DOI: 10.23977/acss.2024.080613
ISSN 2371-8838 Vol. 8 Num. 6

A dual-branch network architecture for sEMG-based gesture recognition

Chenyu Shi[1,a], Yuchun Wang[1,b,*]

[1]School of Information and Electronic Technology, Jiamusi University, Jiamusi, China
[a]shichenyu@stu.jmsu.edu.cn, [b]swanjm@163.com
[*]Corresponding author

Keywords: sEMG, Gesture Recognition, Deep Learning, CNN, BiLSTM, Transformer

Abstract: The surface electromyography (sEMG) signal, as a type of bioelectrical signal, has been widely applied in modern human-computer interaction, especially for gesture recognition. The rapid advancement of deep learning has significantly promoted the development of sEMG-based gesture recognition technology. However, existing studies often face challenges such as insufficient feature extraction from sEMG signals and low differentiation between similar gestures. To address these issues, this study proposes a novel dual-branch model architecture specifically designed for sparse-channel sEMG gesture recognition. The model leverages the strengths of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory-Transformer (BiT) networks to process both the time-frequency representations and raw signals of sEMG data, thoroughly extracting spatiotemporal features. Additionally, the proposed Hybrid Attention Block (HAB) further enhances the feature representation capability of the CNN branch. To verify the model's effectiveness, multiple experiments were conducted on the NinaPro-DB1 dataset. The results demonstrate that the proposed model achieved a classification accuracy of 89.23%, outperforming most mainstream models.

1. Introduction

Technological advancements have fueled growing interest in recognizing human motion intentions across various fields. Gesture recognition, a natural and intuitive interaction method, has found broad applications in medical rehabilitation, virtual reality, and human-computer interaction [1][2]. sEMG, which records muscle activity through electrodes placed on the skin, is widely used for gesture recognition due to its non-invasive nature and convenience. sEMG captures the unique electrical patterns generated by specific muscle groups during gestures, enabling real-time recognition. Unlike vision-based methods or those relying on accelerometers and gyroscopes, sEMG is unaffected by lighting changes or hand occlusion, offering notable advantages.

Traditional gesture recognition methods, such as SVM and KNN [3][4], offer advantages like low computational complexity and fast processing times. However, their accuracy is limited when handling complex gestures and diverse sEMG signal patterns, as they struggle with the high-dimensional, nonlinear nature of the data. Additionally, manually designed feature engineering is inefficient and may miss important information.The advancement of artificial intelligence has enhanced sEMG gesture recognition, with deep learning models like RNN and CNN showing superior accuracy by extracting complex features, particularly through converting sEMG signals into time-frequency maps.

Wei et al.[5] introduced a multi-stream CNN that divides raw sEMG images into blocks based on electrode layout, using a "divide-and-conquer" approach to independently learn muscle and gesture correlations. This method achieved 85% accuracy on the NinaPro database.Sandoval-Espino et al. [6] compared four sets of time-domain features, converting them into images for CNN training. They found that incorporating power spectral information and optimizing image representations significantly improved classification accuracy, with results of 97.61% on NinaPro DB2 and 90.23% on DB3.

LSTM networks, designed to address gradient issues in RNN, are well-suited for capturing the complex temporal dynamics of sEMG signals in time series gesture classification tasks.Samadani A [7] compared different RNN configurations for sEMG-based gesture classification, focusing on LSTM and GRU units, achieving a maximum accuracy of 86.7% on the NinaPro DB2 dataset. Bittibssi et al. [8] proposed three LSTM-based models: standard LSTM, convolutional dilated LSTM, and GRU, with the dilated convolutional LSTM showing the best performance on most datasets, reaching 90% accuracy for 12 gestures on NinaPro DB1. Zhang et al. [9] introduced the LSTM-MSA model, which integrates LSTM with a dual-stage attention mechanism, enhancing feature extraction and improving gesture prediction accuracy. The model achieved 91.36% accuracy for 17 gestures on the NinaPro DB5 dataset.

Although CNN and LSTM have demonstrated outstanding performance in their respective fields, a single model may struggle to fully capture the multi-layered features and dynamic variations of data in complex application scenarios. As a result, researchers have proposed hybrid models that combine different deep learning models, leveraging their respective strengths to further enhance model performance. Prabhavathy T. et al. [10] developed a CNN-LSTM hybrid gesture recognition framework using VMD for frequency pattern identification and spectral analysis, achieving 98.04% accuracy for ten gestures, a 3% improvement over traditional CNN models.

Wang et al. [11] proposed ALCNet, a CNN-LSTM hybrid model that uses stationary wavelet packet transform to decompose time-frequency information, achieving 81.80% accuracy on the NinaPro DB1 dataset. Liu et al. [12] introduced a CNN-Transformer hybrid model, combining CNN with AFB and MFA modules for local feature extraction and a Transformer for global context, achieving 99.02% accuracy on a custom dataset of nine gestures.

Despite the aforementioned progress, several challenges remain. First, most models perform poorly when recognizing similar gestures, as they rarely address the issue of low distinction among such gestures. Second, some deeply stacked network models may experience degradation, and relying solely on one form of sEMG data (either time-frequency maps or raw sEMG signals) makes it difficult to fully leverage the multiple characteristics of sEMG data.

To address these issues, this study proposes a dual-branch network model based on CNN and BiLSTM-Transformer (BiT) for sEMG gesture recognition. The model combines the BiT’s ability to model long-term temporal dependencies with CNN’s strength in detecting local patterns in time-frequency maps, thoroughly exploring and extracting the spatiotemporal features of sEMG signals. By adopting a parallel structure instead of a stacked multi-layer network architecture, the model prevents the degradation problem. The parallel processing of sEMG time-frequency maps and raw data through CNN and BiT allows for the full utilization of the data's diverse characteristics, thereby improving gesture classification accuracy.

2. Methods

2.1. CNN Branch

CNN, widely used in image processing, are also commonly applied in sEMG gesture recognition. In this study, the CNN branch processes time-frequency maps, capturing both temporal and frequency information to highlight subtle muscle activities. While CNN excel at extracting local spatial features, in multi-channel sEMG tasks, feature extraction can be insufficient after several convolutional layers. To address this, a Hybrid Attention Block (HAB) is introduced to assign weights to different channels and spatial positions, enhancing attention to key features and reducing interference from irrelevant information.

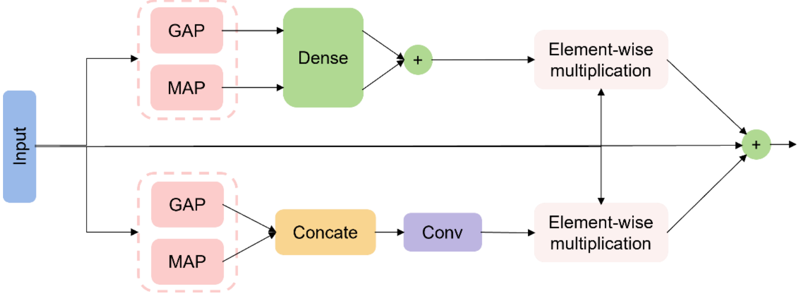

Figure 1: Hybrid Attention Block.

As shown in Figure 1, the HAB uses a parallel approach to compute channel and spatial attention, which are fused through weighted integration to highlight the most important channels in the global information. The specific process is as follows: Suppose $X \in R^{C \times H \times W}$ is the input feature map, where C, H, W denote the number of channels, height, and width, respectively. The input feature map undergoes global average pooling (GAP) and global max pooling (GMP) operations, forming two feature vectors $X_{avg}, X_{max} \in R^{C \times 1 \times 1}$. The calculation process of channel attention is as follows:

$$W_C = \sigma(FC(\mathrm{Re}LU(FC(X_{avg}))) + FC(\mathrm{Re}LU(FC(X_{\max})))) \tag{1}$$

Where $W_C \in R^{C \times 1 \times 1}$ represents the channel attention weights, and σ is the sigmoid activation function. Then, through element-wise multiplication, the channel attention-enhanced feature map $X_C \in R^{C \times H \times W}$is obtained.

$$X_C = W_C \cdot X \tag{2}$$

The feature map enters two parallel paths: channel attention and spatial attention, with spatial attention identifying key regions along spatial dimensions. Like channel attention, the feature map first undergoes GAP and GMP along the channel dimension. However, in this case, the global information of each spatial location is retained, generating two single-channel spatial feature maps $X_{avg\_s}, X_{\max\_s} \in R^{1 \times H \times W}$.

Next, to capture global spatial dependencies and generate attention weights for the spatial dimension, the two feature maps are concatenated along the channel dimension and input into a convolutional layer. Subsequently, the spatial attention weights are utilized on the spatial dimension of the original feature map through element-wise multiplication. The specific formula is presented below:

$$W_S = \sigma(Conv([X_{avg\_s}, X_{\max\_s}])) \tag{3}$$

$$X_S = W_S \cdot X \quad (4)$$

Where $W_S \in R^{1\times H\times W}$ represents the attention weights for the spatial dimension, and $X_S \in R^{C\times H\times W}$ is the feature map enhanced by spatial attention. Finally, as shown in Equation (5), the channel-enhanced feature map and the spatial-enhanced feature map are combined through weighted addition to generate the final output of the HAB.

$$X^{'} = \alpha X_C + \beta X_S \quad (5)$$

Here,α, β are tunable hyperparameters used to control the contribution ratios of channel attention and spatial attention, respectively.

2.2. BiLSTM-Transformer Branch

sEMG gesture recognition is a time series classification problem, and CNN alone struggle to capture long-term dependencies. LSTM networks are effective at modeling temporal dynamics using memory cells, but they rely only on past information. BiLSTM overcomes this by utilizing both past and future information, yet it still faces challenges with longer sequences and high computational complexity due to its step-by-step processing, limiting parallelization and training efficiency. Unlike BiLSTM, the Transformer architecture allows parallel processing of entire input sequences during training and inference, greatly improving computational efficiency, especially with long sequences. Its multi-head attention mechanism captures temporal dependencies at various levels, enhancing adaptability in gesture recognition tasks. However, relying solely on self-attention may cause the Transformer to lose the natural order of sequences, limiting its effectiveness in handling temporal data.

To combine the strengths of both architectures while mitigating their limitations, the BiT branch integrates the advantages of BiLSTM and Transformer. This combination enables complementary strengths in local and global, short-term and long-term feature extraction, allowing the model to consider different patterns and features in sEMG data. This enhances the accuracy and efficiency of sEMG gesture recognition tasks. Gestures typically involve multi-phase dynamic changes, and the BiT branch can effectively model both local and global features of these actions, improving the recognition accuracy of complex gestures.

2.3. Overall Architecture of the Model

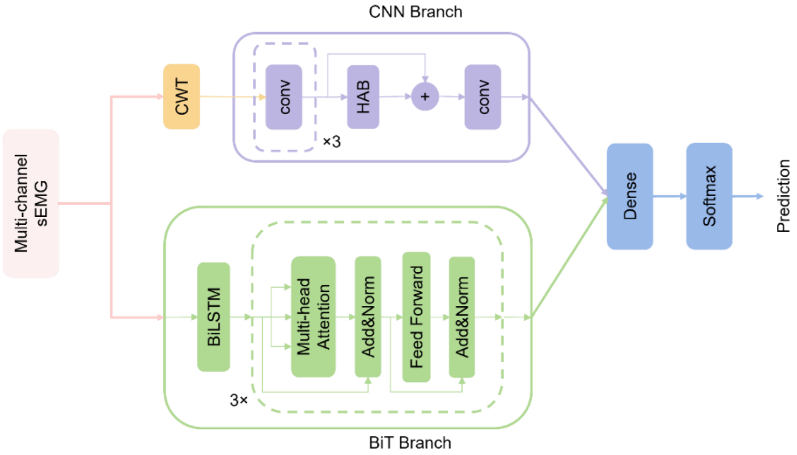

Figure 2: Overall Architecture of the Dual-Branch Network Model.

As shown in Figure 2, the dual-branch CNN and BiT network processes different types of sEMG data. The CNN branch uses sEMG time-frequency maps derived from Continuous Wavelet Transform (CWT) as input, with features extracted through several 3×3 convolutional layers. The number of filters increases from 32 to 256 as the network deepens. ReLU is used as the activation function, and L2 regularization (weight 0.001) is applied to prevent overfitting. After the third convolutional layer, the Hybrid Attention Block (HAB) is introduced with initial α and β values set to 0.5. GAP and GMP compress the feature maps, and batch normalization and a Dropout layer (rate 0.3) are applied before concatenating the pooled feature vectors.

The BiT branch processes preprocessed sEMG data by first using a BiLSTM layer with 128 neurons to extract temporal sequence features. Next, the BiLSTM output is passed through three Transformer encoder layers with four attention heads each to capture global dependencies. The features are then mapped to a higher-dimensional space through a fully connected layer with 512 neurons, followed by a Dropout layer (rate 0.3) and layer normalization to prevent overfitting. Finally, the outputs from both branches are flattened into 1D tensors and concatenated to form a unified feature vector. This concatenated feature vector is passed through a fully connected layer with 64 units, and gesture classification is completed using the SoftMax activation function.

3. Experimental Setup

3.1. Dataset

The Ninapro dataset is an indispensable resource for the development of gesture recognition and prosthetic control technologies. This database was developed and maintained by the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland with the goal of providing high-quality sEMG data. In this study, the first dataset, Ninapro-DB1, was selected to evaluate the proposed model. DB1 contains electromyographic data collected from 27 healthy subjects, recorded using 10 electrodes with a sampling rate of 100 Hz, covering the major muscle groups in the forearm and hand. Subjects followed a predefined experimental protocol, performing 52 different hand movements, including basic gestures, common gestures, and complex daily life activities. Each subject repeated each movement 10 times, ensuring diversity and reliability of the data.

3.2. Data Preprocessing

First, the action labels in the dataset were reassigned. The 52 actions were sequentially assigned labels in the range of 1 to 52, following the order of categories A, B, and C, ensuring that each action corresponds to a unique label. To retain useful components of the sEMG signal while removing DC offset, high-frequency noise, and power line interference, a 20Hz-450Hz fourth-order Butterworth band-pass filter and a 50Hz notch filter were applied to the sEMG signal. Subsequently, a μ-law transformation was used for non-linear processing of the sEMG signal. Many of the useful features of the sEMG signal are concentrated near zero; the μ-law transformation logarithmically amplifies the small-magnitude sensor outputs while keeping them consistent with larger sensor values. The formula for μ-law transformation is as follows:

$$F(x_i) = sign(x_i)\frac{\ln(1+\mu|x_i|)}{\ln(1+\mu)} \tag{6}$$

Here, $\boldsymbol{x_i}$ represents the input at the i-th sampling point, and in this experiment, the parameter μ was set to 256.

To fully extract sEMG features, a sliding window with a window size of 250 ms and a step size of 62.5 ms is applied to segment the sEMG signal. After window segmentation, CWT is applied to each signal segment within the window, converting the time-domain signal into a time-frequency domain representation. The segmented signals are used as inputs for the BiT branch, while the CWT-processed signals are fed into the CNN branch.

3.3. Experimental Parameters

The dual-branch CNN-BiT model was implemented using PyTorch on an Intel Core i5-12400 CPU and NVIDIA RTX 3060 GPU, running on Windows 10. The dataset was split 8:2 into training and test sets. Categorical cross-entropy was used as the loss function, with the Adam optimizer, and a batch size of 64. The model was trained for 200 epochs, starting with a learning rate of 0.001, which was halved every 50 epochs to improve convergence and generalization.

3.4. Evaluation Criteria

After the model training is completed, the classification accuracy of the classifier is calculated based on the recognition results from the test set. Let the average classification accuracy be denoted as *Accuracy*, the number of correctly classified samples as *N*, and the total number of samples as *T*. The formula is as follows:

$$Accuracy = \frac{N}{T} \times 100\% \tag{7}$$

4. Results and Analysis

This study evaluated the contribution of each component to the overall model performance through experiments combining various branches and modules. The results, summarized in Table 1, show that adding the HAB significantly improved the CNN branch's accuracy, increasing it from 81.72% to 83.38%.The experiments showed that the performance of the CNN and BiT branches individually was much lower than the dual-branch model. When both branches were used together, the accuracy increased to 88.06%, highlighting the benefits of multimodal feature fusion. Adding the HAB module further improved accuracy to 89.23%, demonstrating its role in enhancing feature representation and creating a synergistic effect with the fusion architecture. These findings emphasize the importance of each component and their combined impact on improving model performance.

Table 1: Impact of Different Modules on Classification Accuracy.

	CNN	BiT	HAB	Acc/%
1	✓	–	–	81.72
2	✓	–	✓	83.38
3	–	✓	–	84.91
4	✓	✓	–	88.06
5	✓	✓	✓	89.23

As shown in Table 2, the dual-branch CNN-BiT model achieves the best classification accuracy, thanks to its ability to process richer information and more diverse features compared to single-feature models. By using a dual-path architecture—CNN for time-frequency maps and BiT for raw sEMG data—the model combines time-frequency and temporal features effectively. The CNN branch, with its HAB module, enhances critical feature focus through parallel channel and spatial attention, reducing redundancy in high-dimensional data and improving feature distinction. This combination increases the model's ability to differentiate similar gestures, leading to improved classification accuracy.

Table 2: Comparison of Classification Accuracy between the Proposed and Other Models.

Methods	Number of gestures	Acc/%
CNN[13]	50	66.60
RNN with weight loss[14]	53	79.30
ALCNet[11]	52	81.80
MS-CNN[5]	52	85.00
MVCANet[15]	52	87.98
Ours	52	89.23

Figure 3 shows the confusion matrix for the dual-branch model’s recognition of 52 gestures, with rows representing true labels and columns representing predicted labels. The main diagonal indicates correct classifications, where darker colors show higher accuracy. The distinct coloring along the diagonal suggests that the model accurately recognizes most gestures. Notably, the model excels at differentiating between similar gestures, such as gesture 6 (Ring extension) and gesture 20 (Adduction of extended fingers), and gesture 30 (Large diameter grasp) and gesture 49 (Power disk grasp). This high performance is due to the model’s comprehensive extraction of spatiotemporal features from both time-frequency maps and raw signals, allowing it to effectively capture local and global features. The confusion matrix highlights the model’s ability to handle complex classification tasks involving similar gestures by enhancing feature distinction.

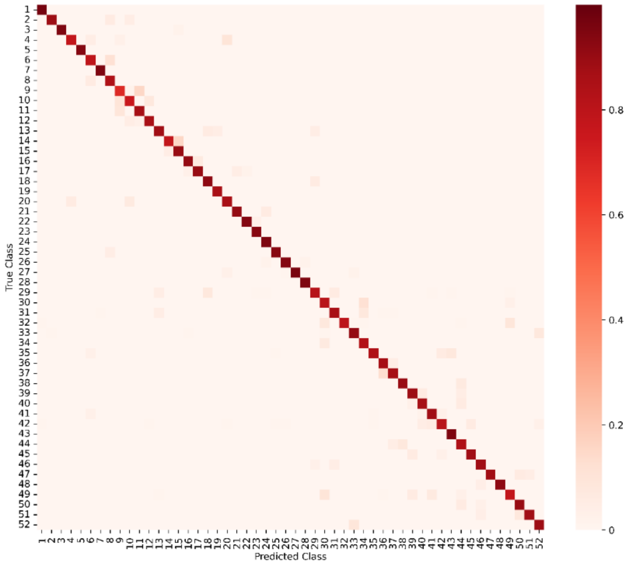

Figure 3: The confusion matrix for the classification of 52 gestures.

5. Conclusion

To fully exploit the diverse characteristics of sEMG signals and improve the differentiation between similar gestures, this paper proposes a dual-branch network model based on CNN and BiT for gesture recognition. By combining the strengths of both CNN and BiT branches, the model thoroughly explores and utilizes the time-frequency and temporal features of sEMG signals. The CNN branch, equipped with a HAB, processes the time-frequency maps generated by Continuous Wavelet Transform, enhancing feature distinction in both channel and spatial dimensions. The BiT branch, integrating BiLSTM and Transformer, captures the long-term temporal dependencies of sEMG data and extracts global features. The parallel structure of the dual branches effectively avoids the degradation problem commonly seen in deep networks while improving the model’s ability to distinguish similar gestures, resulting in a significant increase in gesture classification accuracy. Experimental results demonstrate that the proposed model outperforms most mainstream models in terms of classification accuracy. Comparative experiments on different branch and module combinations further validate the model's effectiveness. In conclusion, the dual-branch network model based on CNN and BiT proposed in this paper offers a novel solution to the task of sEMG gesture recognition, boasting high classification accuracy and powerful feature extraction capabilities, and provides valuable insights for future research and applications.

References

[1] Wei, Z., Zhang, Z. Q., & Xie, S. Q. (2024). Continuous Motion Intention Prediction Using sEMG for Upper-Limb Rehabilitation: A Systematic Review of Model-Based and Model-Free Approaches. IEEE Transactions on Neural Systems and Rehabilitation Engineering. vol. 32, 1487-1504.

[2] Tigrini, A., Al-Timemy, A. H., Verdini, F., Fioretti, S., Morettini, M., Burattini, L., & Mengarelli, A. (2023). Decoding transient sEMG data for intent motion recognition in transhumeral amputees. Biomedical Signal Processing and Control, 85, 104936.

[3] Hye, N. M., Hany, U., Chakravarty, S., Akter, L., & Ahmed, I. (2023). Artificial Intelligence for sEMG-based Muscular Movement Recognition for Hand Prosthesis. IEEE Access, 11, 38850-38863.

[4] Cai, S., Chen, Y., Huang, S., Wu, Y., Zheng, H., Li, X., & Xie, L. (2019). SVM-based classification of sEMG signals for upper-limb self-rehabilitation training. Frontiers in neurorobotics, 13, 31.

[5] Wei, W., Wong, Y., Du, Y., Hu, Y., Kankanhalli, M., & Geng, W. (2019). A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface. Pattern Recognition Letters, 119, 131-138.

[6] Sandoval-Espino, J. A., Zamudio-Lara, A., Marbán-Salgado, J. A., Escobedo-Alatorre, J. J., Palillero-Sandoval, O., & Velásquez-Aguilar, J. G. (2022). Selection of the best set of features for sEMG-based hand gesture recognition applying a CNN architecture. Sensors, 22(13), 4972.

[7] Samadani, A. (2018, July). Gated recurrent neural networks for EMG-based hand gesture classification. A comparative study. In 2018 40th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-4). IEEE.

[8] Bittibssi, T. M., Genedy, M. A., & Maged, S. A. (2021). sEMG pattern recognition based on recurrent neural network. Biomedical Signal Processing and Control, 70, 103048.

[9] Zhang, H., Qu, H., Teng, L., & Tang, C. Y. (2023). LSTM-MSA: A Novel Deep Learning Model With Dual-Stage Attention Mechanisms Forearm EMG-Based Hand Gesture Recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 4749-4759.

[10] Prabhavathy, T., Elumalai, V. K., Balaji, E., & Sandhiya, D. (2024). A surface electromyography based hand gesture recognition framework leveraging variational mode decomposition technique and deep learning classifier. Engineering Applications of Artificial Intelligence, 130, 107669.

[11] Wang, L., Fu, J., Zheng, B., & Zhao, H. (2022, April). Research on sEMG–based gesture recognition using the Attention-based LSTM-CNN with Stationary Wavelet Packet Transform. In 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC) (pp. 1-6). IEEE.

[12] Liu, Y., Li, X., Yang, L., Bian, G., & Yu, H. (2023). A CNN-transformer hybrid recognition approach for sEMG-based dynamic gesture prediction. IEEE Transactions on Instrumentation and Measurement, 72, 1-16.

[13] Atzori, M., Gijsberts, A., Castellini, C., Caputo, B., Hager, A. G. M., Elsig, S., ... & Müller, H. (2014). Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Scientific data, 1(1), 1-13.

[14] Koch, P., Phan, H., Maass, M., Katzberg, F., Mazur, R., & Mertins, A. (2018, September). Recurrent neural networks with weighting loss for early prediction of hand movements. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 1152-1156). IEEE.

[15] Wentao Yuan, Wentao Wei, Demin Gao. (2024).Research on Multiview Convolutional Gesture Recognition with Fusion Attention Mechanism [J]. Computer Engineering, 50(3): 208-215.