STCN-GR: Spatial-Temporal Convolutional Networks for Surface-Electromyography-Based Gesture Recognition

Zhiping Lai[1,2,3,4,5], Xiaoyang Kang[1,2,3,4,5](✉), Hongbo Wang[1,6](✉), Weiqi Zhang[1], Xueze Zhang[1], Peixian Gong[1], Lan Niu[3], and Huijie Huang[5]

[1] Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai Engineering Research Center of AI and Robotics, MOE Frontiers Center for Brain Science, Laboratory for Neural Interface and Brain Computer Interface, Institute of AI and Robotics, Academy for Engineering and Technology, Fudan University, Shanghai, China
{zplai19,xiaoyang_kang,Wanghongbo}@fudan.edu.cn
[2] Yiwu Research Institute of Fudan University, Yiwu City, China
[3] Ji Hua Laboratory, Foshan, China
[4] Research Center for Intelligent Sensing, Zhejiang Lab, Hangzhou, China
[5] Shanghai Robot Industrial Technology Research Institute, Shanghai, China
[6] Shanghai Clinical Research Center for Aging and Medicine, Shanghai, China

Abstract. Gesture recognition using surface electromyography (sEMG) is the technical core of muscle-computer interface (MCI) in human-computer interaction (HCI), which aims to classify gestures according to signals obtained from human hands. Since sEMG signals are characterized by spatial relevancy and temporal nonstationarity, sEMG-based gesture recognition is a challenging task. Previous works attempt to model this structured information and extract spatial and temporal features, but the results are not satisfactory. To tackle this problem, we proposed *spatial-temporal convolutional networks for sEMG-based gesture recognition (STCN-GR)*. In this paper, the concept of the sEMG graph is first proposed by us to represent sEMG data instead of image and vector sequence adopted by previous works, which provides a new perspective for the research of sEMG-based tasks, not just gesture recognition. Graph convolutional networks (GCNs) and temporal convolutional networks (TCNs) are used in STCN-GR to capture spatial-temporal information. Additionally, the connectivity of the graph can be adjusted adaptively in different layers of networks, which increases the flexibility of networks compared with the fixed graph structure used by original GCNs. On two high-density sEMG (HD-sEMG) datasets and a sparse armband dataset, STCN-GR outperforms previous works and achieves the state-of-the-art, which shows superior performance and powerful generalization ability.

Keywords: Gesture recognition · Surface electromyography · Human-computer interaction · sEMG graph · Spatial-temporal convolutional networks

T. Mantoro et al. (Eds.): ICONIP 2021, LNCS 13110, pp. 27–39, 2021.
https://doi.org/10.1007/978-3-030-92238-2_3

1 Introduction

The technology of human-computer interaction (HCI) allows human to interact with computers via speech, touch, or gesture [14], which promotes the prosperity of rehabilitation robots [3] and virtual reality [10]. With the development of HCI, a new technology called muscle-computer interface (MCI), which used surface electromyography (sEMG) to recognize gestures and realized natural interaction with human, has emerged and used in many applications, especially rehabilitation robots. sEMG is a bio-signal derived from the muscle fibers' action potential [11], which is recorded by electrodes placed on the skin. According to the number of electrodes, sEMG can be categorized into sparse sEMG and high-density sEMG (HD-sEMG), both of them record spatial and temporal changes of muscle activities when gestures are performed. Since sEMG signals provide sufficient information to decode muscle activities and hand movements, gesture recognition based on surface electromyography (sEMG) forms the technical core of non-intrusive MCIs [1].

Gesture recognition based on sEMG can be divided into two categories: conventional machine learning (ML) approaches and novel deep learning (DL) approaches. ML approaches (e.g., SVM) depend heavily on hand-crafted features (e.g., root mean square), which limits their wider application. As revolutionary ML approaches, DL approaches have achieved great success on the sEMG-based gesture recognition tasks. In existing DL-based recognition approaches, sEMG data are represented as images ([4,6,17]) or sequences ([13]), and are fed into convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to extract high-level features for gesture classification tasks. Since the superposition effect of muscle fibers' action potential, neither sEMG images nor sequences can reveal this characteristic. Moreover, most of the previous works only focus on spatial information or temporal information and have not considered them together [4,7,13,17]. The potential spatial-temporal information is not fully utilized, which further limits the performance of these approaches.

To bridge the gaps mentioned above, we proposed *spatial-temporal convolutional networks for sEMG-based gesture recognition* called *STCN-GR*, in which spatial information and temporal information are taken into consideration together by using graph convolutional networks (GCNs) and temporal convolutional networks (TCNs). Instead of the representation of images or sequences for sEMG data, we propose the concept of sEMG graph and use graph neural networks for sEMG-based gesture recognition, in which the topology of the graph can be learned on different layers of networks. To our knowledge, it is the first time that graph neural networks have been applied in the sEMG-based gesture recognition tasks. Our work makes it possible for graph neural networks to be used in sEMG-based gesture recognition tasks and provides a new perspective for the research of sEMG-based tasks.

The main contributions of our work can be summarized as:

- We propose the concept of sEMG graph and use graph neural networks to solve the task of sEMG-based gesture recognition for the first time, in which the connectivity of graph can be learned automatically to suit the hierarchical structure of networks.

- We propose spatial-temporal convolutional networks *STCN-GR* which uses spatial graph convolutions and temporal convolutions to capture spatial-temporal structured information for gesture recognition.
- On three public sEMG datasets for gesture recognition, the proposed model exceeds all previous approaches and achieves the state-of-the-art, which verifies the superiority of STCN-GR.

The remainder of this paper is organized as follows. Section 2 provides an overview of the DL approaches for gesture recognition and the neural networks on graph. Section 3 describes the proposed model. Section 4 is the experimental details followed by the conclusion in Sect. 5.

2 Related Work

DL-Based Gesture Recognition. sEMG signals are time-series data with high correlation in spatial and temporal dimensions, which reflect the activities of gesture-related muscles. Given by a sequence (i.e., window) of sEMG data, the object of the gesture recognition task is to determine the gesture corresponding to these data. As a leading technology to solve gesture recognition tasks, the deep learning approaches are categorized into CNN approaches, RNN approaches, and hybrid approaches. CNN approaches describe each frame of sEMG data as an image to extract spatial features and turn the gesture classification task into an image classification task. The recognition result obtains by performing a simple majority vote over all frames of a window [2,4,17]. There also exist works that use CNNs directly on the whole window data of sEMG [11]. RNN approaches treat sEMG data as vector sequences and directly feed them into RNN to obtain the recognition results [7,8], in which the temporal information is mainly utilized. Hybrid approaches use CNN, RNN, or other experiential knowledge, simultaneously. Hybrid CNN-RNN architecture [6] has been used and achieves 99.7% recognition accuracy. By integrating experience knowledge into deep models [16,20], good outcomes also are achieved.

However, all of these works are failed to capture structured spatial-temporal information, especially spatial information. Since the superposition effect of muscle fibers' action potential, correlations exist between majority channels. Graph-based approaches may be more appropriate for sEMG-based gesture recognition.

Graph Convolutional Network. Graph neural network (GNN) is a kind of network used to solve the tasks based on graph structure, such as text classification [19], recommender system [9], point cloud generation [15], and action recognition [12,18]. As a typical GNN, graph convolutional network (GCN) is the most widely used one and follows two streams: the spectral perspective and the spatial perspective. The spectral perspective approaches consider graph convolution operations in the form of spectral analysis in the frequency domain. The spatial perspective approaches define graph nodes and their neighbors, where convolutional operations are performed directly using defined rules. Our work follows the second stream. Graph convolutions are performed on constructed sEMG graph followed by temporal convolutions. More details will be introduced in Sect. 3.

3 Spatial-Temporal Convolutional Networks

When performing gestures, human muscles (e.g., extensor) in the arm are involved at the same time. Motor units (MU) in muscles "discharge" or "fire" and generate "motor unit action potential" (MUAP) [4]. The superposition of MUAPs forms sEMG signals. Usually, a gesture is relevant to a sEMG window that contains a sequence of frames, i.e., the sEMG signal shows temporal and spatial correlation. In tasks such as skeleton-based action recognition, this spatial-temporal feature can be extracted using graph convolutional networks (GCNs) and temporal convolutional networks (TCNs) jointly [12,18]. Motivated by them, we introduce GCNs and TCNs into sEMG-based gesture recognition and propose our STCN-GR model. The pipeline of gesture recognition using STCN-GR is presented in Fig. 1. Given a sEMG window, spatial graph convolution and temporal convolution will be performed several times alternately after graph construction to obtain high-level features. Then the corresponding gesture category will be obtained by the softmax classifier.

Graph Construction. Surface electromyography is usually acquired as muti-channel temporal signals, To model this complex spatial-temporal structured information appropriately, for the first time, we propose the concept of *sEMG graph* and create a sEMG graph $G = (V, E)$ for sEMG signals.

In constructed sEMG graph, the states of sEMG channels are represented as the vertex set $V = \{v_i | i = 1, 2, ..., N\}$, N is the number of sEMG channels , and each frame of gesture windows shares this graph. Particularly, the states of sEMG channels will be referred to as vertices for distinguishing them from the channels of the feature map below. Given a vertex v_i and its neighbor v_j, the connectivity between vertex v_i and vertex v_j can be denoted as a spatial edge e_{v_i,v_j}, and all the edges form the spatial edge set $E = \{e_{v_i,v_j} | v_i, v_j \in V\}$. It is worth noting that every spatial edge $e \in E$ (dark blue solid line in Fig. 1) can be learned and updated using a learnable offset δ with the parameters of networks dynamically. For each graph convolutional network layer, a unique topology is learned to suit hierarchical structure base on the original graph.

Spatial Graph Convolution. In graph convolutional networks (GCNs), vertices (dark blue circles in Fig. 1) are updated by aggregating neighbor vertices' information along the spatial edges. Each vertex in the sEMG graph will go through multiple layers and be updated several times. In the $(m + 1)^{th}$ layer, the process of vertex feature aggregation can be formulated as [12]:

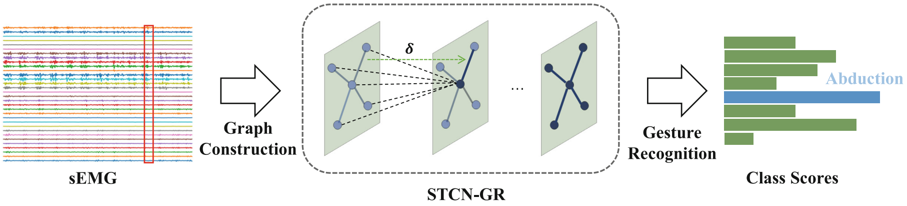

Fig. 1. Pipeline of gesture recognition. Data of a sEMG window (red rectangle) are input to STCN-GR after sEMG graph construction. Then gesture category will be classified by the softmax classifier. The edges of the sEMG graph (dark blue solid lines) are updated using a learnable offset δ with parameters of networks and the vertices (dark blue circles) are also updated according to their neighbors and themselves. (Color figure online)

$$h_i^{m+1} = \sum_{j \in \mathcal{B}_i} \frac{1}{c_{ij}} h_j^m w(l_i(j)) \tag{1}$$

where h_i^{m+1} is the feature representation of vertex i through the aggregation of the $(m+1)^{th}$ layer, $i = 1, 2, ..., N$, $m = 0, 1, 2, ..., M-1)$, N and M are the numbers of vertices and the total number of graph convolution layers. h_i^0 denotes the initial state of vertex i. c_{ij} is a normalization factor. $w(\cdot)$ is the weighting function, which is similar to the original convolution. l_i is a mapping function to map vertex j with a unique weight vector [12]. $\mathcal{B}_i$ is the neighbors of vertex i. It can be considered that neighbors are connected to each other. For a standard 3 × 3 convolution operation, the number of neighbors $|\mathcal{B}|$ can be considered as 9. More generally, the adjacency relationship of vertices can be denoted as an *adjacency matrix*. The spatial graph convolution can be rewritten as [12] in a matrix form:

$$H_{m+1} = \sum_{k=1}^{K} W_{m+1,k}(H_m \tilde{A}_{m+1,k}) \tag{2}$$

where $H_m \in \mathbb{R}^{C_m \times T \times N}$ is the input feature map, $H_{m+1} \in \mathbb{R}^{C_{m+1} \times T \times N}$ is the output feature map after aggregation. C, T and N are the channels of the feature map, length of the window and the number of vertices, respectively. K is the spatial kernel size of the graph convolution, in our work, it is set to 1. $W_{m+1,k} \in \mathbb{R}^{C_{m+1} \times C_m}$ is a weight matrix that can realize a mapping: $\mathbb{R}^{C_m} \rightarrow \mathbb{R}^{C_{m+1}}$. $A_{m\tilde{+}1,k} \in \mathbb{R}^{N \times N}$ is the adjacency matrix, to note that, $\tilde{A}_{m,k}$ denotes the "soft" connectivity of sEMG graph learned by the networks, which is a significant improvement compared with original graph convolutional network that uses "hard" fixed topology. The connectivity of sEMG graph is parameterized and can be optimized together with the other parameters of networks, which increases the flexibility of the networks. $\tilde{A}_{m,k}$ is calculated by:

$$\tilde{A}_{mk} = \bar{A}_{mk} + \varDelta A_{mk} \tag{3}$$

where $\bar{A}_{mk} = \Lambda_{mk}^{-\frac{1}{2}} A_{mk} \Lambda_{mk}^{-\frac{1}{2}}$, A_{mk} is the connection relationship between vertices (includes self loops), and the fully connected relationship is used as the basic topology in our work. $\Lambda_{mk}^{i,i} = \sum_j A_{mk}^{i,j}$ is the normalized diagonal matrix. As for ΔA_{mk}, it can be regarded as a supplement of $\bar{A}_{mk}$, each element $\delta_{i,j}$ of ΔA_{mk} is a learnable parameter that learns an offset for each spatial edge $e_{i,j}$ and captures import information for gesture recognition (illustrated in Fig. 1). After M updates, vertices in the spatial graph include task-related information. Combined with temporal features, networks can obtain high-level features, which is beneficial for gesture classification.

We can find that graph convolution is similar to traditional convolutional operation, but graph convolution is more flexible, its neighbors can be determined according to actual situation or tasks (traditional convolution has only local grid neighbors), that's why it can achieve good performance on the gesture recognition tasks.

Temporal Convolution. For a T-frames data of a sEMG window, a spatial feature map $S \in \mathbb{R}^{C \times T \times N}$ is obtained after graph convolution is finished, and it is input to a temporal convolution network (TCN) to extract the temporal features. At this stage, temporal feature extraction which uses temporal convolution operation is performed on every vertex (i.e., state sequence $s \in \mathbb{R}^{C \times T \times 1}$). In practice, $K \times 1$ convolution kernel is used to perform temporal convolution, K and 1 are the kernel size along the temporal axis and spatial axis, respectively. By changing the kernel size, the receptive field in the temporal dimension can be adjusted, which means that it can process sequences of arbitrary lengths. In our work, $K = 9$, the stride of 1, and zero paddings are utilized. Using dilated convolutions and stacking TCN layers, history information can be seen in the current time step. However, unlike [13] performs standard temporal convolution operations on the overall sequence, for convenience, our temporal convolution operations use simple convolution operations along the time dimension without dilated convolutions or causal convolutions. In this way, temporal convolutions can be simple enough to be embedded anywhere.

Spatial-Temporal Convolutional Networks. Our spatial-temporal convolutional networks for gesture recognition (STCN-GR) follow similar architectures as [12,18]. As shown in Fig. 2, a basic spatial-temporal convolution block (STCB, box with blue dashed line) includes one GCN block and one TCN block to capture spatial and temporal information together. Besides, batch normalization (BN) layers and ReLU layers are followed to speed up convergence and improve the expression ability of networks. Residual blocks (RBs) are used to stabilize the training, which uses 1×1 kernels to match input channels and out channels (if need).

The overall architecture of networks is shown in Fig. 3. The STCN-GR is stacked by M STCBs, in our work, $M = 4$. c_1, c_2, c_3, c_4 denote the number of output channels of STCBs, which are set to $4, 8, 8, G$ (the number of gestures), respectively. A global average pooling (GAP) layer is added after the last STCB

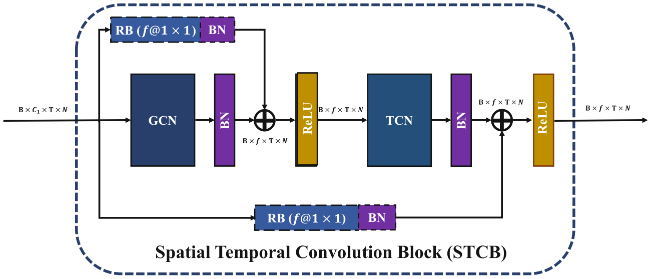

Fig. 2. Illustration of spatial-temporal convolution block (STCB). GCN and TCN are the graph convolutional network and temporal convolutional network, respectively, and both of which are followed by batch normalization (BN) and ReLU. RB stands for residual connection block. (Color figure online)

to improve generalization ability and get the final features, which replaces the full connection layer and reduces the number of parameters. Then, an optional dropout operation is performed. Through a softmax layer, class-conditional probability $\log p(y_j|x_i, \theta)$ can be get to predict gestures. The loss of the i_{th} sample is defined as:

$$\mathcal{L}_i = -\sum_{j=1}^{G} \mathbb{1}_i(y_j) \log p(y_j|x_i, \theta) \tag{4}$$

Where $\mathbb{1}$ is the indicator function, G is the number of gestures, y_j is the j_{th} labels. x_i and θ are the input sEMG signals and parameters of networks, respectively.

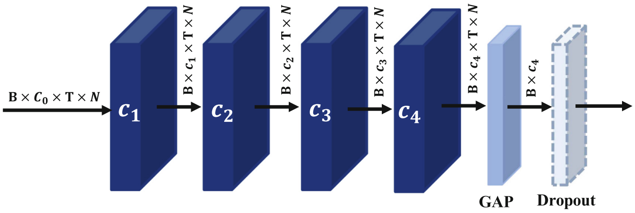

Fig. 3. The overall architecture of STCN-GR. STCN-GR comprises $M(M = 4)$ STCBs, a global average pooling (GAP) layer, and an optional dropout layer. In this architecture, c_1, c_2, c_3, c_4 are 4, 8, 8, G, respectively.

4 Experiments

4.1 Datasets and Settings

To evaluate the performance of STCN-GR, experiments are conducted on three sEMG datasets for gesture recognition: CapgMyo DB-a, CapgMyo DB-b and BandMyo. Different experiments are performed on these datasets to illustrate the superior performance of the proposed STCN-GR.

CapgMyo. CapgMyo [1] is a high-density surface electromyography (HD-sEMG) database for gesture recognition, which is recorded by two-dimensional arrays (8 × 16, total 128) of closely spaced electrodes. This dataset sampled 1000 Hz has three sub-datasets: DB-a, DB-b, and DB-c. 18, 10 and 10 subjects are recruited for DB-a, DB-b and DB-c, respectively. DB-a is designed for evaluating the intra-session performance and fine-tuning hyper-parameters of models, DB-b and DB-c are used for inter-session and inter-subject evaluation [1]. In this work, to compare performance with most existing works, DB-a and DB-b are used. DB-a and DB-b both contain 8 isometric and isotonic hand gestures, each gesture in them is held for 3–10 s and 10 trials are performed for each gesture. We followed the pre-processing procedure like [1,7,17] and used the preprocessed data which use a 45–55 Hz second-order Butterworth band-stop filter to remove the power-line interference and only include the middle one-second data, 1000 frames of data for each trial.

BandMyo. BandMyo dataset [20] is a sparse armband dataset collected by a Myo armband wore on the forearm. This dataset is comprised of finger movement and wrist movement, other movements, a total of 15 gestures. Significantly, the Myo armband just has 8 channels, which means much fewer channels compared with HD-sEMG. 6 subjects are recruited to perform all of 15 gestures by following video guidance, and all of 15 gestures are performed one by one in a trial. When a trial is finished, the armband is taken off and participants will be given a short rest. Briefly, participants wear the armband again and the acquisition process is repeated 8 times (i.e., 8 trials). From the acquisition process, we can know that domain shift [7] exists in one subject's data for the slight change of armband position. No preprocessing is used on this dataset, which is different from CapgMyo datasets.

Experimental Settings. All experiments were conducted on a Linux server (16 Intel(R) Xeon(R) Gold 5222 CPU @ 3.80 GHz) with a NVIDIA GeForce RTX 3090 GPU. All the details in this paper are implemented by using the PyTorch deep learning framework.

For all the experiments, STCN-GR was trained using Adam optimizer, and a weight decay of 0.0001. The base learning rate was set to 0.01 and was divided by 10 after the 5th, 10th, and 25th epochs on three datasets. For CapgMyo DB-a and DB-b, the number of epochs and the batch size were 30 and 16, respectively. 30 epochs and a batch size of 32 were utilized on BandMyo. To get enough samples, the sliding window strategy is used like most works [6,7,11,20]. The window size and window step are 150 ms and 70 ms [7] on CapgMyo DB-a and DB-b, respectively. Since the detailed parameters and training details are not clear [20], 150 and 10 are set as window size and window step on BandMyo. Following the same evaluation method [1,2,6,16,17] on CapgMyo, the model is trained on the odd trials and tested on the even trials. The same evaluation method is used on BandMyo like [20]. Before training, all sEMG signals were normalized in the temporal dimension. No pre-training process has been adopted in our experiments, which is different from other works [2,4,6,17].

4.2 Comparison Results

To evaluate the overall performance of the proposed model in this paper, we compare it with existing approaches on all three datasets. The best results reported literatures are summarized in Table 1, which are all from the latest approaches or the existing state-of-the-art approaches that can be found. Comparison results show that our STCN-GR achieves state-of-the-art performance on all three sEMG datasets, which verifies the superiority of the proposed model.

Table 1. Comparison results on three datasets. The results of the other approaches are the best results reported in literatures. The results in bold show that STCN-GR achieves the best performances on all three datasets.

	Accuracy (%)		
	CapgMyo DB-a	CapgMyo DB-b	BandMyo
GengNet [4]	99.5	98.6	57.8
DMA [1]	99.5	98.6	–
SSL-GR [2]	99.6	98.7	–
CNN-RNN [6]	99.7	–	–
MS-CNN [17]	99.7	–	–
2sRNN [7]	97.1	97.1	–
SA-CNN [5]	96.1	–	–
SVM [20]	71.0	70.8	59.4
RF [20]	83.2	76.2	68.1
STF-GR [20]	91.7	90.3	71.7
STCN-GR (ours)	**99.8**	**99.4**	**75.8**

The experimental results in Table 1 show that, on three public sEMG datasets for gesture recognition (CapgMyo DB-a, CapgMyo DB-b and BandMyo), the results of STCN-GR are superior to that of previous best approaches by **0.1%**, **0.7%** and **4.1%**, respectively. Since accuracy on CapgMyo DB-a is almost saturated [5], it is a significant improvement on this dataset, though the improvement is only 0.1%. The same reason can be found on CapgMyo DB-b. Since the results of the other approaches shown in Table 1 are the best ones in their reports, which means the window size may be 200 ms, 300 ms, even the entire trail. However, just 150 ms is used as window size on CapgMyo in our STCN-GR, and the best performance is achieved. In other words, better performance is achieved using less data, which shows advantages both in accuracy and speed. Due to inter-session domain shift [7] which is a very common phenomenon in practical applications, gesture recognition can be more difficult on the BandMyo dataset compared with the CapgMyo dataset, and our STCN-GR also achieves the state-of-the-art, which shows the powerful generalization ability.

4.3 Ablation Study

We examine the effectiveness of the proposed components by conducting experiments on the first subject of three datasets. As is shown in Fig. 4. "stcn-gr" is the proposed complete model STCN-GR, "tcn" and "gcn" are the models that remove GCN and TCN components, respectively. Particularly, 1×1 convolutions are applied to replace the GCNs to match the output channels. As seen from Fig. 4, the "stcn-gr" always outperforms the other two models. As the number of STCBs (Fig. 2) increases, the performance of "gcn" gradually approaches "stcn-gr", which indicates the core role of GCNs based on the learnable graph. What's more, the performance of "stcn-gr" decreases on some datasets (e.g., CapgMyo DB-b) with the number of layers deepens, and the reason for it may be overfitting. From this ablation experiment, it can be seen that STCN-GR with 4 STCBs performs well on three datasets, which achieves superior performance while maintains the uncomplicated structure of the networks.

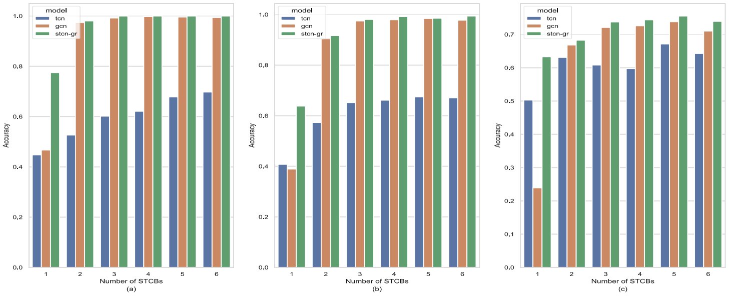

Fig. 4. Ablation study on three datasets. $(a) \sim (c)$ are the results on CapgMyo DB-a, Capgmyo DB-b and BandMyo, respectively. The "stcn-gr" is the proposed complete model, the "tcn" and the "gcn" are the models that remove GCNs and TCNs, respectively.

4.4 Visualization of the Learned Graphs

Figure 5 gives an illustration of learned adjacency matrices by our model based on the first subject of BandMyo. The far left is the original graph, which are followed by learned graphs of 4 STCBs. The darker color represents the stronger connectivity. The visualization of learned graphs indicates that the connectivity with significant vertices (i.e., sEMG channels) will be strengthened, e.g., the vertex 0 and vertex 6, while the connectivity with insignificant vertices will be weakened with the deepening of network layers. Hence, important information will be gathered on a small number of vertices, which is significant for the following classification task.

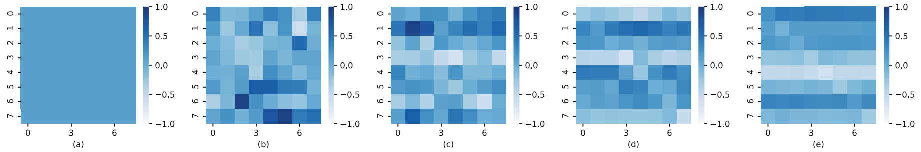

Fig. 5. Visualization of the learned graph. (a) is the original adjacency matrix, $(b) \sim (e)$ are adjacency matrices learned by 4 STCB layers of STCN-GR. The darker color indicates stronger connectivity. (Color figure online)

5 Conclusion

In this paper, we propose the spatial-temporal convolutional networks for sEMG-based gesture recognition. The concept of sEMG graph is first proposed by us to describe structured sEMG information, which provides a new perspective for the research of sEMG-based tasks. The novel learnable topology of graph can adjust the strength of connectivity between sEMG channels and gathers important information on a small number of vertices. Spatial graph convolutional are performed on the constructed sEMG graph followed by temporal convolution. The proposed networks can fully utilize spatial-temporal information and extract task-related features. The experimental results show that our model outperforms the other approaches and achieves the state-of-the-art on all three datasets. In our feature work, we will concentrate on solving the domain adaptation problem [1,7], includes the adaptation of inter-session domain shift and inter-subject domain shift. Based on this work, the self-supervised and semi-supervised learning framework will be also taken into our consideration.

Acknowledgments. This research is supported by the National Natural Science Foundation of China, grant no. 61904038 and no. U1913216; National Key R&D Program of China, grant no. 2021YFC0122702 and no. 2018YFC1705800; Shanghai Sailing Program, grant no. 19YF1403600; Shanghai Municipal Science and Technology Commission, grant no. 19441907600, no.19441908200, and no. 19511132000; Opening Project of Zhejiang Lab, grant no. 2021MC0AB01; Fudan University-CIOMP Joint Fund, grant no.FC2019-002; Opening Project of Shanghai Robot R&D and Transformation Functional Platform, grant no. KEH2310024; Ji Hua Laboratory, grant no. X190021TB190; Shanghai Municipal Science and Technology Major Project, grant no. 2021SHZDZX0103 and no. 2018SHZDZX01; ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology.

References

1. Du, Y., Jin, W., Wei, W., Hu, Y., Geng, W.: Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation. Sensors **17**(3), 458 (2017)
2. Du, Y., et al.: Semi-supervised learning for surface EMG-based gesture recognition. In: IJCAI, pp. 1624–1630 (2017)

3. Fan, Y., Yin, Y.: Active and progressive exoskeleton rehabilitation using multi-source information fusion from EMG and force-position EPP. IEEE Trans. Biomed. Eng. **60**(12), 3314–3321 (2013)
4. Geng, W., Du, Y., Jin, W., Wei, W., Hu, Y., Li, J.: Gesture recognition by instantaneous surface EMG images. Sci. Rep. **6**(1), 1–8 (2016)
5. Hao, S., Wang, R., Wang, Y., Li, Y.: A spatial attention based convolutional neural network for gesture recognition with HD-sEMG signals. In: 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM), pp. 1–6. IEEE (2021)
6. Hu, Y., Wong, Y., Wei, W., Du, Y., Kankanhalli, M., Geng, W.: A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition. PloS one **13**(10), e0206049 (2018)
7. Ketykó, I., Kovács, F., Varga, K.Z.: Domain adaptation for sEMG-based gesture recognition with recurrent neural networks. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2019)
8. Koch, P., Brügge, N., Phan, H., Maass, M., Mertins, A.: Forked recurrent neural network for hand gesture classification using inertial measurement data. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2877–2881. IEEE (2019)
9. Monti, F., Bronstein, M.M., Bresson, X.: Deep geometric matrix completion: a new way for recommender systems. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6852–6856. IEEE (2018)
10. Muri, F., Carbajal, C., Echenique, A.M., Fernández, H., López, N.M.: Virtual reality upper limb model controlled by EMG signals. In: Journal of Physics: Conference Series, vol. 477, p. 012041. IOP Publishing (2013)
11. Rahimian, E., Zabihi, S., Atashzar, S.F., Asif, A., Mohammadi, A.: Xceptiontime: independent time-window xceptiontime architecture for hand gesture classification. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1304–1308. IEEE (2020)
12. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019)
13. Tsinganos, P., Cornelis, B., Cornelis, J., Jansen, B., Skodras, A.: Improved gesture recognition based on sEMG signals and TCN. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1169–1173. IEEE (2019)
14. Turk, M.: Perceptual user interfaces. In: Earnshaw, R.A., Guedj, R.A., Dam, A., Vince, J.A. (eds.) Frontiers of Human-Centered Computing, Online Communities and Virtual Environments, pp. 39–51. Springer, Heidelberg (2001). https://doi.org/10.1007/978-1-4471-0259-5_4
15. Valsesia, D., Fracastoro, G., Magli, E.: Learning localized generative models for 3D point clouds via graph convolution. In: International Conference on Learning Representations (2018)
16. Wei, W., Dai, Q., Wong, Y., Hu, Y., Kankanhalli, M., Geng, W.: Surface-electromyography-based gesture recognition by multi-view deep learning. IEEE Trans. Biomed. Eng. **66**(10), 2964–2973 (2019)
17. Wei, W., Wong, Y., Du, Y., Hu, Y., Kankanhalli, M., Geng, W.: A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface. Pattern Recogn. Lett. **119**, 131–138 (2019)

18. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
19. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7370–7377 (2019)
20. Zhang, Y., Chen, Y., Yu, H., Yang, X., Lu, W.: Learning effective spatial-temporal features for sEMG armband-based gesture recognition. IEEE Internet Things J. **7**(8), 6979–6992 (2020)