Original Article

DOI: 10.5582/bst.2024.01370

Multimodal optimal matching and augmentation method for small sample gesture recognition

Wenli Zhang[1,*], Bo Liu[1], Tingsong Zhao[1], Shuyan Qie[2]

[1] Faculty of Information Science and Technology, Beijing University of Technology, Beijing, China;
[2] Department of Rehabilitation, Beijing Rehabilitation Hospital Capital Medical University, Beijing, China.

SUMMARY: In human-computer interaction, gesture recognition based on physiological signals offers advantages such as a more natural and fast interaction mode and less constrained by the environment than visual-based. Surface electromyography-based gesture recognition has significantly progressed. However, since individuals have physical differences, researchers must collect data multiple times from each user to train the deep learning model. This data acquisition process can be particularly burdensome for non-healthy users. Researchers are currently exploring transfer learning and data augmentation techniques to enhance the accuracy of small-sample gesture recognition models. However, challenges persist, such as negative transfer and limited diversity in training samples, leading to suboptimal recognition performance. Therefore, We introduce motion information into sEMG-based recognition and propose a multimodal optimal matching and augmentation method for small sample gesture recognition, achieving efficient gesture recognition with only one acquisition per gesture. Firstly, this method utilizes the optimal matching signal selection module to select the most similar signals from the existing data to the new user as the training set, reducing inter-domain differences. Secondly, the similarity calculation augmentation module enhances the diversity of the training set. Finally, the Modal-type embedding enhances the information interaction between each mode signal. We evaluated the effectiveness on Self-collected Stroke Patient, the Ninapro DB1 dataset and the Ninapro DB5 dataset and achieved accuracies of 93.69%, 91.65% and 98.56%, respectively. These results demonstrate that the method achieved performance comparable to traditional recognition models while significantly reducing the collected data.

Keywords: Neuro-robotics, gesture recognition, small sample, rehabilitation therapy, signal similarity

1. Introduction

With the advancement of sensor and deep learning technology, gesture recognition based on physiological signals has been widely applied in human-computer interaction. It finds applications in various areas, such as sign language recognition, robot control, virtual reality, and prosthetic control (*1-4*). Utilizing physiological signals to capture gestures is a more natural and immersive interaction, offering advantages such as low latency and computational requirements. Among physiological signals, surface electromyography (sEMG) signals are particularly suitable for capturing muscle activities. By placing electrodes on the skin's surface in the region of interest, muscle action potentials can be measured without causing harm to the human body. Consequently, acquiring and recognizing gestures from sEMG signals have become a hot research topic in related fields (*5*).

In previous research, we proposed a long-short-term feature fusion network called LST-EMG-Net for sEMG gesture recognition (*6*). The network is entirely designed based on attention mechanisms. LST-EMG-Net extracts long-term and short-term features separately and employs a feature cross-attention module to fuse them, addressing the mismatch between the extracted feature information and the information required for gesture recognition. On the DB2 E2, Ninapro DB5 E3, and CapgMyo DB-C datasets, LST-EMG-Net achieved accuracies of 81.47%, 88.24%, and 98.80%, respectively. Compared to the state-of-the-art methods in the literature, it improved the accuracy by 2.70%, 4.49%, and 0.42% for the respective datasets, enhancing the accuracy and stability of gesture recognition across various classes. However, the challenge of a heavy user data collection burden still exists:

The datasets used by traditional gesture recognition models such as LST-EMG-Net are obtained under ideal conditions, where each participant has sufficient data to train individual recognition models. For example, in the Ninapro public dataset (*7-10*), each sub-dataset gesture typically requires patients to repeat it 6 to 10 times. Each participant must perform continuous arm movements for approximately half an hour when collecting over a dozen gestures. Prolonged data collection leads to muscle fatigue in the participants and affects the data quality (*11*). This data collection burden, particularly in stroke or disabled patients, imposes significant physical and time costs on the patients. Moreover, there exist differences among participants in terms of height, weight, body mass index (BMI), and the amount of fat in the superficial muscles. Even when performing the same gesture, individuals exhibit significant variations in signal distribution (*12*). As a result, personalized small-sample signal-based models and models trained on signals from other individuals struggle to achieve the desired recognition accuracy.

Related researchers generally propose small-sample gesture recognition methods in terms of either transfer learning (TL) or signal augmentation to reduce the user's acquisition burden.

Transfer learning Small-sample gesture recognition methods typically use existing user data as the source domain and the new user data as the target domain. Researchers design transfer strategies from the perspectives of data, features, or models to effectively recognize the target domain (*13-20*).

Kanoga *et al*. (*13*) proposed a transfer framework that projects the source domain data onto the target domain data distribution through linear projection. Azab *et al*. (*14*) introduced a data transfer method based on Kullback-Leibler (K-L) divergence measurement. Wang *et al*. (*15*)presented a multi-source integration transfer learning (MSITL) approach to explore cross-user gesture recognition. It involves training recognition models for each source domain (user) and fine-tuning them using the target domain (new user) data evaluation scores. Colli Alfaro *et al*. (*16*) introduced IMU data to the existing EMG signals of subjects. Multiple pre-trained prediction models are created for each source data and fine-tuned using an adaptive least squares support vector machine (LS-SVM) to select the model with the highest accuracy. Sheng *et al*. (*17*) proposed a general framework called the common spatial-spectral analysis (CSSA) framework. Campbell *et al*. (*18*) introduced an Adaptive Domain Adversarial Neural Network (ADANN), which freezes certain layers and fine-tunes others when adding new subjects. Tsinganos *et al*. (*19*) proposed a new convolutional neural network (TSNet) that combines both temporal and spatial features, as well as an improved version of AtzoriNet denoted as AtzoriNet*. They trained the network models using data from multiple participants (source domain) and then fine-tuned the model weights using data from the target domain to recognize gestures performed by new users. Yu *et al*. (*20*) employed a similar approach, utilizing the source domain data to train an improved CNN model and fine-tuning the fully connected layers with target domain data for recognizing gestures from new subjects. We summarize the relevant studies on transfer learning for small sample recognition from the perspectives of methodology, datasets, number of subjects, types of gestures, signal types, and accuracy, as shown in Table 1.

In summary, applying transfer learning to small sample gesture recognition can somewhat improve recognition performance. However, due to the inherent differences in the physical characteristics of physiological signals such as sEMG signals among different individuals and the influence of factors such as environmental noise and body posture, there are significant differences in the feature distributions of signals between the source and target domains. Consequently, knowledge learned in the source domain may not be applicable to the target domain, resulting in a performance decline when the knowledge or model learned from the source domain data is applied to the target domain data. This can lead to the problem of negative transfer (*21*), which affects the model's recognition performance.

In addition to the methods mentioned above, in tasks such as emotion recognition (*22-25*), researchers have employed multimodal data fusion methods to combine data from different types of sensors. The method aims to obtain comprehensive, accurate, and reliable physiological information, thereby improving the recognition accuracy and generalization of the models. Although the application of multimodal data fusion methods in small-sample gesture recognition is currently limited, this approach provides valuable insights. It allows us to complement the general physiological information of gesture movements, reduce inter-user domain differences, and enhance the accuracy of small-sample gesture recognition.

Based on the above problems and research status, in order to maximize the accuracy of small-sample gesture recognition we propose a multimodal optimal matching and augmentation method for Small sample gesture recognition. The method can effectively address the negative transfer problem and enhance the diversity of signals in the training set. With only one data collection for each gesture, this method achieves comparable accuracy to traditional gesture recognition models based on individual data. The main contributions of this paper are as follows: In response to the negative transfer problem caused by the significant domain differences between the source and target domains, we analyze the characteristics of signals from different individuals for the same gesture. We propose an optimal matching signal selection module that calculates the similarity between existing and new user signals from a time-frequency perspective. This module selects high-similarity signals to form the optimal matching signal training set, enhancing the similarity between the source

Table 1. Summary of Related Research on Transfer Learning-Based Small-Sample Gesture Recognition

References	Methods	Dataset	Number of Participants	Types of Gestures	Signal Types	Accuracy
Kanoga *et al.* (*13*)	SVM	1-DoF	25 Intact Subjects	8	EMG	91.96 ± 5.75%
		2-DoFs		14		63.28 ± 10.43%
Wang *et al.* (*15*)	MSITL	NinaPro DB1	27 Intact Subjects	52	EMG	69.93% ± 4.33%
		CapgMyo DB-a	18 Intact Subjects	8		86.62% ± 5.68%
		CapgMyo DB-b	10 Intact Subjects	8		88.35% ± 4.67%
		CapgMyo DB-c	10 Intact Subjects	12		74.61% ± 5.45%
Colli Alfaro *et al.* (*16*)	LS-SVM	Self-collected dataset	22 Intact Subjects	7	EMG + IMU	84.6% ± 7.3%
Sheng *et al.* (*17*)	CSSA	Self-collected dataset	7 Intact Subjects	11	EMG	63.83%
				9		72.84%
				7		86.03%
Campbell *et al.* (*18*)	ADANN	Intact-limb and amputee datasets	10 Intact Subjects	10	EMG	86.8 – 96.2%
			5 amputees	10	EMG	64.1 – 84.2%
Tsinganos *et al.* (*19*)	TSNet	NinaPro DB7	20 Intact Subjects	17	EMG	91.93% ± 4.29%
	AtzoriNet*					90.57% ± 4.43%
Yu *et al.* (*20*)	CNN	CapgMyo DB-a	18 Intact Subjects	8	EMG	86.7%
		CapgMyo DB-c	10 Intact Subjects	12		84.47%
		NinaPro DB1	27 Intact Subjects	12		74.7%

and target domains and avoiding the occurrence of negative transfer phenomena.

2. Materials and Method

The overall framework of the proposed multimodal optimal matching and augmentation method is depicted in Figure 1. The framework consists of three main components: the optimal matching signal selection module, the similarity calculation augmentation module, and the multimodal LST-EMG-Net.

Optimal matching signal selection module: This module calculates the similarity between new user-calibrated gestures signals and the database's signals. After adaptive selection based on multimodal signals, it outputs the optimal matching signals as part of the training set.

Signal augmentation module: This module utilizes Variational Autoencoder (VAE) to expand the multimodal signal samples of new user calibrated gestures by using a reconstruction loss based on time-domain similarity calculation to compute the difference between the generated samples and the original samples to optimize the generation of samples. The enhanced samples are then outputted as another part of the training set.

Multimodal LST-EMG-Net: This network inputs the training set consisting of the optimal matching signal and augmentation samples. The sEMG and motion signals in each sample are separated with Modal-type embedding and fed into the LST-EMG-Net to extract features for gesture recognition.

The application process of the multimodal optimal matching and augmentation gesture recognition method consists of two stages:

Model training stage: The stage involves organizing the multimodal signals of existing users into a multimodal signal database based on the number of gesture repetitions. A single repetition of the gesture signal from a new user is collected as the calibration gesture signal. Once the data collection is complete, the new user's calibration gesture dataset and the multimodal signal database are inputted into the optimal matching signal screening module and the signal enhancement module based on signal similarity calculation. The obtained training set is then inputted into the Multimodal LST-EMG-Net for training, constructing a gesture recognition model.

Online gesture recognition stage: The user wears the collection device to capture real-time multimodal signals. These signals are then input into the trained Multimodal LST-EMG-Net to obtain the gesture category.

2.1. Multimodal Datasets

We evaluate our small-sample gesture recognition using

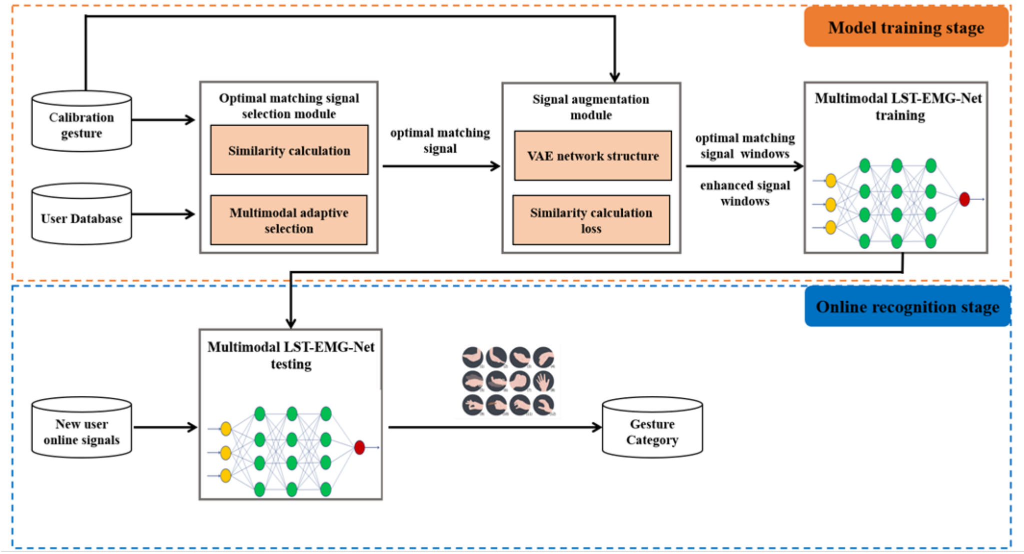

Model training stage
Calibration gesture
User Database
Optimal matching signal selection module
Similarity calculation
Multimodal adaptive selection
optimal matching signal
Signal augmentation module
VAE network structure
Similarity calculation loss
optimal matching signal windows
enhanced signal windows
Multimodal LST-EMG-Net training
Online recognition stage
New user online signals
Multimodal LST-EMG-Net testing
Gesture Category

Figure 1. Overall framework of multimodal optimal matching and augmentation method.

the self-collected stroke patient dataset, the public dataset Ninapro DB1 and the public dataset Ninapro DB5.

Self-collected Stroke Patient Dataset: This dataset was obtained from 6 stroke patients (5 males, 1 female, aged 57–68 years) at the Beijing rehabilitation hospital. Under the guidance of professional rehabilitation physicians, the patients used an MYO armband (as shown in Figure 2) to collect multimodal signals of 7 commonly used hand gestures on their unaffected side, which are beneficial for muscle recovery in daily activities. Each hand gesture was recorded six times, with a duration of 5 seconds per repetition and a 3-second rest period. A 30-second rest interval was provided between consecutive hand gestures. Before data collection, participants received detailed explanations about the experiment, and their informed consent was obtained. The study adhered to the principles of the Helsinki Declaration and obtained ethical approval from the ethics committee (*This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Beijing Rehabilitation Hospital Affiliated with Capital Medical University No. 2022bkky-048*).

Ninapro DB1 Dataset: We selected Ninapro DB1 Exercise C's 7 dynamic gestures (as shown in Figure 3(b)) that are beneficial for muscle recovery training. The DB1 dataset was collected from 27 participants using ten electrodes (Otto Bock MyoBock 13E200) and a data glove, capturing 10 channels of sEMG signals and 22 channels of hand and finger joint motion information. The sampling rate for each electrode was 100Hz. In Ninapro DB1 Exercise C, each hand gesture was repeated 10 times, with a 5-second duration for each active signal collection and a 3-second rest interval between each collection.

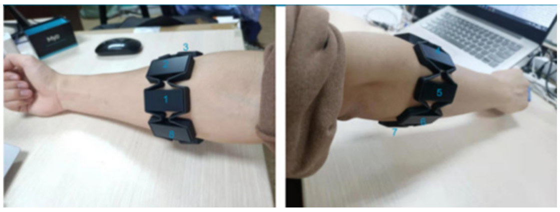


Figure 2. MYO ring acquisition signal schematic. The 7 hand gestures include fist grip, holding a cellphone, palm-to-palm exercise (cupping), cylindrical grip (holding a water cup), finger opposition exercise (thumb and index finger gripping a pen), single finger extension (extending index finger to touch the screen), and lateral thumb pinch (pinching a key, *etc.*) (as shown in Figure 3(a)). The multimodal signals include 8-channel sEMG signal,3-channel arm acceleration signal,3-channel angular velocity signal,4-channel quaternion signal. Acceleration, angular velocity, and quaternion are inertial measurement units (IMU).

Ninapro DB5 Dataset: We selected Ninapro DB5 Exercise C's 7 dynamic gestures (as shown in Figure 3(c)) which differ from those selected in Ninapro DB1 Exercise Cg. The DB5 dataset was collected from 10 healthy participants using two Myo armbands (Thalmic Labs Myo) and a data glove, capturing 16 channels of sEMG signals and 22 channels of hand and finger joint motion information. The sampling rate for each electrode was 200Hz. In Ninapro DB5 Exercise C, each hand gesture was repeated six times, with a 5-second duration for each active signal collection and a 3-second rest interval between each collection.

In each dataset experiment, for the sEMG signals, a fourth-order Butterworth bandpass filter (20 Hz–500 Hz) was first applied to remove motion artifacts and high-frequency noise, preserving useful motion information. Subsequently, the sEMG signals were standardized using a min-max normalization algorithm. For motion signals, the min-max normalization method was similarly used to map the data to the range of 0–1, facilitating subsequent processing.

2.2. Optimal matching signal selection module

In gesture recognition tasks, when using existing signals to recognize new user signals in the target domain, it is important to select training data from existing signals similar to the target domain. This enables knowledge transfer learning. However, determining the similarity between signals directly through observation can be challenging. When two individuals have different muscle activity patterns, their sEMG and other physiological signals often differ significantly. In such cases, selecting inappropriate source domain signals can lead to negative transfer, resulting in a decline in the performance of the recognition model.

Therefore, we propose an optimal matching signal selection module to select highly similar signals to the target domain, enabling better transfer learning between users. The optimal matching signal selection module consists of signal similarity calculation and multimodal signal adaptation selection, as shown in Figure 4. Specifically, this method first constructs a multimodal signal database, "user1, user2, ..., user n," based on existing user data. Then, the calibration gestures of the new user obtained are compared with the signals in the database using the signal similarity calculation part to calculate the similarity of each modality. Finally, multimodal signal adaptation selection combines highly similar modal data to form the optimally matched signals, creating a new user training set. As shown in Figure 4.

Signal Similarity Calculation: Signal similarity assessment typically involves describing the differences

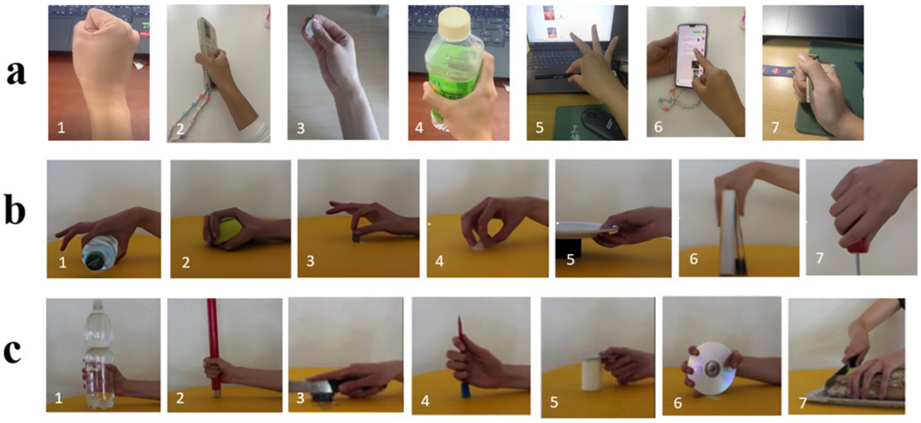


Figure 3. Types of gestures in this paper: (**a**) 7 gestures in the self-harvested stroke patient; (**b**) 7 dynamic gestures in the Ninapro DB1 Exercise C. (**c**) 7 dynamic gestures in the Ninapro DB5 Exercise C.

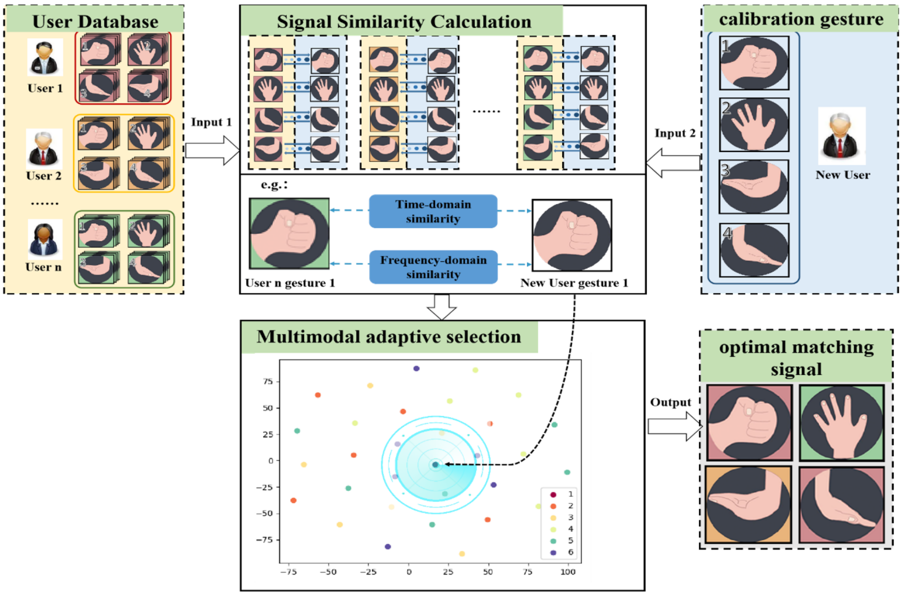


Figure 4. Optimal matching signal selection module.

in shape, spectrum, amplitude, and other features between two signals. Traditional methods mainly focus on time-domain calculations (*26*), but this approach has some limitations. Firstly, it lacks sufficient consideration of the differences in signal spectrum features. Secondly, traditional similarity measurements often use Euclidean distance or Pearson correlation coefficient, which may overlook some important features. Additionally, traditional methods cannot effectively compute the similarity between signals of inconsistent lengths. To overcome these issues, this paper proposes a new method for signal similarity calculation. This method comprehensively and accurately describes the similarity between two signals. Specifically, we consider the similarity between gesture single-cycle signals from both time and frequency domains to comprehensively evaluate their shape, spectrum, and other features. Figure 5 illustrates the detailed similarity calculation process.

Firstly, in terms of time-domain similarity calculation, this paper adopts the Dynamic Time Warping (DTW) algorithm (*27*) to compute the similarity d_{DTW} between two-time series. The time-domain similarity calculation d_{DTW} between two-time series $x = \{x_1,x_2,x_3,\ldots,x_n\}$ and $y = \{y_1,y_2,y_3,\ldots,y_m\}$ is calculated as shown in Equation (1):

$$d_{DTW} = -\gamma\log\left(\sum_{A\epsilon A_{m,n}} e^{-\langle A,\Delta(x,y)\rangle/\gamma}\right) \quad (1)$$

Here, γ represents the smoothing parameter used to select the smoothness of the path, in this study, the default value of γ is set to 1. $\gamma = 1$ indicates a moderate strength of the smoothing factor, which makes the path selection of DTW more focused on the optimal path while still allowing a certain degree of suboptimal paths to contribute to the weight calculation, and $A_{m,n}$ represents the set of alignment matrices, indicating all possible paths that can be selected. $\Delta(x,y)$ is the cost matrix composed of distance values between corresponding points of the two time series. This method effectively overcomes the potential issues caused by inconsistent lengths of time series in describing signal time-domain similarity.

In terms of frequency-domain similarity calculation, we first apply Fast Fourier Transform (FFT) to transform two time series into the frequency domain. Then, we compute the Mean Squared Error (MSE) between the amplitudes of the two frequency-domain signals to calculate the frequency-domain similarity d_{MSE}. The calculation of frequency-domain similarity d_{MSE} between two signals is as shown in Equation (2):

$$d_{MSE} = \frac{1}{len}\sum_{l=0}^{len-1}\left|FFT(x_j) - FFT(y_j)\right|^2 \quad (2)$$

Here, *len* = min(n,m), representing the length of the shorter sequence.

Next, since we aim to fully consider both the time-domain and frequency-domain information of the signals in the final similarity value, it is necessary to scale the obtained time-domain similarity value d_{DTW} and frequency-domain similarity value D_{MSE}. The scaled time-domain similarity value is denoted as D_{TD} and the scaled frequency-domain similarity value is denoted as D_{FD}. The scaling process is illustrated in Equation (3) and Equation (4):

$$D_{TD} = \frac{d_{DTW_i} - avg(d_{DTw_1},\ldots,d_{DTw_i},\ldots,d_{DTw_{N\times L}})}{\max(d_{DTw_1},\ldots,d_{DTw_i},\ldots,d_{DTw_{N\times L}}) - \min(d_{DTw_1},\ldots,d_{DTw_i},\ldots,d_{DTw_{N\times L}})} \times \sigma \quad (3)$$

$$D_{FD} = \frac{d_{MSE_i} - avg(d_{MSE_1},\ldots,d_{MSE_i},\ldots,d_{MSE_{N\times L}})}{\max(d_{MSE_1},\ldots,d_{MSE_i},\ldots,d_{MSE_{N\times L}}) - \min(d_{MSE_1},\ldots,d_{MSE_i},\ldots,d_{MSE_{N\times L}})} \times \sigma \quad (4)$$

σ is the scaling factor used to measure the importance of time-frequency domain similarity, in this study, the scaling factor σ is set to 0.5, indicating that temporal information and frequency information are considered equally important for similarity calculation.

Finally, the scaled time-domain similarity value D_{TD} and frequency-domain similarity value D_{FD} are combined using Equation (5) to obtain the final signal similarity value, denoted as *SSV*. The signal similarity value (SSV) represents the temporal and spectral distance between two signals, with smaller values indicating a smaller temporal and spectral distance and higher similarity between the two signals.

$$SSV = \sqrt{{D_{TD}}^2 + {D_{FD}}^2} \quad (5)$$

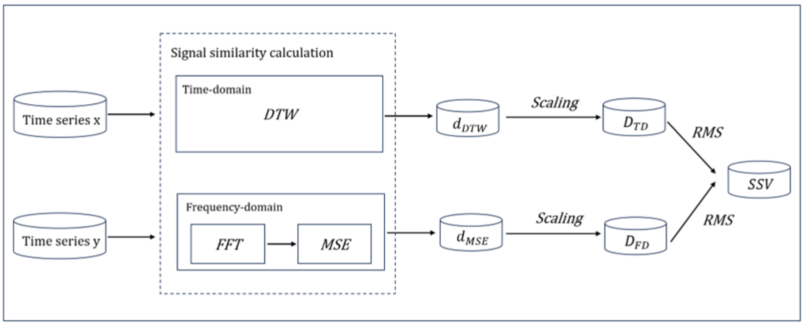


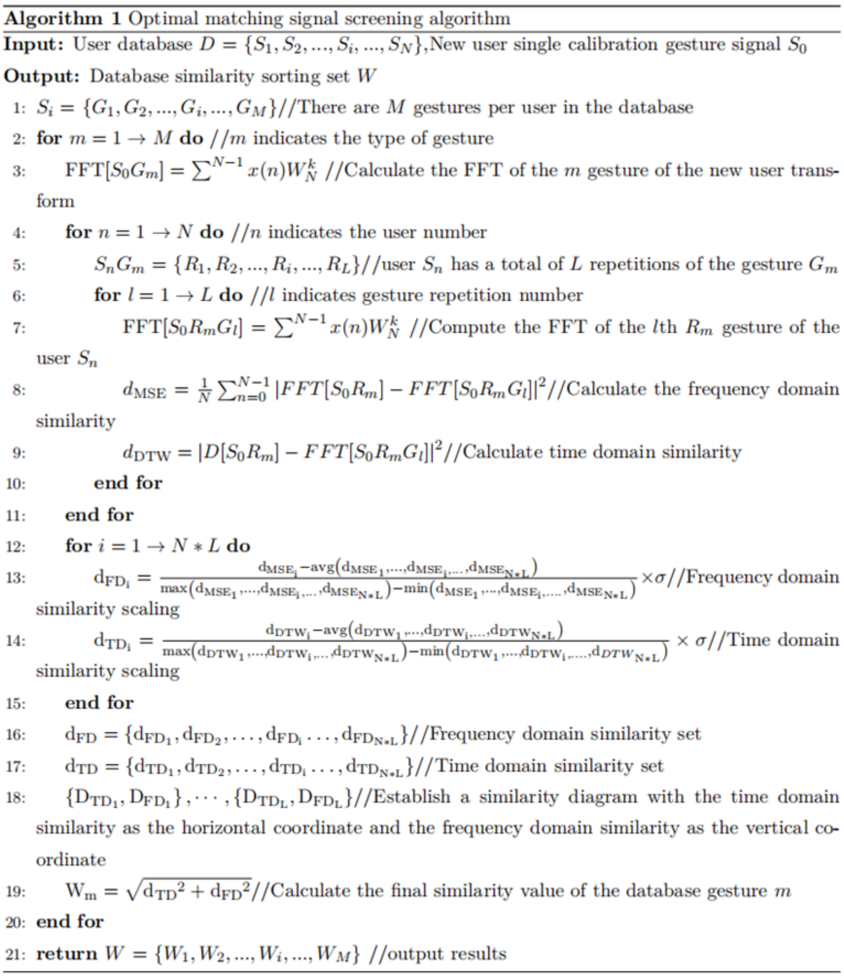
Figure 5. Signal similarity calculation process.

The input of the entire signal similarity calculation section consists of two parts. The first part is the existing user database D, which contains N users, each with M gestures, and each gesture repeated L times. $D = \{G_1 \{R_{11},...,R_{ij},...,R_{NL}\},...,G_k \{R_{11},...,R_{ij},...,R_{NL}\},...,G_M \{R_{11},...,R_{ij},...,R_{NL}\}\}$, where G_k represents the k_{th} gesture, and R_{ij} in G_k represents the j_{th} repetition data of the ith user. The second part is the calibration gesture set D_0 for the new user, which includes only one new user with M gestures, each repeated once. $D_0 = \{G_1 \{R_{01}\},... ,G_k \{R_{01}\},...,G_M \{R_{01}\}\}$, where R_{01} in G_k represents the single calibration gesture data of the new user.

The final output is a set W containing the similarity values between each calibration gesture data of the new user and all the repetitive data for that gesture in the database. $W = \{G_1 \{S_{11},...,S_{ij},...,S_{NL}\},...,G_k \{S_{11},...,S_{ij},... ,S_{NL}\},...,G_M \{S_{11},...,S_{ij},...,S_{NL}\}$, where S_{ij} in G_k represents the similarity value between the single calibration gesture data R_{01} of the new user and the jth repetition data R_{ij} of the ith user in the database. The pseudocode for the overall signal similarity calculation process is illustrated in Figure 6.

In the above pseudocode, the similarity calculation is performed on the signals corresponding to the complete execution of a single gesture. This approach captures the global information within the entire time series of the gesture, accurately and intuitively reflecting the similarity between signals from both the time and frequency domains. It enables sorting the similarity between the signals in the database and the calibration gesture.

Multimodal signal adaptation selection: After completing the signal similarity calculation section and obtaining the similarity value set W for each modality signal in the database, the next step is to select the optimal matching signals to construct the training set based on the sorting of similarities in W. Considering the influence of data quality in the database, relying solely on the average or median of similarity indicators as a threshold and selecting all signals below this threshold as the optimal matching signals cannot guarantee the accurate selection of the optimal signals. Especially when the behavioral patterns of patients in the database are close to those of the new user, there may be a large amount of similar data, while such data may be scarce when the behavioral patterns are



```
Algorithm 1 Optimal matching signal screening algorithm
Input: User database $D = \{S_1, S_2, ..., S_i, ..., S_N\}$,New user single calibration gesture signal $S_0$
Output: Database similarity sorting set $W$
$S_i = \{G_1, G_2, ..., G_i, ..., G_M\}$//There are $M$ gestures per user in the database
for $m = 1 \to M$ do //$m$ indicates the type of gesture
    $\mathrm{FFT}[S_0 G_m] = \sum^{N-1} x(n) W_N^k$ //Calculate the FFT of the $m$ gesture of the new user transform
    for $n = 1 \to N$ do //$n$ indicates the user number
        $S_n G_m = \{R_1, R_2, ..., R_i, ..., R_L\}$//user $S_n$ has a total of $L$ repetitions of the gesture $G_m$
        for $l = 1 \to L$ do //$l$ indicates gesture repetition number
            $\mathrm{FFT}[S_0 R_m G_l] = \sum^{N-1} x(n) W_N^k$ //Compute the FFT of the $l$th $R_m$ gesture of the user $S_n$
            $d_{\mathrm{MSE}} = \frac{1}{N} \sum_{n=0}^{N-1} |FFT[S_0 R_m] - FFT[S_0 R_m G_l]|^2$//Calculate the frequency domain similarity
            $d_{\mathrm{DTW}} = |D[S_0 R_m] - FFT[S_0 R_m G_l]|^2$//Calculate time domain similarity
        end for
    end for
    for $i = 1 \to N * L$ do
        $\mathrm{d_{FD_i}} = \frac{\mathrm{d_{MSE_i}} - \mathrm{avg}(\mathrm{d_{MSE_1}},...,\mathrm{d_{MSE_i}},...,\mathrm{d_{MSE_{N*L}}})}{\max(\mathrm{d_{MSE_1}},...,\mathrm{d_{MSE_i}},...,\mathrm{d_{MSE_{N*L}}}) - \min(\mathrm{d_{MSE_1}},...,\mathrm{d_{MSE_i}},....,\mathrm{d_{MSE_{N*L}}})} \times \sigma$//Frequency domain similarity scaling
        $\mathrm{d_{TD_i}} = \frac{\mathrm{d_{DTW_i}} - \mathrm{avg}(\mathrm{d_{DTW_1}},...,\mathrm{d_{DTW_i}},...,\mathrm{d_{DTW_{N*L}}})}{\max(\mathrm{d_{DTW_1}},...,\mathrm{d_{DTW_i}},...,\mathrm{d_{DTW_{N*L}}}) - \min(\mathrm{d_{DTW_1}},...,\mathrm{d_{DTW_i}},....,\mathrm{d}_{DTW_{N*L}})} \times \sigma$//Time domain similarity scaling
    end for
    $\mathrm{d_{FD}} = \{\mathrm{d_{FD_1}}, \mathrm{d_{FD_2}}, \ldots, \mathrm{d_{FD_i}} \ldots, \mathrm{d_{FD_{N*L}}}\}$//Frequency domain similarity set
    $\mathrm{d_{TD}} = \{\mathrm{d_{TD_1}}, \mathrm{d_{TD_2}}, \ldots, \mathrm{d_{TD_i}} \ldots, \mathrm{d_{TD_{N*L}}}\}$//Time domain similarity set
    $\{\mathrm{D_{TD_1}}, \mathrm{D_{FD_1}}\}, \cdots, \{\mathrm{D_{TD_L}}, \mathrm{D_{FD_L}}\}$//Establish a similarity diagram with the time domain similarity as the horizontal coordinate and the frequency domain similarity as the vertical coordinate
    $\mathrm{W_m} = \sqrt{\mathrm{d_{TD}}^2 + \mathrm{d_{FD}}^2}$//Calculate the final similarity value of the database gesture $m$
end for
return $W = \{W_1, W_2, ..., W_i, ..., W_M\}$ //output results
```

Figure 6. Pseudo-code for signal similarity calculation.

further apart. Therefore, to adaptively select similar data and reduce the domain gap between the training set and the new user, this paper proposes an adaptive selection method for the optimal signals. This method can automatically select high-similarity signals from the database as the optimal matching signals adaptively, denoted as Q representing the number of adaptively selected optimal matching signals. The specific steps are as follows:

Selecting and combining the top Q similar signals: After obtaining the similarity rankings of various modality signals from the signal similarity calculation module, the value of Q is set. Q is initialized to 1, and Q signals are selected from each modality in ascending order of similarity values. After corresponding with each modality signal one by one, they are combined to create the training set.

Train model: The training set obtained from the previous step trains LST-EMG-Net. The model's average recognition accuracy is calculated.

Determination of the optimal matching signals: Increment Q by 1, then repeat steps (1) and (2). The range of Q is from 1 to N×L. Additionally, to reduce the time required to determine the optimal Q value during model training, we stipulate that if $Q = n$, and the recognition accuracy Accuracy_n is greater than Accuracy_{n+1} and Accuracy_{n+2}, then it is considered that the signal similarity is higher at this point, and n is considered the optimal value. The first n similar data points are considered the optimal matching data.

The above steps demonstrate that the adaptive selection of multimodal signals allows for identifying and filtering data from the database, improving recognition accuracy. This approach mitigates the negative transfer caused by significant differences between the source and target domain signals.

2.3. Signal Augmentation Module

Considering the impact of the size of the database on the recognition accuracy of the model by the optimal Match Signal Selection Module, to maximize the recognition rate of the model, this paper adopts the Variational Autoencoder (VAE) (*28*) as an augmentation network to generate new signal samples and enrich the training set.

The basic architecture of the VAE encoder-decoder consists of three parts: the encoder, latent variable generation, and decoder, as shown in Figure 7. The encoder uses a fully connected layer to map the input data to a latent space distribution, which is used to calculate the low-dimensional mean μ and variance σ of each input sEMG signal. The latent variable generation part computes the probability density function $Z = \{Z_1, Z_2, Z_3, \ldots, Z_n\}$ by performing mathematical operations with random noise e and the mean μ and variance σ. Finally, the decoder generates more diverse sEMG signals. The initial learning rate of the VAE is set to 0.0001, with the Adam optimizer used and a batch size of 100. The VAE loss function consists of two components: the reconstruction error loss and the KL divergence loss. To minimize the reconstruction error, this study employs Soft-DTW, a gradient-calculable variation of DTW similarity, as the reconstruction loss function to measure the difference between the original input data and the generated data, thereby enhancing the learning of signal temporal dependencies.

This signal augmentation method used in this study effectively doubles the quantity of sEMG signals, thereby enriching the multi-modal training samples and further contributing to improving the model's recognition accuracy.

2.4. Multimodal LST-EMG-Net

In the linear projection of the model LST-EMG-Net (*6*) that we studied previously, the sEMG segments were transformed into patch tokens and combined with Position Embedding and classification token as input to the sub-encoder, as shown in Figure 8. However, when dealing with multi-modal tasks, the sub-encoder fails to differentiate the modalities of the patch tokens. The self-attention makes capturing the continuity between different modalities challenging and hampers the information interaction between modalities.

Therefore, for multimodal tasks, we introduce a Modal-type embedding as shown by the gray vector in Figure 9. Modal-type embedding was initially

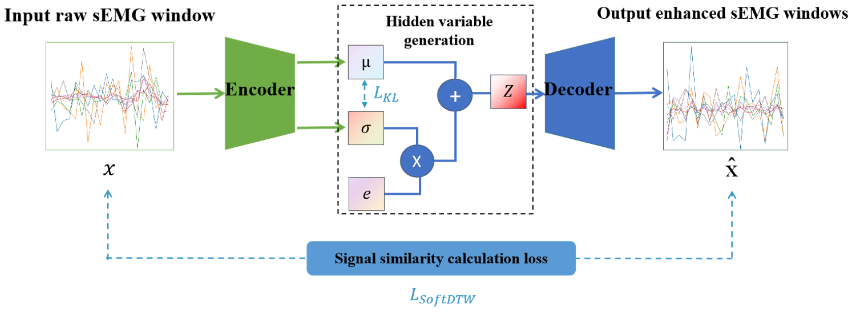

Input raw sEMG window
Encoder
Hidden variable generation
μ
L_{KL}
σ
e
+
X
Z
Decoder
Output enhanced sEMG windows
x
x̂
Signal similarity calculation loss
$L_{SoftDTW}$

Figure 7. Structure and loss function of variational autoencoder network.

introduced by Kim *et al.* (*29*) in 2021 as part of a multimodal model to separately label image and text tokens, enhancing interaction between images and text. In this paper, Modal-type embedding is utilized to label sEMG and motion signals, concatenating with slice and position marker vectors. This allows the embedding of two modalities into different vector spaces of the same dimensionality, facilitating the interaction learning of category information by the encoder. The value "1" represents that the patch token is from the sEMG signal, while "2" represents that the patch token is from the motion signal.

The Modal-type embedding allows distinguishing which type of signal the patch token originates from. The Modal-type embedding is concatenated with the patch tokens and Position Embedding and then input to the sub-encoder. Multimodal LST-EMG-Net facilitates the learning of temporal characteristics within the same type of signals and promotes interaction between different modal signals.

3. Results

This study utilized deep learning frameworks on a computer platform for model training and testing. The computer hardware configuration used is shown in Table 2: Intel Core i7-10700K CPU (64GB RAM), GeForce GTX 3090 GPU (24GB VRAM), operating system Ubuntu 18.04.5 LTS, and programming language Python 3.6.5. The network model was built, trained, and validated using the PyTorch 1.8.0 deep learning framework.

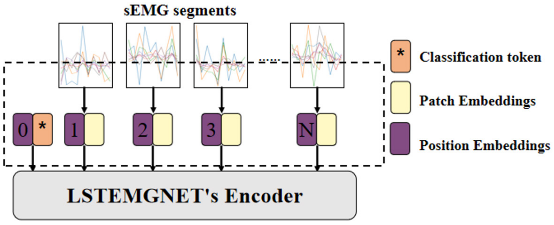


Figure 8. LST-EMG-Net Linear Projection module.

In the validation of the algorithm, each patient in the dataset is taken in turn as a new user and remaining data from other users in the dataset as the database. By repeating this process, we obtained recognition accuracy for each user and calculated the average recognition accuracy. The experiments were divided into three parts:

Optimal matching signal selection experiment : This part presents the results of signal similarity in the time-frequency domain for each modality and demonstrates the change in recognition accuracy when selecting the top N similar signals. Taking the multimodal dataset of patients as an example, it explains the process of selecting the optimal matching signals.

Comparison experiments: This part of the experiment compares the method in this paper with several of the more effective small-sample gesture recognition algorithms that we have summarised, in order to demonstrate the effectiveness of the method proposed in this paper by exploring the differences in the performance of the various algorithms when dealing with small-sample gesture recognition problems.

Ablation experiment: This part of the experiment is divided into two parts, the first part of the experiment are divided into two scenarios based on the composition of the training set. The first scenario involves not using new user data in the training set, while the second scenario involves using new user calibration gestures in the small sample data. The effectiveness of the optimal matching signal selection module, multimodal LST-EMG-Net, and similarity calculation augmentation module are validated sequentially. In the second part of the experiments, we use three different types of data, namely sEMG, IMU, and sEMG + IMU, to verify that multimodal signals yield better results compared to

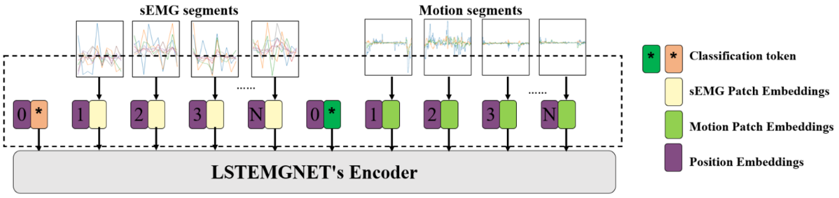


Figure 9. Multimodal LST-EMG-Net Linear Projection module.

Table 2. Computer development environment

Hardware environment	Software environment
CPU: Intel(R) Core(TM) i7-10700K CPU 3.8GHz RAM: 64.00GB System: Ubuntu 18.04.5 LTS GPU: NVIDIA Geforce GTX 3090	Programming language: Python 3.6.5 Deep learning framework: Pytorch 1.8.0 Development tool: JetBrains Pycharm

single-modal signals.

3.1. Optimal matching signal selection experiment

First, we evaluated the effectiveness of the optimal matching signal selection module on the self-collected stroke patient dataset:

The 6 patients are denoted as Subject 1 to Subject 6 (S1-S6). Each patient performed 7 gestures labeled Gesture 1 to Gesture 7 (G1-G7). Each gesture had 6 repetitions labeled as Repetition 1 to Repetition 6 (R1- R6). One of the repetitions R1, R3, R4, or R6 of a gesture was randomly selected as the calibration gesture, and the repetitions R2 and R5 of the same gesture were used as the test set.

Taking S1 as the new user and R1 as the calibration gesture as an example, we used DTW (Dynamic Time Warping) and FFT (Fast Fourier Transform) to calculate the time-frequency domain distances between the user's gestures in the database and the sEMG and IMU (Inertial Measurement Unit) signals of Gesture 1 as the calibration gesture. These distances represented the similarity values between the signals, as shown in Table 3.

Table 3 shows the DTW and MSE distance as time-frequency domain similarities for each modality that Gesture 1 corresponds to a fist-clenching action. Except for Subject 5, the sEMG signals of the other subjects show similar time-frequency domain similarities to the sEMG signal of Subject 1's calibration gesture. However, there are significant differences in the similarity of the IMU signals. This demonstrates that the muscle activation patterns are similar among the subjects for Gesture 1, but there are significant variations in the movement trajectories during fist-clenching. Therefore, cross-user recognition may rely more on the information contained in the sEMG signals. The time-frequency domain similarities from the graph above are then scaled using the multimodal signal-adaptive selection module, resulting in the sEMG and IMU similarity maps shown in Figure 10.

Each point in the graph represents one repetition of Gesture 1, and the point's color indicates the patient's identifier. The point labeled "1" represents the calibration gesture of the new user, while the other points represent the data of each patient in the database. Each patient has six points corresponding to the six repetitions of Gesture 1.

We sequentially select the data based on the distance between each point and the calibration gesture point. The point with the shortest distance is considered the most similar sEMG/IMU signal, the second closest point is the second most similar sEMG/IMU signal, and so on. We select the top N similar data points

Table 3. Time-frequency domain similarity of sEMG and IMU for Gesture 1

Calibration gesture	Patient ID	Database	sEMG time domain similarity	sEMG frequency domain similarity	IMU time domain similarity	IMU frequency domain similarity
S1G1R1	S2	G1R1	19857	87764	24820	136687
		G1R2	17113	81258	18848	113217
		G1R3	17664	87163	28720	163913
		G1R4	15901	76296	3658	27274
		G1R5	14336	79550	1679	16066
		G1R6	14099	72952	2336	20250
	S3	G1R1	15355	70516	5442	19040
		G1R2	14253	67061	12605	100635
		G1R3	14375	66950	2339	8545
		G1R4	14664	70103	1566	7007
		G1R5	15043	69187	1047	11653
		G1R6	14684	69802	844	4373
	S4	G1R1	16142	71191	-522	4498
		G1R2	13512	59287	-342	5393
		G1R3	13201	57734	-202	6011
		G1R4	12773	52954	900	10988
		G1R5	12829	54230	-505	2837
		G1R6	12899	51904	1064	12723
	S5	G1R1	34228	209840	10158	14845
		G1R2	37123	227054	8867	30776
		G1R3	32275	194394	10997	38618
		G1R4	32814	208064	8233	24182
		G1R5	24403	141499	10582	49617
		G1R6	37392	235632	7279	20191
	S6	G1R1	15439	69035	4422	11588
		G1R2	14413	64632	3776	12296
		G1R3	15617	64293	3036	10492
		G1R4	14855	64446	3006	14249
		G1R5	14084	60028	2666	14510
		G1R6	15025	61445	2246	15820

for each gesture, N refers to the number of the top N most similar data points to the current calibration gesture data, selected after sorting all the data based on similarity, repeating this process until all gestures have been used to create the training set. The recognition accuracy is evaluated using the LST-EMG-Net model, and the results are shown in Table 4 as N increases.

Table 4 shows that the average recognition accuracy initially increases and decreases as N increases. When N is less than 3, the training dataset is not saturated, and the selected data is closest to the new user. In this case, increasing N significantly improves the accuracy. As N increases to 3-5, the selected data has moderate similarity with the new user, and the average recognition accuracy fluctuates within a certain range. When N is greater than 5, the selected data have lower similarity with the new user, resulting in negative transfer effects and decreased recognition accuracy. Based on the table, in this experiment on the self-collected multimodal dataset of stroke patients, selecting N=5 achieves the highest average accuracy of 75.20%. This indicates that the selected data at this N value represents the optimal matching data.

In addition, we visualized the best-match signal screening experiments on Ninapro DB5 following the above procedure, as shown in Figure 11.

The graph illustrates that the value of N and the maximum average accuracy are related to the amount of user data in the database. In theory, as the data volume increases, more similar data are in the database, and the N value that achieves the maximum recognition accuracy will shift to the right. Therefore, for the Ninapro DB5, when N is equal to 9, the average recognition accuracy is 91.00%, both N=10 and N=11 have accuracies lower than 91.00%. Hence, for the Ninapro DB5 dataset, N=9 is considered the optimal matching signals.

3.2. Comparison experiments

In Table 1, we have quantitatively summarized existing literature and selected ADANN, TSnet, and AtzoriNet*

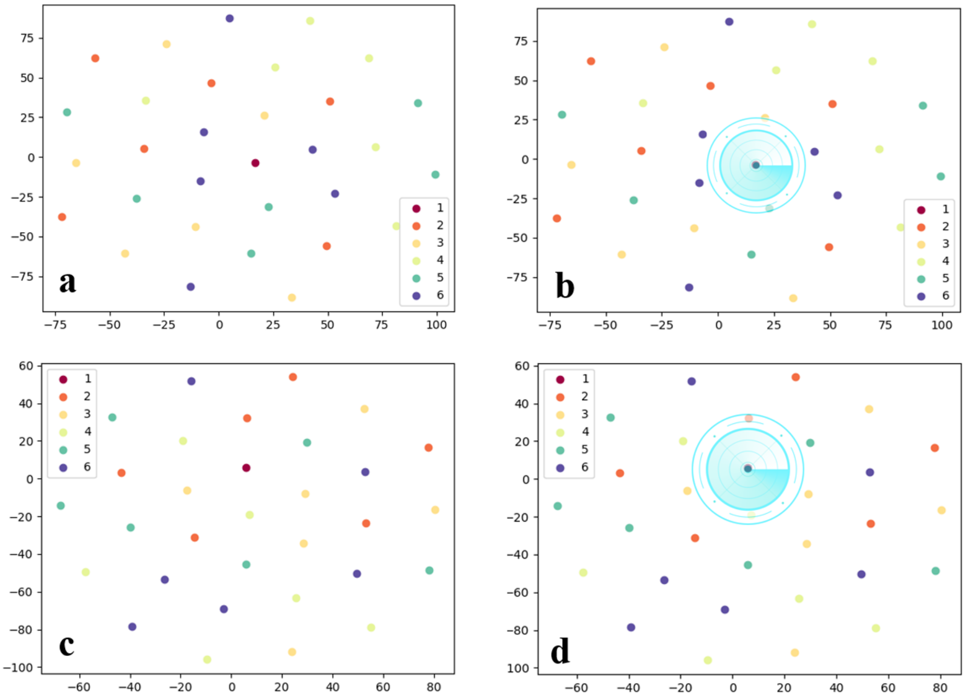


Figure 10. Gesture 1 Similarity Graphs: (**a**) sEMG Similarity Graph (**b**) Selection of Nearest Gestures based on sEMG Calibration Gesture (**c**) IMU Similarity Graph (**d**) Selection of Nearest Gestures based on IMU Calibration Gesture.

Table 4. Recognition Accuracy of Subjects at Different N Values

Selecting Signals N	S1	S2	S3	S4	S5	S6	Average recognition
1	51.97%	64.21%	67.05%	48.56%	65.90%	80.46%	63.03%
2	63.25%	71.30%	76.38%	61.63%	79.37%	79.27%	71.87%
3	74.09%	71.80%	89.89%	55.84%	71.46%	83.09%	74.36%
4	74.43%	63.16%	87.73%	61.99%	68.04%	88.99%	74.06%
5	85.95%	61.03%	89.48%	64.01%	79.26%	77.18%	75.20%
6	76.09%	61.35%	85.97%	58.28%	68.09%	81.60%	71.90%
7	75.32%	59.08%	82.25%	55.34%	65.20%	78.03%	69.20%

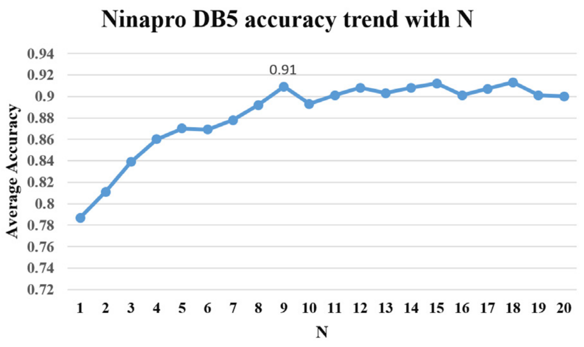


Figure 11. Average recognition accuracy of Ninapro DB5 dataset with N value.

Table 5. Average Accuracy Values of Comparative Algorithms on Each Dataset

Model	Dataset	Average recognition
ADANN	Stroke patients	0.7527
	Ninapro DB5	0.9282
	Ninapro DB1	0.8523
AtzoriNet*	Stroke patients	0.7428
	Ninapro DB5	0.6516
	Ninapro DB1	0.7369
TSNet	Stroke patients	0.6632
	Ninapro DB5	0.6381
	Ninapro DB1	0.8704
Ours	Stroke patients	0.9369
	Ninapro DB5	0.9856
	Ninapro DB1	0.9165

as the three best-performing methods for comparison. In this section, we conduct comparative experiments on our self-collected stroke patient dataset, Ninapro DB1 dataset, and Ninapro DB5 dataset. Table 5 presents the average accuracy values of each comparative algorithm across the three datasets.

From Table 5, it can be observed that our method achieved an average accuracy of 93.69% on the self-collected stroke patient dataset, 98.56% on the Ninapro DB5 dataset, and 91.56% on the Ninapro DB1 dataset. This represents an improvement over ADANN by 18.42%, 5.74%, and 6.42%, respectively, over TSNet by 27.37%, 34.75%, and 4.61%, respectively, and over AtzoriNet* by 19.41%, 33.4%, and 17.96%, respectively. Compared to the comparative algorithms ADANN, TSNet, and AtzoriNet*, our method consistently maintains a relatively high and stable level of average accuracy across all three datasets. This is primarily attributed to the optimal signal matching module, which selects signals most similar to the new user from the existing data as the training set, thereby reducing domain differences to a large extent. Additionally, our proposed LST-EMG-Net effectively learns features from sEMG signals, resulting in excellent performance in small-sample gesture recognition.

We believe the poorer performance of the comparison algorithms compared to the proposed method may be due to two main factors. First, regarding the network architecture design, the proposed method uses a Transformer architecture, while the comparison algorithms use convolutional architectures. Compared to convolutional architectures, Transformers have significant advantages, particularly in modeling long-range dependencies in sequential data, dynamic feature extraction, efficient parallelization, and context-awareness. These advantages are especially evident when dealing with complex, multi-dimensional, long-time series data. In contrast, convolutional architectures focus on local feature extraction, limiting their ability to model complex global interactions. Secondly, in terms of data volume, the proposed method utilizes data augmentation techniques, and the enriched data volume is crucial for improving accuracy.

3.3. Ablation experiments

The aim is to verify whether the proposed multimodal optimal matching and augmentation method can achieve effective recognition with reduced data collection. In this section, two approaches are evaluated for recognizing new users: (1) not using new user data in the training set (Experiments 1-4) and (2) using small-sample data of the new user's calibration gesture (CG) (Experiments 5-8). Furthermore, the effectiveness of the optimal matching signal screening module (OMSS), the MM-LSTEMGNet, and the similarity calculation augmentation module (SCA) is gradually evaluated in both approaches using the self-collected stroke patient and the Ninapro DB5 dataset, as shown in Table 6.

Experiment 2 demonstrated that using the Optimal Matching Signal Screening module allows for selecting data from the database similar to the new user, effectively avoiding negative transfer. This resulted in an improvement of 19.42% and 18.77% in accuracy on the two datasets, respectively, compared to Experiment 1, where the entire database was used as the training set. The similarity calculation augmentation module effectively utilized the signals' temporal characteristics. Particularly, in Experiment 4, where calibration gesture data (CG data) was not used, there was a 9.42% increase in accuracy on the stroke patient dataset, significantly increasing the diversity of signals.

Experiment 4 showed that a model trained only on the optimal matching data achieved an accuracy of 85.85% and 93.76% on the two datasets, respectively, meeting basic rehabilitation needs. Experiment 8 demonstrated that our method using only the single repeat calibration gesture data, the accuracy reached 93.69% and 98.56% on the two datasets. This achieved results comparable to models trained on individual data while greatly reducing the burden of data collection. It makes the intelligent rehabilitation device more user-friendly and beneficial for practical application.

Table 6. Multimodal optimal matching and augmentation method ablation experiments

Experiment	CG data	OMSS	MM-LSTEMGNet	SCA	Stroke patients dataset	Ninapro DB5 dataset
Experiment 1					55.78%	72.53%
Experiment 2		√			75.20%	91.30%
Experiment 3		√	√		76.41%	92.00%
Experiment 4		√	√	√	**85.83%**	**93.76%**
Experiment 5	√				88.28%	95.68%
Experiment 6	√	√			89.62%	97.17%
Experiment 7	√	√	√		91.08%	97.06%
Experiment 8	√	√	√	√	**93.69%**	**98.56%**

To validate whether using multimodal data can improve gesture recognition accuracy, in this section, we conducted ablation experiments using two different types of data: sEMG alone + IMU alone and sEMG + IMU (IMU signals include 3 channels of arm acceleration signals, 3 channels of angular velocity signals, and 4 channels of quaternion signals). These experiments were conducted on the self-collected stroke patient dataset, Ninapro DB1 dataset, and Ninapro DB5 dataset. The results of the ablation experiments on the three datasets are shown in Table 7.

Table 7 shows that on our self-collected stroke patient dataset, using only sEMG signals resulted in an average accuracy increase of 3.15% compared to using only IMU signals. Similarly, on the Ninapro DB1 dataset and Ninapro DB5 dataset, using only IMU signals led to average accuracy improvements of 12.91% and 7.89%, respectively, compared to using only sEMG signals. Furthermore, utilizing multimodal signals achieved an average accuracy increase of 24.02% and 18.74% compared to using only sEMG signals and only IMU signals, respectively, across all three datasets. These results indicate that using multimodal signals yields higher accuracy compared to using single-modal signals.

Table 7. Results of Ablation Experiments

Dataset	Sensor	Average recognition
Stroke patients	sEMG	0.7958
	IMU	0.7463
	sEMG + IMU	0.9369
Ninapro DB5	sEMG	0.6938
	IMU	0.7727
	sEMG + IMU	0.9856
Ninapro DB1	sEMG	0.6287
	IMU	0.7578
	sEMG + IMU	0.9165

4. Discussion

To alleviate the burden of data collection in gesture recognition, we propose a new approach to address small-sample gesture recognition. This method selects the optimal matching signals with high similarity to the new user from the existing users' multimodal data, which are then used as the training set. This reduces the domain differences between the signals of the target user and the training data, thereby avoiding the negative transfer issue that can affect the model's recognition accuracy. Additionally, the method generates enhanced data, which expands the diversity of the training set signals.

Currently, research teams have publicly released large-scale datasets(*30,31*) such as Ninapro, Csl-hdemg, and Capgmyo, which contain multimodal information, including sEMG signals, IMU, and motion information collected from various devices. These datasets also include a substantial number of subjects and a wide variety of gestures. Therefore, it is relatively easy to obtain a large amount of multimodal public data to build small-sample databases, providing strong support for the portable use of our method.

However, in the adaptive selection of the optimal signal, the method proposed in this paper still relies on evaluating the model accuracy to screen the optimal matching signals, which requires a certain amount of computational resources. Because the time required for the optimal matching signal selection process is influenced by the number of users in the database. The more users in the database, the more similarity calculations are needed between the new user's data and the existing user data, ultimately increasing the time required for the optimal matching signal selection, we are considering, as a potential avenue for future research, the development of a method for extracting common features of user gestures. This method would aim to extract common features from all users in the database for a specific action. By comparing the gesture data features of new users with the common features of relevant actions in the database, this approach can identify the type of gesture performed by the new user. Such a method would reduce the time overhead associated with an increasing number of users in the database, thereby making the gesture recognition method more effective in practical applications.

In addition, an increase in the number of users does not necessarily lead to more training cycles, as it depends on whether there is beneficial similar data in the dataset. Currently, the entire system in practical applications consists of three main components: user data collection, model training, and model usage. The time required to collect user samples has been reduced from 6 collections in traditional algorithms to just 1, resulting in an efficiency improvement of 83.3%. The entire model training process takes approximately 15 minutes, and these two steps only need to be performed once. After that, the main focus during model usage is the inference time. The model inference time proposed in this paper is only 4-6 milliseconds, which fully meets the real-time usage needs of new users.

As this study is conducted in the context of hand rehabilitation training for stroke patients, the gestures used are predefined as part of the rehabilitation program. In contrast to random gestures performed in daily life, the gestures in this study exhibit low uncertainty. However, in real-life scenarios, patients' movements entail randomness and uncertainty. To address these issues, we plan to harness the potential of graph structure learning, such as leveraging methods like " EGNN: Graph structure learning based on evolutionary computation " (*32*), for further improvement and enhancement. We believe that graph-based learning can be applied to small-sample cross-user recognition for sEMG signals. In cross-user recognition tasks, the nodes in the graph network can represent the signal features or muscle activity patterns of different users, while the edges represent the similarity or dependency of signals between different users. This approach creates a shared feature space across users, enabling model transfer learning or knowledge sharing between users.

In summary, the multimodal optimal matching and augmentation method effectively improves small-sample gesture recognition accuracy. When using only a single calibration gesture, it achieves 93.69%, 91.65% and 98.56% accuracy on the multimodal dataset of stroke patients, the publicly available Ninapro DB1 dataset and the publicly available Ninapro DB5 dataset, respectively, comparable to the performance of traditional recognition models trained on personal data. In the future, our method will be applied to active hand rehabilitation treatment for stroke patients.

Funding: This work was supported by a grant from the Special Project on Scientific Research and Development of Beijing Rehabilitation Hospital Affiliated with Capital Medical University (2022-001). Beijing Municipal Science and Technology Program: Research and Translational Application of Clinical Characteristic Diagnostic and Treatment Techniques in the Capital City (Project No. Z221100007422113).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Amin MS, Rizvi STH, Hossain MM. A comparative review on applications of different sensors for sign language recognition. Journal of Imaging. 2022; 8:98.
2. Yadav D, Veer K. Recent trends and challenges of surface electromyography in prosthetic applications. Biomedical Engineering Letters. 2023; 13:353-373.
3. Tong L, Zhang M, Ma H, Wang C, Peng L. sEMG-based gesture recognition method for coal mine inspection manipulator using multistream CNN. IEEE Sensors Journal. 2023; 23:11082-11090.
4. Mahmud A, Heidari O, Schoen MP. sEMG based Real-Time Motion Classification using Virtual Reality and Artificial Neural Networks. Journal of Electrical Engineering. 2022; 4:241-256.
5. Li W, Shi P, Yu H. Gesture recognition using surface electromyography and deep learning for prostheses hand: state-of-the-art, challenges, and future. Frontiers in neuroscience. 2021; 15:621885.
6. Zhang W, Zhao T, Zhang J, Wang Y. LST-EMG-Net: Long short-term transformer feature fusion network for sEMG gesture recognition. Frontiers in Neurorobotics. 2023; 17:1127338.
7. Atzori M, Gijsberts A, Castellini C, Caputo B, Hager A-GM, Elsig S, Giatsidis G, Bassetto F, Müller H. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Scientific data. 2014; 1:1-13.
8. Atzori M, Gijsberts A, Kuzborskij I, Elsig S, Hager A-GM, Deriaz O, Castellini C, Müller H, Caputo B. Characterization of a benchmark database for myoelectric movement classification. IEEE transactions on neural systems and rehabilitation engineering. 2014; 23:73-83.
9. Gijsberts A, Atzori M, Castellini C, Müller H, Caputo B. Measuring movement classification performance with the movement error rate. IEEE Trans Neural Syst Rehabil Eng. 2014; 89621:735-744.
10. Atzori M, Gijsberts A, Heynen S, Hager A-GM, Deriaz O, Van Der Smagt P, Castellini C, Caputo B, Müller H. Building the Ninapro database: A resource for the biorobotics community. In: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob) (IEEE, 2012; pp. 1258-1265.
11. Rampichini S, Vieira TM, Castiglioni P, Merati G. Complexity analysis of surface electromyography for assessing the myoelectric manifestation of muscle fatigue: A review. Entropy. 2020; 22:529.
12. He J, Jiang N. Biometric from surface electromyogram (sEMG): Feasibility of user verification and identification based on gesture recognition. Frontiers in bioengineering and biotechnology. 2020; 8:58.
13. Kanoga S, Hoshino T, Asoh H. Semi-supervised style transfer mapping-based framework for sEMG-based pattern recognition with 1-or 2-DoF forearm motions. Biomedical Signal Processing and Control. 2021; 68:102817.
14. Azab AM, Mihaylova L, Ang KK, Arvaneh M. Weighted transfer learning for improving motor imagery-based brain–computer interface. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2019; 27:1352-1359.
15. Wang K, Chen Y, Zhang Y, Yang X, Hu C. Multi-source integration based transfer learning method for cross-user semg gesture recognition. In: 2022 International Joint Conference on Neural Networks (IJCNN) (IEEE, 2022; pp. 1-8.

16. Colli Alfaro JG, Trejos AL. User-independent hand gesture recognition classification models using sensor fusion. Sensors. 2022; 22:1321.
17. Sheng X, Lv B, Guo W, Zhu X. Common spatial-spectral analysis of EMG signals for multiday and multiuser myoelectric interface. Biomedical Signal Processing and Control. 2019; 53:101572.
18. Campbell E, Phinyomark A, Scheme E. Deep cross-user models reduce the training burden in myoelectric control. Frontiers in Neuroscience. 2021; 15:657958.
19. Tsinganos P, Cornelis J, Cornelis B, Jansen B, Skodras A. Transfer learning in semg-based gesture recognition. In: 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA) (IEEE, 2021; pp. 1-7.
20. Yu Z, Zhao J, Wang Y, He L, Wang S. Surface EMG-based instantaneous hand gesture recognition using convolutional neural network with the transfer learning method. Sensors. 2021; 21:2540.
21. Novick LR. Analogical transfer, problem similarity, and expertise. Journal of Experimental Psychology: Learning, memory, and cognition. 1988; 14:510.
22. Abdullah SMSA, Ameen SYA, Sadeeq MA, Zeebaree S. Multimodal emotion recognition using deep learning. Journal of Applied Science and Technology Trends. 2021; 2:73-79.
23. Liu W, Qiu JL, Zheng WL, Lu BL. Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis. 2019.
24. Yan M, Deng Z, He B, Zou C, Wu J, Zhu Z. Emotion classification with multichannel physiological signals using hybrid feature and adaptive decision fusion. Biomedical Signal Processing and Control. 2022; 71:103235.
25. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion. 2023; 91:424-444.
26. Singh MK, Singh N, Singh A. Speaker's voice characteristics and similarity measurement using Euclidean distances. In: 2019 International Conference on Signal Processing and Communication (ICSC) (IEEE, 2019; pp. 317-322.
27. Müller M. Information Retrieval for Music and Motion. Information Retrieval for Music and Motion. 2007.
28. Kingma DP, Welling M. An introduction to variational autoencoders. Foundations and Trends® in Machine Learning. 2019; 12:307-392.
29. Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision. In: International conference on machine learning (PMLR, 2021; pp. 5583-5594.
30. Geng W, Du Y, Jin W, Wei W, Hu Y, Li J. Gesture recognition by instantaneous surface EMG images. Scientific reports. 2016; 6:36571.
31. Dai Q, Li X, Geng W, Jin W, Liang X. CAPG-MYO: a muscle-computer interface supporting user-defined gesture recognition. In: Proceedings of the 9th International Conference on Computer and Communications Management (2021; pp. 52-58.
32. Liu Z, Yang D, Wang Y, Lu M, Li R. EGNN: Graph structure learning based on evolutionary computation helps more in graph neural networks. Applied Soft Computing. 2023; 135:110040.

Received November 23, 2024; Revised January 5, 2025; Accepted January 21, 2025.

**Address correspondence to:*
Wenli Zhang, Faculty of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China
E-mail: zhangwenli@bjut.edu.cn

Released online in J-STAGE as advance publication January 25, 2025.