Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/74/20240493

Dynamic gesture recognition model based on FMCW radar and multidimensional feature fusion with dual-channel CNN and GRU

Xiaowei Ye

School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, 611730, China

1412638537@qq.com

Abstract. With the development of computers, gestures have also become a commonly used human-computer interaction, so it becomes important to recognize gestures correctly. The objective of this paper is to propose a deep learning model to improve the accuracy of low performance radar in dynamic gesture recognition.We use a dataset acquired by CAL60S244-AB radar to classify ten gestures. The model consists of a dual-channel convolutional neural network(CNN), Gate Recurrent Unit(GRU) and a classification module. The inputs to the model are feature vectors fused in four dimensions of distance, speed, receiving antenna channel and time, from which the dual-channel CNN can extract shallow and deep spatial information, and GRU can extract temporal information, and the combination of these two parts makes the model's classification accuracy improve. Through experiments, the accuracy of the model on the dataset reaches 95.1%, and the results show that the model proposed in this paper can improve the accuracy of dynamic gesture classification for low-precision radar, which verifies the effectiveness of the model.

Keywords: FMCW radar, Dual-channel CNN, GRU, gesture recognition.

1. Introduction

The ways of human-computer interaction become more and more with the development of computers, and hand gesture recognition has become one of the most important ways of human-computer interaction [1] due to the flexibility of its joints and the indispensable part of daily communication. Gesture recognition is used in many aspects, such as smart phones, automotive assisted control, sign language translation, etc. The gesture recognition model designed in this paper is mainly used in automotive assisted driving. At present, there are three main ways of gesture recognition, the traditional data glove-based gesture recognition, but it is not convenient to carry and to consider the problem of power; computer vision-based gesture recognition method, which is affected by the environment of light and background is relatively large, and requires a high cost of high-precision cameras, and now people attach importance to privacy, the visual method may lead to the leakage of the user's photo; based on wireless signals, gesture recognition method, mainly radar, which is the most popular method of gesture recognition, the main reason is that it is not easy to use. recognition method, mainly radar, has the

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/74/20240493

advantages of small working energy consumption, low cost, and is favorable to privacy protection. Based on the above, radar is chosen as the hardware for gesture recognition.

Currently, both machine learning and deep learning based radar gesture recognition algorithms are important research directions. The methods based on machine learning are support vector machine, k-nearest neighbor(KNN) method, hidden Markov model and so on. For example, WANG et al. used Doppler frequency domain and time features as well as physical property change features for gesture recognition using KNN algorithm [2]. The traditional approach relies more on radar signal analysis and processing techniques, and the use of deep learning shifts the research focus from radar signals to recognition and classification algorithms, which are more robust and accurate than traditional machine learning methods, and different structures of the network can meet the needs of different scenarios. For example, Choi et al. used a long short-term memory network (LSTM) for gesture recognition with a recognition accuracy of 97.48% [3]; safa et al. proposed a pulsed neural network (SNN) for radar gesture recognition, which achieved 91% recognition accuracy [4]; Saiwen et al. used Soli sensors based on a combination of deep convolutional neural networks and LSTM to achieved 87% recognition rate on 11 gestures [5]. However, after experiments, it was found that the accuracy of the same neural network model obtained in radar gesture datasets with different resolutions has some differences, and the difference in radar performance has a greater impact on the accuracy of the final model, and the real-time recognition time needs to be taken into account when building the model. Therefore, how to build a model that can reduce the effect of radar performance on accuracy and also make the running time less deserves more attention.

In this paper, we propose a dual-channel CNN combined with GRU for radar dynamic gesture classification model, which extracts the information in the radar signals as a feature vector, and uses this feature vector for the classification of gestures.

Because of the need for the model to have better applicability to low performance radars and to extract more features while controlling the number of model parameters, this paper utilizes two CNN channels to extract the spatial information in the shallow and deep layers, and then combines them with the GRU and experiments with the effect of different deep networks on the accuracy, and finally this paper adopts the deformed resnet18 network as the deep network. The experimental results show that this model can indeed extract more information in the radar gesture feature vector, which is useful for increasing the generalization and accuracy of the model, and the gesture recognition accuracy reaches 95.1%.

2. Method

2.1. dataset collection and preprocessing

The dataset for this work was produced by self-collection, using a 60Ghz CAL60S244-AB Frequency modulated continuous wave (FMCW) radar. The performance of this radar is not as good as some of the radars used in the public datasets, but it is cheaper, for example, the soli radar [6]\cite{Lien10.1145} has a distance resolution of 0.3mm compared to this radar's 25mm, and this was chosen as the radar for the acquisition given the cost implications at deployment. The collected dataset contains 10 categories of gestures, which are waving left, waving right, pulling up, pushing down, waving left and right, waving right and left, ticking, drawing a fork, drawing a circle clockwise, and drawing a circle counterclockwise. With 300 samples for each category of gesture, there are a total of 3000 samples in the dataset, consisting of 10 people each doing 30 samples for each gesture.

Two Fast Fourier Transforms are applied to the radar gesture signals to obtain a Range-Doppler Map (RDM) of the gesture, with an image size of 32x32. For each sample 30 frames are selected as a sequence to represent a complete gesture, and each frame consists of the RDM of the four receiving antennas. The dataset is divided into training and validation sets in the ratio of 7:3. Before inputting into the model, the RDM need to be preprocessed, as the body movements other than the hand and the background form noise when acquiring radar signals, a noise floor is tested that retains the most RDM features and also

DOI: 10.54254/2755-2721/74/20240493

removes the noise, and then the processed dataset is fed into the designed model for gesture classification as well as accuracy effect detection.

The RDM of the ten gestures in this dataset is shown in Figure 1:

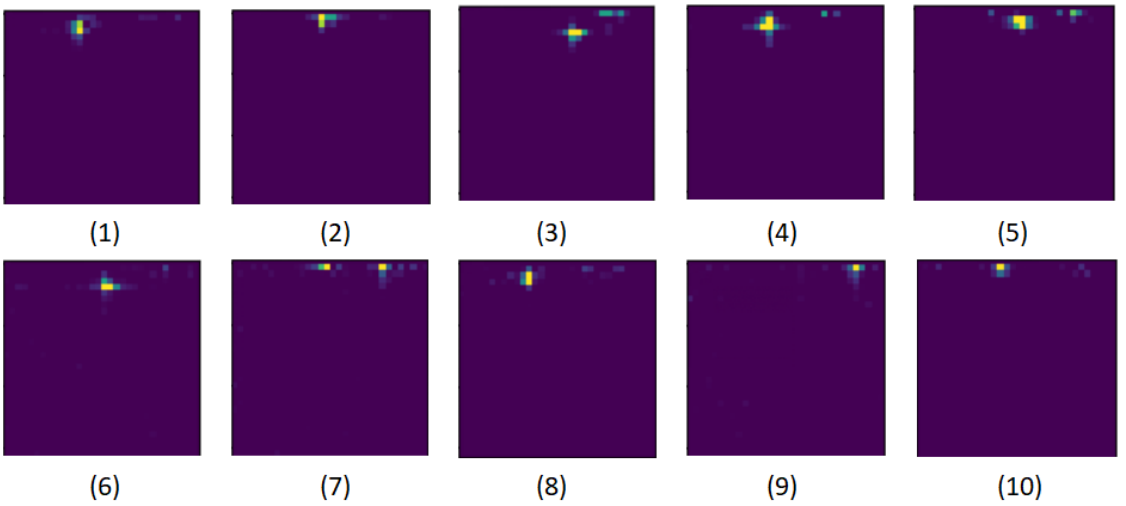

Figure 1. RDM of the 15th frame of the 1st receiving antenna for each gesture in the dataset. (1) left wave (2) right wave (3) pull up (4) push down (5) left-right wave (6) right-left wave (7) tick (8) cross (9) clockwise circle (10) counterclockwise circle.

2.2. Proposed approach

Deep learning is one of the most effective and applicable methods in the field of classification currently, and deep learning solves the dependence of radar dynamic gesture recognition on radar signal feature processing relative to traditional methods. In this paper, a new model for radar dynamic gesture recognition based on deep learning is proposed, which consists of three modules: a dual-channel CNN, GRU, and a classification module. The data input to the model is a series of RDM continuous frames, which are four-dimensional feature vectors consisting of distance, velocity, receiving antenna, and time. The parameters of these four dimensions are fused as inputs to the model to provide more information for subsequent model predictions.

Dual channel convolutional neural network is used to extract shallow and deep features in RDM continuous frames. The module consists of two parts: shallow CNN and deep CNN.Shallow CNN has only three convolutional layers and the convolutional kernel sizes are all 3x3.Each convolutional layer is followed by batch normalization(BN), which normalizes each batch to speed up the convergence of the model.After the BN layer, LeakyReLU is chosen as the activation function, which can solve the problem of neuron death.The BN layer is followed by the Dropout layer, which alleviates the overfitting of the model. This process is repeated in three convolutional layers. The deep CNN is changed from the resnet18 network by making the number of channels of the original resnet18 model smaller, from 64, 128, 256, and 512 to 18, 24, 32, and 24, which greatly reduces the number of parameters of the model and the running time; a dropout layer is added to prevent the model from overfitting; and the Maxpool layer is removed because the size of the RDM is 32x32, the Maxpool operation will reduce the features in the RDM. The final feature vectors of shallow CNN and deep CNN are input into a linear layer respectively, converted to the dimension size needed by the GRU module, and then these two feature vectors are spliced as the input of the next module.

GRU is an extension of RNN, which is advantageous for processing sequence data, traditional RNN has the problem of gradient vanishing or explosion, GRU introduces gate function, which effectively alleviates this problem. And GRU has only two gates, update gate and reset gate, which is one less gate than LSTM, the number of parameters is less and easier to converge, and the performance is the same

DOI: 10.54254/2755-2721/74/20240493

on the gesture recognition task.The input of GRU module is the feature vector spliced with the outputs of two convolutional modules, and the temporal features of the radar's dynamic gestures are obtained through the computation of the module, which is used as an input to the final classification module.

The classification module consists of a fully connected layer, and an activation function layer. The fully connected layer converts the output channels of the GRU module to 10, outputs the prediction of the class to which the gesture belongs based on the results of the fully connected layer, and determines the prediction probability for each class using the Softmax activation function. The classification module finally gives the prediction of the class to which the gesture belongs. The model uses cross entropy as a loss function and Adam as an optimizer and mainly uses accuracy as an evaluation metric.

The structure of the model is shown in Figure 2:

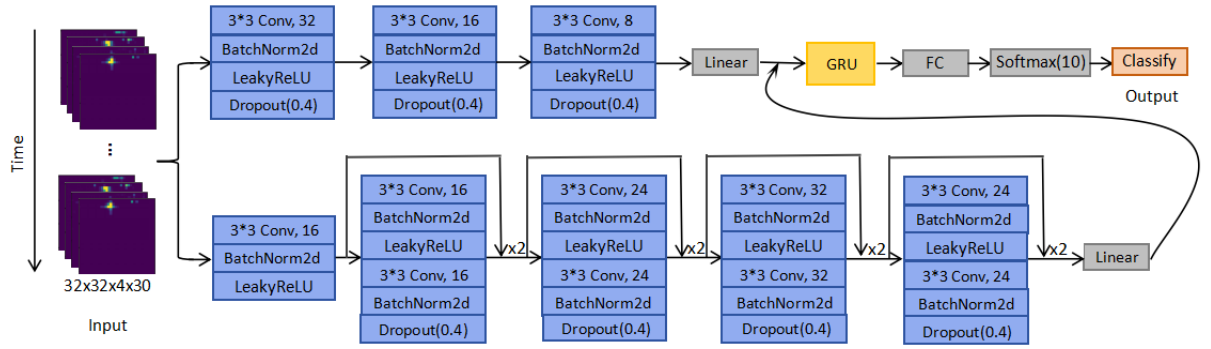

Figure 2. The structure of the proposed model.

3. Experimental result and discussion

After experiments to adjust the hyperparameters to a suitable size, the learning rate of the model was finally determined to be 0.001, batch to be 64, and epoch to be 100, at which point the model reached its highest accuracy. After training the model for 20 batches, the accuracy of both the training set and the test set reaches more than 90%, which slowly increases between 20 and 100 batches, and the final accuracy on the training set is 99.7% with a loss value of 0.03083, and the accuracy on the test set is 95.1% with a loss value of 0.22612.The results of the model on the test set shows that gestures can be classified with higher accuracy on a low performance radar.

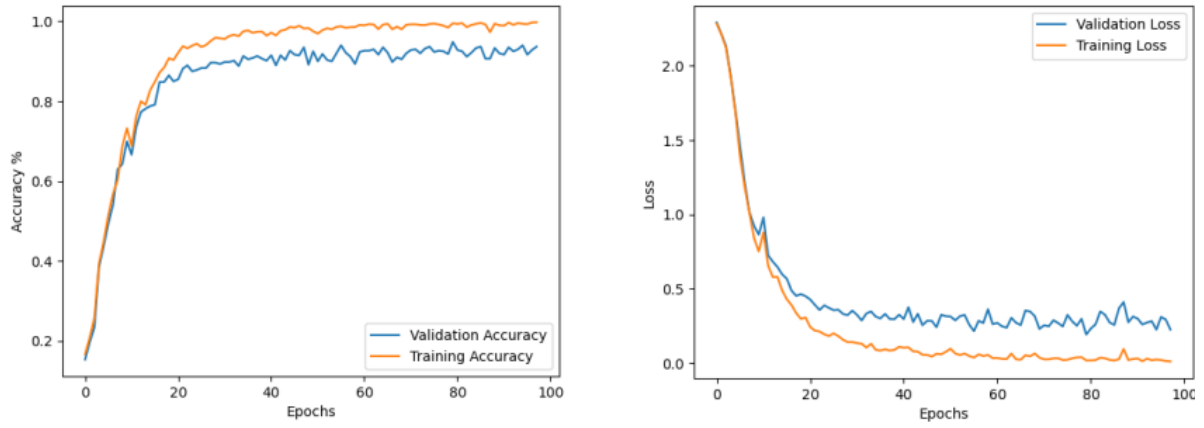

Figure 3. Training accuracy and losses vs. number of epochs.

DOI: 10.54254/2755-2721/74/20240493

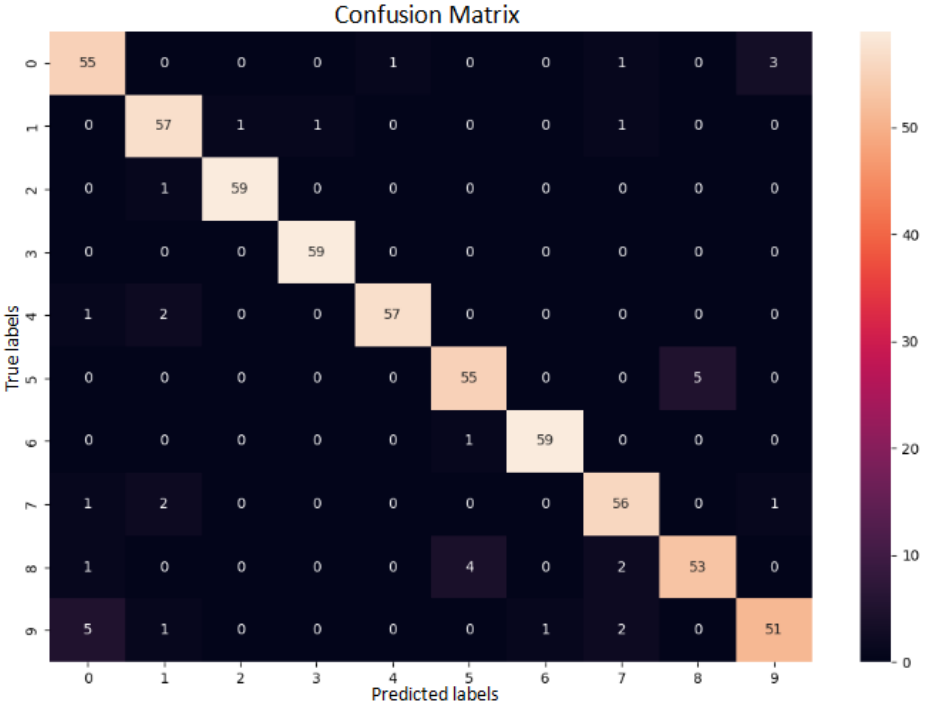

Figure 4. Confusion Matrix.

In the confusion matrix in Figure 4, the vertical axis represents the true label of the gesture, the horizontal axis represents the predicted category of the model, and the labels 0 to 9 represent the ten gestures in order in the description of the dataset. We can conclude from the confusion matrix that there is a slight difference in the accuracy of each gesture, and the model performs poorly on the "circle clockwise" and "circle counterclockwise" gesture categories, and best on the "push down" category, the accuracy is 100%, followed by "pull up" and "draw a fork", both with an accuracy of 98%.

The confusion matrix shows that there is some degree of difference in the accuracy of different gestures. The lack of radar accuracy and performance leads to some gestures having very similar RDM features, and the model still cannot completely separate these two gestures when extracting features. However, compared to the model of CNN and RNN alone, the model in this paper has increased the accuracy by about 10% on the collected dataset.

4. Conclusion

In this work, a deep learning model is proposed to implement the gesture classification problem for low performance FMCW radar. The dual-channel CNN extracts more deep and shallow information in the RDM image, where the deep CNN selects the residual network to alleviate the gradient vanishing problem as well, and the radar dynamic gesture recognition is a time-series problem, and the GRU is able to efficiently mine the temporal information in the data, which is very suitable for dealing with this aspect of the problem, and the combination of the two modules results in improved model performance.

Future work may use this radar to capture datasets with more scenes and more gestures, further accuracy improvements are needed for samples with similar features, and the model needs to be able to achieve the same high level of accuracy for radars with different performance. For these two aspects that can be improved, there is still a need to further explore how to extract more features, and it can be considered to continue to add an attention mechanism to this model to effectively focus on certain features in the radar gesture signals to further improve the generalization and accuracy of the model.

DOI: 10.54254/2755-2721/74/20240493

References

[1] Zhou Y, Jiang G and Lin Y 2016 A novel finger and hand pose estimation technique for real-time hand gesture recognition Pattern Recognition 102-114
[2] Swan Q, Li Y, Li C, Pal and Ranadip 2014 Gesture recognition for smart home applications using portable radar sensors 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 6414-6417
[3] Choi J W, Ryu S J and Kim J H 2019 Short-range radar based real-time hand gesture recognition using LSTM encoder IEEE Access 33610-33618
[4] Safa A, Bourdoux A, Ocket I and Catthoor F 2021 91% accuracy spiking neural network approach for radar gesture recognition arXiv:2108.02669
[5] Wang S, Song J, LIEN J, Poupyrev I and Hilliges O 2016 Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum Proceedings of the 29th Annual Symposium on User Interface Software and Technology 851-860
[6] Lien J, Gillian N, Karagozler M E, Amihood P, Schwesig C and Olson E 2016 Soli: Ubiquitous gesture sensing with millimeter wave radar ACM Trans. Graph. 35 L4