IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2024, vol. 2, no. 2, pp: 5–12

Improved Long Short-term Memory Network for Gesture Recognition

Yuchang Si[1]

Software College, Shenyang Normal University
Shenyang 110034, China
siyuchang@163.com

Abstract. Surface EMG contains a lot of physiological information reflecting the intention of human movement. Gesture recognition by surface EMG has been widely concerned in the field of human-computer interaction and rehabilitation. At present, most studies on gesture recognition based on surface EMG signal are obtained by discrete separation method, ignoring continuous natural motion. A gesture recognition method of surface EMG based on improved long short-term memory network is proposed. sEMG sensors are rationally arranged according to physiological structure and muscle function. In this paper, the finger curvature is used to describe the gesture state, and the gesture at every moment can be represented by the set of different finger curvature, so as to realize continuous gesture recognition. Finally, the proposed gesture recognition model is tested on Ninapro (a large gesture recognition database). The results show that the proposed method can effectively improve the representation mining ability of surface EMG signal, and provide reference for deep learning modeling of human gesture recognition.

Keywords: Surface EMG, Human-computer interaction, Gesture recognition, Long short-term memory network.

1. Introduction

With the continuous development and popularization of human-computer interaction technology, gesture recognition 11, as a natural, intuitive and non-contact interaction method, has attracted more and more people's attention and attention [1,2]. In gesture recognition technology, gesture recognition based on surface EMG has high practicability and wide application scenarios [3]. This technology can recognize the muscle activities of fingers, palms, wrists and other parts of the human body through the acquisition and processing of surface electromyographic signals, so as to realize the recognition and control of gestures. Compared with traditional gesture recognition technologies based on vision or inertial sensors [4], gesture recognition technology based on surface EMG has better robustness and stability, and can achieve efficient and accurate gesture recognition under different illumination and posture conditions. Continuous gesture recognition technology based on surface EMG has been widely applied and promoted in the fields of smart home, smart medical treatment [5] and prosthetic limb control [6], and has become one of the important research directions in the field of human-computer interaction [7,8].

However, most studies focus on the discrete classification of predefined gestures. For example, Nguyen et al. [9] used one-dimensional CNN to classify gestures, and the average recognition rate of 18 gestures was 78.86%. However, with the increase of gesture types and the mutation of signals when switching actions, the recognition accuracy will decrease. In view of these shortcomings, continuous motion estimation based on surface EMG can better improve the application scenarios in practice [10]. At present, continuous motion estimation based on surface EMG signals can be divided into two categories: mapping based on physiological model and regression model. Among them, Hill model is the most commonly used physiological model [11], which uses Hill contraction mechanics, musculoskeletal geometry and joint forward dynamics to establish the mapping relationship between EMG and joint Angle [12]. In the regression model approach, the mapping between the surface EMG signal and the output creates redundancy. In response to these problems, more and more people use deep learning methods to achieve continuous motion estimation [13]. The deep learning methods do not need complex physiological knowledge as the background, and can extract the spatial correlation of surface EMG well. For example, Karatay et al. [14] proposed a multi-output convolution to estimate 16 joints of a finger with an average correlation coefficient (cc) of less than 0.8. However, as the degree of freedom increasing, the accuracy of finger motion estimation decreased.

Aiming at finger curvature, the mapping relationship between surface EMG and finger curvature is established by an improved LSTM model to realize continuous gesture recognition. The experimental results show that the method is correct.

2. Methods

The overall framework is shown in Figure 1. First, the corresponding surface EMG of the forearm is collected while performing the task gesture set. After pretreatment, it is input into the improved STM model to obtain the corresponding finger curvature, and finally the corresponding continuous gesture is obtained.

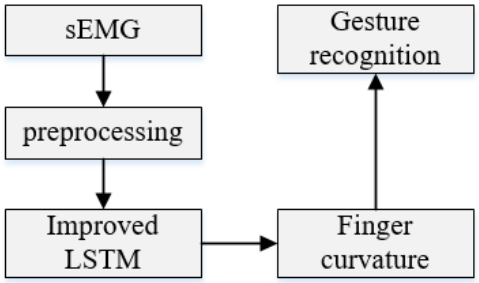

Fig. 1. Overall framework for continuous gesture recognition

2.1. Finger Curvature

Fingers play an almost decisive role in gestures, so the analysis of gestures can be transformed into the analysis of finger states. The finger movement can be regarded as the movement of the metacarpophalangeal joints. Metacarpophalangeal joint is the main control point of finger movement, and the interphalangeal joint bends and extends around the metacarpophalangeal joint. It describes this linkage as finger curvature. Finger curvature C_i, $i = 1, 2, 3, 4, 5$, representing the curvature of 5 fingers respectively. $C_i \in [0, 1]$. Gesture $G = [C_1, C_2, C_3, C_4, C_5]$. The gesture consists of the set of finger curvatures of each finger.

2.2. Improved LSTM

The improved LSTM model is designed in two parts. First, feature extraction is carried out on the input sEMG data through the CNN layer, and a series of feature representations are obtained. These features are then passed to the LSTM layer in chronological order [15-18]. LSTM links the output of the previous CNN layer with the input of the current moment to form a coherent time series. This design makes LSTM better able to deal with time series data processing tasks. Figure 2 shows the overall structure of improved LSTM model.

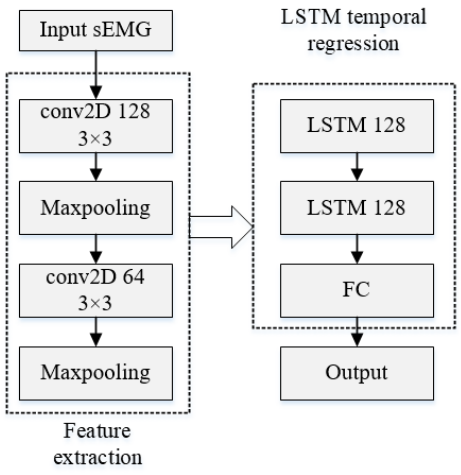

Fig. 2. Overall structure of improved LSTM model

First, the surface EMG obtained is divided into several parts by sliding window method, and each signal is rearranged into EMG $x \in R^{L \times C}$ as the input of CNN. The size of the EMG input data is 20×5, where 20 represents the active window width of 200ms and 5 represents five EMG acquisition channels. In order to determine the most suitable window length, experiments under different window lengths are carried out and comparative analysis is carried out. The CNN layer consists of the two convolution layers, two pooling layers and two fully connected layers. The first and second convolution layers contain 128 and 64 3×3 filters, respectively, using ReLU as the activation function. The first and second pooling layers use 2×2 maximum pooling, respectively. After the first pooling layer, a Dropout layer with parameter 0.5 is added to reduce the risk of over-fitting the model. Adding a Batch Normalization layer after each convolutional layer helps normalize the output of each layer, improving the convergence speed and generalization performance of the model. The eigenvector of CNN output is $f = [f_1, f_2, \cdots, f_n]$.

LSTM [19] constitutes the lower layer of the proposed model. The eigenvector obtained from the CNN layer is used as the input of the layer, and the finger curvature is predicted by the LSTM layer. The basic unit of LSTM network consists of memory gate, input gate and output gate [20]. With forget and input gates, LSTM structures can more effectively determine what information should be forgotten and what should be retained. The update of the LSTM unit at time step t can be described as:

$$f_t = \sigma(W_f \cdot [h_{t-1} \cdot x_t] + b_f). \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1} \cdot x_t] + b_i). \tag{2}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1} \cdot x_t] + b_c). \tag{3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t. \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1} \cdot x_t] + b_o). \tag{5}$$

$$h_t = o_t \cdot \tanh(c_t). \tag{6}$$

Where f_t represents the forgetting gate, i_t is the input gate, o_t is the output gate, c_t is the storage unit, σ is the sigmoid function, and b is the corresponding bias vector. The LSTM network has two LSTM layers and one fully connected layer. The number of LSTM layer is 128.

3. Experimental Results and Analysis

3.1. Parameter Setting

All gesture recognition models in the experiment are implemented based on Python language, built by TensorFlow2.3 library, trained and tested on Intel i7-10750H processor and NVIDIA RTX2070 graphics card. The total training rounds of the gesture recognition model are set to 40, the training batch is set to 64, Adam is selected as the optimizer of the experimental training, the cross entropy loss function (Softmax) is set as the loss function, and the activation function is set as ReLU by default. The initial learning rate of the training stage is 0.0001, and the learning rate drops to half of the original every 10 rounds. In order to ensure the reliability of the experiment, the same hyperparameters are set on the comparison models, and the recognition models are trained three times to obtain the optimal recognition model.

3.2. Results and Discussion

The total number of samples obtained in this paper is 183560, including 52 kinds of actions, and the number of samples of each gesture is 3530. In the experiment, the sample data of electromyography after sliding sampling are divided into training set and test set according to 7:3. In this paper, the sliding window size is set to 200ms, which can achieve the best accuracy. At the same time, considering the network computing cost and the real-time consideration that the sample window length should be less than 300ms. In addition, five comprehensive evaluation indexes are used to evaluate the classification effect of each model: accuracy rate (Acc), precision (Pre), recall rate (Rec), F1-Score (F1) and ROC curve, respectively [21,22].

$$Acc = \frac{Predictedcorrectnumberofsamples}{Totalnumber}. \tag{7}$$

$$Pre = \frac{1}{n}\sum_{i=1}^{n} Pre_i. \tag{8}$$

$$Rec = \frac{1}{n}\sum_{i=1}^{n} Rec_i. \tag{9}$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec}. \tag{10}$$

Where Pre_i is the accuracy rate of each type of sample, and Rec_i is the recall rate of each type of sample. Sample type n is 52.

The evaluation of the model mainly includes two aspects. The first objective is to verify the validity of CNN-Net by selecting the module with the optimal receptive field parameters through detailed ablation experiments, and to verify that the channel attention mechanism can enhance the model accuracy. Secondly, the multi-scale feature fusion network structure proposed in this paper is compared with the traditional feature pyramid network structure to check the contribution of the proposed model components to the model recognition performance.

A. CNN-Net module evaluation

The evaluation of CNN-Net module consists of two parts: one is to test the validity of the spatio-temporal feature extraction module; the other is to explore the influence of different receptive field parameters in the branch of spatial feature extraction on gesture recognition. This section designs the following experiment. Under the removed attention mechanism based on the NinaproDB5 dataset, it explores the performance of S-Net module (only multi-channel CNN feature extraction branch), T-Net module (only bidirectional LSTM feature extraction branch) and ST-Net module (combination of multi-channel CNN and bidirectional LSTM). In addition, convolution kernels of different sizes are set on the network module, and the parameters of convolution kernels with the best performance of the model are selected. Specific steps of the experiment: (1) The original data are converted into EMG sample data by the pretreatment method; (2) As in previous works [23,24], the EMG samples are divided into training sets and test sets as 7:3; (3) Design a gesture recognition model for ablation experiment, which is specifically one-layer feature extraction module (such as S-Net, T-Net and ST-Net feature extraction models, that is, based on STA-Net module, CNN branch or LSTM branch is deleted) series decoding module; (4) The size parameters of convolutional kernel in T-Net and ST-Net are set to [3,5,7], [3,7,11] and [3,11,15] respectively, while there is no convolutional branch in S-Net. (5) The same training set is used to train the above 7 models for 40 rounds, and the test set is used to evaluate the classification effect of the models in various aspects.

The experimental results are shown in Table 1. ST-Net module has the best performance and the highest recognition accuracy reaches 86.91%. The performance of T-Net is significantly better than that of S-Net, which verifies that LSTM network is more able to extract strong discriminant features (that is, time-dependent relationships) from sEMG signals, and when T-Net is integrated into the time-feature extraction branch, it can be found that the recognition accuracy of ST-Net is significantly improved. According to the analysis of different convolution kernel parameters, S-Net with large convolution kernel is superior to S-Net with small convolution kernel in four comprehensive evaluation indexes, but ST-Net has the highest accuracy when the convolution kernel parameters are [3,7,11]. Meanwhile, the convolution kernel parameters of STA-Net are determined in subsequent experiments based on Pre, Rec, $F1$ and network computing costs are [3,7,11].

Table 1. Gesture recognition accuracy rate, precision rate, recall rate and F1 with different convolutional kernel sizes

Methods	kernel	Acc/%	Pre/%	Rec/%	F1
T-Net	none	86.41	92.22	81.11	86.31
S-Net	[3,5,7]	83.07	90.32	76.53	82.85
S-Net	[3,7,11]	84.04	90.74	78.32	84.07
S-Net	[3,11,15]	84.96	91.39	79.66	85.12
ST-Net	[3,5,7]	86.70	92.82	81.19	86.62
ST-Net	[3,7,11]	86.91	92.61	81.42	86.65
ST-Net	[3,11,15]	86.83	92.13	81.73	86.62

On the basis of determining ST-Net with convolution kernel parameter [3,7,11] as the optimal model, this paper will explore the role of channel attention mechanism, that is, comparative analysis of ROC curve performance. The experimental steps are the same as above, but the channel attention module is added to the gesture recognition model, and the convolution kernel parameters are fixed as [3,7,11]. The experimental results are shown in Figure 3. The AUC values of ST-Net and STA-Net (STA-Net is 0.9921, ST-Net=0.9923) are not significantly different. Therefore, this paper chooses to compare and analyze the Loss and Acc curves of the two models to further verify the role of the channel attention mechanism module. As shown in Figure 4, although the Loss curve of STA-Net oscillates significantly at the initial stage of training, the amplitude of oscillation is significantly smaller than that of ST-Net when it reaches a stable state. By observing the Loss curve, it can be found that both models converge at about 26 turns, and the Loss is lower when STA-Net converges. In addition, the Acc curve of STA-Net still oscillates when Loss converges, but the Acc value of STA-Net is significantly higher than that of ST-Net, which verifies that the channel attention mechanism can effectively enhance the model's recognition accuracy of gestures.

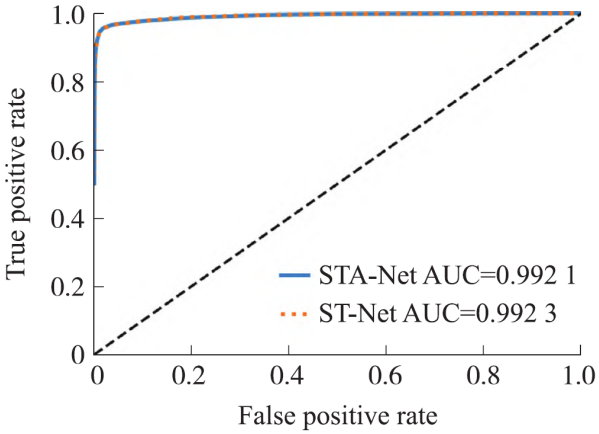

Fig. 3. ROC curve of attention ablation experiment

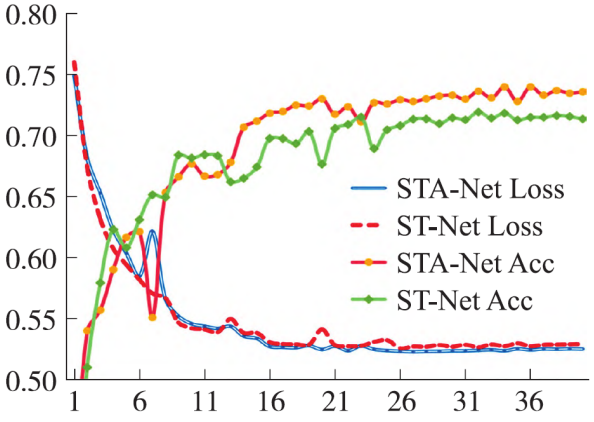

Fig. 4. Loss and Acc curves of attention ablation experiment

B. LSTM-Net model evaluation

On the basis of the above, the evaluation of LSTM-Net model includes two parts: one is to test the effectiveness of multi-scale feature fusion structure; the other is to compare the accuracy of LSTM-Net and advanced gesture recognition models. The following experiments are designed in this section to explore the influence of three kinds of feature pyramid networks on the multi-scale fusion of space-time features, which are STA-Net+FPN, STA-Net+BiFPN and LSTM-Net.

The structure of the LSTM-Net+FPN model is based on the LSTM-Net model, and the feedback loop branch is deleted (that is, 4-layer LSTM-Net module in series+FPN module+decoding module, in which the FPN feature

fusion module performs top-down weighted operations on multi-scale spatio-temporal features). Similarly, the LSTM-Net+BiFPN model replaces the feature fusion module FPN in LSTM-Net+FPN with BiFPN (that is, the 4-layer LSTM-Net module in series+BiFPN module+decoding module). In this model, the BiFPN feature fusion module carries out bidirectional weighted operations on multi-scale spatio-temporal features to improve feature fusion efficiency).

Similar to the gesture model based on sEMG signal proposed in this paper, some existing research results have also achieved accurate recognition of gesture movements. Here, the performance comparison between the proposed space-time deep neural network model and some advanced models is shown in Table 2. Comparison methods include CNN+Attention+IMU [25], CNN+LSTM [26], MV-CNN [27], Compact CNN[28], CNN+transfer learning(CNN+TL) [29], and CNN+CviT [30]. Although the above methods have achieved good results, the model proposed in this paper can provide more accurate identification results than the mainstream methods.

Table 2. Comparison with different models

Method	Window size/ms	Acc/%
LSTM-Net	200	88.5
CNN+Attention+IMU	260	87.1
CNN+LSTM	260	84.57
MV-CNN	200	75.9
Compact CNN	200	70.1
CNN+TL	260	69.08
CNN+CviT	200	76.93

4. Conclusion

In this paper, an improved gesture recognition method based on LSTM network is proposed, which includes CNN-Net to extract spatiotemporal features, attention module to enhance features and multi-scale fusion method based on spatiotemporal features. In each STA-Net module of the network model, multi-channel CNN with different convolution kernel sizes are used to obtain spatial features of different receptor fields, and the time features extracted by bidirectional LSTM are combined to obtain more discriminative spatiotemporal features. Finally, channel attention mechanism and residual structure are added between CNN and LSTM to enhance the characterization ability of features. In particular, a multi-scale combination method is designed to fuse multilevel information of spatio-temporal characteristics. Based on the feature pyramid structure, both resolution and semantic information are taken into account, and the spatial and temporal feature attention module and feedback loop branch are constructed to act as the bridge of spatial and temporal features during iterative updating, which strengthens the complementarity of features. The CNN-Net module, multi-scale fusion structure and decoding module constitute the proposed gesture recognition model LSTM-Net, which reaches the most advanced classification level on the large dataset Ninapro.

Future work focuses on reducing the number of sEMG sensors without degrading model performance, and the information contained in the signal is lost as the number of sEMG sensors decreases. Therefore, accurate recognition of gestures with fewer sEMG sensors remains a challenge. In addition, the robustness of the model is also a difficult problem. It is considered to change the data set from the division between different repetitive actions to the division between different subjects, and it is possible to introduce generative adversarial networks to further enhance the data in future work.

5. Conflict of Interest

All authors disclosed no relevant relationships.

Acknowledgments. None.

References

1. Kosch T, Karolus J, Zagermann J, et al. A survey on measuring cognitive workload in human-computer interaction[J]. ACM Computing Surveys, 2023, 55(13s): 1-39.

2. Azofeifa J D, Noguez J, Ruiz S, et al. Systematic review of multimodal humanCcomputer interaction[C]//Informatics. MDPI, 2022, 9(1): 13.
3. Yin S, Li H, Laghari A A, et al. An Anomaly Detection Model Based On Deep Auto-Encoder and Capsule Graph Convolution via Sparrow Search Algorithm in 6G Internet-of-Everything[J]. IEEE Internet of Things Journal, 2024. doi:10.1109/JIOT.2024.3353337.
4. Jiang Y, Yin S. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment[J]. Computer Science and Information Systems, 2023, 20(4): 1869-1883.
5. O'Brien H L, Roll I, Kampen A, et al. Rethinking (Dis) engagement in human-computer interaction[J]. Computers in human behavior, 2022, 128: 107109.
6. Lv Z, Poiesi F, Dong Q, et al. Deep learning for intelligent humanCcomputer interaction[J]. Applied Sciences, 2022, 12(22): 11457.
7. Whig P, Velu A, Ready R. Demystifying federated learning in artificial intelligence with human-computer interaction[M]//Demystifying federated learning for Blockchain and industrial internet of things. IGI Global, 2022: 94-122.
8. Fan Y, Li H, Sun B. Cycle GAN-MF: A Cycle-consistent Generative Adversarial Network Based on Multifeature Fusion for Pedestrian Re-recognition: Cycle GAN-MF[J]. IJLAI Transactions on Science and Engineering, 2024, 2(1): 1-9.
9. Nguyen P T T, Kuo C H. A Novel Surface Electromyographic Gesture Recognition Using Discrete Cosine Transform-Based Attention Network[J]. IEEE Signal Processing Letters, 31: 266-270, 2023.
10. Sun X, Liu Y, Niu H. Continuous Gesture Recognition and Force Estimation using sEMG signal[J]. IEEE Access, 11: 118024-118036, 2023.
11. Liu Y, Zhang S, Gowda M, et al. Leveraging the properties of mmwave signals for 3d finger motion tracking for interactive iot applications[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2022, 6(3): 1-28.
12. Yin S. Object Detection Based on Deep Learning: A Brief Review[J]. IJLAI Transactions on Science and Engineering, 2023, 1(02): 1-6.
13. Lu Y, Wang X, Gong J, et al. Classification, application, challenge, and future of midair gestures in augmented reality[J]. Journal of Sensors, 2022, 2022.
14. Karatay B, Betepe D, Sailunaz K, et al. CNN-Transformer based emotion classification from facial expressions and body gestures[J]. Multimedia Tools and Applications, 2024, 83(8): 23129-23171.
15. Li Q, Langari R. Myoelectric human computer interaction using CNN-LSTM neural network for dynamic hand gesture recognition[J]. Journal of Intelligent & Fuzzy Systems, 2023, 44(3): 4207-4221.
16. Meng X, Wang X, Yin S, et al. Few-shot image classification algorithm based on attention mechanism and weight fusion[J]. Journal of Engineering and Applied Science, 2023, 70(1): 14.
17. Toro-Ossaba A, Jaramillo-Tigreros J, Tejada J C, et al. LSTM recurrent neural network for hand gesture recognition using EMG signals[J]. Applied Sciences, 2022, 12(19): 9700.
18. Yu J, Li H, Yin S L, et al. Dynamic gesture recognition based on deep learning in human-to-computer interfaces[J]. Journal of Applied Science and Engineering, 2020, 23(1): 31-38.
19. L*ó*pez L I B, Ferri F M, Zea J, et al. CNN-LSTM and post-processing for EMG-based hand gesture recognition[J]. Intelligent Systems with Applications, 2024, 22: 200352.
20. Karnam N K, Dubey S R, Turlapaty A C, et al. EMGHandNet: A hybrid CNN and Bi-LSTM architecture for hand activity classification using surface EMG signals[J]. Biocybernetics and biomedical engineering, 2022, 42(1): 325-340.
21. Wang L, Shoulin Y, Alyami H, et al. A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images[J]. 2022. doi:10.1002/gdj3.162.
22. Teng L, Qiao Y. BiSeNet-oriented context attention model for image semantic segmentation[J]. Computer Science and Information Systems, 2022, 19(3): 1409-1426.
23. Luqman H, ELALFY E. Utilizing motion and spatial features for sign language gesture recognition using cascaded CNN and LSTM models[J]. Turkish Journal of Electrical Engineering and Computer Sciences, 2022, 30(7): 2508-2525.
24. Patel P, Patel N. Dynamic Hand Gesture Recognition for Indian Sign Language using Integrated CNN-LSTM Architecture[J]. International Journal of Next-Generation Computing, 2023, 14(4).
25. Josephs D, Drake C, Heroy A, et al. sEMG gesture recognition with a simple model of attention[C]//Machine Learning for Health. PMLR, 2020: 126-138.
26. Khushaba R N, Phinyomark A, Al-Timemy A H, et al. Recursive multi-signal temporal fusions with attention mechanism improves EMG feature extraction[J]. IEEE Transactions on Artificial Intelligence, 2020, 1(2): 139-150.
27. Wei W, Dai Q, Wong Y, et al. Surface-electromyography-based gesture recognition by multi-view deep learning[J]. IEEE Transactions on Biomedical Engineering, 2019, 66(10): 2964-2973.
28. Chen L, Fu J, Wu Y, et al. Hand gesture recognition using compact CNN via surface electromyography signals[J]. Sensors, 2020, 20(3): 672.
29. C*ô*t*é*-Allard U, Fall C L, Drouin A, et al. Deep learning for electromyographic hand gesture signal classification using transfer learning[J]. IEEE transactions on neural systems and rehabilitation engineering, 2019, 27(4): 760-771.
30. Shen S, Wang X, Mao F, et al. Movements classification through sEMG with convolutional vision transformer and stacking ensemble learning[J]. IEEE Sensors Journal, 2022, 22(13): 13318-13325.

Biography

Yuchang Si is with the Software College, Shenyang Normal University. His research direction is image processing, computer application and AI.