China Instrument and Control Society
Instrumentation
INSTRUMENTATION, Vol. 11, No. 4, December 2024 https://doi.org/10.15878/j.instr.202400228

Article

A Two-Stream Hybrid Spatio-Temporal Fusion Network For sEMG-Based Gesture Recognition

Ruiqi Han[1], Juan Wang[1,*], Jia Wang[2]

[1] School of Mechanical and Electrical Engineering, Xi'an University of Architecture and Technology, Xi' an, Shaanxi 710005, China

[2] Shanxi Key Laboratory of Nanomaterials and Nanotechnology, Xi 'an University of Architecture and Technology, Xi' an 710005, China

[*] Corresponding author email: juanwang618@126.com

Abstract: With the advancement of human-computer interaction, surface electromyography (sEMG) -based gesture recognition has garnered increasing attention. However, effectively utilizing the spatio-temporal dependencies in sEMG signals and integrating multiple key features remain significant challenges for existing techniques. To address this issue, we propose a model named the Two-Stream Hybrid Spatio-Temporal Fusion Network (TS-HSTFNet). Specifically, we design a dynamic spatio-temporal graph convolution module that employs an adaptive dynamic adjacency matrix to explore the spatial dynamic patterns in the sEMG signals fully. Additionally, a spatio-temporal attention fusion module is designed to fully utilize the potential correlations among multiple features for the final fusion. The results indicate that the proposed TS-HSTFNet model achieves 84.96% and 88.08% accuracy on the Ninapro DB2 and Ninapro DB5 datasets, respectively, demonstrating high precision in gesture recognition. Our work emphasizes the importance of extracting spatio-temporal features in gesture recognition and provides a novel approach for multi-source information fusion.

Keywords: gesture recognition; deep learning; two-stream spatio-temporal; feature fusion; dynamic neighbor matrix

Citation: Ruiqi Han, Juan Wang, Jia Wang."A Two-Stream Hybrid Spatio-Temporal Fusion Network For sEMG-Based Gesture Recognition. " Instrumentation 11, no.4 (December 2024). https://doi.org/10.15878/j.instr.202400228

1 Introduction

Human-computer interaction is essential for enhancing the efficiency of human-machine collaboration[1,2]. Hand gesture recognition (HGR), as an efficient and intuitive interaction, has become an essential branch in the field of human-computer interaction[3,4] . HGR research primarily relies on techniques such as visual images, surface electromyography (sEMG), and inertial signals[5-7] . However, vision-based HGR methods are susceptible to environmental factors like illumination changes, complex backgrounds, and occlusions, which limit their recognition accuracy. On the other hand, inertia-based HGR methods exhibit advantages under varying lighting conditions, but their limitations in capturing physiological characteristics and potential latency issues seriously affect their ability to respond to rapid or continuous gestures.

In contrast, sEMG-based HGR has demonstrated significant application potential due to its excellent signal-to-noise ratio and rapid response time[8-10] . As anon-invasive electrophysiological signal, sEMG contains rich movement and physiological information[11-13], which can reveal muscle activities and accurately capture intrinsic differences between similar gestures. Consequently, sEMG has become a crucial tool for understanding and decoding gesture movement intentions, occupying a unique position in human-computer interaction.

Analyzing and extracting representative sEMG features is challenging due to the abundant temporal and spatial information in sEMG signals. Existing characterization learning methods primarily focus on both temporal and spatial domains. In gesture recognition research, capturing temporal features typically relies on

recurrent neural networks (RNNs)[14,15], while extracting spatial features is primarily achieved through convolutional neural networks (CNNs) [16,17] . These networks play crucial roles in exploring the temporal dynamics and spatial distribution characteristics of sEMG signals. Atzori et al. [18] designed a four-layer shallow CNN model to extract spatial representations of sEMG for recognizing 52 gestures, but its average recognition accuracy was relatively low at 66.59%. However, sEMG is a signal with significant temporal dependence, which poses a challenge for CNN models in effectively capturing spatial representations. Since CNN excel at handling static spatial features, the time-series characteristics of sEMG signals necessitate the model that not only comprehends the features of the signal at a single time point, but also capture the dynamic patterns as the signal evolves. Therefore, Zhang et al.[19] introduced an extended short-term transformer feature fusion network called LST-EMG-Net, which achieved an 81.47% recognition accuracy for 12 gestures using the Ninapro DB2 dataset. However, RNN models have difficulty capturing local temporal patterns in sEMG signals, which limits their effectiveness in improving gesture recognition accuracy.

Furthermore, understanding the relationship between hand muscle groups and gesture intent is equally crucial for extracting sEMG features. Graph convolutional networks (GCN) excel at modelling spatial information. Xiong et al. [22] proposed a fusion of biological feature information-based graph convolutional network (FBI-GCN), which constructs adjacency matrices using channel node information and focuses on the potential spatial information between neighbouring nodes of individuals. However, this method of constructing the adjacency matrix relies on prior graph knowledge and struggles to capture spatial dynamic information induced by gesture changes adaptively. In summary, current methods are inadequate in decoupling the multi-level spatio-temporal representations and spatial dynamic information from sEMG signals.

How to effectively integrate multi-level spatio-temporal and spatial dynamic information is another key challenge in improving the accuracy of gesture recognition. Feature fusion technology is an essential part of integrating multi-feature domain information, aiming to fully explore potential relevance and complementary advantages through information interaction among multi-feature domains. Existing feature fusion methods primarily include concatenation fusion[23], additive fusion[24], and attention fusion[21] .For example, Xu et al.[23] proposed a concatenation feature fusion recursive convolutional neural network (CFF- RCNN), which aggregates temporal and spatial features through CNN and Long Short-Time Memory (LSTM), not only retaining spatial information but also obtaining temporal information from the context. Wu et al. [24] proposed a narrow kernel dual-view feature fusion convolutional neural network (NKDFF-CNN), which employs a narrow kernel to extract temporal features of each channel adequately and fuses different features by additive fusion to avoid overfitting. Duan et al. [21] proposed a novel alignment-enhanced interactive fusion model that inputs sEMG feature maps with concatenated channel dimensions into a self-attention mechanism for gesture classification to facilitate effective feature fusion. Although the above three feature fusion methods provide beneficial strategies for integrating information from multi-feature domains, they still face challenges in capturing complex feature details and may miss complementary information provided by multi-feature domains. Consequently, it is crucial to develop a model capable of highly integrating spatio-temporal features. This requires the model to not only focus on the interconnectivity of features, but also handle their heterogeneity, thereby significantly improving the model's representation and generalization capabilities.

To address these challenges, we propose a novel two-stream hybrid spatio-temporal fusion network (TS-HSTFNet), designed to enhance the accuracy and robustness of sEMG-based gesture recognition. The network employs MSTCM and DSTGCN to extract discriminative spatio-temporal features from sEMG data. TS-HSTFNet aims to exploit highly discriminative spatio-temporal feature information from sEMG data using MSTCM and DSTGCN, respectively. Additionally, the STAFM learns and integrates latent correlations between different features to capture deeper feature details. The fused features are then fed into the CM to generate accurate results. Furthermore, the model's effectiveness is validated through comparisons with state-of-the-art methods and ablation experiments. The primary contributions of this paper are as follows:

1. We propose a dynamic graph structure that adaptively capture spatial correlations and dynamics relevant gesture recognition, thereby enhancing the spatial feature extraction capability of GCN.

2. We designed a feature fusion module combining spatio-temporal attention and convolutional neural networks. This module effectively fuses spatio-temporal features from different perspectives and integrates complementary information of sEMG, significantly improving the model's gesture recognition performance.

2 Methodology

2.1 Network structure for TS-HSTFNet

The Network structure of the proposed TS-HSTFNet is illustrated in Fig. 1. TS-HSTFNet mainly consists of MSTCM, DSTGCN, STAFM, and CM. Specifically, MSTCM is designed to capture global and local spatio-temporal information at different levels in the

signal. Additionally, DSTGCN employs a dynamic graph approach to mine potential spatial topological relationships among sEMG multi- channels and extracts spatio-temporal features through temporal convolutional network (TCN) blocks. Then, the two-stream hybrid spatio-temporal features are fed into the STAFM, aiming to capture inter-domain correlation and complementarity, and then extract more discriminative fusion features. Finally, the fused features are passed into the CM to predict the category of the gesture, and the joint loss function is applied to train the model. Next, the proposed TS-HSTFNet framework's individual modules are described as follows:

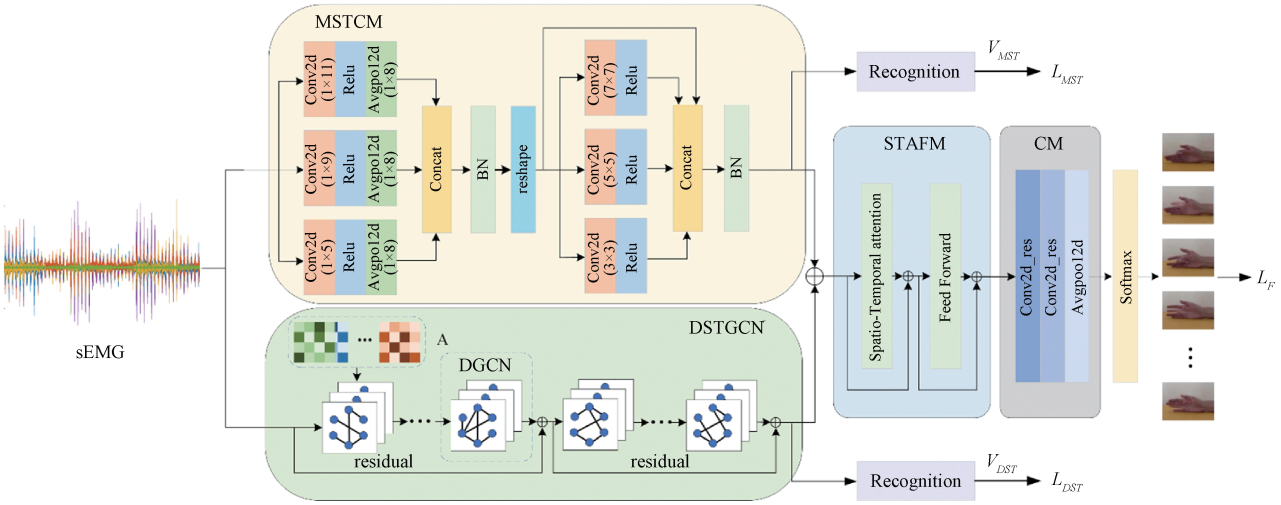

Fig.1 Network of the structure of TS-HSTFNet

1) MSTCM: The model acquires spatio-temporal features across various sensory domains through the implementation of convolutional blocks tailored to distinct scales. Specifically, suppose the input raw sEMG signal is denoted as $X = [x1, \cdots, xN] \in R^{N\times V\times T\times C}$. Here, N represents the number of samples, V is the feature dimension (initially set to 1), T denotes the time dimension, and C is the number of channels. Initially, convolutional kernels of varying sizes are applied to perform multi-scale convolutional operations Fi on the reshaped signal $X \in R^{N\times 1\times C\times T}$ to extract temporal features at distinct levels:

$$\begin{cases} F_1: X \to F_{T1} \in R^{N\times V\times C\times T_1} \\ F_2: X \to F_{T2} \in R^{N\times V\times C\times T_2} \\ F_3: X \to F_{T3} \in R^{N\times V\times C\times T_3} \end{cases} \quad (1)$$

where F_{T1}, F_{T2}, F_{T3} are the temporal features obtained from three different scales of convolution operations. T_1, T_2, T_3 represent the temporal dimension of the output, and $V(V=64)$ is the feature dimension of the output. Subsequently, the ReLU activation function and an average pooling layer are utilized to augment the network's nonlinear capabilities. Ultimately, the three temporal feature F_{T1}, F_{T2}, F_{T3} are fused via a concatenation operation to yield the temporal features $F_{MT} \in R^{N\times V_t\times C\times T}$, $T=T_1+T_2+T_3$, which is then subjected to batch normalization:

$$F_{MT}=BN(Concat(F_{T1}, F_{T2}, F_{T3})) \quad (2)$$

nodes V. Each element a_{ij} in A signifies the strength of the coupling between nodes i and j .

In order to evaluate the functional connectivity between any two electrode channels, we propose a new method for adaptive dynamic learning of relationships between neighboring nodes. This mechanism assigns weights to all edges in the graph so that the model can pair information from between different edges and learn here, Concat (.) denotes the concatenated temporal feature vector, which improves generalization by integrating features across various time scales. BN (.) denotes batch normalization operation.

Next, we reshape the shape of F_{MT} to $F_{MST} \in R^{N\times V_t\times C\times T}$ and then apply three convolution kernels with a kernel size of (k, k) and dilation rate d=2 to perform dilated convolution and ReLU operations on it to further expand the sensory field and obtain the spatio-temporal features $F^1_{MST} \in i^{N\times V_1\times T\times C}$, $F^2_{MST} \in i^{N\times V_2\times T\times C}$ and $F^3_{MST} \in i^{N\times V_3\times T\times C}$. Finally, the output spatio-temporal and temporal features F_{MT} are concatenated and batch normalized to obtain the final spatio- temporal feature $F_{MST} \in R^{N\times V\times T\times C}$, $V=V_1+V_2+V_3$:

$$F_{MST}=BN(Concat(F^1_{MST}, F^2_{MST}, F^3_{MST}, F_{MT})) \quad (3)$$

2) DSTGCN: Recognizing that the movement of gestures relies heavily on the spatial connectivity between electrodes. This study introduces an adaptive dynamic graph-based DSTGCN to capture this dynamic spatial connectivity information. Specifically, an undirected weighted graph $G=(V, A)$ is utilized, with $V=\{v_1, v_2, \cdots, v_n\}$ representing the set of vertices comprising n nodes. The adjacency matrix $A=(a_{ij})_{n\times n}$ delineates the weights of the edges connecting these their weights through a back-propagation mechanism during training. First, an adjacency matrix $A \in i^{C\times C}$ is randomly initialized, where the value of the (i, j) -th element indicates the coupling strength between the i -th electrode and the j -th

electrode channel. The adjacency matrix A indicates the correlation between each channel, taking into account strength. The matrix A is then encoded using the Tanh nonlinear activation function to simulate the dependencies between the different channels as shown below:

$$A\%_{dd}=\sigma(W_2\delta(W_1A\%)) \tag{4}$$

where $A\% \in i^{C\times C}$ is vectorized by A, $W_1 \in i^{\left(\frac{C\times C}{r}\right)\times(C\times C)}$ and $W_2 \in i^{(C\times C)\times\left(\frac{C\times C}{r}\right)}$ are the weight matrices, $\delta(\cdot)$ and $\sigma(\cdot)$ are the ELU and Tanh functions, respectively, and r is the reduction ratio. Thus, the tight neighbor matrix $A_{dd} \in i^{C\times C}$ is obtained by reshaping $A\%_{dd} \in i^{(C\times C)\times 1}$, where the (i, j) -th element value is learnable at the weight update and maps the dependencies between the i -th electrode and j -th electrode channel. We then employ the rectified linear unit (ReLU) to penalize weak channel coupling to obtain the nonnegative adjacency matrix A_{ds}. Ultimately, we construct an autonomously learnable graph structure that accurately captures critical node information through dynamic adaptive tuning.

To fully utilize the temporal information in the data, we incorporate TCN blocks and multiple residual connections based on the dynamic graph convolution. This design strategy not only enables DSTGCN to capture the local dependencies of the data in the temporal dimension but also accurately captures the dynamic evolution characteristics of the data in the spatial dimension. Assuming the input is $X=[x_1, \cdots, x_N] \in R^{N\times V\times T\times C}$, the update process of each layer of DSTGCN can be defined as:

$$F_{DST}=TCN(\sigma(A_{ds}XW))+X, F_{STD}\in R^{N\times V'\times T\times V} \tag{5}$$

where $TCN(\cdot)$ denotes TCN layer, $\sigma(\cdot)$ denotes the Relu activation function, W is the weight of the convolution layer. In this paper, we set the number of DSTGCN layers to 6 to extract spatio-temporal features from the preprocessed sMEG signal.

3) STAFM: In order to effectively fuse global and local spatio-temporal features and integrate complementary information between different feature domains, we propose the STAFM module. The STAFM module addresses the issue of pattern collapse, which can arise from the inconsistency in the distributions of multi-feature domain, thereby ensuring the adequate fusion of information from these domains. Specifically, we extract local spatio-temporal features F_{MST} from the MSTCM and global spatio-temporal features F_{DST} from the DSTGCN. Integration of F_{MST} and F_{DST} yields the

$$F_{ST} = F_t \text{ e } A_s \tag{6}$$

where F_{ST} denotes the outputspatio-temporal feature.

Finally, the spatio-temporal attention-weighted feature F_{ST} is added to the original fusion feature F_{fused} and normalized, and then fed into the feed-feature F_{fused} as input, and subsequently spatio-temporal attention is applied to F_{fused} to enhance the characterization capability of the feature.

First, we use adaptive maximum pooling and adaptive average pooling to independently obtain the maximum and average values for each channel in the feature map Ffused. These two values are then fed into the convolutional layer to learn the channel weights Wmax and Wmax:

$$\begin{cases} W_{max} = Conv(\delta(Conv(Maxpool(F_{fused})))) \\ W_{avg} = Conv(\delta(Conv(Avgpool(F_{fused})))) \end{cases} \tag{7}$$

where $Conv(\cdot)$ denotes the convolution operation, $Maxpool(\cdot)$ denotes the adaptive maximum pooling operation, $Avgpool(\cdot)$ denotes the adaptive average pooling operation and $\delta(\cdot)$ denotes the Relu activation function. After that, the two learned weights are summed and the final channel attention weights A_C are generated by the Sigmoid function and applied to the original feature map F_{fused} to adjust the contribution of each channel:

$$A_C = \sigma(W_{max} + W_{avg}) \tag{8}$$

$$F_t = F_{fused} \text{ e } A_C \tag{9}$$

where F_t denotes the output temporal feature map and e denotes the element-by-element multiplication.

Next, the maximum and average values of each channel are extracted from F_t and spliced over the channel dimensions. Then, the spatial attention weights are learned by a 2D convolutional layer Ws:

$$Ws=Conv(Concat(max(F_t), avg(F_t))) \tag{10}$$

where $max(\cdot)$ and $avg(\cdot)$ denotes the maximum and average values of each channel are extracted from F_t, respectively.

The final spatial attention weights AS were subsequently obtained by a Sigmoid function and applied to the temporal feature map Ft to adjust its spatial dimensions:

$$AS = \sigma(W_S) \tag{11}$$

forward network (FFN) to obtain the fusion features F_f:

$$Ff = (FFN(F_{ST} \boxplus F_{fused})) \boxplus (F_{ST} \boxplus F_{fused}) \tag{12}$$

where $\boxplus$ denotes element-wise addition.

4) CM: Finally, we use two residual convolutional classification prediction. The final high-level feature FCM is obtained by feeding Ff into the CM. We set the filter size of the residual convolution layer to 1×1 and strides of 2, and the kernel size of the average pooling layer to [10,4]:

$$F_{CM} = Conv\_res(Avgpool(F_f)) \tag{13}$$

where $Conv\_res(\cdot)$ denotes the convolution operation with residual structure and $Avgpool(\cdot)$ denotes the average pooling operation.

2.2 Recognition Model

Since we perform spatio-temporal feature extraction from different angles of sEMG, single peak recognition is performed first to ensure that the features extracted by the encoder contribute to the recognition and thus support further feature fusion in the fusion module. The recognition process uses cross-entropy (CE) to obtain the

final loss:

$$V_{MST} = \text{view}\ (\text{Avgpool}\ (\text{Conv}\ (F_{MST}))) \quad (14)$$

$$L_{MST} = \text{CE}\ (V_{MST}, Y) \quad (15)$$

$$V_{DST} = \text{view}\ (\text{Avgpool}\ (\text{Conv}\ (F_{DST}))) \quad (16)$$

$$L_{DST} = \text{CE}\ (V_{DST}, Y) \quad (17)$$

where V_{MST} and V_{DST} are the features extracted by MSTCM and DSTGCN, respectively, and Y is the true label. For the fused advanced features F_{CM}, the recognition process is shown in Eq. (13):

$$L_F = \text{CE}\ (F_{CM}, Y) \quad (18)$$

2.3 Total Loss Function

To further improve the generalization ability and classification accuracy of the model, we used joint loss to train the model and updated our final loss function to:

$$L = \alpha L_{MST} + \beta L_{DST} + L_F \quad (19)$$

where L_{MST} , L_{DST} and L_F are defined in Eq. (15), Eq. (17) and Eq. (18), respectively. α and β are the interacation weights that determine the contribution of each regularization component to the overall loss L.

2.4 Algorithm For TS-HSTFNet

Suppose we are given a labeled gesture recognition dataset $\{X, Y\} = \{xi, yi\}_{i=1}^{N}$, where $X = [x_1, \cdots, x_N] \in R^{N\times1\times T\times V}$ represents multi-channel sEMG signals and $Y = [y_1, \cdots, y_N] \in R^N$ represents true labels. First, we obtain the convolutional features F_{MST} by passing the input x_i through MSTCM according to Eq. (1)-(3). Subsequently, the feature F_{DST} after DSTGCN is obtained according to Eq. (4)-(5). Then, F_{MST} and F_{DST} are integrated into fusion features based on Eq. (6) - (11). Finally, the final high-level feature F_{CM} is obtained by feeding F_f into CM according to Eq. (13). In addition, we calculate the total loss of training L based on Eq. (14) - (19) and use this to guide the parameter updates. Algorithm 1 provides the detailed steps of the model optimization process.

In this paper, we use the cross-entropy loss function as a loss function to measure the difference between the model's predicted label and the true label. The formula is as follows, here we assume that the true label is $Y = [y_1, \cdots, y_N] \in R^N$:

$$L_{ce} = -\Sigma y_{i,c} \log\ (p_{i,c}) \quad (20)$$

where M is the total number of categories, $y_{i,c}$ is the one-hot encoding of the true label of the i -th sample if the sample belongs to category c, $y_{i,c} = 1$, otherwise $y_{i,c} = 0$. $p_{i,c}$ is the probability that the model predicts that the i -th sample belongs to category c. The training process in this paper utilizes the adam optimizer containing 30 epochs initialized with a learning rate of 0.001 and a batch size of 64. All networks in this paper were implemented using Python 3.9 and PyTorch 2.0.1. Training and evaluation were performed on NVIDIA GeForce GPUs (RTX 4060).

Algorithm 1 The Optimizing Procedure of the TS- HSTFNet

Input: the labeled gesture recognition dataset based on sEMG signals $\{X, Y\} = \{x_i, y_i\}_{i=1}^{N}$, the number of training sessions E_p, the batch size for each training session B_a, and the model hyper parameters θ .
Output: optimal set of model parameters θ.
Initialize: Parameters in the proposed TS- HSTFNet θ.
For $e = 1: Ep$ do
  for $b = 1: Ba$ **do**
    Extract a batch of samples $x_{e,b}$ and $y_{e,b}$ from $\{X, Y\}$.
    Based on Eq. (1)-(3) input $x_{e,b}$ into MSTCM to compute the multi-level spatio-temporal features F_{MST}.
    The dynamic spatio-temporal features F_{DST} are computed by inputting $x_{e,b}$ into DSTGCN according to Eq. (4)-(5).
    According to Eq.(6)-(12), F_{MST} and F_{DST} are integrated and fed to STAFM to obtain the fusion feature F_{fused}.
    The final high-level feature F_{CM} is computed by feeding F_{fused} into the CM according to Eq.(13).
    Calculate the final loss L according to Eq.(14)-(20).
    Based on the loss function, the model parameters are updated by the Adam optimizer θ.
  end for
end for

3 Experimental data and Processing

3.1 Experimental data

The Ninapro database, widely utilized in gesture recognition, uses up to four multi-channel physiological signal acquisition devices to gather sEMG data and consists of 10 distinct sub-datasets. In this paper, sub-datasets Ninapro DB2 and DB5 are selected to evaluate the classification performance of the proposed TS-HSTFNet model. Ninapro DB2[26] records 12-channel sEMG signals at a sampling frequency of 2000 Hz using a Delsys Trigno wireless system. It contains 49 gestures from 40 healthy subjects (as shown in Fig. 2 of Exercise B, C, D), with each gesture repeated six times. Ninapro DB5[27] uses two Thalmic Labs MYO armbands for sEMG acquisition at a sampling frequency of 200 Hz. It contains 52 gestures from 10 healthy subjects (as shown in Fig. 2 of Exercise A, B, C), with each gesture repeated ten times.

3.2 Data Processing

Due to the low amplitude, low frequency, high noise, and instability of sEMG signals, precise preprocessing steps are essential. To accurately extract the envelope of the sEMG signal, a first-order Butterworth low-pass filter with a cutoff frequency of 1 Hz is used to filter out high-frequency noise, and a 50- Hz notch filter is applied to reduce industrial frequency interference. Additionally, to further suppress noise, the moving average method is used to smooth the sEMG signal. Finally, Min-Max normalization is applied to normalize the data and reduce gradient explosion during training.

Given the real-time requirements of gesture recognition, a window length of no more than 300ms is

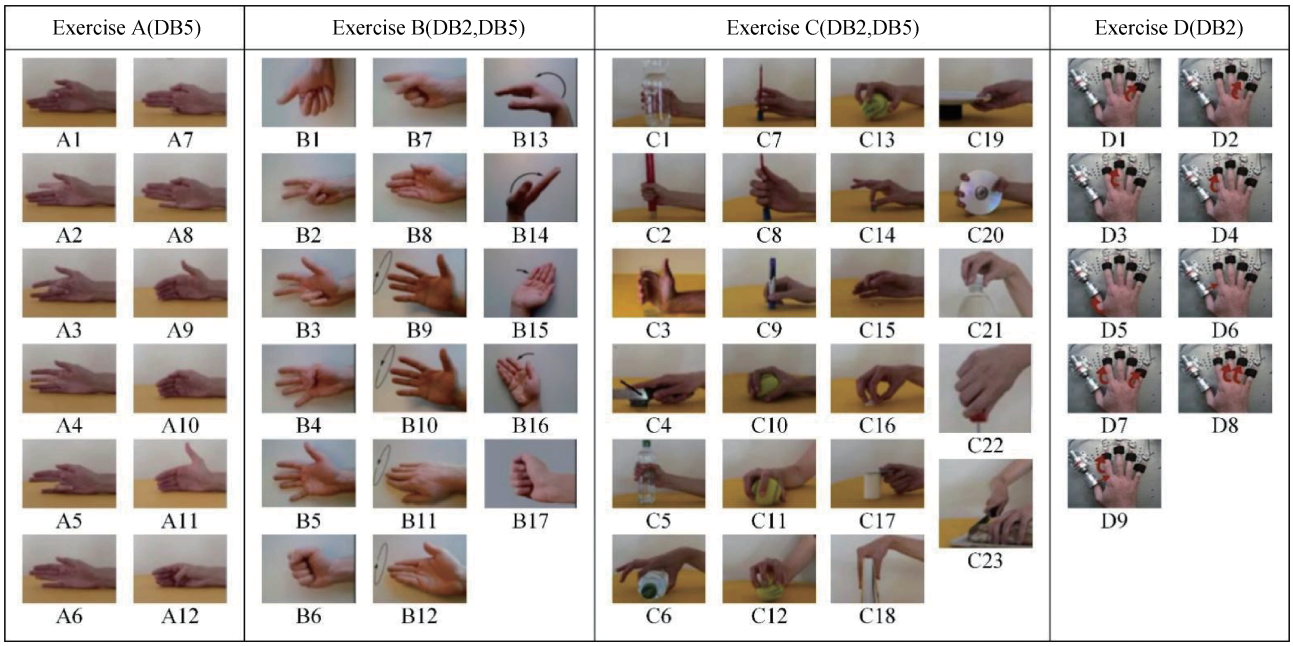

Fig.2 Gesture types in Ninapro DB2 and DB5 datasets.A,B:finger.C:wrist.D: grip

typically chosen for analysis[28]. sEMG signals are divided into active and resting segments. Considering the varying lengths of active segment waveforms, a sliding window is used to calculate the energy and identify the maximum energy window (MEW) in the active segment on multi-channel sEMG signals. Additionally, we use a sliding window sampling method to construct time series samples from the active segment region, increasing the sample size. In this paper, we select a window length of T milliseconds, sliding once every S milliseconds, as shown in Fig.3. In this paper, we set T to 50 ms, 100 ms, 150 ms, 200 ms, 250 ms and S to 25 ms.

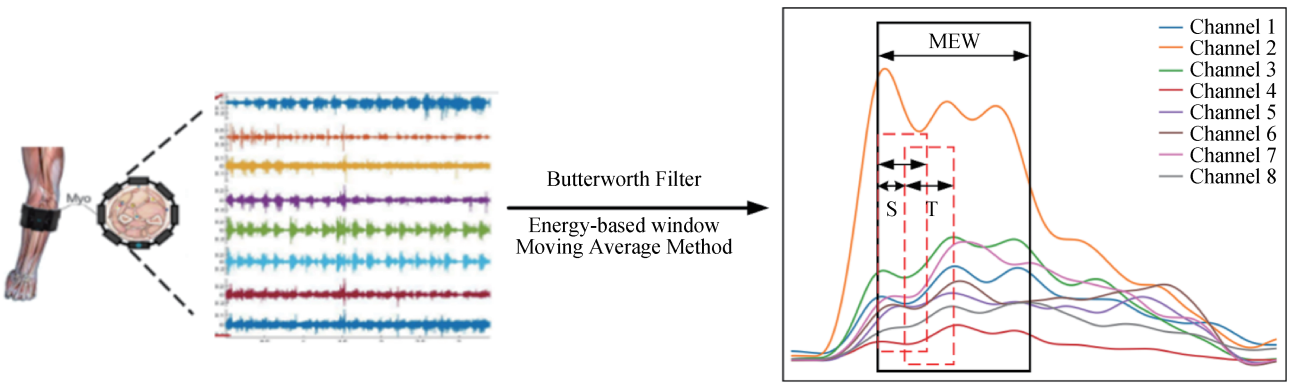

Fig.3 Pretreatment process. MEW denotes the maximum energy window

Additionally, a sample size balancing operation is performed for each category to mitigate the impact of sample size imbalance on classification results. Following the data partitioning strategy outlined in the study[29], repetitions 1, 3, 4, and 6 of each gesture are used as the training set for the NinaPro DB2 and DB5 datasets, while repetitions 2 and 5 are used as the test set.

To evaluate the performance of the proposed method , the accuracy (Acc), precision (Pre), recall (Rec), and F1-score (F1) of the gesture recognition were calculated as follows.

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{21}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{22}$$

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{23}$$

$$\mathrm{F1}=\frac{2\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{24}$$

where TP, TN, FP, and FN correspond to true positive, true negative, false positive, and false negative, respectively.

4 Results and discussion

4.1 Feasibility analysis to TS-HSTFNet

4.1.1 Ablation experiments

In this subsection, we conduct ablation studies on Ninapro DB2 and DB5 datasets to verify the validity and

significance of the TS-HSTFNet model's components. The experimental results are presented in Table 2. The models compared include: (1) TS- HSTFNet w/o DSTGCN: with the DSTGCN module removed. (2) TS-HSTFNet w/ o MSTCM: with the MSTCM module deleted. (3) TS-HSTFNet w/ Add: with the fusion method changed from STAFM to additive fusion. (4) TS-HSTFNet w/ Concat: with the fusion method changed from STAFM to cascade fusion. (5) TS-HSTFNet: the proposed model.

The results of the ablation experiments from Table 1 show that the TS-HSTFNet model has significant enhancement compared to other models on both Ninapro DB2 and Ninapro DB5 datasets. Compared to TS-HSTFNet w/o DSTGCN and w/o MSTCM, our model shows significant improvement, indicating the effectiveness of the dual-stream spatio- temporal feature framework. Furthermore, comparing two primary feature-level fusion strategies demonstrates that traditional fusion has a limited ability to capture complex correlations among multiple features. In contrast, STAFM facilitates spatio- temporal feature interactions, capturing potential dependencies among features more effectively.

Table 1 Experimental study ofablation of TS-HSTFNet model on Ninapro DB2 and Ninapro DB5 datasets

Method	Dataset	Acc (%)	Pre(%)	Re(%)	F1 (%)
TS-HSTFNet w/o DSTGCN		81.84	81.79	81.78	81.74
TS-HSTFNet w/o MSTCM		76.51	76.23	75.82	75.76
TS-HSTFNet w/ Add	Ninapro DB2	82.02	82.06	82.01	81.97
TS-HSTFNet w/ Concat		83.08	82.13	82.17	82.09
TS-HSTFNet		84.96	85.21	84.95	84.93
TS-HSTFNet w/o DSTGCN		81.72	81.48	81.28	81.23
TS-HSTFNet w/o MSTCM TS-HSTFNet w/ Add	Ninapro DB5	83.79 83.38	83.48 83.33	83.34 83.16	83.25 83.12
TS-HSTFNet w/ Concat		85.49	85.59	85.41	85.39
TS-HSTFNet		88.08	88.24	88.07	88.05

To further explore the performance of different ablation models, we analyzed the ROC curves of the five ablation models, as shown in Fig. 4. On the Ninapro DB2 dataset, our model achieved the highest AUC value of 0.9861, with improvements of 0.0776, 0.0019, 0.0009, and 0.0002 over TS-HSTFNet w/o DSTGCN, w/o MSTCM, DGTC-MFNe w/ Add, and w/ Concat, respectively. Similarly, on the Ninapro DB5 dataset, the TS-HSTFNet model demonstrates its superiority with an AUC value of 0.9928, improving by 0.0062, 0.0059, 0.003, and 0.004, respectively, compared to the aforementioned ablation models. These results indicate that the TS-HSTFNet model performs excellently in gesture recognition tasks. The three functional modules, DSTGCN, MSTCM, and STAFM, all significantly contribute to the model's performance. This confirms the TS-HSTFNet model's ability to capture and fuse complex sEMG signal features and its potential in gesture recognition applications.

4.1.2 Feature Visualization (t-SNE)

In this paper, t-SNE technique is used to visualize the high-level features extracted from TS-HSTFNet model on Ninapro DB2 and Ninapro DB5 datasets, aiming to visually assess the spatial distribution of these features and their discriminative ability. According to the principle of feature distribution, a greater distance among different feature clusters indicates stronger uniqueness of the features, while a tighter cohesion within the same feature cluster reflects better consistency of the features[29]. Fig. 5 depicts the spatial distribution of feature clusters for the 4th, 13th, and 34th gestures in the Ninapro DB2 dataset and the 2nd, 15th, and 32nd gestures in the Ninapro DB5 dataset, respectively. It can be seen that there is a significant overlap among different feature clusters when no fusion strategy is introduced, as shown in Fig. 5(a) and 5(c). Introducing the STAFM strategy significantly increases the distance between different feature clusters and improves clustering within the same feature cluster, as depicted in Fig.5(b) and 5(d). These observations highlight the crucial role of the STAFM strategy in enhancing feature fusion capability.

4.2 Performance analysis of TS-HSTFNet

4.2.1 Different datasets

TS-HSTFNet demonstrates outstanding recognition performance on both the Ninapro DB2 and DB5 datasets, achieving accuracies of 84.96% and 88.08%, respectively. This indicates that the model maintains and improves its performance with more complex datasets, as shown in Table 2. In the Ninapro DB2 dataset, the TS-HSTFNet model achieves an accuracy of 84.96%, a precision of 85.21%, a recall of 84.95%, and an F1-score of 84.93%. Similarly, in the Ninapro DB5 dataset, the TS-HSTFNet model achieves an accuracy of 88.08%, aprecision of 88.2%, a recall of 88.07%, and an F1-score of 88.05%. The results show that the model's performance in terms of accuracy, precision, recall, and F1-score is consistent across both datasets, demonstrating

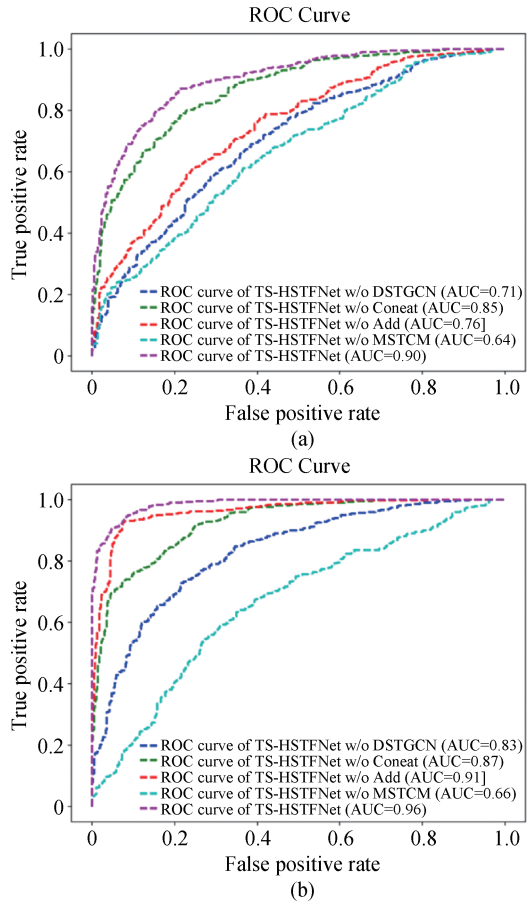

Fig.4 Comparison of AUC curves of ablation experimental studies on different datasets. (a) Ninapro DB2 dataset (b) Ninapro DB5 dataset

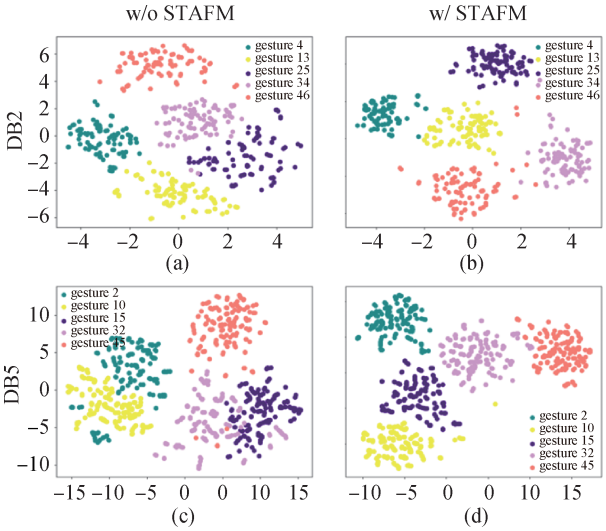

Fig.5 Spatial distribution of features. Different colors represent different categories

good generalization performance.

To comprehensively evaluate the TS-HSTFNet model's ability to classify different gestures, we calculate confusion matrices for the Ninapro DB2 and DB5 datasets, as shown in Fig.6. For the Ninapro DB2 dataset, our model's classification accuracy is above 80% for most gestures. In the Ninapro DB5 dataset, our model achieves

Table 2 Performance metrics of the TS-HSTFNet model for different datasets

Dataset	Classes	Subjects	Acc (%)	Pre (%)	Re (%)	F1 (%)
Ninapro DB2	49	40	84.96	85.21	84.95	84.93
Ninapro DB5	52	10	88.08	88.24	88.07	88.05

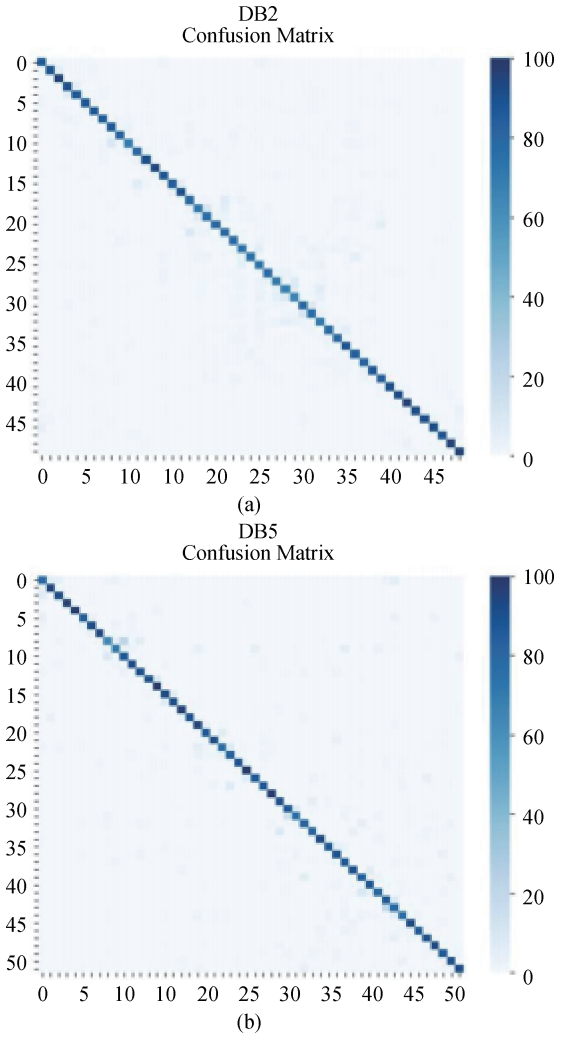

Fig.6 (a) Confusion matrix of TS-HSTFNet model under Ninapro DB2 dataset; (b) Confusion matrix of TS-HSTFNet model under Ninapro DB5 dataset

classification accuracies above 85% for most gestures. This indicates that the TS-HSTFNet model accurately classifies most gestures. However, analysis of the confusion matrix reveals that recognition accuracies for the 10th, 28th, and 29th gestures in the Ninapro DB2 dataset are below 70%. Similarly, the 8th and 9th gestures in the Ninapro DB5 dataset have accuracies below 75%.

4.2.2 Different models

In order to evaluate the performance advantages of our model, we designed a series of comparison experiments and the results are shown in Table 3. In these experiments we compare them with four models on the Ninapro DB2 database, which include TDCT[29], TC-HGR[30], MLP-Mixer[31], and CviT[32]. In addition, we

also compare them with five models on the Ninapro DB5 database, TDCT[29], TC-HGR[30], MLP- Mixer[31], CviT[32], and SE-CNN[33] compared. These results fully demonstrate that the model presented in this paper aims to be significantly competitive in gesture recognition tasks. Our model significantly outperforms the feature extraction-based machine learning modeling approach[31] in an end-to-end manner. Additionally, some SOTA methods[29,30,33] mainly focus on extracting temporal or spatial features from sEMG signals. They fail to fully exploit multi- feature fusion advantages and neglect higher-order semantic information by not integrating different feature information. In particular, a convolutional visual transformer (CviT) with stacked ensemble learning proposed in the literature[32] extracts temporal and spatial features of sEMG signal sequences and fuses them with the Transformer for parallel training. This model is closest to ours. However, the CviT model merely utilizes convolution and Transformer modules for extracting temporal and spatial features without considering the spatial dynamics of gestures. In addition, the CviT model only utilizes the cascade fusion method to fuse multiple features, and does not consider the significant effect of the fusion strategy on the model performance. In contrast, our proposed method not only integrates the temporal dynamics and spatial dependence of sEMG but also pays attention to the potential correlation between features.

Table 3 Comparison with other models

methodologies	timing	data set	window size	Number of gestures	accuracy
TC-HGR[30]	2022	Ninapro DB2	200ms	49	77.43%
CviT[32]	2022		200ms	49	80.02%
TDCT[29]	2024		200ms	49	80.68%
MLP-Mixer[31]	2024		200ms	49	80.74%
Ours	-		200ms	49	84.96%
TC-HGR[30]	2022	Ninapro DB5	200ms	52	68.61%
CviT[32]	2022		200ms	12/17	76.83%/73.23%
SE-CNN[33]	2023		260ms	53	87.42%
TDCT[29]	2024		200ms	52	72.83%
MLP-Mixer[31]	2024		200ms	52	73.39%
Ours	-		200ms	52	88.08%

4.3 Key influencing factors in TS-HSTFNet

4.3.1 Impact of different-sized convolutional kernels

In the multi-scale spatio-temporal convolution module (MSTCM), the different sizes of convolution kernels determine theirrespective receptive field sizes, which are decisive for spatio-temporal feature capturing. Choosing the appropriate size of the convolution kernel is key to improving the model's performance in spatio-temporal feature extraction. To investigate the role of different convolution kernel sizes in feature extraction, we designed experiments with three kernel size combinations: [7, 5, 3], [9, 7, 5], and [11, 9, 5]. Subsequently, we conducted experiments with these convolutional kernels on two publicly available datasets, Ninapro DB2 and Ninapro DB5, to reveal how different kernel scales affect the model's ability to capture spatio-temporal information.

The experimental results are shown in Fig. 7, for both Ninapro DB2 and DB5 datasets, the model achieves optimal performance when the convolution kernel size is [9, 7, 5]. The results show that larger convolutional kernel sizes help capture local spatial features more comprehensively, significantly enhancing the model's ability to recognize gesture features. However, a convolutional kernel that is too large may introduce unnecessary information, while one that is too small may limit the model's expressive

4.3.2 Impact of sliding window length

In addition, this experiment delves into the specific impact of sliding window length on performance in a gesture recognition system. Given that window lengths longer than 300 ms increase parameter counts and processing delays, we focus on analyzing lengths shorter than 300 ms. After referring to previous studies [30,31] and weighing the computational cost against real-time requirements, we selected 50 ms, 100 ms, 150 ms, 200 ms, and 250 ms as the key parameters for our experiments. The experimental results in Fig. 8 demonstrate that recognition accuracy increases with longer window lengths, suggesting that longer windows capture richer sEMG signal features and enhance model recognition capability. However, we also observe that the recognition accuracy instead decreases when the window length exceeds 200 ms. A window length of 200 ms yields the highest recognition accuracies of 84.96% and 88.08% for

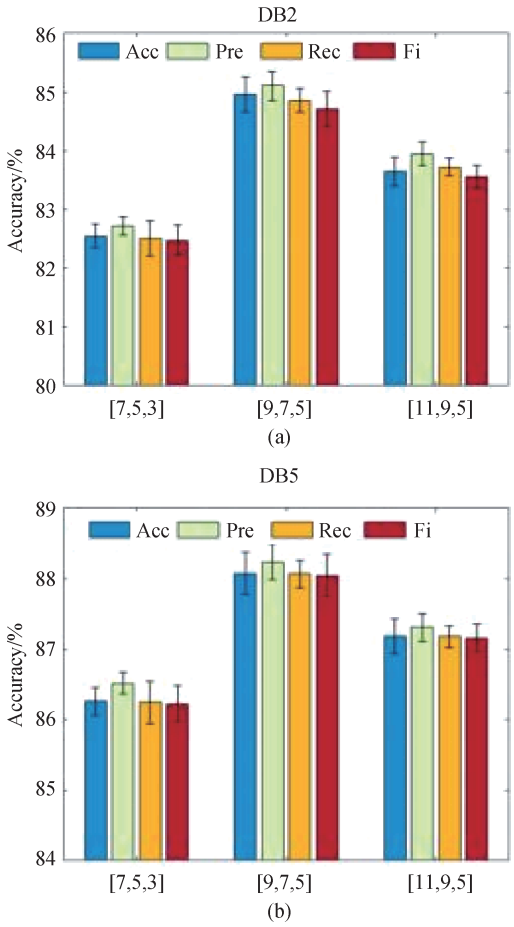

Fig.7 Comparison of different convolution scales gesture recognition accuracy, Precision, Recall, and F1-Score for Models with different convolutional kernel sizes

the Ninapro DB2 and Ninapro DB5 datasets, respectively. This phenomenon may be attributed to excessive window length introducing unnecessary noise, which interferes with the precise feature extraction of key features and affects model recognition accuracy.

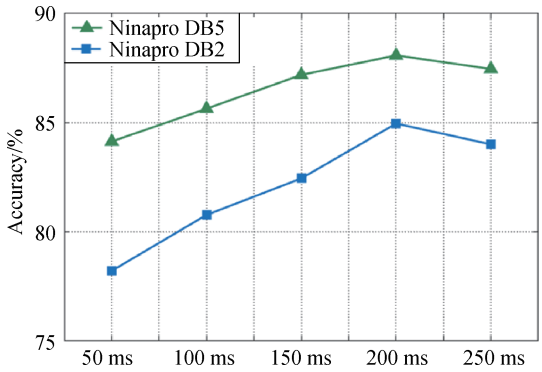

Fig.8 Effect of sliding window length on model performance

The two-stream spatio-temporal features to further enhance the performance of the model. Through extensive experiments on Ninapro DB2 and Ninapro DB5 datasets, we demonstrate the superior recognition capability of the TS-HSTFNet model relative to various ablation models and state-of-the-art methods. The results validate that our model excels in extracting informative features from sEMG signals and enhancing gesture recognition performance. Our findings clearly demonstrate the effectiveness of jointly considering spatio-temporal features for gesture recognition. This discovery opens up new approaches for future research in the field of gesture recognition.

5 Conclusion

In this paper, we propose a novel two-stream hybrid spatio-temporal feature fusion network named TS-HSTFNet. Our model integrates two-stream spatio-temporal features of sEMG signals effectively to improve the accuracy of gesture recognition. The DSTGCN module learns correlations among sEMG nodes and captures dynamic connectivity patterns over time in an adaptive manner. Meanwhile, MSTCM focuses on the temporal dependence of sEMG and employs multi-scale convolution to explore both the local and global importance of each period of sEMG and extract more representative spatio-temporal features. Finally, STAFM integrates ability, affecting recognition accuracy. Based on these results, this study adopts the convolution kernel size configuration of [9, 7, 5] in all experiments to ensure that the model's generalization ability is maintained while enhancing recognition accuracy.

Author Contribution:

Ruiqi Han: Software, Writing, Visualization. Juan Wang and Jia Wang: Regulation, Data.

Funding Information:

This research obtained fund project: Funding from the Key Research and development plan of Shaanxi Province "Human robot interaction technology and implementation of bionic robotic arm based on remote operation" (2023-ZDLGY-24).

Data Availability:

The authors declare that the main data supporting the findings of this study are available within the paper and its Supplementary Information files.

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 17 July 2024; Accepted 17 December 2024; Published online 31 December 2024

References

[1] Liu Z, Wu C, Ye W. Category-extensible human activity recognition based on Doppler radar by few-shot learning[J]. *IEEE Sensors Journal* , **2022**, 22(22): 21952-21960.

[2] Qureshi M F, Mushtaq Z, ur Rehman M Z, et al. Spectral image-based multiday surface electromyography

classification of hand motions using CNN for human-computer interaction[J]. *IEEE Sensors Journal* , **2022**, 22 (21): 20676-20683.

[3] Song X, Van De Ven S S, Liu L, et al. Activities of daily living-based rehabilitation system for arm and hand motor function retraining after stroke[J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , **2022**, 30: 621-631.

[4] Li D, Kang P, Zhu K, et al. Feasibility of wearable PPG for simultaneous hand gesture and force level classification[J]. *IEEE Sensors Journal* , **2023**, 23(6): 6008-6017.

[5] Jiang S, Kang P, Song X, et al. Emerging wearable interfaces and algorithms for hand gesture recognition: a survey[J]. *IEEE Reviews in Biomedical Engineering*, **2021**, 15: 85-102.

[6] Tchantchane R, Zhou H, Zhang S, et al. A review of hand gesture recognition systems based on noninvasive wearable sensors[J]. *Advanced Intelligent Systems*, **2023**, 5(10): 2300207.

[7] Zheng M, Crouch M S, Eggleston M S. Surface electromyography as a natural human-machine interface: a review[J]. *IEEE Sensors Journal*, **2022**, 22(10): 9198-9214.

[8] Xu X, Zhou Y, Shao B, et al. GestureSurface: VR sketching through assembling scaffold surface with non-dominant hand [J]. *IEEE Transactions on Visualization and Computer Graphics*, **2023**, 29(5): 2499-2507.

[9] Fang Y, Guo W, Sheng X. Toward a wireless wearable system for bidirectional human-machine interface with gesture recognition and vibration feedback [J]. *IEEE Sensors Journal*, **2022**, 22(10): 9462-9472.

[10] Fathima Shafana A R, Silpasuwanchai C. Investigating the role of gesture modalities and screen size in an AR 3D game [J]. **2023**.

[11] Zhou X, Wang C, Zhang L, et al. Continuous estimation of lower limb joint angles from multi- stream signals based on knowledge tracing[J]. *IEEE Robotics and Automation Letters* , **2023**, 8(2): 951-957.

[12] Yang B, Shi C, Liu Z, et al. Fingertip proximity- based grasping pattern prediction of transradial myoelectric prosthesis[J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , **2023**, 31: 1483-1491.

[13] Vecchiato G, Del Vecchio M, Ambeck-Madsen J, et al. EEG-EMG coupling as a hybrid method for steering detection in car driving settings [J]. *Cognitive neurodynamics* , **2022**, 16 (5): 987-1002.

[14] Zhang Z, He C, Yang K. A novel surface electromyographic signal-based hand gesture prediction using a recurrent neural network[J]. *Sensors* , **2020**, 20(14): 3994.

[15] Wang Y, Wu Q, Dey N, et al. Deep back propagation-long short-term memory network based upper-limb sEMG signal classification for automated rehabilitation[J]. *Biocybernetics and Biomedical Engineering* , **2020**, 40(3): 987-1001.

[16] Chollet F. Xception: deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* . **2017**: 1251-1258.

[17] Xu L, Zhang K, Yang G, et al. Gesture recognition using dual-stream CNN based on fusion of sEMG energy kernel phase portrait and IMU amplitude image[J]. . *Biomedical signal processing and control* , **2022**, 73: 103364.

[18] Atzori M, Cognolato M, Miller H. Deep learning with convolutional neural networks applied to electromyography data: a resource for the classification of movements for prosthetic hands[J]. *Frontiers in neurorobotics* , **2016**, 10: 9.

[19] Zhang W, Zhao T, Zhang J, et al. LST-EMG-Net: Long short-term transformer feature fusion network for sEMG gesture recognition[J]. *Frontiers in Neurorobotics*, **2023**, 17: 1127338.

[20] Peng F, Chen C, Lv D, et al. Gesture recognition by ensemble extreme learning machine based on surface electromyography signals[J]. *Frontiers in human neuroscience*, **2022**, 16: 911204.

[21] Duan S, Wu L, Liu A, et al. Alignment-enhanced interactive fusion model for complete and incomplete multimodal hand gesture recognition[J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , **2023**, 31: 4661-4671.

[22] Xiong B, OuYang Y, Chang Y, et al. A fused biometrics information graph convolutional neural network for effective classification of patellofemoral pain syndrome[J]. *Frontiers in Neuroscience* , **2022**, 16: 976249.

[23] Xu P, Li F, Wang H. A novel concatenate feature fusion RCNN architecture for sEMG-based hand gesture recognition [J]. *PloS one* , **2022**, 17(1): e0262810.

[24] Wu H, Jiang B, Xia Q, et al. A Convolutional Neural Network with Narrow Kernel and Dual- View Feature Fusion for sEMG-Based Gesture Recognition[C]// Asian-Pacific Conference on Medical and Biological Engineering. Cham: Springer Nature Switzerland, **2023**: 353-362.

[25] Duan S, Wu L, Liu A, et al. Alignment-enhanced interactive fusion model for complete and incomplete multimodal hand gesture recognition[J]. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* , **2023**, 31: 4661-4671.

[26] Atzori M, Gijsberts A, Castellini C, et al. Electromyography data for non-invasive naturally-controlled robotic hand prostheses[J]. *Scientific data* , **2014**, 1(1): 1-13.

[27] Pizzolato S, Tagliapietra L, Cognolato M, et al. Comparison of six electromyography acquisition setups on hand movement classification tasks[J]. *PloS one* , **2017**, 12(10): e0186132.

[28] 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.

[29] Wang Z, Yao J, Xu M, et al. Transformer-based network with temporal depthwise convolutions for sEMG recognition[J]. *Pattern Recognition* , **2024**, 145: 109967.

[30] Rahimian E, Zabihi S, AsifA, et al. Hand gesture recognition using temporal convolutions and attention mechanism[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, **2022**: 1196-1200.

[31] Shen S, Li M, Mao F, et al. Gesture Recognition Using MLP-Mixer With CNN and Stacking Ensemble for sEMG Signals [J]. *IEEE Sensors Journal* , **2024**.

[32] 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.

[33] Xu Z, Yu J, Xiang W, et al. A novel SE-CNN attention architecture for sEMG-based hand gesture recognition[J]. *CMES-Computer Modeling in Engineering & Sciences* , **2023**, 134(1): 157-177.