Dynamic Gesture Recognition Based on LSTM-CNN

Yuheng Wu

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences

Changchun University of Science and Technology

Chongqing, China

wuyuheng21@163.com

Bin Zheng

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences

Chongqing, China

zhengbin@cigit.ac.cn

Yongting Zhao

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences

Chongqing, China

zhaoyongting@cigit.ac.cn

***Abstract*—The current research on using surface electromyography (sEMG) for gesture recognition mainly focuses on designing EMG signal features, decent feature designs can significantly improve the result. Nevertheless, the process of designing and selecting features can be complicated, as well as the precision of recognition for different features will be largely different even for the same model. Therefore, in this paper, we take advantage of the complementarity of Long Short-term Memory (LSTM) and Convolution Neural Networks (CNNs) by combining them into one unified architecture, which we call LSTM-CNN (LCNN). This model can directly input pre-processed EMG signal into the network for dynamic recognition of gestures. The LSTM model is used to extract timing information in signals. The CNN model can perform a secondary feature extraction and signal classification. In the experiment stage, the average recognition accuracy of LCNN can achieve 98.14%. As the experiment showed, LCNN model is feasible on dynamic gesture recognition based on sEMG signal.**

Keywords— Surface Electromyography (sEMG); Convolution Neural Networks (CNNs); Long Short-term (LSTM); Hand Gesture Recognition.

Ⅰ. INTRODUCTION

Robotics and artificial intelligence can be leveraged to improve the ability of disabled people to complete their daily lives independently. By using artificial intelligence algorithm to analyse the collected sEMG signals from the residual limb, the user's motion can be monitored, which will assist the prosthesis to better help the disabled people to complete some basic interactions in daily life.

The sEMG signals, which are non-stationary, represent the sum of subcutaneous motor action potentials generated through muscular contraction[1], and is one of the main physical signals for intelligent algorithm to recognize motion intention. However, in the application of using sEMG signals, the key problem is how to develop an algorithm to distinguish sEMG signals in different states. At present, the literature on sEMG-based gesture recognition or prosthesis control primarily focuses on feature engineering[2], which aims to distinguish sEMG signals by feature recognition. In years of studies, some effective feature combinations have been proposed in time domain and frequency domain. For example, K. Englehart [1] et al.[3] proposed time-domain feature combination (TD), and R. N. Khushaba et al.[4] proposed enhanced TD feature combination, which had good performance in their respective data sets. However, for the same model, selected the different features, the classification accuracy will be significantly different. Therefore, for traditional methods, it is particularly important to choose a feature that can easily distinguish different gestures, but maintain a small variance between the same gestures. Nevertheless, the process of designing and selecting features can be complicated and the combinations of features are diverse, leading to the increasing of workload and dissatisfied result. So far, some scholars have begun to consciously use deep learning for sEMG signals recognition to avoid artificially design features and to reduce workload.

CNN has powerful feature extraction function. Some researchers have used CNN to distinguish different EMG signals. Because CNN is mainly used to process image information, U. C. Allard et al. first converted the sEMG signals into a myoelectric spectrum by fast Fourier transform[5], then fed the spectrograms of eight channel to the neural network. To solve the problem that training networks requires a lot of data, U. C. Allard et al. extended the content in the literature[5], by using transfer learning[6] to reduce the data for single person. They made further extension later in use of wavelets that fit better to the signals to perform wavelet transform[7] on the sEMG signals, which further improved the classification accuracy. However, using CNN to classify the sEMG signal cannot take advantage of the timing information of the signals. The signals need to be preprocessed by wavelet transform rather than simple filtering, so that it cannot achieve the purpose of directly inferring gestures based on the sEMG signals.

Aiming at the above problems, in this article, we came up with the combination of LSTM and CNNs, called as LCNN. At the model verification stage, we also introduced CNN-SLTM model and single LSTM model. We proved the feasibility of the LCNN network in application of gesture recognition applications and its superiority compared with the other two models.

This paper is organized as follows: LCNN architecture is

Supported by Chongqing Science and Technology Commission

introduced in section Ⅱ, the acquisition and processing of the dataset in details in section Ⅲ experiment and results in section Ⅳ, the conclusion in section Ⅴ.

Ⅱ. LCNN ARCHITECTURE

LSTMs and CNNs are complementary in their modeling capabilities, as LSTMs are good at temporal modeling, while CNNs are more outstanding in feature extraction. In this paper, we take advantage of the complementarity of CNNs and LSTMs by combing them into one unified architecture, which we call LCNN. The LSTMs model is used to extract timing information in signals. The CNNs model can perform a secondary feature extraction and signal classification. Also, we analyze the effect of adding CNN before the LSTM, which is inspired by literature[8]. In view of the better performance of the LCNN model in the actual measurement, this section mainly introduces the LCNN model.

A. LCNN

As the architecture shown in Figure 1, the LCNN model consists of 2 LSTM layers, 2 one-dimensional convolution layers and 1 output layer. First, we extract time information by passing the input signals through a LSTM layers. We use 2 LSTM layers, and each LSTM layer has 52 cells, and every cell has 64 hidden layers. The output signals of LSTM layers is inputted into the CNN layer that is good at feature extraction. Specifically, we use 2 convolutional layers, the first layer uses 65 of 3*1 convolution kernel, followed by 30 of 3*1 filter for second convolution layer. The non-linear activation function is PReLU[9]. ADAM[10] is utilized for the optimization of the LCNN.

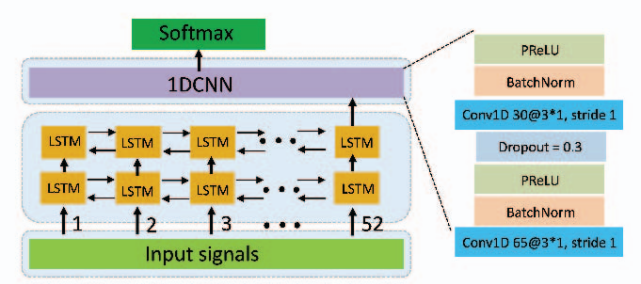

Fig.1. LSTM-CNN Architecture

B. Network optimization

Due to the small data volume used in the training of network, optimation methods are needed to prevent network from over-fitting.

1) Dropout

Dropout is a regularization method, which avoids the over-fitting by reducing the co-adaptation of hidden units[11]. During training, for each sample of the mini-batch, hidden units are randomly deactivated with probability p. In this article, the Dropout in the layers of LSTM and CNN is set to 0.3.

2) Batch Normalization

Batch Normalization (BN) is a regularization technique aiming to maintaining the standard distribution of hidden layer activation values throughout training[12], which can accelerate the convergence of the network. BN accomplishes this goal by normalizing the mean and variance of each dimension of a batch of examples.

Ⅲ. SEMG DATASET ACQUISITION AND PREPROCESSING

One of the main contributions of this article is to provide a EMG dataset, named MyoDataset. It contains 5 gestures of four healthy participants. The gestures include rest, hand close, hand open, wrist flexion, and wrist extension, as shown in Figure 2.

A. Myo armband

In this article, the Myo armband made by Thalmic Labs is selected as EMG signal measurement sensor, as shown in Figure 2. Myo is an 8-channel, dry-electrodes, low-sampling rate (200Hz), low-cost consumer grade sEMG armband, which is convenient to wear and easy to use.

Fig.2. The 5 hand/wrist gestures and Myo armband

B. Time window length

For real-time control, input latency is an important factor to be considered. B. Hudgins et al.[13] first proposed that 300ms was the maximum allowable latency. Although recent studies have shown that the input latency is best maintained within 100-125ms, the performance of the classifier should take priority over speed[14, 15]. Due to the low sampling frequency of Myo armband, too short sampling time will not be enough to predict the gesture, In order to get a reasonable number of samples, we selected the time window length of 300ms to conduct experiment according to the time window length used in literature[7].

C. Recoding labeled data

In this paper, we collected a total of 3 sets of data, two of them are used as training dataset and the rest one is test dataset. For sub-dataset, each participant is asked to hold every gesture for 5 s. The recording of full five gestures for 5 s is referred to as a cycle, with four cycle forming a round.

Signal processing must be applied to efficiently train a classifier on the data recorded by Myo armband. We used a sliding window of 260 ms (52 sampling points) and an incremental window of 25 ms (5 sampling points) to segment the signals, which can increase the training data volume. Based on sectionⅢ-B mentioned, employing windows of 260 ms allows 40 ms for the preprocessing and LCNN classification. The schematic diagram of signal segmentation is shown in Figure 3. The data input format for LCNN is

*batch*seq_len*n_ch,* which *batch* represents the size of data volume inputted into the network; *seq_len* represents the sequence length, which is 52 in this article; *n_ch* represents the number of channels, and 8 channels are used in this article.

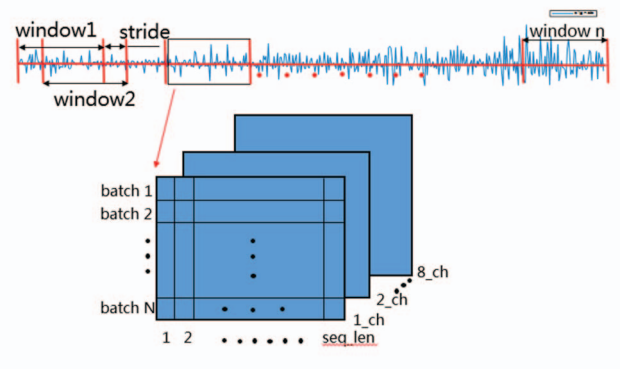

Fig.3. Signal segmentation and the format for LCNN

Ⅳ. EXPERIMENTS AND RESULTSE

We design two experiments using the self-collected MyoDataset dataset and DB5 dataset[16, 17] in NinaPro. The results show the feasibility and performance advantage of the LCNN model, and the accuracy of LCNN and standard methods are compared based on the same dataset.

A. DB5 dataset

The DB5 dataset contains the EMG data of 10 participants, with each participant's data recorded by two Myo armbands. The first Myo armband is placed closer to the elbow with the first electrode on the radio humeral joint, and the second Myo armband is placed just below the first, closer to the hand[16]. The dataset of each participant contains two exercises, namely, Exercise A and B. Exercise A contain 12 basic movements of the fingers; Exercise B contain 8 iosmetric and isotonic hand configurations and 9 basic movements of the wrist. Each movement lasts 5 seconds, repeats 6 times.

B. Contrast experiment of LCNN, LSTM and CNN-LSTM

The contrast experiment conducted in this section is mainly to compare the performance of the LCNN, CNN-LSTM and the single LSTM model based on MyoDataset, and obtain the deep learning model most suitable for gesture recognition according to quantitative indicators.

All three models have a good performance in MyoDataset, and the average accuracy of gesture recognition can reach more than 96%. The best results are obtained on model LCNN (98.14%±1.56% considering the matched group of 4 subjects). Comparable average accuracies are obtained on model CNN-LSTM (98.03%±1.5%) and LSTM (97.99%±1.7%). The results of the classification procedure are shown in figure 4.

From the results, we know that LCNN model perform better, but the accuracies of all three models are very close. Therefore, we makes further comparisons between the models from convergence rate and the accuracy number of gestures from the actual measurement. We just take the dataset of candidate 1 as an example to illustrate the advantages and disadvantages of the models.

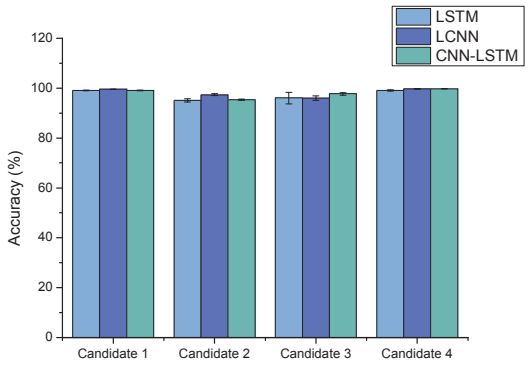

Fig.4. Classification results

Three models are set the same parameters during training, as shown in table 1:

TABLE.1 TRAINING PARAMETERS

Epoch	Batch	Dropout	Samples
30	128	0.3	7600

Three models tend to be stable after 30 cycles, but the convergence rate is different, LCNN has the fastest training time with little fluctuation while CNN-LSTM has the slowest convergence rate with small fluctuation in the whole process, as shown in figure 5. In terms of the convergence speed, LCNN performs better.

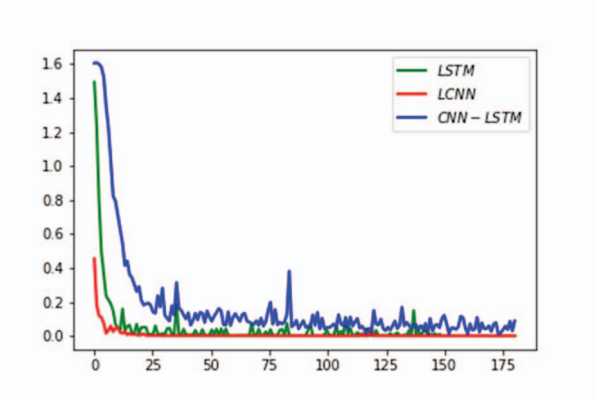

Fig.5 Loss value

In order to further explore the advantages and disadvantages of three models, we shows the actual test results performed by the three models, each gesture has been tested for 20 times, and the actual accuracy number of each gesture are shown in table 2.

TABLE.2 ACCURACY NUMBER OF GESTURES

	rest	Hand open	Hand close	Wrist flexion	Wrist exte-nsion	Average run time (ms)
LSTM	15	17	19	19	18	33
LCNN	**16**	**19**	**20**	**18**	**20**	**27**
CNN-LSTM	15	18	18	19	20	21

Table 2 shows that LCNN model performs best in 5 gestures and the average run time is 27ms, meets the real time requirements. All three models probably may have mistake recognition when a rest gesture follows a longer period of tightly grip (3-4 s), At this moment, the rest gesture is recognized as hand open, which is also the reason that the

accuracy number of rest gesture in the above table is less than 20. When the hand is suddenly released after a long period of grip and restored to the rest state, the arm muscle movement is similar to the open action.

LCNN model first uses the LSTM module to extract the features of the sEMG signal, which can retain the signals's timing information, and the subsequent CNN module can further extract the feature information in the signal. Therefore, LCNN model retains the signal timing information, and also gives play to the advantages of the CNN model to extract the features. On the other hand, the CNN-LSTM model first uses the CNN module to extract features for the signals, which will lose most of the timing information in the signals, and the subsequent LSTM model can only extract the few remaining timing information in the signals, so the performance of CNN-LSTM is relatively poor compared with LCNN.

In summary, LCNN model has better performance in terms of test accuracy, convergence rate, and actual measurement, so we will use LCNN model for follow-up experiments.

From the network optimization in section Ⅱ, the probability of Dropout may affect the performance of the model, so the influence of several different Dropout probabilities on LCNN are compared, as shown in table 3,

TABLE.3 INFLUENCE OF DROPOUT VALUE ON MODEL

Dropout prob.	0	**0.3**	0.5
LCNN	97.96%	**98.14%**	98.02%

It shows that when the Dropout probability is 0.3, the model performs relatively well on the test dataset, so all the Dropout probabilities of the LCNN model used in this article are set to 0.3.

C. Comparison between LCNN and standard methods

In this section, it takes the DB5 dataset to complete the experiment. To keep consistent with the literature[16], the standard classifier adopts LDA and SVM, and the EMG features adopt the feature sets (RMS, TD[3], HIST, mDWT) with best performance in the selected literature[16].

DB5 dataset is collected by two Myo armbands with 16 channels, so the LCNN model's parameter *n_ch* needs to be modified to 16.

We test LCNN and standard methods on Exercise A of DB5 first. When using the data from two Myo armband, the average accuracy of the LCNN model on 10 experimenters was 71.66±5.42%, LDA and SVM are 69.49±5.1%, 67.9±4.7% respectively. The accuracy of each model is also tested for only using one of the Myo armband. The test results are shown in figure 6.

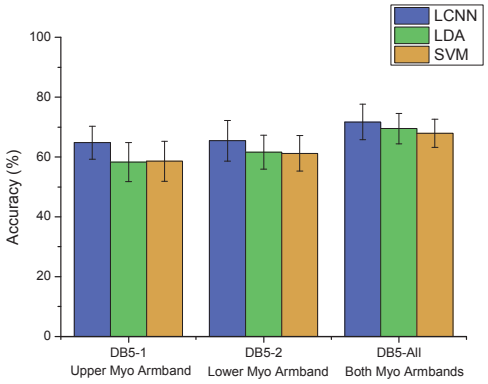

Fig.6 Classification results of exercise A

It shows that LCNN model performs better on this dataset than standard methods.

DB5 dataset provides Exercise B with a larger number of gestures. The accuracy of all models are reduced compare with Exercise A, LCNN (61.4±4.49%), LDA (61.75±3.1%) and SVM (58.27±3.1%). The results are shown in figure 7.

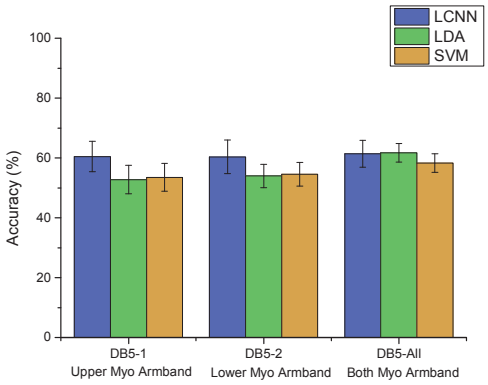

Fig.7 Classification results of exercise B

The results shown in Fig.6 and Fig.7 suggest that LCNN might be better suited to classify sEMG signals. Unsurprising, increasing the amount of gestures systematically degraded the performance of all tested methods. This is likely due to increase the gestures but the samples of each gesture are not changed, and the gestures similarity in Exercise A and Exercise B is very high, so, LCNN model is overfitting.

Ⅴ. CONCLUSION

This paper presents a new model, LCNN, based on LSTM and CNN for gesture recognition. LCNN model does not need to carry out feature design for sEMG signals as the traditional classifiers (LDA, SVM) do. Instead, LCNN architecture uses LSTM to perform temporal modeling, and then passes to CNN layers to extract features, which avoid the complicated feature design process. Two experimental results in this paper prove that the LCNN model is feasible in the application of dynamic gesture recognition.

Future works will focus on the extraction of the commonality between sEMG signals corresponding to different users and the same behavior based on the work done in this article.

ACKNOWLEDGEMENT

The work in this paper is supported by Chongqing Science

and Technology Commission (Grant Nos. cstc2017zdcy-zdyfX0036). Professor Bin Zheng is the corresponding author.

REFERENCE

[1] M. A. Oskoei and H. Hu, "Myoelectric control systems—A survey," *Biomedical Signal Processing & Control,* vol. 2, no. 4, pp. 275-294, 2007.

[2] A. Phinyomark, S. Hirunviriya, C. Limsakul, and P. Phukpattaranont, "Evaluation of EMG feature extraction for hand movement recognition based on Euclidean distance and standard deviation," in *International Conference on Electrical Engineering /electronics Computer Telecommunications and Information Technology*, 2010, pp. 856-860.

[3] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Trans on Biomedical Engineering,* vol. 50, no. 7, p. 848, 2003.

[4] R. N. Khushaba and S. Kodagoda, "Electromyogram (EMG) feature reduction using Mutual Components Analysis for multifunction prosthetic fingers control," in *International Conference on Control Automation Robotics & Vision*, 2012, pp. 1534-1539.

[5] U. C. Allard *et al.*, "A convolutional neural network for robotic arm guidance using sEMG based frequency-features," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2016, pp. 2464--2470.

[6] U. Côté-Allard *et al.*, "Transfer Learning for sEMG Hand Gesture Recognition Using Convolutional Neural Networks," in *IEEE International Conference On Systems, Man and Cybernetics*, 2017.

[7] U. Côtéallard *et al.*, "Deep Learning for Electromyographic Hand Gesture Signal Classification by Leveraging Transfer Learning," 2018.

[8] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4580-4584.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," pp. 1026-1034, 2015.

[10] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Computer Science,* 2014.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[12] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," pp. 448-456, 2015.

[13] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Trans Biomed Eng,* vol. 40, no. 1, pp. 82-94, 1993.

[14] B. Peerdeman *et al.*, "Myoelectric forearm prostheses: state of the art from a user-centered perspective," *Journal of Rehabilitation Research & Development,* vol. 48, no. 6, pp. 719-737, 2011.

[15] T. R. Farrell and R. F. Weir, "The optimal controller delay for myoelectric prostheses," *IEEE Trans Neural Syst Rehabil Eng,* vol. 15, no. 1, pp. 111-118, 2007.

[16] S. Pizzolato, L. Tagliapietra, M. Cognolato, M. Reggiani, H. Müller, and M. Atzori, "Comparison of six electromyography acquisition setups on hand movement classification tasks," *Plos One,* vol. 12, no. 10, p. e0186132, 2017.

[17] M. Atzori *et al.*, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Nature,* vol. 1, pp. 605-10, 2014.