Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/119/2025.21590

Human Activity Recognition Algorithm Based on Bidirectional Multi-channel Feature Fusion

Rui Wang[1,a,*]

[1]*University of Electronic Science and Technology, Xi yuan Street, Chengdu, China*
a. 13472987016@163.com
**corresponding author*

Abstract: This paper designs a bidirectional spatiotemporal feature fusion algorithm for human activity recognition based on frequency modulated continuous wave radar. The algorithm takes the three-dimensional point cloud data of human activity collected by the radar as input, and adopts a dual channel feature extraction method in spatial feature extraction. The voxelated point cloud data is put into a convolutional neural network for extracting coarse-grained spatial information. At the same time, a multi-layer perceptron is used to extract fine-grained spatial information from individual points in the input point cloud data. Then, the coarse-grained and fine-grained spatial feature information extracted from two channels are fused as subsequent inputs. In terms of extracting temporal features, this algorithm uses Bi-LSTM to extract temporal feature, which can simultaneously link temporal information before and after for feature extraction. This network model can not only extract local and global features from point cloud data, but also consider the information before and after when extracting temporal information, which can greatly improve the accuracy of human activity recognition.

Keywords: FMCW radar, human activity recognition, neural network, multi-channel feature fusion

1. Introduction

In recent years, human activity recognition, has made significant progress in research over the past decade. The purpose of human activity recognition is to identify user behavior, enabling computing systems to actively provide assistance to users[1]. Due to its convenience and non-contact control characteristics, It has been widely applied in many fields[2-5].Unlike traditional methods of visual recognition using cameras and other devices, using millimeter wave radar sensors for human activity recognition can better protect users' privacy and security. At the same time, millimeter wave radar has strong robustness to various lighting and weather conditions, and can work normally in various environments , such as rainy, foggy, and dark environments, making it applicable to more complex scenarios. In addition, sparse point clouds generated from millimeter wave radar raw data can better describe the motion and shape of the human body in three-dimensional space, with good interpretability[6], while traditional Doppler features and their derived features are limited to the velocity domain and have limited interpretability, making it difficult to distinguish different parts of the body. Compared with traditional Doppler image recognition models, using point cloud data for human activity recognition has higher accuracy, stronger interpretability, and more significant

DOI: 10.54254/2755-2721/119/2025.21590

generalization ability. Therefore, research on human activity recognition based on millimeter wave radar point clouds is of great significance.

2. Related Work

In 2015, Kim's team [7] converted radar collected data into micro Doppler images as input and used convolutional neural networks (CNN) to recognize seven types of human activity behaviors, achieving a recognition accuracy of 90.9%; In 2016, Wang's team[8] integrated an end-to-end human activity recognition system that used CNN and RNN networks to extract spatial and temporal information from human activity data for final result judgment. The system achieved an average recognition accuracy of 87% for human activity recognition; In 2019, Singh's [9] team constructed a point cloud dataset MMACTVITY obtained from low resolution radar acquisition, and voxelized the point cloud data as input for the model. The model was trained and tested on support vector machines (SVM), multi-layer perceptons (MLP), long short-term memory neural networks (LSTM), and models combining convolutional neural networks (CNN) and long short-term memory networks (LSTM). Among them, the model combining time distributed CNN and bidirectional LSTM not only learned the spatial features of the data but also retained the temporal dependence of the data. Its classification performance was the best in the experiment, with a recognition accuracy of 90.47%; In 2020, the Shrestha team [10] inputted micro Doppler images as time-series signals into a Long Short Term Memory (LSTM) network to classify continuous human activity behaviors with an accuracy rate of over 90%; In 2022, Zheng Zhiyuan's[11] team simultaneously concatenated distance time maps and micro Doppler maps as inputs, and used the Long Short Term Memory (LSTM) network to extract time dependencies, achieving a recognition accuracy of 94.5%;

Although there have been numerous excellent research achievements in the field of deep learning based human activity recognition, we have also found that there are still various problems that urgently need to be solved in this field: (1) In terms of point cloud datasets, most models are currently based on dense point cloud data collected by high-resolution sensors for research. Dense point cloud data itself can carry enough information, and the accuracy is generally not too poor. However, based on practical considerations, low resolution and low-cost millimeter wave sensors often have better practicality. However, due to the sparse and uneven point cloud data generated by such radars, higher requirements are put forward for recognition models; (2) In terms of data input format, current recognition models mostly use a single input format, such as directly inputting point clouds or voxelizing point clouds before inputting. This method is simple, but has disadvantages such as low efficiency and memory waste. How to choose the appropriate data input mode based on the characteristics of different data input formats for different types of point cloud data will have a crucial impact on the model; (3) In terms of the balance between model recognition time and recognition accuracy, it can be found through the above research that existing models can be divided into two categories: those with long recognition time but excellent recognition accuracy, and those with short recognition time but slightly inferior recognition accuracy. How to improve the overall recognition time of the model on the basis of high accuracy is also a major problem that urgently needs to be solved in this field. (4) In terms of processing time series, determining current human activity requires both reference to past information and consideration of future information. Therefore, how to extract temporal information is equally important for human activity recognition tasks, as it involves considering both the preceding and following information in both directions. The following model algorithm will mainly be proposed based on the existing problems mentioned above.

DOI: 10.54254/2755-2721/119/2025.21590

3. Methodology

3.1. Motivation and Approach for Algorithm Design

Based on the analysis of existing human activity recognition algorithms in the previous section, summarize and generalize the current problems: 1. how to design algorithms to achieve high-precision recognition of sparse and uneven point cloud data; 2. How to choose the appropriate data input format based on actual needs; 3. How to balance the relationship between recognition time and recognition accuracy; 4. How to design a network to achieve bidirectional reference in temporal information extraction. The solution to problem 1 in this article is to expand the sparse point cloud data to make it relatively dense, that is, to accumulate the point cloud data in a very short time slice (20ms). This method can ensure that all point cloud data is real and reliable, and can also densify the sparse point cloud. At the same time, considering the research background of this article, the human activities of the elderly will not produce continuous and various forms of changes in a very short time slice of 20ms. For problem 2, consider voxelizing the point cloud to extract local features, while considering the loss of global information during voxelization, a method of extracting single point cloud information from the original point cloud data can be used. Then, the features obtained from the two processing channels are fused and used for subsequent temporal feature extraction; For question 3, appropriate choices should be made based on actual user needs; For problem 4, this article adopts a bidirectional temporal information extraction model, which can ensure that the model can both refer to past information and make final predictions based on future information.Overall architecture

The algorithm in this chapter takes the three-dimensional point cloud time series of human activity as input, and extracts features from the input data in both the direct point cloud channel and the voxelization channel. The direct point cloud channel uses multi-layer perceptrons to extract local features for each point in the point cloud data, and outputs the corresponding local feature sequence. The voxelization channel uses convolutional neural networks to extract global features from the voxelized point cloud data, and performs inverse voxelization on the extracted global feature sequence, which is then fused with the local feature sequence extracted from the direct point cloud channel. The fused information is used as input for the Bi-LSTM to extract temporal features. Finally, the fully connected layer and the softmax layer for classification ultimately outputs the recognition result. The overall architecture of the algorithm proposed in this chapter is shown in the figure 1.

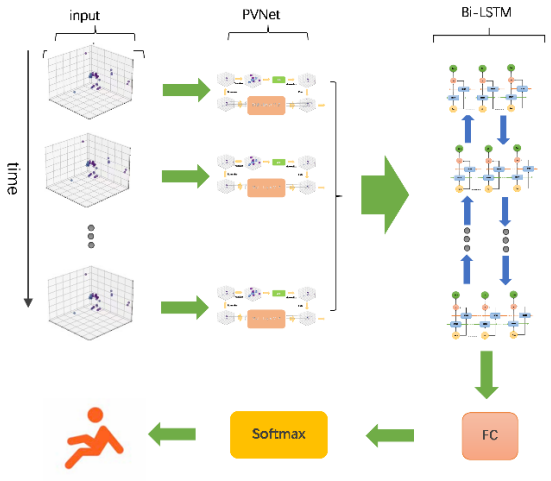

Figure 1: The overall architecture.

DOI: 10.54254/2755-2721/119/2025.21590

3.2. Design of Dual Channel Network Structure

The dual channel network combines the advantages of point based and voxel based methods. The voxel based branch first converts points into a low resolution voxel grid, then aggregates adjacent points through voxel based convolution, and finally converts them back to points through voxelization. Voxel or de voxel both require a scan of all points, which makes storage costs very low. Extract the features of each individual point based on the branch of the following points. Because it does not gather information from neighbors, it can provide very high resolution. The Voxel method is used to extract local features, the Point method is used to extract global features, and then the two features are fused. The network structure is shown in the following figure 2.

Based on the regularity of voxels, this article chooses to perform specialized aggregation in voxels. The scale of different point clouds may vary greatly, therefore, before converting point clouds into voxels, the coordinates need to be normalized. After converting points into voxel grids, 3D voxel convolution is applied to aggregate features. As information needs to be fused with point based feature transformation branches, voxel based features need to be transformed back into the point cloud domain. Since both voxelization and de voxelization are differentiable, the entire voxel based feature aggregation network can be optimized end-to-end. The voxel based feature aggregation branch fuses neighborhood information at a coarse-grained level. However, in order to model fine-grained single point features, methods based solely on low resolution voxels may not be sufficient. For this, we directly operate on each point and use MLP to extract the features of individual points. Although the method is simple, each point output by MLP has distinct distinguishing features. This high-resolution single point information is crucial for supplementing information based on bold pixels.

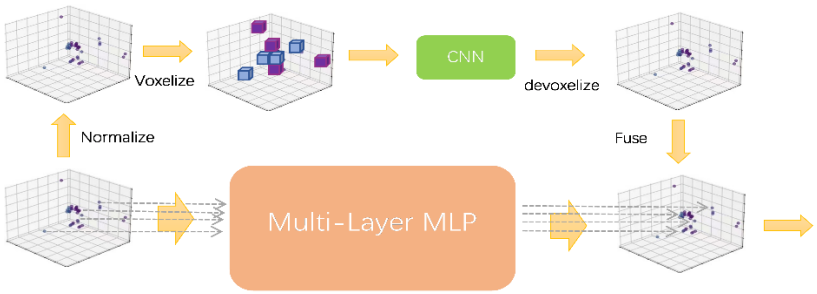

Figure 2: Spatial feature extraction module.

4. Evaluation Metrics

There are many evaluation metrics for classification models. The confusion matrix is a typical tool for evaluating classification models. For a k-class classification model, the confusion matrix is a matrix of k × k size, with the horizontal axis representing the model prediction results and the vertical axis representing the true labels. This matrix is used to record the classification results of the model. When k=2, a special binary confusion matrix can be obtained, this matrix will be used to explain some of the evaluation indicators mentioned above.

5. Experiments

After the model is established, an important step is to adjust the hyperparameters. Parameters are the goals that the network needs to learn and adjust, and the model can achieve optimal parameters

DOI: 10.54254/2755-2721/119/2025.21590

through training. However, hyperparameters are manually set, and different values of hyperparameters will result in different outcomes for the model. Therefore, experiments need to be conducted on the values of various hyperparameters to select the optimal combination of hyperparameters. This section mainly selects learning rate, batch size, and epochs for experiments.

From the experimental results, it can be seen that different hyperparameter settings have a significant impact on the accuracy of the model. When exploring the impact of batch size on model results, the learning rate was set to 0.01, and batches were set to 32 and 64. After training, the accuracy of gesture classification was 87.88% and 94.44%. It was found that the classification accuracy was highest when the batch size was 64. Comparing the effects of different learning rates on model performance, the learning rates were set to 0.1, 0.01, 0.005, 0.0001, and the initial learning rate was set to 0.01, which decreased to half of the original every 20 epochs. Through model training, among these three learning rates, the classification accuracy obtained when the learning rate was 0.005 was the highest. When exploring the required number of epochs for the model, two values were set: 50 and 100. The highest accuracy is achieved when epoch is 100.

After the above analysis, the hyperparameters with the best model performance were selected as shown in Table 1. Choose Adam algorithm for optimization algorithm and cross entropy loss function commonly used in classification tasks for loss function.

Table 1: Accuracy of different hyperparameters.

hyperparameters	value
batch size	64
learning rate	0.01
epochs	100
optimization algorithm	Adam
loss function	Cross entropy loss
activation function	Leaky ReLU

According to the optimal hyperparameters obtained from the previous experiment, the network was trained and tested. Through the experiment, it was found that within 20 epochs, the classification accuracy of the training and testing of the network in this chapter reached over 90%; Between 20-100 epochs, the training accuracy remained stable around 1, with the initial testing accuracy fluctuating between 90% and 95%, and the latter remaining stable at around 95%. Testing was conducted on the dataset in this chapter to recognize five different human postures, and the recognition results are shown in Table 2 below. The algorithm achieved 100% accuracy in recognizing walking and standing, 99.89% accuracy in recognizing running, 99.12% accuracy in recognizing sitting, and 97.83% accuracy in recognizing falls. Among them, the algorithm performed the best in recognizing walking and standing, mainly because these two actions have simple structures and obvious features, making them easier to recognize. The accuracy in recognizing falls is relatively low but slightly higher than other recognition models in the same category. The main reason is that falls are complex human activities that involve changes in the complex state of multiple parts of the human body, and there are many ways to fall, which also brings great challenges to the recognition of falls. Overall, the algorithm proposed in this article considers both global and local features of human activity, resulting in high recognition accuracy and excellent model performance for the five types of actions mentioned above.

Table 2: Accuracy of human posture recognition.

posture	Walk	stand	run	sit	fall
accuracy	100.00%	100.00%	99.89%	99.12%	97.83%

DOI: 10.54254/2755-2721/119/2025.21590

6. Conclusion

The human activity recognition algorithm based on FMCW radar proposed in this article can fully extract deep temporal features between dynamic point cloud data, while avoiding data redundancy caused by unified data specifications, effectively enhancing the recognition accuracy of the algorithm on self-collected datasets. Experimental verification has shown that the algorithm model has good robustness and generalization ability in different environments, which has important reference value and guiding significance for the further development of human activity recognition algorithms in the direction of millimeter wave radar point clouds in the future.

References

[1] *Andreas Bulling, Ulf Blanke, Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors[J]. ACM Computing Surveys (CSUR), 2014. 46(3):1–33.*

[2] *Qian Wan, Yiran Li, Changzhi Li, Ranadip Pal. Gesture recognition for smart home applica tions using portable radar sensors[C]//2014 36th annual international conference of the IEEE engineering in medicine and biology society. IEEE, 2014:6414–6417.*

[3] *Pavlo Molchanov, Shalini Gupta, Kihwan Kim, Jan Kautz. Hand gesture recognition with 3d convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015:1–7.*

[4] *Theodore Alexandrov, Katja Ovchinnikova, Andrew Palmer, Vitaly Kovalev, Artem Tarasov, LachlanStuart, RenatNigmetzianov, DominikFay, KeyMETASPACEcontributors,Mathieu Gaudin, et al. Metaspace: A community-populated knowledge base of spatial metabolomes in health and disease[J]. BioRxiv, 2019:539478.*

[5] *MoenessGAmin,YiminDZhang,FauziaAhmad,KCDominicHo. Radarsignalprocessing for elderly fall detection: The future for in-home monitoring[J]. IEEE Signal Processing Magazine, 2016. 33(2):71–80.*

[6] *David J Brenner, Richard Doll, Dudley T Goodhead, Eric J Hall, Charles E Land, John B Little, Jay H Lubin, Dale L Preston, R Julian Preston, Jerome S Puskin, et al. Cancer risks at tributable to low doses of ionizing radiation: assessing what we really know[J]. Proceedings of the National Academy of Sciences, 2003. 100(24):13761–13766.*

[7] *Youngwook Kim, Taesup Moon.Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks[J]. IEEEgeoscience and remote sensing letters, 2015. 13(1):8–12.*

[8] *Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, Otmar Hilliges. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spec trum[C]//Proceedings of the 29th Annual Symposium on User Interface Software and Tech nology. 2016:851–860.*

[9] *Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, Mani Srivastava. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar[C]//the 3rd ACMWorkshop. 2019.*

[10] *Aman Shrestha, Haobo Li, Julien Le Kernec, Francesco Fioranelli. Continuous human ac tivity classification from fmcw radar with bi-lstm networks[J]. IEEE Sensors Journal, 2020. 20(22):13607–13619.*

[11] *Zheng Zhiyuan, Zhu Wenzhang. FMCW Radar Human Activity Recognition Method Based on Domain Fusion and LSTM [J]. Journal of Xiamen University of Technology, 2022, 30(5): 15-21.*