dual-view pictures as inputs to the Feature Fusion Network. The main contributions of this research are:

a) Texas Instruments' IWR6843AOP device was used to collect data on human activities. A total of 10 human activities were collected from 12 participants.

b) Two preprocessing methods for point cloud data were proposed in this study.

c) A feature fusion neural network that integrates two data inputs was designed, achieving a precision of 95.30%. This proves the positive significance of multi-feature inputs for the human activity recognition problem using millimeter wave radar.

II. Related Work

In experiments to recognize human activity based on millimeter wave radar, researchers have used a variety of methods. For example, [13] proposes a framework called RadHAR, which converts point clouds into voxels as inputs to its neural network, and evaluates different classifiers on its point cloud MMActivity dataset. The experimental results show that the Time-distributed CNN + Bi-directional LSTM classifier has the best classification results. [14] A feature extraction framework is proposed, which consists of a key voxel region framework and two-view image construction. The proposed new method to learn each type of activity using a dual-view convolutional neural network (DVCNN).

At present, several challenges persist in the study of Human Activity Recognition (HAR) implementation based on millimeter wave radar.

DBSCAN is widely used as a noise reduction technique to distinguish human bodies from noise, but the parameter settings of DBSCAN vary greatly across different authors, as they are dependent on various factors such as the radar model, experimental environment, and type of data set. As a result, the traditional DBSCAN algorithm exhibits considerable variability in noise reduction performance.

Currently, many datasets lack well-defined application scenarios. while there are studies dedicated to fall monitoring, but no comprehensive examination of other potentially abnormal activities such as crawl and wave hand at a long time.

Supervisor: Prof. Runhe Huang

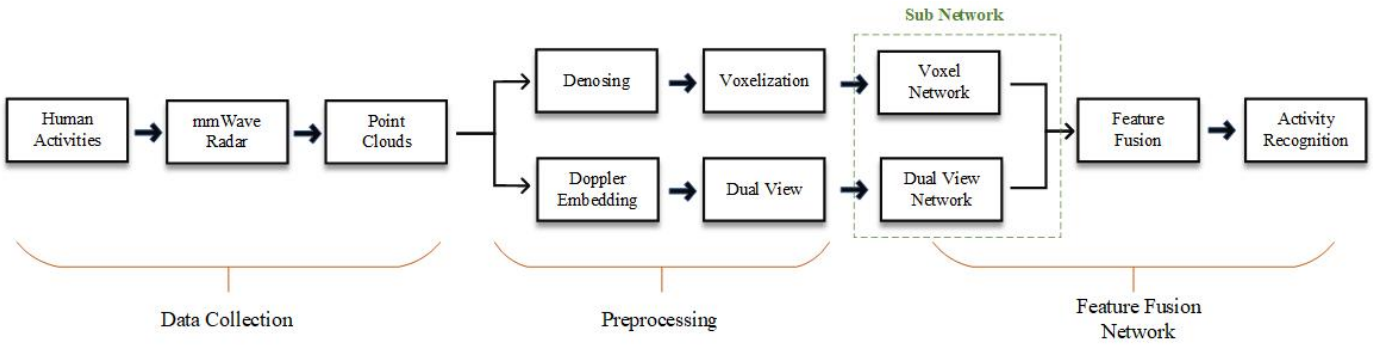

Fig. 1. Overview of mmWave Radar based HAR System

The point cloud complexity makes it challenging to implement real-time systems based on voxel research. A more efficient and low-consumption neural network is needed.

III. Proposed Mmwave Radar Based Har System

Figure 1 gives the overview of mmWave Radar Based HAR system. As it shows, the system consists of three main components: collection of point cloud data from millimeter wave radar, pre-processing of point cloud data, and fusion neural network that integrates two sub-networks. The following three sections will elaborate on the details of each components block.

In the function of radar point cloud data collection, the raw point cloud data is collected using a millimeter wave radar device, the specific radar parameters and experimental environment will be described in later section. The radar dataset collected is categorized into 10 classes. The specific number of samples and the number of frames for each sample are organized into a table.

In the function of point cloud data pre-processing, radar collected point cloud data will undergo different preprocessing to generate 3D and 2D data respectively, which will be input into two sub-networks.

In the function block of feature fusion network, it fuses the features obtained from two sub-networks, and inputs the new features into multi-layer for classification.

IV. Radar Point Cloud Data Collection

In this section, the data collection experimental environment was described, including the placement position and angle. Finally, the dataset collected by the millimeter wave radar was specifically introduced.

A. Radar Configurations and Experiment Setting

Radar Equipment Settings: the radar device used in this research is IWR6843AOP, and the operating frequency band is 60-64GHz. The number of antennas is three transmitting antennas and four receiving antennas. The experimental environment at 5m × 5m indoor.

B. Radar Point Cloud Dataset description

The proposed system aims to provide health monitoring for elderly individuals, particularly those who live alone.

The dataset is divided into two categories: abnormal activity and normal activity. About the abnormal activities, falls are of utmost concern as they pose a significant threat to any individual. The other abnormal activities include convulsion, crawling on the ground, and waving hand for help.

To have a more comprehensive understanding of human activities, it is equally important to distinguish normal activities. Common activities such as walking, jumping, running, sitting, lying down, and standing up were selected. Each sample consists of 25 continuous millimeter wave radar point cloud frames. The distribution of the dataset was shown in Tab 1.

Table 1. Summary of Dataset

Activity	Samples	Frames
Wave hands	763	25
convulsion	728	25
Crawl	748	25
Fall	731	25
Jump	770	25
Lie down	643	25
Run	762	25
Sit down	767	25
Stand	765	25
Walk	770	25

To facilitate the data collection process, we designed a corresponding point cloud data visualization software that can achieve real-time data trimming while collecting data. This ensures the authenticity and effectiveness of the data.

V. Data Preprocessing of Point Cloud

Machine learning algorithms typically require low-dimensional, sparse feature vectors. However, millimeter wave point cloud data cannot be directly used as input for machine learning due to its complex and noisy nature. Therefore, preprocessing is necessary to extract useful features and convert them into a form suitable for machine learning.

In this section, two different pre-processing techniques were described that were applied to convert point clouds into 3D voxels and 2D planes.

A. Denoising: RF-DBSCAN

Although millimeter-wave radar has the advantage of sensitivity to detect weak vibrations caused by the human body,

at the same time, the data measured by millimeter-wave radar is susceptible to a variety of factors, such as reflection and scattering caused by the complex environment (especially indoor), the multipath effect and other external interference, which will lead to noise in radar data. Therefore, point cloud data usually contains a certain amount of noise.

DBSCAN algorithm is widely used in HAR tasks based on millimeter-wave radar. However, it has the following problems:

- Raw DBSCAN requires *MinPts* and *Eps* parameters to be entered beforehand. However, for each frame of millimeter wave radar data, the best choice of these two parameters is different.
- The point clouds' distribution on the z-axis is more discrete than on the x and y-axis due to the point clouds' formation principle, making DBSCAN clustering even more challenging.
- Point clouds generated from the Doppler effect show a distinct difference in activity level between moving and stationary parts, leading to DBSCAN clustering human bodies into multiple clusters.

A new adaptive DBSCAN model was proposed to overcome these issues.

The model adapts to the current point cloud frame by selecting DBSCAN parameters using a Random Forest, performs cluster merging and noise re-judgement, optimizing DBSCAN clustering on millimeter-wave radar data. The model achieves an average clustering accuracy of 92.6% in 9 class.

B. Point cloud voxelization with denoising

Voxelization is a technique that transforms point cloud data into cubic voxels with a fixed size. It works by dividing the point cloud data into small blocks of equal size, then finding the average value for each block and transforming the point cloud data into fixed-sized voxels.

The length, width, and height of the cube that confines the voxels (*Rl*, *Rw*, *Rh*) can be determined from the coordinates of the n points in the frame. The maximum value of x, y, and z of the points are subtracted from the minimum value of x, y, and z of the points, respectively. The voxel resolutions along the X, Y , and Z axes are given by

$$x_{res} \stackrel{\text{def}}{=} \frac{R_l}{x_{cut}} \tag{1}$$

$$y_{res} \stackrel{\text{def}}{=} \frac{R_w}{y_{cut}} \tag{2}$$

$$z_{res} \stackrel{\text{def}}{=} \frac{R_h}{z_{cut}} \tag{3}$$

A voxel size of 32 * 32 * 32 was chosen based on prior research [14] and experimental results. However, voxelization also has some drawbacks such as loss of accuracy, redundancy, high computational complexity, poor directionality, and boundary problems, and when isolated distant points are present, they usually represent far-field noise in millimeter-wave radar, and the cube will be enlarged to cover the point. If there are very few points, the cube mask will be shrunk.

Therefore, traditional voxelization methods are very sensitive to the distribution of points and whether far-field noise is present. The proposed RF-DBSCAN method is able to effectively extract the human body point cloud and automatically filter out noise points besides the target, solving the above problems and making the shape of the voxels more similar to the human body shape, as shown in Fig.2. There are two voxels, the first one is the result of proposed RF-DBSCAN, and the rest on is the result of traditional DBSCAN.

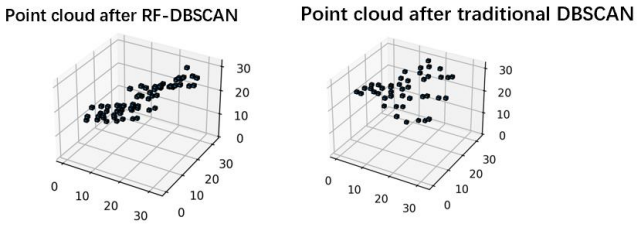

Fig. 2. Voxel shape compare

C. Doppler information embedded in the image

The purpose of compressing 3D point cloud data into a 2D plane is to facilitate subsequent processing, such as image recognition and feature extraction. The commonly used planes are the *xoy* plane and *xoz* plane, which respectively represent the mapping of point clouds on the *x*, *y*, and *z* axes. Through the mapping of these two planes, more spatial information can be obtained, and it is more convenient to identify and analyze point cloud data.

Different from the projection of voxel onto a 2D plane in previous studies [14], Two 2D images based on the original point cloud were chosen to be generated, namely the bird's eye view of xoy and the front view of xoz.

The doppler information of the point cloud was embedded into the 2D images by assigning Doppler values to pixel values. The image generated after embedding the doppler information was shown in Fig.3.

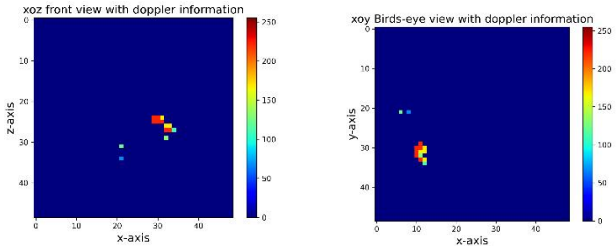

Fig. 3. Bi-views about fall

The action taking place in the image is a fall, and a point cloud with a higher Doppler value will have a larger pixel value, it will be closed to red (255) in the image. The more intense the activity, the greater the percentage of red.

For the two subnetworks, different pre-processing was adopted and the generated 3D data and 2D image data were respectively input into the two subnetworks for training to get features.

VI. SUBNET OF FEATURE FUSION NET

The feature fusion network proposed in this research is mainly composed of two sub-networks, a voxel network, and a dual-view network.

In the dataset, the shortest activity lasted 28 frames, approximately 1.5 seconds, so 25 frames, approximately 1.3 seconds, were used to capture information about activities.

There is specific input size shape in these two network. For Dual-view network, there are two pictures for learning.

Voxel data: After denoising and voxelization, the sample length of shapes 25*32*32*32 was obtained.

Dual-view data: After pre-processing, Doppler information is embedded into pixels of the image to obtain the sample tensor of 25*48*48.

A. 3DCNN-LSTM Voxel net

The network structure is shown in Fig.4. below. For the convolution module, a three-layer cov3d-cov3d-max-pooling structure is used to select 32 convolution kernels with the size of (3,3,3), and the stride size is set as (1,1,1). The last layer of each layer is the Max-Pooling layer. Finally, it is transmitted through a fully connected layer from shape2048 into 64 features.

Data of size (25,64) is generated every 25 frames and input to an LSTM network, wherein the number of hidden layers of LSTM network is set to 128. Finally, the output results of LSTM are input into the full-connection layer to obtain 800-dimensional features for activities classification as an input for the Feature fusion network.

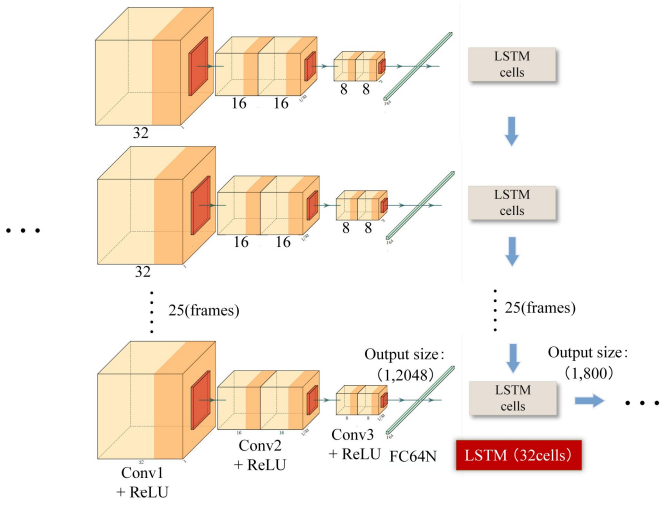

Fig. 4. 3DCNN-LSTM Voxel Network Architectures

A. 2DCNN-LSTM Dual-view net

The 3d point cloud was mapped to two images with different perspectives of 48x48 and the doppler information was embedded into the pixel values. The higher the doppler value of the point cloud, the higher the pixel value of a pixel when projected into the picture.

Compared with the voxel network, the preprocessing of this network is very fast. The down-sampling method greatly reduces the complexity and parameter number of the model and can provide enough feature information for the full connection layer. The proposed model is shown in Fig. 5.

The data shape of the dual view network is 25 * 48 * 48, which inputted into two parallel convolution modules, each consisting of two layers of convolution and one layer of pooling.

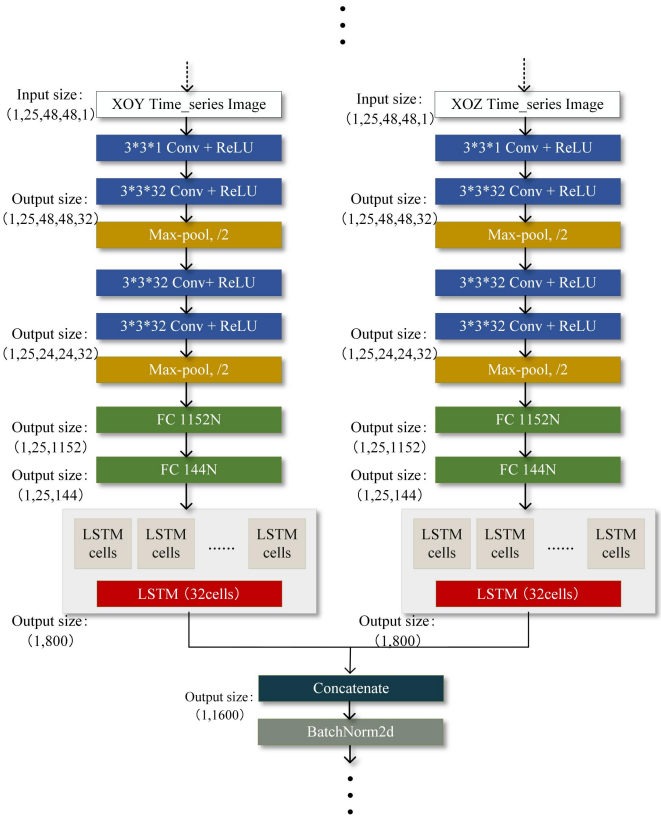

Fig. 5. 2DCNN-LSTM Dual-view Network Architecture

After the results of the two convolutional modules are input to their respective fully connected layers, two 800-dimensional feature vectors are obtained respectively, and these two vectors

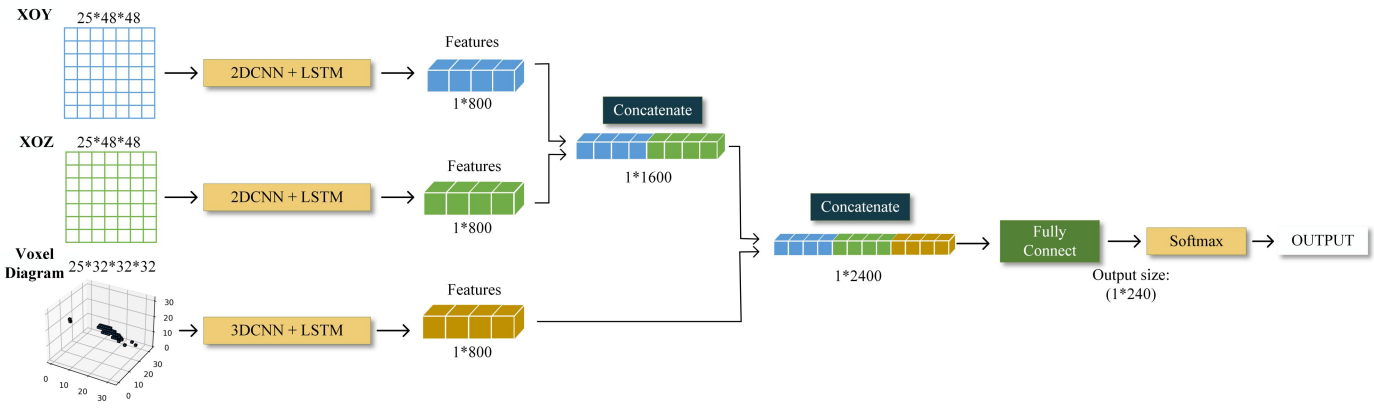

Fig. 6. Feature Fusion Network Architecture

enter different LSTM networks respectively to extract the time series information of the dual-view. Finally, the features of these two parallel networks are concatenated into 1600-dimensional vectors, which are input into the final full-connection layer after regularization. The output of the full connection layer waits to be fused as the input of the feature fusion network.

VII. Feature Fusion Network

In this research, the deep network is utilized to fuse features at the layer level, resulting in a more comprehensive feature representation that enhances classification accuracy. This approach is widely applied in image and video classification, as well as other tasks.

The Feature Fusion Neural Network is shown in Fig. 6. By combining the advantages of both types of data, the network performs human activity classification tasks through feature fusion. The network consists of two parallel networks, the voxel network and the dual-view network. The voxel and two 2D images are inputted into each network, respectively. The voxel network learns the shape changes of point clouds over time, while the dual-view 2DCNN network learns the Doppler information and spatiotemporal changes of the human body. The outputs of the two networks are two-dimensional feature vectors, which are fused before entering the fully connected layer. The regularized new feature vector is finally classified through Softmax in the fully connected layer.

VIII. Evaluation

In this section, the model was evaluated. Firstly, the evaluation metrics and the setting of neural network learning parameters were introduced. Afterwards, comparative experiments and analysis on the two pre-processing methods proposed were conducted. Finally, the best parameters of the two sub-networks were selected for individual learning and compared with the fused neural network in an experiment.

A. Machine Learning Evaluation Indicators

In this study, the performance of the model was evaluated using 5-fold cross-validation. We used cross-entropy loss as the evaluation metric for all model. Adam was chosen as the optimizer and the learning rate was set to 0.001. The model was trained for 20 epochs per fold. The Adam optimizer automatically adjusts the learning rate to suit each parameter, making the training process smoother. Point Cloud Pre-processing Comparison Experiment

In order to validate the effectiveness of preprocessing, we set up different control groups for both sub-networks. For the voxel network, experiments were conducted comparing the Fixed-parameter DBSCAN and the proposed RF-DBSCAN.

The RF-DBSCAN outperformed the fixed-parameter DBSCAN, shown as increasing the accuracy from 80.62% to 86.83%.

However, the all the DBSCAN preprocessing decreased the accuracy of the dual-view network. This may be because the sparse point cloud in a single frame and the relationship between environmental noise and human activities. The noise itself has a shape and information similar to the target and can enrich the image data when the image contains it. In the case of voxel data, removing distant noise is important to preserve the size and shape of the voxels, allowing for accurate extraction of human shape information during feature extraction.

For the Dual-view network, experiments were conducted comparing 2D images without Doppler information and those with Doppler information embedded.

The experiment shown in Fig.7 shows that incorporating Doppler information into pre-processing improves the accuracy of the model in recognizing similar action shapes by 2.5%, reaching 92.6% compared to when no doppler information was embedded.

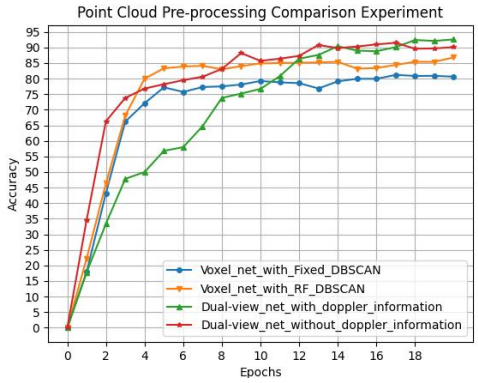

Fig. 7. Comparison Experiment Result

B. *Accuracy of Sub-Network and Fusion Neural Network*

As shown in Figure 9. It can be seen that the fusion neural network that integrates voxel information and dual-view information achieved the best results 95.30% among the three models, which proves the effectiveness of feature fusion model. However, whether the improvement brought by the huge amount of computation consumed is worthwhile needs to be further discussed.

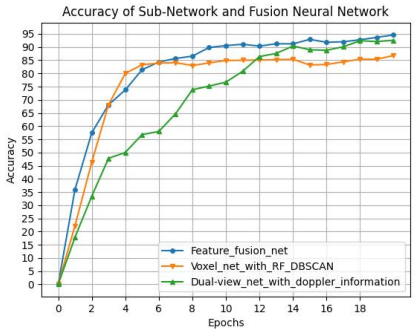

Fig. 8. Feature Fusion Network Experiment Result

The confusion matrix of the feature fusion neural network is shown in Fig. 9.

After fusing the features of two sub-networks, the accuracy of the proposed fusion neural network for misclassified classes, running and walking, has also significantly improved, reaching 94.8% and 87.9%, respectively, indicating higher potential than the two sub-networks.

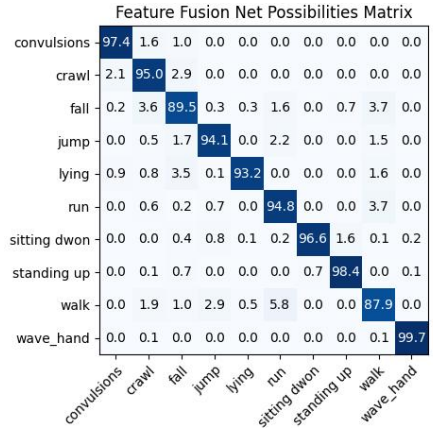

Fig. 9. Feature Fusion Network Experiment Result

VIII. Conclusion

This research focuses on experimenting and optimizing two main neural networks for Human Activity Recognition (HAR) using millimeter wave radar point clouds. A new clustering algorithm was proposed for pre-processing of voxel data and its effectiveness was demonstrated. A dual-view image data embedded doppler information was designed to reduce the size of data while maintaining high accuracy. Finally, a novel feature fusion network was designed and proved to be effective with an accuracy of 95.30%, higher than the two sub-networks, after optimization of each sub-network. However, The robustness of the trained Random Forest network for all household environments still needs to be tested. The dual-view approach showed better recognition accuracy without noise reduction, which may be a key factor in improving radar recognition accuracy in the future. The new dataset created has a unique feature and categorizes common abnormal and normal activities for home monitoring, with an average accuracy of 95.40% for abnormal activity recognition.

Acknowledgements

I would like to express my sincere gratitude to all the people who have supported and encouraged me throughout my graduation journey. First and foremost, I would like to extend my heartfelt thanks to my supervisor, Professor Runhe Huang, for her invaluable guidance, encouragement, and support.

And I also like to extend my gratitude to the other members of the Huang laboratory, Lin Kong, Guochen Zhang, Eric, who have provided me with invaluable support and assistance throughout my research.

References

[1] Attal, Ferhat, et al. "Physical human activity recognition using wearable sensors." Sensors 15.12 (2015): 31314-31338.

[2] Jalal, Ahmad, Shaharyar Kamal, and Daijin Kim. "A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring systems." (2017).

[3] Schrader, Lisa, et al. "Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people." Journal of Population Ageing 13 (2020): 139-165.

[4] Suma, Evan A., et al. "Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit." Computers & Graphics 37.3 (2013): 193-201.

[5] Zou, Han, et al. "Poster: WiFi-based device-free human activity recognition via automatic representation learning." Proceedings of the 23rd annual international conference on mobile computing and networking. 2017.

[6] Yan, Huan, et al. "WiAct: A passive WiFi-based human activity recognition system." IEEE Sensors Journal 20.1 (2019): 296-305.

[7] Subasi, Abdulhamit, et al. "Human activity recognition using machine learning methods in a smart healthcare environment." Innovation in health informatics. Academic Press, 2020. 123-144.

[8] Mekruksavanich, Sakorn, and Anuchit Jitpattanakul. "Smartwatch-based human activity recognition using hybrid lstm network." 2020 IEEE SENSORS. IEEE, 2020.

[9] Lee, Junwoo, and Bummo Ahn. "Real-time human action recognition with a low-cost RGB camera and mobile robot platform." Sensors 20.10 (2020): 2886.

[10] Luo, Xiaomu, et al. "Simultaneous indoor tracking and activity recognition using pyroelectric infrared sensors." Sensors 17.8 (2017): 1738.

[11] Roche, Jamie, et al. "A multimodal data processing system for LiDAR-based human activity recognition." IEEE Transactions on Cybernetics 52.10 (2021): 10027-10040.

[12] Wang, Yuheng, et al. "m-activity: Accurate and real-time human activity recognition via millimeter wave radar." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[13] Singh, Akash Deep, et al. "RADHAR: Human activity recognition from point clouds generated through a millimeter-wave radar." Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. 2019.

[14] Yu, Chengxi, et al. "Noninvasive human activity recognition using millimeter-wave radar." IEEE Systems Journal 16.2 (2022): 3036-3047.