Current Science
CS 5 (6), 5519-5537 (2025)
Received 05 November 2025 | **Accepted 25 November** 2025 | **Online Available 15 December** 2025
https://doi.org/10.52845/CS/2025-5-6-19

Current Science

Current Science

OPEN ACCESS

ISSN (O) 2795-8639

Original Article

Smooth Multiscale Convolutional Attention Transformer Network for Continuous Motion Estimation of Hand Knuckle Angle using Surface EMG Signals

Shenghua Chen[1], Ming Dai[2]*, Yue He[3]*

[1]The School of Applied Artificial Intelligence, Guangxi Technical College of Mechanical and Electrical Engineering, Guangxi 530007, China

[2]The School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen 518055, China

[3]*The School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

***Corresponding Author: Ming Dai, Yue He**

Abstract:
As a key technology in the field of human-computer interaction and intelligent perception, continuous hand motion estimation has made significant progress in recent years driven by deep learning. Its core goal is to accurately capture the dynamic trajectory of hand joints through the modeling of high-dimensional time-series data, so as to realize natural and low-latency human-machine cooperative operation. Although current deep learning algorithms have achieved more satisfactory prediction results in the field of continuous motion estimation, they still need to be improved in terms of prediction accuracy, compatible action diversity and prediction robustness.In order to solve the problems of few compatible actions and poor robustness of existing deep learning models, based on the research in the previous part, this part proposes a smooth multiscale convolutional attention Transformer network that improves the generalization and the number of compatible actions for continuous motion estimation. The model consists of a serial connection between a multiscale convolutional channel attention network and an improved Transformer network, which integrates the advantages of the two network architectures and is capable of extracting features in multiple scale dimensions of the sEMG, in addition to the addition of a smoothing algorithm to further improve the accuracy and noise immunity of the model. This section focuses on the design of the experimental scheme and the analysis of the results of the smoothed multiscale convolutional attention Transformer network. Firstly, the components of the network are introduced: the multiscale convolutional attention module, the improved Transformer network module, and the smoothing algorithm module. The proposed network is then evaluated in the Ninapro dataset and compared with several existing state-of-the-art algorithms to demonstrate the effectiveness of the proposed method.

Keywords: Transformer, multiscale convolutional attention, deep learning, multi-head attention mechanism, multi-feature sEMG.

1. Introduction

Intelligent mechanical systems technology has achieved multidimensional application breakthroughs, and their scope of service is being integrated exponentially into the basic units of society. Through the synergistic architecture of integrated human-computer interfaces and heterogeneous communication protocols, such systems can collaborate with human operators to realize the precise execution of complex task flows. As a core topic of cross-disciplinary

research, the technical framework of human-computer collaborative systems has demonstrated significant adaptability in a variety of vertical domains, such as intelligent management of industrial manufacturing processes, optimization of teleoperation terminals, iterative upgrading of automated driving systems, construction of immersive education and interaction platforms, implementation of neurorehabilitation projects, and development of minimally invasive surgical robots. Empirical studies have shown that such intelligent systems have significant advantages in terms of key performance indicators, such as optimization of workflow efficiency, standardization of operational precision, and intensive allocation of human resources [1].

Electromyogram (EMG) signal is the sum of Motor Unit Action Potential (MUAP), which is generated by muscle fibers and contains the neural code behind the movement[2] . It can be obtained either by non-invasive electrodes on the skin or by invasive electrodes inserted into the muscle during muscle contraction[3] . The EMG signal can reflect intuitive motor intentions even when no physical movement is occurring[4]. There are a large number of studies that have utilized EMG signals to predict human movement intentions such as gestures, joint angles, and joint torques. It has good potential in human-centered robotics and has been widely studied in prosthetic control [5,6], exoskeleton robot control [7], and rehabilitation training [8].

Suface Electromyography (sEMG), as a biometric signal reflecting the electrophysiological activity of muscles, has been used in many fields such as rehabilitation medicine, clinical diagnosis and treatment, and motor function analysis. This technology has become a key sensing tool in human-computer interaction research by virtue of its non-invasive detection advantages and convenient acquisition characteristics. It is worth noting that sEMG has a unique pre-activation property, which can detect the neural electrical activity signal 50-100 ms before the action occurs, and this timing advantage provides a theoretical basis for the early recognition of motor intention. In the process of constructing the sEMG human-machine cooperative system, the research focuses on the development of intelligent parsing algorithms with high robustness to realize the fast and accurate recognition of motor intention. This technological breakthrough plays a decisive role in improving the smoothness of operation and naturalness of interaction of human-machine cooperative system, which directly affects whether the system can realize the interaction response that is more in line with human kinematic characteristics [9,10].

In general, EMG signal recognition tasks can be categorized into two types: classification tasks and regression tasks. Typical examples of classification problems are gesture classification, limb movement discrimination, sign language processing, etc., which predefine labels into several classes. Regression problems, also known as continuous motion estimation, such as real-time joint angle, force and torque estimation, etc., in which the labels vary in real time over a fixed range. Although classification methods based on predefined movement categories can achieve basic motor function reconstruction, their application is limited to a finite set of established movement patterns, and they are unable to completely characterize the continuous biomechanics of human movement. In contrast, continuous motion estimation methods are able to construct continuous motion spatial mapping models by analyzing the time-varying characteristics of EMG signals, which can more accurately capture the dynamic change process of the user's intention. This technical path not only enhances the naturalness of interaction, but also improves the smoothness of operation, thus becoming an important research direction in the field of human-computer collaboration, which is also the core focus of this study.

Existing research on continuous motion estimation focuses on large joint motion units (e.g., shoulder and elbow joints and lower limb kinematic chain), and less on hand function units. As the most functionally valuable structure in the biological motion control system, the hand joint system shows unique kinematic advantages - its multi-degree-of-freedom synergistic mechanism can realize sub-millimeter operation accuracy, and this characteristic has irreplaceable application value in the field of neurorehabilitation engineering and human-computer collaborative manipulation system. Human hand movement is complex, and its hand joints involve more than twenty degrees of freedom, making the problem of continuous estimation of hand joint angles challenging.

Paradigm innovations in the field of artificial intelligence have given rise to the breakthrough development of deep learning technology, a data-driven approach that has triggered structural changes in the industrial effectiveness paradigm in multidisciplinary dimensions. The core advantage of this technology system stems from its autonomous feature extraction mechanism, which is capable of automatically capturing the spatial distribution pattern of potential features from massive heterogeneous data by constructing a multilayer nonlinear mapping architecture. This feature enables deep neural networks to show unique advantages in parsing high-dimensional sparse data and solving complex system problems, especially in processing unstructured data streams and building multimodal association models. The development of continuous motion estimation algorithms for multiple degrees of freedom based on deep learning is of great significance in the fields of robot control and human prosthesis. Due to the complexity of hand joint motion and continuous multi-degree-of-freedom estimation, current deep learning algorithms applied in this field have room for improvement in terms of accuracy and efficiency. Therefore, it is important to investigate how to develop deep learning algorithms that can accurately and efficiently decode the intention of hand joint motion.

In this study, we systematically explore the application of deep learning in hand continuous motion parsing, focusing on solving the key problems such as insufficient accuracy, limited real-time and weak generalization ability in the existing continuous motion estimation of EMG signals. By optimizing the signal decoding algorithm architecture, the flexibility and coordination of human-machine cooperative operation are significantly improved. The modeling research on continuous hand motion estimation based on sEMG has a dual value: in medical rehabilitation, the technique provides a new method for muscle function assessment and motor function reconstruction; in the field of human-computer interaction, it pushes forward the development process of naturalized interaction systems. In addition, this algorithmic framework has demonstrated significant technical value in various application scenarios, such as intelligent robot manipulation, assisted diagnostic and therapeutic devices, virtual reality interactive interfaces, and the development of wearable devices, which is both academically innovative and practically significant in engineering.

Methods

Dataset

In order to evaluate the performance of the newly introduced algorithm and to compare it fairly with other traditional deep learning algorithms, this study chooses the Ninapro open dataset. The Ninapro data set records the motion data of multiple intact subjects and amputee subjects in the form of electromyography, and is widely used in the development and testing of motion recognition control algorithms (Atzori and Muller, 2015). Ninapro is divided into 10 sub-datasets according to the type of experiment. In

this article we use the second one, referred to as DB2. This dataset contains motion data of 40 complete subjects.

During the process of obtaining hand movement data, subjects must follow the prompts to make corresponding actions. Perform each movement for five seconds, alternating with rest positions for three seconds. Twenty-two joint angles were accurately measured at a sampling rate of 20 Hz using the CyberGlove II data glove. Surface EMG signals are measured and acquired through two models of dual differential surface EMG signal electrodes. One of the electrode combinations, called Delsys Trigno Wireless System, samples raw surface electromyography signals at a frequency of 2 kHz. It contains 12 wireless surface electromyography signal electrodes and a base station. In order to synchronize the frequency of data collected by the above two devices, the joint angles are resampled to 2kHz.

In this paper, we choose to use 12-channel EMG to estimate 10 joint angles. As shown in Figure 1(A), the ten joint angles selected, including the proximal interphalangeal and metacarpo - phalangeal joint points, are the primary active joints in the grasping movement. To ensure the generalizability of the algorithm, we selected 10 representative subjects from Ninapro DB2. Their height ranges from 169-187cm and their weight ranges from 58-75kg. 12 different grasping actions were chosen for each participant, as shown in Figure 1(B), since the focus of this paper is on the continuous estimation of joint angles for grasping actions. In order to verify the robustness of the model proposed in this paper, we also tested the same six grasping actions in Ninapro DB7 with a wide selection of 10 subjects.

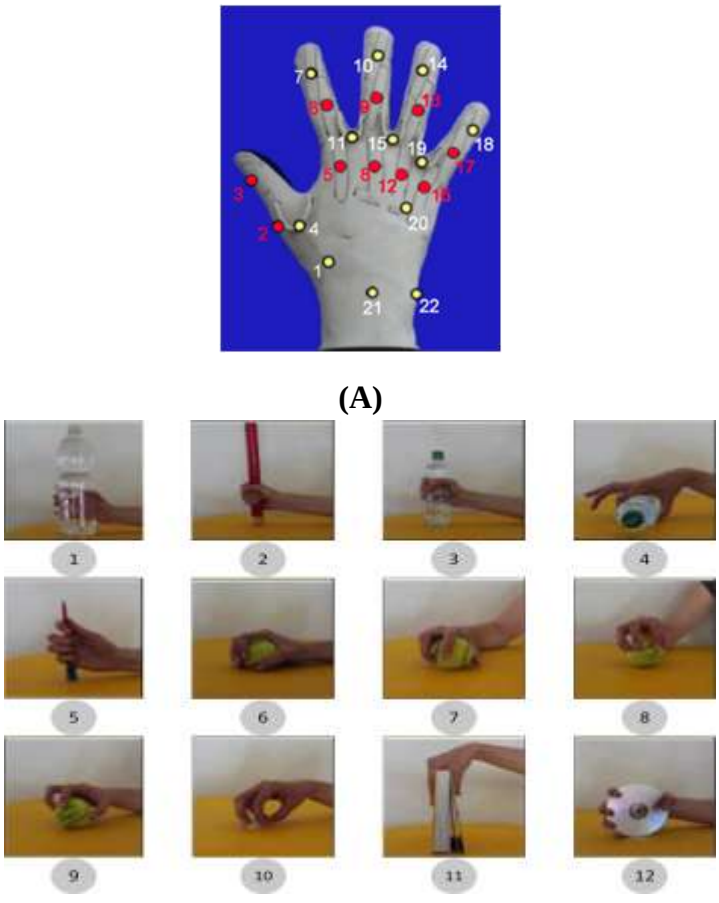


Fig. 1. Dataset Description. (A) 10 selected joint angles. (B) 12 grasping actions.

Data Processing

Joint angle continuous motion is motion across multiple degrees of freedoms (DoFs). It represents a higher dimensional and more complex target space and requires more comprehensive feature extraction methods. However, conventional deep learning feature extraction methods struggle to provide sufficient information for complex motion estimation across multiple degrees of freedom. Therefore, we used and fused multiple surface EMG signal feature extraction methods oriented to deep learning models. The extracted features such as Peak Stress (PS) and Shake Expectation (SE) are closely related to the strength and period of muscle contraction, which determines the accuracy of joint angle estimation.

In order to more completely extract data features, a sliding window of 100ms, with a step size of 0.5ms, is segmented from the surface electromyography signal and the resampled hand movement joint angle signal. Then we extract the following features from the segmented data intervals.

Peak Stress (PS)

Peak stress (PS) is defined as a measurement that quantifies the amount of stress that becomes the peak in each sliding window. This feature is described using zeroth, second, and fourth-order moments. The calculation formulas are as follows:

$$\mathbf{PS} = \frac{\boldsymbol{m_4}}{\boldsymbol{m_2 m_0}}$$

where $\boldsymbol{m_0}$ represents the intensity of muscle contraction, $\boldsymbol{m_2}$ can be seen as a formula describing changes in surface electromyographic signals, and $\boldsymbol{m_4}$ describes the changes in $\boldsymbol{m_2}$. The quotient of $\boldsymbol{m_4}$ and $\boldsymbol{m_2}$ represents the stress change of the surface electromyographic signal, and then divided by $\boldsymbol{m_0}$, the stress of each unit can be obtained.

Shake Expectation (SE)

The second derivative can explain the curvature of the surface EMG signal. Therefore, the amplitude changes speed expectation of each sliding window, that is, Shake Expectation (SE), can be described by the average absolute second derivative of the surface electromyographic signal, and its calculation formula is:

$$\mathbf{SE} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=0}}^{\boldsymbol{N-1}} \left| \Delta^{\mathbf{2}} \boldsymbol{x_i} \right|$$

where $\Delta^{\mathbf{2}}$ represents the second derivative.

Unbiased Standard Deviation (USTD)

A sliding window of surface EMG signals can be viewed as a sample of movement progression. Unbiased Standard Deviation (USTD) uses unbiased estimators to evaluate the distribution of sample values. We assume that $\overline{\boldsymbol{x}}$ is the prediction of the surface electromyographic signal amplitude in the sliding window, then its calculation formula is:

$$\mathbf{USTD} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N-1}} \sum_{\boldsymbol{i=0}}^{\boldsymbol{N-1}} |\boldsymbol{x_i} - \overline{\boldsymbol{x}}|^{\mathbf{2}}}$$

Long Short Term Memory Network

The LSTM model was proposed to address the issue of vanishing gradients in conventional recurrent neural networks (RNNs). As a result, it demonstrates exceptional performance in capturing dependencies over longer distances. BiLSTM is an LSTM variant that connects in both directions. BiLSTM can effectively improve the long-term dependence of learning and further improve the model prediction accuracy (Siami-Namini, Tavakoli and Namin, 2019). Thus, the BiLSTM model finds extensive usage in natural language processing (Xu et al., 2019) and regression forecasting (Ma et al., 2021b).

The fundamental component of the LSTM network consists of three gate structures, the

forget gate, the input gate and the output gate, as shown in Figure 2. The forget gate usually uses the Sigmoid function to achieve the purpose of retaining or deleting existing information. The input gate is made up of Sigmoid and Tanh layers that determine how much new information is being stored in memory. The output gate passes the information to the next LSTM unit. The specific mathematical formulas of the gates control unit are as follows:

$$\boldsymbol{f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)}$$

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

$$\boldsymbol{\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)}$$

$$\boldsymbol{C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t}$$

$$\boldsymbol{o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)}$$

$$\boldsymbol{h_t = o_t * \tanh(C_t)}$$

where $\boldsymbol{\sigma}$ and tanh are the Sigmoid and Tanh activation functions respectively. The $\boldsymbol{W}$ item is the weight matrix, and its value range is 0-1. The subscripts $\boldsymbol{f}$, $\boldsymbol{i}$, $\boldsymbol{C}$, and $\boldsymbol{o}$ represent the forget layer, input layer, hidden layer and output layer respectively. The $\boldsymbol{b}$ item is the bias matrix of each layer. $\boldsymbol{f_t}$, $\boldsymbol{i_t}$, $\boldsymbol{o_t}$ represent the status of the forget layer, input layer and output layer at time $\boldsymbol{t}$ respectively. $\boldsymbol{C_t}$ and $\boldsymbol{\widetilde{C}_t}$ represent the unit state vector. $\boldsymbol{h_t}$ is the final unit state vector at time $\boldsymbol{t}$.

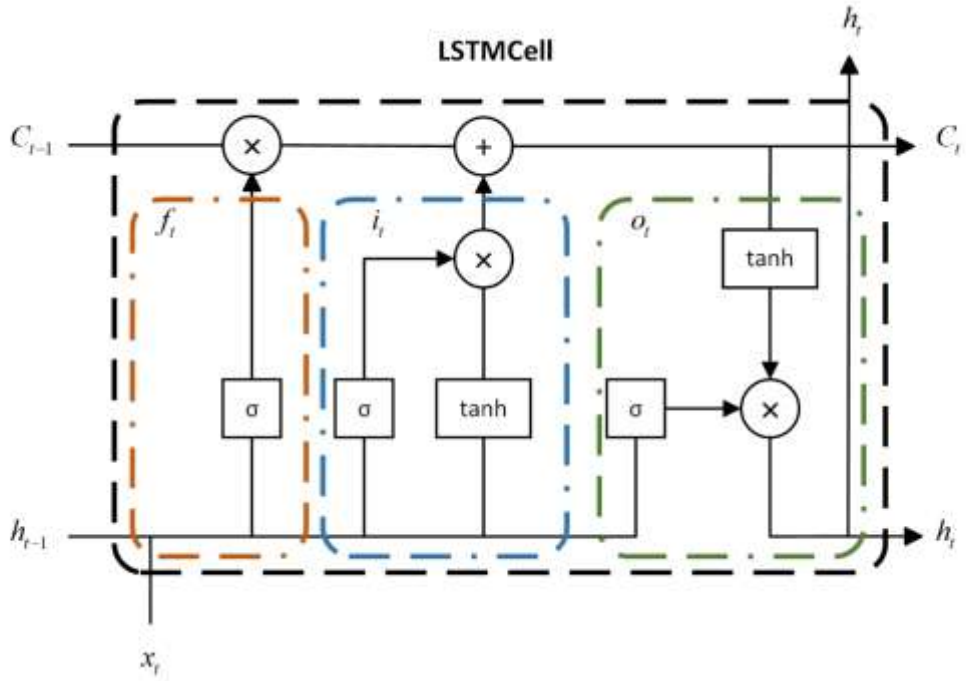


Fig. 2. Structure of LSTMCell.

In this article, we utilized a Bidirectional Long Short Term Memory network consisting of two layers, with a predetermined number of 64 hidden units. Our tests showed that this model structure performs well in terms of both training speed and accuracy. The output is generated via a fully connected layer.

Temporal Convolutional Network

It is well known that recurrent neural networks (RNN) have excellent performance in processing time series problems, but they can only handle one time step at a time and cannot be processed in large-scale parallelism like convolutional neural networks. A Temporal Convolutional Network (TCN) using a convolutional architecture with extended causal convolutions and residual connections was introduced by Bai et al. (2018).

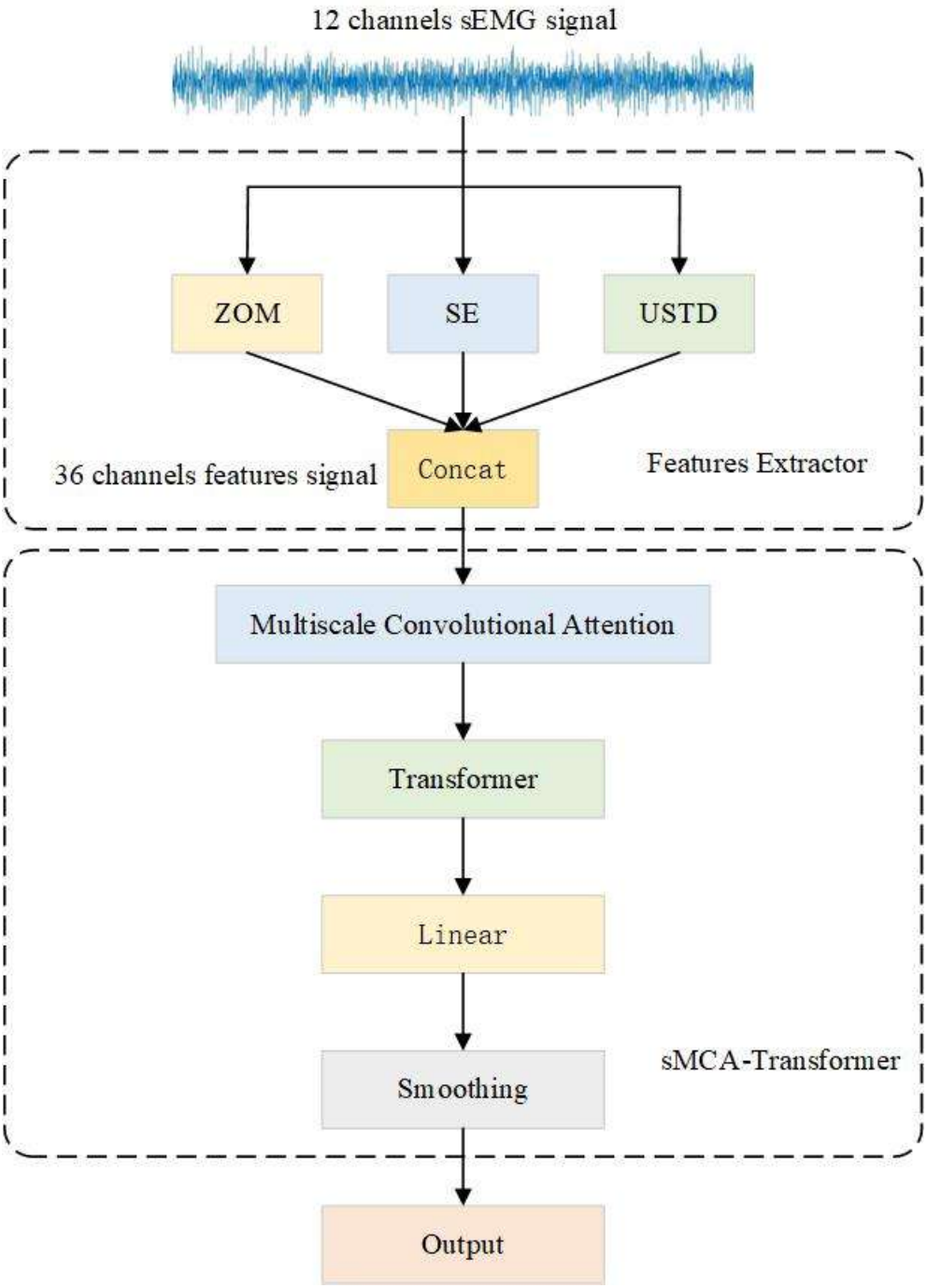
Due to its excellent performance on time series problems, TCN is widely used in fields such as signal processing (Zanghieri et al., 2020) and regression prediction (Song et al., 2020).

Causal convolution allows the network to only see future data and is a strictly time-constrained model. The size of the convolution kernel limits the time available for modelling by simple causal convolution. This requires a very deep network when processing tasks with a large workload, which will occupy a very large amount of computing resources. The extended causal convolution of TCN effectively expands the receptive field of the convolutional neural network and solves this problem. The TCN network incorporates a residual connection component that allows gradient transmission across layers to effectively prevent gradient disappearance.

In this article, we constructed a 5-layer TCN network with 48 convolution channels and a convolution kernel size of 3. Finally, the output utilizes a fully connected layer. The above configuration has been experimentally verified to achieve a relatively optimistic accuracy rate.

Smooth Multiscale Convolutional Attention Transformer Network

This section introduces the smooth multiscale convolutional attention Transformer network, firstly, in order to extract the features of sEMG in all aspects, the multiscale convolutional attention module is proposed in this section to extract the features of sEMG at different scales in an adaptive mode, to improve the generalization ability and to do the normalization operation for the inputs of the subsequent Transformer model. After that, an improved Transformer model is used to capture remote dependencies with its multi-head attention mechanism to further improve the prediction accuracy. Finally, after a smoothing algorithm module, a small number of historical finger joint angles are utilized to smooth out some prediction points with large errors, thus improving the robustness of the model. The overall network architecture of sMCA-Transformer is shown in Fig. It includes the multi-scale convolutional attention module, the improved Transformer network module, and the smoothing algorithm module, which will be introduced in turn below.It is worth noting that in terms of data preprocessing, following the multi-feature fusion strategy of Chapter 3 and adapting it, the 12-channel surface EMG signals were transformed into 36-channel feature signals by extracting the three features of the Zero order moment (ZOM), the expectation of the rate of change of the EMG amplitude, and unbiased standard deviation, respectively, and fusing them in a dimensionally additive manner.SE and USTD The two features are consistent with the previous description; the ZOM can be used to represent the muscle contraction intensity, the mathematical expression of which is given in the previous section.



Multiscale Convolutional Attention Module

The channel attention mechanism, as an important research direction in the field of computer vision, aims to improve the representation ability of deep neural networks by dynamically adjusting the weights of feature channels. The early representative work SENet significantly improves the model performance by introducing a fully connected layer to construct the channel attention module. However, the dimensionality reduction strategy adopted by SENet may destroy the direct correlation between channels, and the increase in the number of parameters limits its application in lightweight scenarios. To address these issues, Wang et al [68] proposed an efficient channel attention module, which achieves a balance between performance and efficiency by improving cross-channel interactions.

In this study, based on the multiscale convolutional module, an efficient channel

attention mechanism is added in parallel to it in order to efficiently extract features in the sEMG space and across channels. The structure of the multiscale convolutional attention module is shown in Fig. The feature signals first enter the multiscale convolutional layer to extract the spatial dimension information, and then enter the three parallel efficient channel attention layers to capture the inter-channel information, respectively, and finally the outputs of the three branches are spliced by dimension and go through a fully connected layer to get the output of this module.

The core idea of the Efficient Channel Attention Module is to capture inter-channel dependencies through one-dimensional convolution. Compared with traditional attention mechanisms, the Efficient Channel Attention Module avoids the complex process of dimensionality reduction and upgrading, thus realizing the properties of high efficiency and lightweight. Specifically, the Efficient Channel Attention Module first adaptively computes the kernel size of the one-dimensional convolution based on the number of channels, k. The kernel size is computed as follows:

$$k = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \#$$

where C is the number of channels of the input feature, and γ and b are hyperparameters, here set to 2 and 1. Taking absolute values and rounding down to the nearest odd number is to ensure that the kernel size is odd.

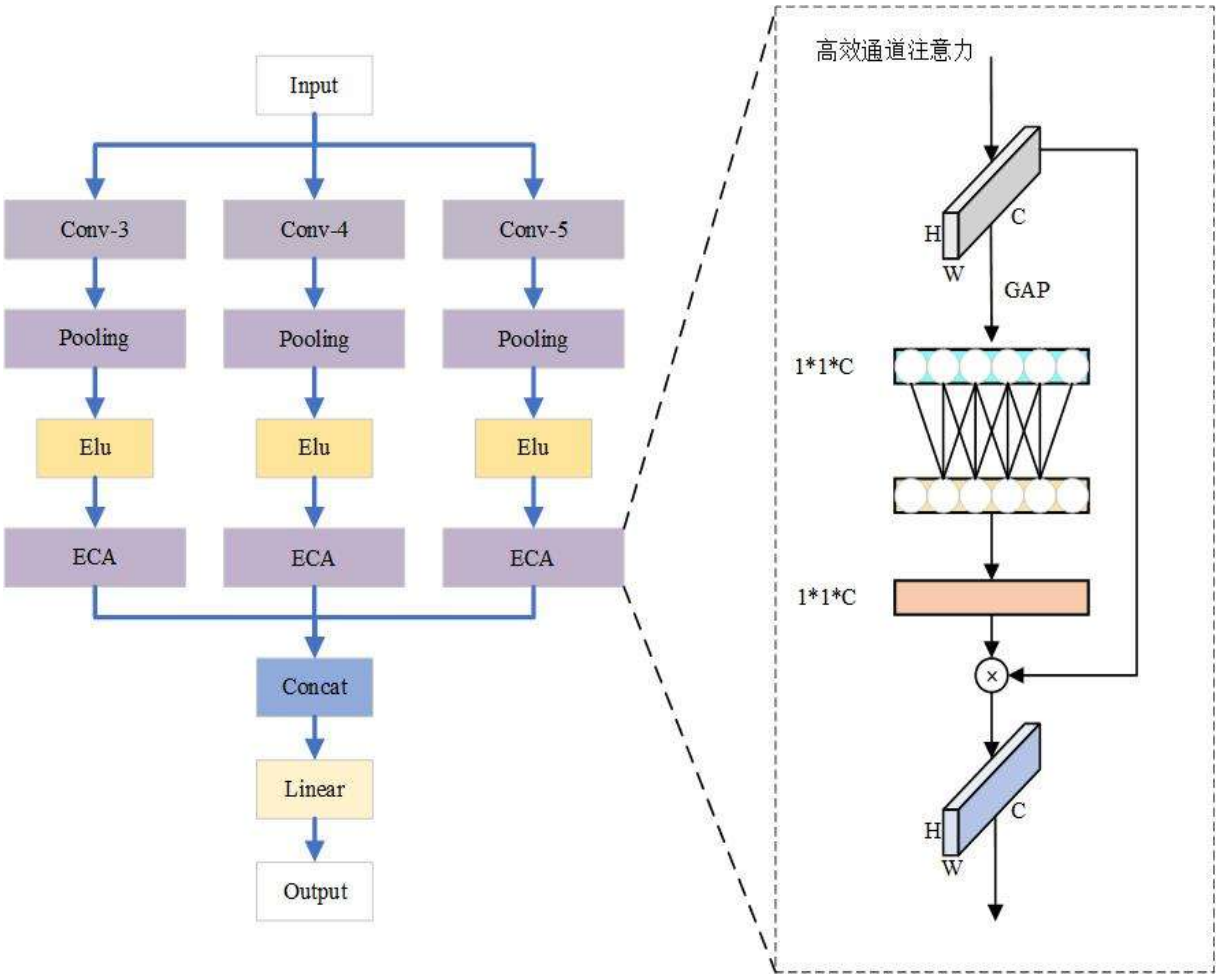


Fig. 3. Structure of multi-scale convolution.

Improved Transformer Module

In the Transformer architecture, feedforward neural network is one of its core components. The FFN module in each Transformer layer usually consists of two linear transformation layers and a nonlinear activation function. However, FFNs process the features at each position independently and cannot explicitly model the local relationships between neighboring positions, and additional convolution or sliding window mechanisms need to be introduced in tasks that require local dependencies [69].

Therefore, as shown in Fig. This paper introduces a LEFFN layer [70] and improves its activation function to improve the prediction accuracy of continuous motion estimation. First, the feature vectors are mapped to a high-dimensional space by 1×1 convolution. Since downsampling loses some local information while expanding the receptive field, 1×3 deep convolution is used to extract local features in the channel to compensate. Deep convolution is a special case of group convolution where the number of groups is equal to the number of channels, which reduces the amount of computation compared to normal convolution. Residual concatenation is applied to stabilize the data and the ELU activation function is used to improve the nonlinear representation of the model. Finally, the feature vectors are mapped back to the low-dimensional space by 1×1 convolution.LEFFN can enhance the ability of the model to extract local features and combine the advantages of CNN to extract local features and Transformer to extract global features. Its mathematical expression is as follows:

$$z' = Dropout\big(ELU(zW_1 + b_1)\big)\#$$

$$z'' = z' + DWConv(z')\#$$

$$LEFFN(z) = Dropout\big(ELU(z''W_2 + b_2)\big)\#$$

where z denotes the input signal, W and b represent the weight matrix and bias values, respectively, and DWConv represents the deep convolution operation.

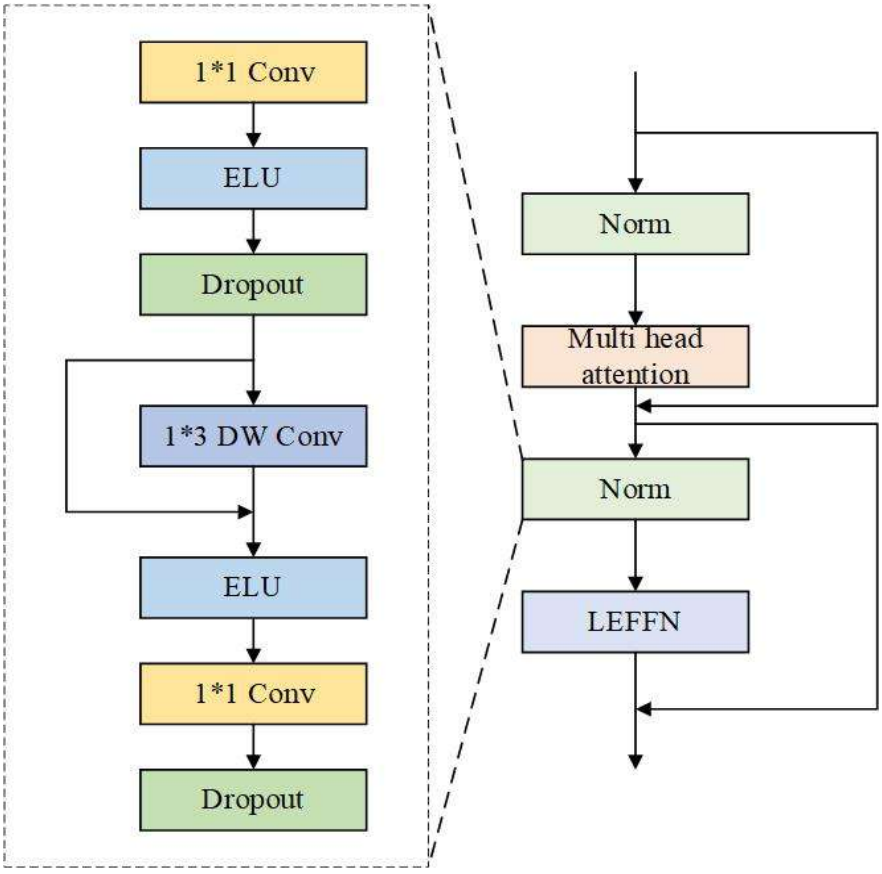


Fig. 4 Improved Transformer structure diagram

Smoothing Algorithm Module

The smoothing algorithm can use a small number of historical finger joint angles to smooth some prediction points with large errors. Experiments show that the predicted joint angles are often unstable, in order to improve the robustness of the algorithm, an exponential smoothing algorithm module is added at the end of the model, which makes the joint angles predicted by the model more closely match the real motion trajectory of the hand. Its mathematical expression is as follows:

$$out_0 = z_1 + z_2 + z_3\#$$

$$out_t = \alpha z_t + (1-\alpha) out_{t-1}\#$$

Where out_t is the output of the smoothing algorithm module at moment t, z_t is the output of the model at moment t, and α is the smoothing coefficient. In this paper, the smoothing coefficient is set to 0.30.

Result

Evaluation Metrics

In this study, we use the Pearson Correlation Coefficient (PCC) to assess the association between the expected and actual joint angles, and thus the accuracy and effectiveness of the algorithm. The closer the PCC is to 1, the more accurate the algorithm is. Its mathematical expression is as follows:

$$\boldsymbol{PCC} = \frac{\sum_{i=1}^{N}(\boldsymbol{\theta_{est}} - \overline{\boldsymbol{\theta_{est}}})\,(\boldsymbol{\theta_{real}} - \overline{\boldsymbol{\theta_{real}}})}{\sqrt{\sum_{i=1}^{N}(\boldsymbol{\theta_{est}} - \overline{\boldsymbol{\theta_{est}}})^2}\sqrt{\sum_{i=1}^{N}(\boldsymbol{\theta_{real}} - \overline{\boldsymbol{\theta_{real}}})^2}}$$

among them, $\boldsymbol{\theta_{est}}$, $\overline{\boldsymbol{\theta_{est}}}$, $\boldsymbol{\theta_{real}}$, and $\overline{\boldsymbol{\theta_{real}}}$represent joint angle estimates, average joint angle estimates, actual joint angle values, and actual joint angle averages respectively.

The deviation between the prediction and the actual value is the root mean square error (RMSE). This article uses the RMSE to indicate the amount of error in degrees (°) between the predicted and actual joint angle.

R2 objectively assesses the overall accuracy of the algorithm as a comprehensive evaluation index. The coefficient of determination represents the percentage of variance in the predicted value that can be explained by the prediction. The indicator range lies between 0 and 1. The better the estimation performance of the algorithm, the higher the value of the indicator. The calculation formula is:

$$\boldsymbol{R^2} = \mathbf{1} - \frac{\sum_{i=1}^{N}(\boldsymbol{\theta_{real}} - \boldsymbol{\theta_{est}})^2}{\sum_{i=1}^{N}(\boldsymbol{\theta_{real}} - \overline{\boldsymbol{\theta_{real}}})^2}$$

Experimental Parameters and Statistical Analysis

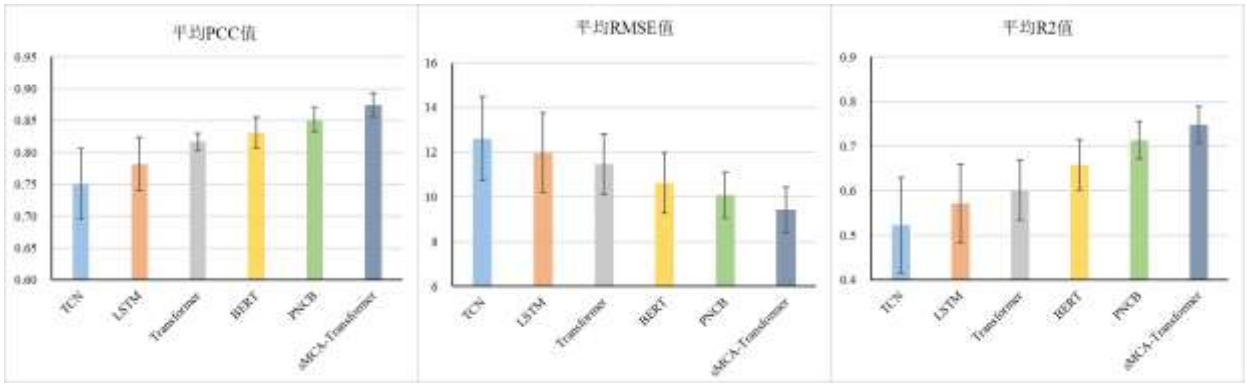
We maintain consistency in the training parameters, including the type of loss function, the type of optimizer and the learning rate, when training a neural network in this study. The loss function used is mean squared error, with Adam optimization as the optimizer, and the learning rate is uniformly set at 0.0002. In regression prediction problems, mean squared error is a commonly used loss function. As an effective deep learning optimizer, Adam optimization has been often used in the fields of signal processing and image processing in recent years. It is based on the Stochastic Gradient Descent Technique (Kingma and Ba, 2014). We keep the network structure parameters in the control experimental algorithms CNN, CNN-Attention, CNN-BiLSTM and BiLSTM consistent with those of the corresponding modules of PNCB. All neural networks are built on the Pytorch framework.

The t-test uses the theory of the t-distribution to infer the probability of a difference occurring, and thus to compare whether the difference between two sets of data is significant or not. It is applicable to normal distributions with small sample size and unknown overall standard deviation. In this article we use it to examine the differences between the comparison algorithm

used in the experiment and the model proposed in the article. The dependent variables used in this study were PCC, RMSE and R2 of evaluation indicators. In this research, the statistical significance threshold was established at p value < 0.05.

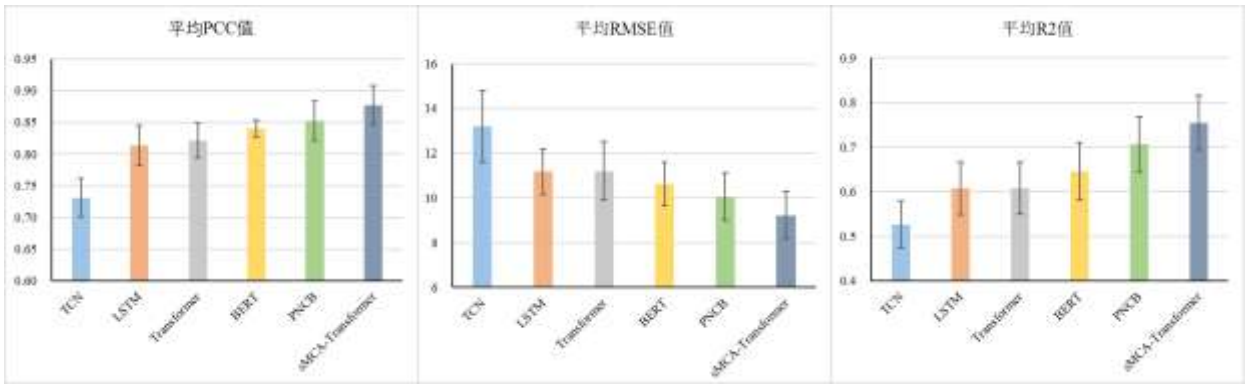
Experimental Results

Fig. 5 and Fig. 6 demonstrate the result averages of the three metrics for different continuous motion estimation models for 20 subjects in both DB2 and DB7 datasets. From the experimental results, it is easy to see that the sMCA-Transformer proposed in this paper performs optimally, with the three evaluation metrics PCC, RMSE and R2 averaged over the subjects in the two datasets of DB2 and DB7 as 0.8746±0.0187, 9.4162±1.0043, 0.7479±0.0413 and 0.8773± 0.0309, 9.2249±1.0553, and 0.7539±0.0614, which were significantly better than PNCB (DB2: 0.8518±0.0196, 10.0770±1.0142, and 0.713 6± 0.0418; DB7: 0.8525±0.0309, 10.0587±1.0512, and 0.7065 ± 0.0628), BERT (DB2: 0.8518 ± 0.0 196, 10.0770 ± 1.0142, 0.7136 ± 0.0418; DB7: 0.8404 ± 0.0134, 10.6100 ± 0.9725, 0.6460 ± 0.0 641), Transformer (DB2: 0.8169±0.0135, 11.4722 ±1.3514, 0.6020±0.0674; DB7:0.8212±0.0277,11 .1992±1.3016, 0.6085±0.0581), demonstrated their strong generalization ability and performance. As for TCN and LSTM, the two classical single network model architectures, their performance on the two datasets is relatively poor because they can only extract part of the features of sEMG. Meanwhile, sMCA-Transformer and the other five comparative algorithms are statistically tested to be significantly different, and the model outperforms the other comparative algorithms in all three evaluation metrics (p-values are less than 0.05).



(A) **(B)** **(C)**

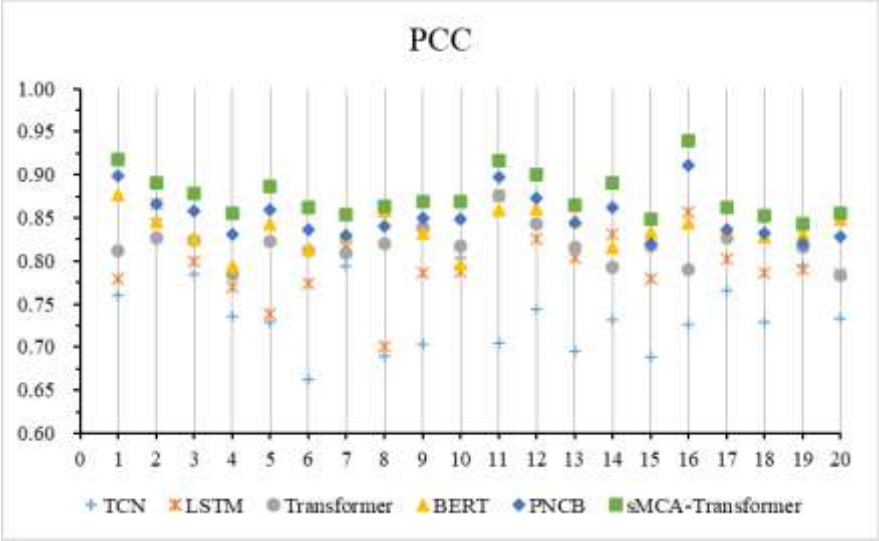
Fig. 5. Plot of the average results of the three metrics on DB2 subjects for the different methods



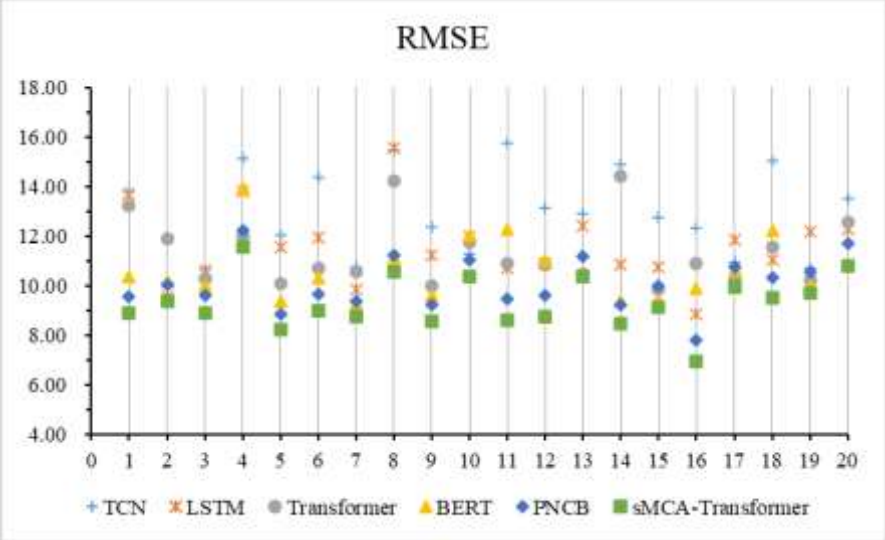
(A) **(B)** **(C)**

Fig. 6. Plot of the average results of the three metrics on DB7 subjects for the different methods

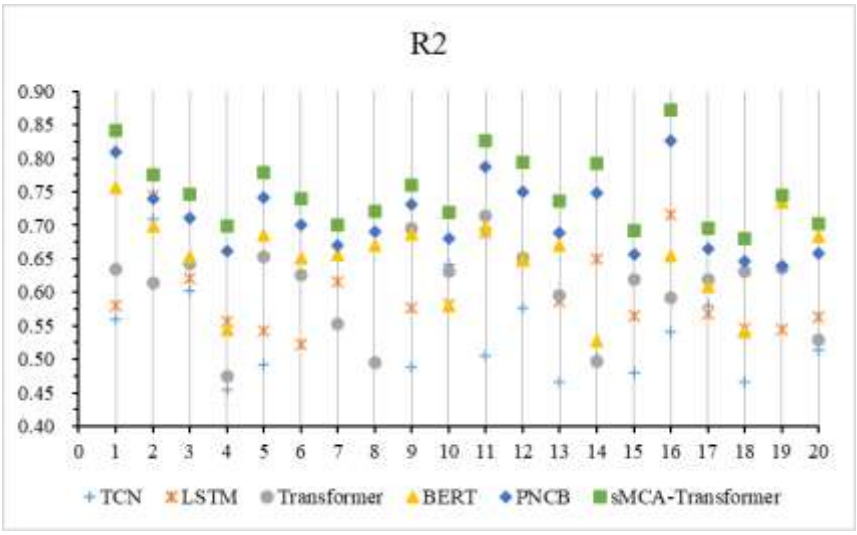
Figures 7(A)-(C) show the accuracy performance of sMCA-Transformer and the other five comparison algorithms in 20 subjects.The subjects numbered 1-10 are those in DB2 and 11-20 are those in DB7.It is easy to see that the PCC values of sMCA-Transformer remain at a high level in all subjects, which are all greater than 0.8430, reaching a strong correlation level.Meanwhile, the sMCA-Transformer model maintained a low RMSE (less than 11.5472) and a high R2 (greater than 0.6812) in all 20 study subjects, with PNCB, BERT and Transformer performing next best, while TCN and LSTM performed significantly worse.The above results indicate that sMCA-Transformer has better performance and generalizability than the other five comparison algorithms, and is able to maintain a high level of accuracy across subjects with different datasets.



(A)



(B)

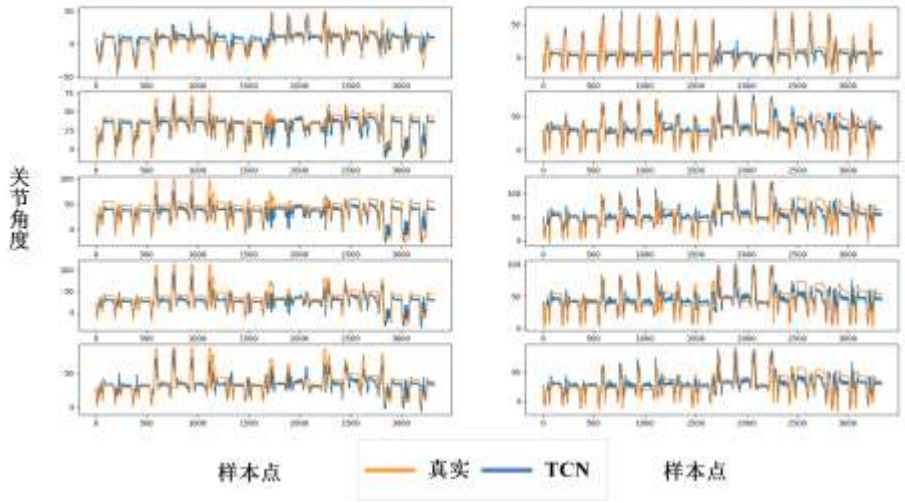


(C)

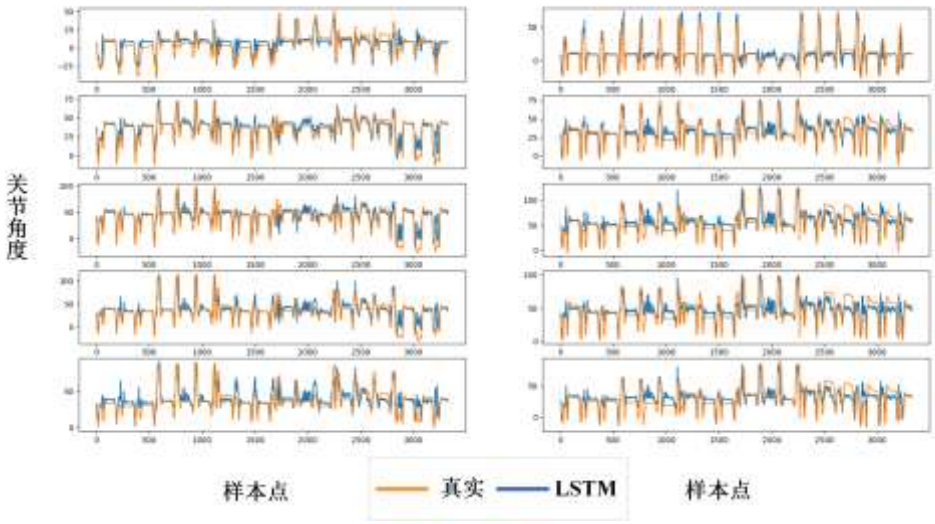
Fig. 7. Comparison of the results of the three evaluation metrics for different methods on 20 subjects

Figures 8(A)-(C) show the comparison between the actual joint angle curves of the first subject and the joint angle curves predicted by TCN, LSTM and sMCA-Transformer, respectively, which are used to visually evaluate the prediction performance of each model.

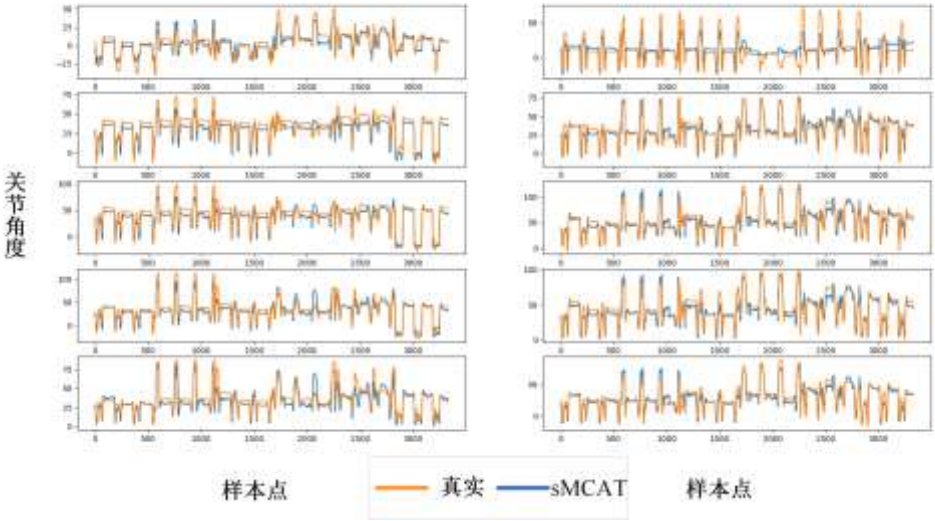
As can be seen from the figure, the sMCA-Transformer predicted joint angle curves have the best consistency with the actual joint angle curves, showing higher stability and accuracy, and closer to the fluidity of the human body movement process; LSTM is the second best, predicting the signal trend basically accurately, but with fluctuation; and TCN can also predict the signal trend, but with higher fluctuation and more obvious error.obvious.In contrast, the joint angle prediction curve of sMCA-Transformer is smoother and closer to the actual movement posture of the hand, which is attributed to the strong generalization ability of the combination of convolutional neural network and Transformer, as well as the smoothing effect of the smoothing algorithm module.



(A)



(B)



(C)

Fig. 8 Actual joint angle profiles and the joint angle profiles predicted by (A) TCN, (B) LSTM and (C) sMCA-Transformer

The inference time is the time for the model to estimate the hand motion from the sEMG, too long inference time may lead to poor user experience in practice and difficult to deploy the application on the edge device, considering the human muscle reaction time is usually between 200-300ms, so if the model can be deployed on the edge device and provide a more consistent interaction, the inference time of the continuous motion estimation model should be no higher than 200 ms.

In this paper, we do a quantitative evaluation of the average inference time of the proposed models and the compared methods.

The results show that TCN has the smallest inference delay (less than 50ms), which is slightly better than sMCA-Transformer (less than 100ms), followed by PNCB and Transformer.These models can be deployed at the edge devices and provide more coherent interactions, which satisfy the real-time performance requirements for continuous motion estimation.However, the inference times of both BERT and LSTM are

greater than 200ms, making it difficult to meet the real-time requirements in practical applications.

Tab. 1 Performance of different models on edge devices

Models	parameter number (mb)	Inference (ms)	energy consumption (w)
sTransformer[Error! Reference source not found.]	3.43	116.43	3.10
sTransformer-EMSA[Error! Reference source not found.]	3.43	102.55	3.00
sBERT[Error! Reference source not found.]	59.63	336.99	2.20
sMCA-Transformer	2.94	96.37	2.60

Table 2 demonstrates the performance evaluation of the model proposed in this study with the advanced deep learning algorithm models proposed in recent years in the field of continuous motion estimation of hand joint angles, choosing the ARM architecture device as the benchmark test platform, and focusing on the parameter scale, inference time, and energy consumption of the model. The results show that the improvement of the Transformer model makes its parameter number drop to 2.94 mb and its inference time drop to less than 100 ms. Although the power consumption rises compared to sBERT, its inference time is obviously faster than sBERT, which indicates that it sacrifices a small amount of power consumption in exchange for a significant increase in performance.

Tab. 2 Inference speed of different models on edge devices

Models	Inference(ms)
TCN	52.36
LSTM	356.67
Transformer	125.58
BERT	303.22
PNCB	102.68
sMCA-Transformer	96.37

Discussion

This chapter first introduces the multiple features extracted in the preprocessing stage of surface EMG data and the fusion strategy; then it introduces the overall architecture of the sMCA-Transformer model and describes its components in turn.

The model consists of a multiscale convolutional attention network and a modified Transformer network serially connected, which integrates the advantages of the two network architectures and is capable of extracting features in multiple scale dimensions of the sEMG, in addition to the addition of a smoothing algorithm to further improve the accuracy and noise immunity of the model.

After that, the proposed algorithmic model is validated on two datasets, Ninapro DB2 and DB7, and the experimental results show that the algorithmic model proposed in this part significantly outperforms the other five state-of-the-art comparative algorithmic models. In addition, new simulations are added in this paper to verify the effect of noise that may occur in practice, and the results show that sMCA-Transformer has good noise immunity and robustness.

Conclusion

This study focuses on the development of continuous hand motion estimation algorithms based on sEMG. By analyzing the limitations of the existing algorithms and integrating the advantages of multiple neural networks, a better continuous hand motion estimation model is constructed, which will ultimately show the potential for applications in the fields of motion control, human-computer interaction and biomedicine. Aiming at the characteristics of sEMG signals, this study implements key improvements in terms of algorithm accuracy, computational efficiency and generalization ability, aiming to achieve efficient and accurate continuous hand motion estimation. The main research content of this paper is summarized as follows:

In order to solve the problems of few compatible movements and poor robustness of existing deep learning models, this study proposes a smooth multi-scale convolutional attention Transformer network. The model consists of a serial connection between a multi-scale convolutional channel attention network and an improved Transformer network, which integrates the advantages of the two network architectures and is capable of extracting features in multiple scale dimensions of the sEMG, in addition to the addition of a smoothing algorithm to further improve the accuracy and noise immunity of the model. Notably, compared to previous experiments, the network expands the number of predicted grasping action types to 12, which satisfies most grasping scenarios in daily life.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: [https://ninapro.hevs.ch/] (15 March 2024).

References

1. Artemiadis, P. (2012). EMG-based Robot Control Interfaces: Past, Present and Future. Advances in Robotics & Automation, 01(02), pp.1–3. doi:https://doi.org/10.4172/2168-9695. 1000e107.
2. Atzori, M. and Muller, H. (2015). The Ninapro database: A resource for sEMG naturally controlled robotic hand prosthetics. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.7151–7154. doi: https://doi.org/10.1109/embc.2015.7320041.
3. Bai, S., Kolter, J.Z. and Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. [online] arXiv.org. Available at: https://arxiv.org/abs/1803.01271
4. Benalcázar, M.E., González, J., Jaramillo-Yánez, A., Anchundia, C.E., Zambrano, P. and Segura, M. (2020). A Model for Real-Time Hand Gesture Recognition Using Electromyography (EMG), Covariances and Feed-Forward Artificial Neural Networks. 2020 IEEE ANDESCON, pp.1–6. doi: https://doi.org/10.1109/ANDESCON50619.2020.9271979.
5. Bi, L. and Guan, C. (2019). A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. Biomedical Signal Processing and Control, 51, pp.113–127. doi: https://doi.org/10.1016/j.bspc.2019.02.011.
6. Bodaghi, M., Hosseini, M. and Gottumukkala, R. (2024). A Multimodal Intermediate Fusion Network with Manifold Learning for Stress Detection. arXiv (Cornell University). doi: https://doi.org/10.48550/arxiv.2403.08077.
7. Chen, J., Bi, S., Zhang, G. and Cao, G. (2020). High-Density Surface EMG-Based Gesture Recognition Using a 3D Convolutional Neural Network. Sensors, 20(4),p.1201. doi: https://doi.org/10.3390/s20041201.
8. Chen, Y., Dai, C. and Chen, W. (2020). Cross-Comparison of EMG-to-Force Methods for Multi-DoF Finger Force Prediction Using

One-DoF Training. IEEE Access, 8, pp. 13 958–13968. doi: https://doi.org/10.1109/acess .2020. 2966007.

9. Chen, Y., Yu, S., Ma, K., Huang, S., Li, G., Cai, S., et al. (2019). A Continuous Estimation Model of Upper Limb Joint Angles by Using Surface Electromyography and Deep Learning Method. IEEE Access, 7, pp.174940–174950. doi: https://doi.org/10.11 09/access.2019.2956951.
10. Gautam, A., Panwar, M., Biswas, D. and Acharyya, A. (2020). MyoNet: A Transfer-Learning-Based LRCN for Lower Limb Movement Recognition and Knee Joint Angle Prediction for Remote Monitoring of Rehabilitation Progress From sEMG. IEEE Journal of Translational Engineering in Health and Medicine, 8, pp.1–10. doi: https://doi.org /10.1109/jtehm.2020.2972523.
11. Hakonen, M., Piitulainen, H. and Visala, A. (2015). Current state of digital signal processing in myoelectric interfaces and related applications. Biomedical Signal Processing and Control, 18, pp.334–359. doi: https://doi.org/10.1016/j.bspc.2015.02.009.
12. Jandaghi, E., Chen, X. and Yuan, C. (2023). Motion Dynamics Modeling and Fault Detection of a Soft Trunk Robot. 2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp.1324–1329. doi: https://doi.org/10.1109/ aim463 23.2023.10196206.
13. Jang, G., Kim, J.-H., Lee, S. and Choi, Y. (2016). EMG-Based Continuous Control Scheme With Simple Classifier for Electric-Powered Wheelchair. IEEE Transactions on Industrial Electronics, 63(6), pp.3695–3705. doi:https://doi.org/10.1109/tie.2016.2522385.
14. Kaplan, K.E., Nichols, K.A. and Okamura, A.M. (2016). Toward human-robot collaboration in surgery: Performance assessment of human and robotic agents in an inclusion segmentation task. 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 20 16,pp.723–729. doi: https://doi.org/10.110 9/icra.2016.7487199.
15. Kingma, D.P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. [online] arXiv.org. Available at: https://arxiv.org/abs /1412.6980.
16. Liu, P., Liu, L. and Clancy, E.A. (2015). Influence of Joint Angle on EMG-Torque Model During Constant-Posture, Torque-Varying Contractions. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 23(6),pp.1039–1046. doi: https:// doi.org/10.1109/tnsre.2015.2405765.
17. Ma, C., Guo, W., Zhang, H., Samuel, O.W., Ji, X., Xu, L., et al. (2021a). A Novel and Efficient Feature Extraction Method for Deep Learning Based Continuous Estimation. IEEE robotics and automation letters, 6(4), pp. 7341–7348. doi:https://doi.org/10.1109/lra.20 21.3097257.
18. Ma, C., Lin, C., Samuel, O.W., Guo, W., Zhang, H., Greenwald, S., et al. (2021b). A Bi-Directional LSTM Network for Estimating Continuous Upper Limb Movement From Surface Electromyography. IEEE robotics and automation letters, 6(4), pp.7217–7224. doi:https://doi.org/10.1109/lra.2021.3097272.
19. Park, K.H. and Lee, S.W. (2016). Movement intention decoding based on deep learning for multiuser myoelectric interfaces. 2016 4th International Winter Conference on Brain-Computer Interface (BCI), pp.1–2. doi: https: //doi.org/10.1109/iww-bci.2016.7457459.
20. Patel, G.K., Castellini, C., Hahne, J.M., Farina, D. and Dosen, S. (2018). A Classification Method for Myoelectric Control of Hand Prostheses Inspired by Muscle Coordination. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26 (9), pp.1745–1755. doi: https://doi.org/10. 1109/tnsre.2018.2861774.
21. Siami-Namini, S., Tavakoli, N. and Namin, A.S. (2019). The Performance of LSTM and

BiLSTM in Forecasting Time Series. 2019 IEEE International Conference on Big Data (Big Data), pp.3285–3292. doi: https://doi.org/10.1109/bigdata47090.2019.9005997.

22. Song, Y., Gao, S., Li, Y., Jia, L., Li, Q. and Pang, F. (2020). Distributed Attention-Based Temporal Convolutional Network for Remaining Useful Life Prediction. IEEE Internet of Things Journal, 8(12), pp.9594–9602. doi: https://doi.org/10.1109/jiot.2020.3004452.
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. (2017). Attention Is All You Need. [online] arXiv.org. Available at: https://arxiv.org/abs/1706.03762.
24. Wang, L. and Buchanan, T.S. (2002). Prediction of joint moments using a neural network model of muscle activations from EMG signals. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10(1),pp.30–37. doi:https://doi.org/10.1109/tnsre.2002.1021584.
25. Xiong, D., Zhang, D., Zhao, X. and Zhao, Y. (2021). Deep Learning for EMG-based Human-Machine Interaction: A Review. IEEE/CAA Journal of Automatica Sinica, 8(3),pp.512–533. doi:https://doi.org/10.1109/jas.2021.1003865.
26. Xu, G., Meng, Y., Qiu, X., Yu, Z. and Wu, X. (2019). Sentiment Analysis of Comment Texts Based on BiLSTM. IEEE Access, 7, pp.51522–51532. doi:https://doi.org/10.1109/access.2019.2909919.
27. Zanghieri, M., Benatti, S., Conti, F., Burrello, A. and Benini, L. (2020). Temporal Variability Analysis in sEMG Hand Grasp Recognition using Temporal Convolutional Networks. 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), pp.228–232. doi: https://doi.org/10.1109/aicas48895.2020.9073888.
28. Zhang, J., Chen, X., Jandaghi, E., Zeng, W., Zhou, M. and Yuan, C. (2023). Dynamics Learning-Based Fault Isolation for A Soft Trunk Robot. 2023 American Control Conference (ACC), pp.40–45. doi: https://doi.org/10.23919/acc55779.2023.10156314.
29. Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., et al. (2021). Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. Medical Image Analysis, 70, p.101918. doi: https://doi.org/10.1016/j.media.2020.101918.