sensors

MDPI

Article

A Static Sign Language Recognition Method Enhanced with Self-Attention Mechanisms

Yongxin Wang [1,*], He Jiang [1], Yutong Sun [2] and Longqi Xu [1]

[1] School of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin 150080, China; 2320600016@stu.hrbust.edu.cn (H.J.); 2220600076@stu.hrbust.edu.cn (L.X.)
[2] School of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin 150080, China; 2303010617@stu.hrbust.edu.cn
* Correspondence: wangyongxin@hrbust.edu.cn

Abstract: For the current wearable devices in the application of cross-diversified user groups, it is common to face the technical difficulties of static sign language recognition accuracy attenuation, weak anti-noise ability, and insufficient system robustness due to the differences in the use of users. This paper proposes a novel static sign language recognition method enhanced by a self-attention mechanism. The key features of sign language gesture classification are highlighted by the weight function, and then the self-attention mechanism is combined to pay more attention to the key features, and the convolutional neural network is used to extract the features and classify them, which realizes the accurate recognition of different types of static sign language under standard gestures and non-standard gestures. Experimental results reveal that the proposed method achieves an average accuracy of 99.52% in the standard static sign language recognition task when tested against the standard 36 static gestures selected within the reference American Sign Language dataset. By imposing random angular bias conditions of ±(0°–9°] and ±(9°–18°], the average recognition rates in this range were 98.63% and 86.33%. These findings indicate that, compared to existing methods, the proposed method not only maintains a high recognition rate for standard static gestures but also exhibits superior noise resistance and robustness, rendering it suitable for static sign language recognition among diverse user populations.

Keywords: static gesture recognition; high robustness; self-attention enhanced; CNN

check for updates

Citation: Wang, Y.; Jiang, H.; Sun, Y.; Xu, L. A Static Sign Language Recognition Method Enhanced with Self-Attention Mechanisms. *Sensors* **2024**, *24*, 6921. https://doi.org/10.3390/s24216921

Academic Editor: Dawid Połap

Received: 20 September 2024
Revised: 16 October 2024
Accepted: 19 October 2024
Published: 29 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As an indispensable means of communication for people with hearing and speech impairments, sign language has a profound impact on all aspects of their lives [1]. According to the World Health Organization's Global Report on Health Equity for People with Disabilities in 2022, the total number of people with disabilities in the world's population of 7.7 billion is 1.3 billion, with hearing and speech impaired people accounting for a significant proportion of the total, highlighting the urgent need for universal access to sign language and assistive technology.

With the leap of Artificial Intelligence (AI) and sensor technology, the integration of wearable devices and human-computer interaction has become the focus of scientific research, which not only broadens the application fields of VR [2] and health monitoring [3], but also shows great prospects in the field of sign language recognition [4]. This technology realizes sign language recognition and conversion by accurately capturing hand movements, opening up a new communication and learning channel for people with disabilities.

In the field of sign language recognition using wearable devices [5,6], researchers have made numerous invaluable contributions. Table 1 summarizes eight different methods for static sign language recognition.

Table 1. Comparison of different sign language recognition methods.

Researcher	Device	Method	Accuracy
Medhanit Y. Alemu [7]	Artificial Skin	MLP	91.13%
Chaithanya Kumar Mammadli [8]	IMU	MLP	92.00%
Paolo Sernani [9]	sEMG	LSTM	97.00%
Liufeng Fan [10]	Flex Sensor	CNN + BiLSTM	98.21%
Yinlong Zhu [11]	Bending Sensor	BP Network	98.50%
Weixin Deng [12]	Bending Sensor	SVM	99.20%
Jungpil Shin [13]	Camera	SVM	99.39%
C.K.M. LEE [14]	Leap Motion	RNN	99.44%

From Table 1, it is evident that gesture recognition technology in the field of machine vision has garnered significant attention due to its convenience, high efficiency, and real-time processing capabilities. Medhanit Y. Alemu's team [7] employed artificial skin and a Multi-Layer Perceptron (MLP) method, achieving an accuracy of 91.13%. However, this approach is associated with high manufacturing costs and complex data processing. Similarly, Chaithanya Kumar Mammadli's team [8] utilized Inertial Measurement Unit (IMU) data gloves with MLP, attaining a recognition accuracy of 92%. While these results are promising, they also highlight challenges related to data handling and equipment expenses.

To address the impact of environmental factors, such as lighting and skin color variations, Paolo Sernani's team [9] explored surface electromyography (sEMG) signals and employed a Long Short-Term Memory (LSTM) network, achieving a recognition rate of 97%. Despite the advantages of sEMG, its complex data processing may hinder its applicability in real-time scenarios.

In contrast to the aforementioned methods, Liufeng Fan and his research team [10] leveraged flexible bending sensors in combination with Convolutional Neural Networks (CNN) and Bidirectional LSTM (BiLSTM), achieving a recognition accuracy of 98.21%. This technique has gained favor among researchers due to its low cost and straightforward data processing.

Furthermore, scholars such as Yinlong Zhu [11] and Weixin Deng [12] attained recognition accuracies of 98.5% and 99.2%, respectively, using BP Neural Networks and Support Vector Machines (SVM). Jungpil Shin's team [13] utilized a camera with SVM to achieve a recognition accuracy of 99.39%. Meanwhile, C.K.M. LEE's team [14] employed Leap Motion controllers integrated with Recurrent Neural Networks (RNN), attaining a remarkable recognition rate of 99.44%. However, this method faces limitations in cross-platform applications.

These studies underscore the substantial potential of various technologies in gesture recognition, particularly highlighting the advantages of portable devices in terms of flexibility and real-time processing, which have made them highly regarded in the field of gesture recognition.

Except for this research on recognition accuracy, all of the above methods have not explored in detail the problem of reduced recognition accuracy directly caused by the difficulty of ensuring that each movement achieves the precision of a standard gesture when a person with speech impairment performs a sign language gesture while wearing a wearable device in a real-world application.

On this basis, this paper proposes a static sign language recognition method based on self-attention enhancement, which highlights the key features of sign language gesture classification through the weight function and then combines the self-attention mechanism to assign higher attention to the key features and use a convolutional neural network to extract features and classification. In this paper, the proposed method effectively reduces the low recognition accuracy caused by the unstandardized sign language movement of beginners and then improves the accuracy and robustness of the system.

Contributions

This paper makes the following contributions to the field of static gesture recognition:

1. Developed a data glove integrated with 5 flex sensors and a 6-axis gyroscope for gesture data collection.
2. Proposed a gesture recognition method capable of identifying 36 static gestures, with two main improvements:
   - A data augmentation method based on noise injection to increase the diversity of the dataset.
   - A feature enhancement technique utilizing a weighted function and self-attention mechanism.
3. Created a dataset of 36 static gestures based on the data collected using the developed glove.

Table 2 presents a comprehensive comparison of the contributions made by the proposed method in this paper against various state-of-the-art approaches introduced by other scholars.

Table 2. Comparison of the Contributions Between the Proposed Method and Existing State-of-the-Art Methods.

Paper	Contributions	Method
[10]	Intelligent Data Glove	Five-Channel Flexible Capacitive Sensor
	Amphibious Recognition Model	AHGR Model
[11]	Intelligent Data Glove	Doped with Carbon Black and Carbon Nanotubes
	Gesture Recognition System	BP Neural Network
[12]	Proposed Method for Guiding Body Movements	Template Matching Method
		SVM
		Kinematic Modeling
Our Method	Proposed Method for Static Sign Language Recognition	Noise Injection
		Self-Attention Mechanism

The remainder of this paper is structured as follows: Section 2 explains the overall system architecture, the data collection process, and the composition of the dataset. Section 3 provides a detailed description of the proposed methodology. Section 4 presents and analyzes the experimental results, while Section 5 offers the conclusions of this study.

2. Experiment

2.1. Overall System Architecture

The overall structure of the gesture analysis system used in this paper is shown in Figure 1, which integrates four key stages of data acquisition, preprocessing, feature extraction, and gesture recognition.

The data collection phase relies on the data glove shown in Figure 2, the technical specifications of which are detailed in Table 3.

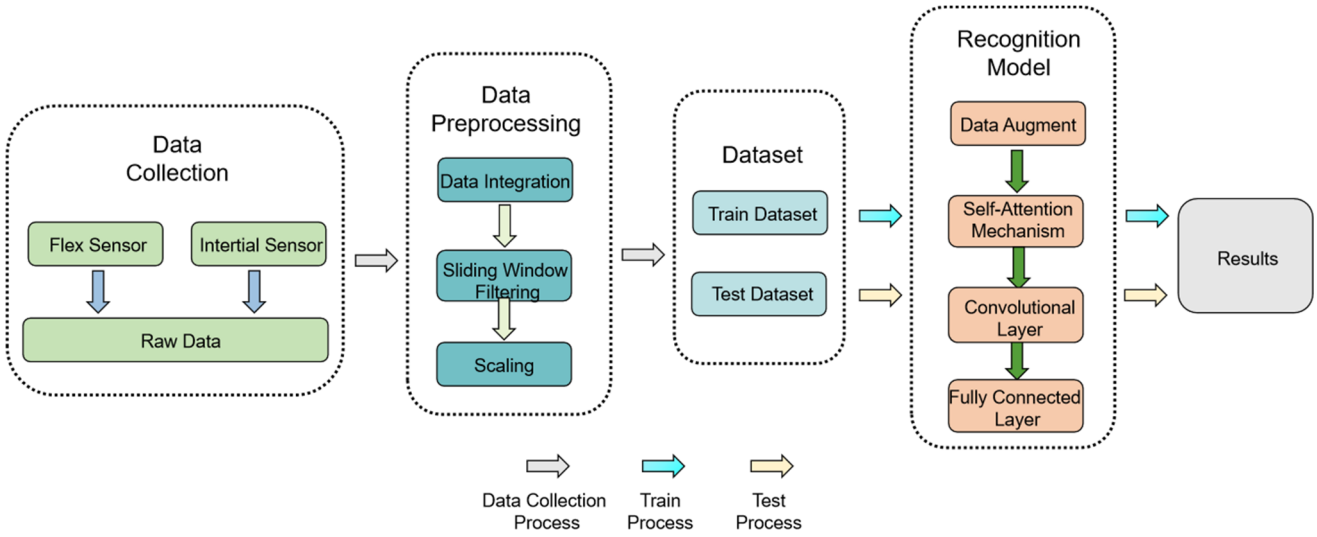
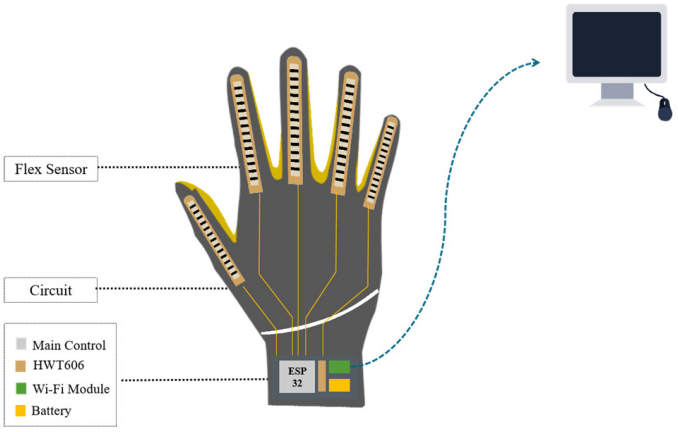
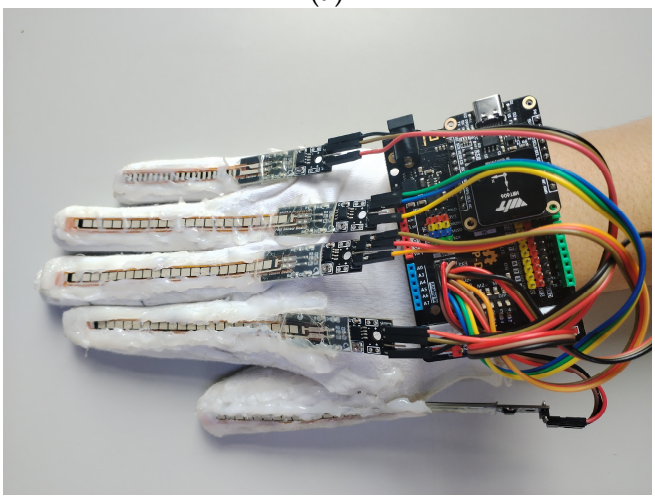


Figure 1. System process flow diagram.



(**a**)



(**b**)

Figure 2. (**a**) Schematic diagram of data glove structure; (**b**) Image of actual data gloves.

Table 3. Data Glove Details.

Core Components	Details
Chip	ESP32-S3
Bending Sensor	Flex Sensor
Attitude Sensor	HWT606
Data Transfer Method	WIFI

In this data collection phase, the Esp32 is used as the main control chip to collect the sensor data of five finger curvatures, noted as input matrix X, where the dimension of matrix X is 1 row and 5 columns. The data from the inertial sensor, which focuses on comprehensively monitoring the overall attitude of the hand in space, including the information of the angular changes in the x, y, and z axes, is also collected, noted as the attitude matrix A. In order to enhance the comprehensive efficacy and analytical value of the data, this paper implements a data integration strategy to integrate the raw data from the two types of sensors into a data integration, which is consolidated into a matrix F, where the matrix F can be expressed by Equation (1), where the dimension of matrix F is 1 row and 8 columns.

$$F = [X|A] = \left[C_{thumb}, C_{index}, C_{middle}, C_{ring}, C_{little}, x_A, y_A, z_A\right] \tag{1}$$

In order to enhance the applicability of the data, this paper further implements the data scaling process. In this process, the maximum value (denoted as y_{max}) and the minimum value (denoted as y_{min}) of the scaled data were set, where the input data are denoted as x, the minimum value of the input data are denoted as x_{min}, and the maximum value as x_{max}. The scaled data range according to the method of Equation (2), which was used to map the raw data of the Flex sensors to the range of 1500–2500, corresponding to the angular values of 0°–180°, to ensure the uniformity of the data and the convenience of analysis.

$$y = y_{min} + \frac{(x - x_{min}) \times (y_{max} - y_{min})}{x_{max} - x_{min}} \tag{2}$$

2.2. Dataset

In this paper, experiments were conducted to construct a new dataset, which is based on the American Sign Language (ASL) dataset [15] shown in Figure 3, and the gesture data were collected with a data glove. The new dataset contains 36 standard static gestures, which not only includes 10 numeric gestures from 0 to 9, but also covers 26 alphabetic gestures from A to Z.

In this experiment, a total of 14,400 samples were collected for 36 gestures during the data acquisition phase. To ensure effective model training, 300 samples for each of the 36 gestures were randomly selected as the training set, totaling 10,800 samples, which covers the diversity of gestures. The remaining 100 samples for each gesture were used as a test set, totaling 3600 samples, to evaluate the model's performance. The waveforms of these 36 static gestures are presented in Figure A1 of Appendix A.

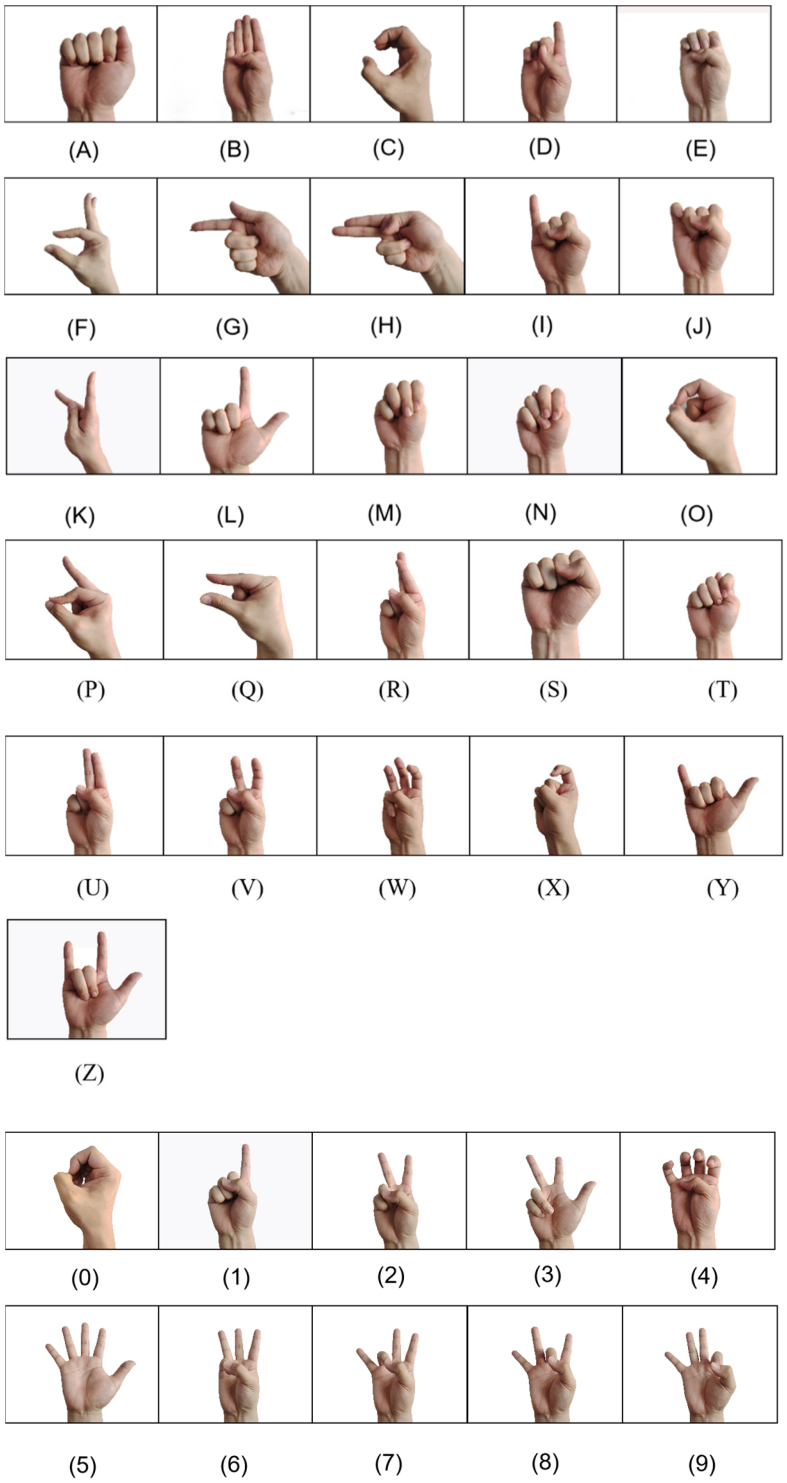


Figure 3. Numerical and alphabetic sign language data set.

2.3. Data Acquisition Process

This experiment recruited 20 volunteers to collect hand gesture data. Each volunteer was instructed to perform 36 gestures, as shown in Figure 3, and maintain each gesture for 30 s, during which 20 data samples were collected per gesture. To mitigate the effects of muscle fatigue between different gestures, a 10 s rest period was provided between gestures for each volunteer. Additionally, to ensure consistency in the data, an electronic goniometer was used to measure finger angles between gestures. This helped ensure that

the same gesture was performed with the same angle across different volunteers, thereby guaranteeing the precision of the standardized gestures.

2.4. The Design of Experiments

In order to evaluate the performance of the proposed method in the gesture recognition task, standard gesture experiments and robustness experiments are designed in this paper. The standard gesture experiment aims to evaluate the performance of the proposed method in ideal conditions, while the robustness experiment examines the robustness of the proposed method to non-standard gestures. Table 4 details the hyperparameters of the model constructed by the proposed method. A customized dataset containing 36 standard gestures as shown in Section 2.2 was used for the experiments, and recognition accuracy was used as the evaluation metric. The standard gesture experiments were performed by conducting five independent experiments on the standard gesture test set and calculating the mean value of its recognition accuracy, precision, recall, and F1 score. The robustness experiments, on the other hand, are based on the standard gesture test sets, and four robustness test sets introducing different levels of random angular bias ranges, which are ±(0°–9°], ±(9°–18°], ±(18°–27°], and ±(27°–36°], are constructed to simulate the angular bias in the actual use. The specific construction method is detailed in Section 2.5. For each level of robustness test sets, five experiments were conducted independently, and the mean value of the recognition accuracy, precision, recall, and F1 score were calculated and recorded as a quantitative assessment of the performance of the proposed method within that random angular bias range.

Table 4. Hyperparameter Configuration.

Parameter Type	Parameter
Training Size	21,600
Test Size	3600
Training Epochs	40
Optimizer	Adam
Learning Rate	0.0001
Batch Size	150
Loss Function	Cross-Entropy

2.5. Design of Robustness Test Sets

In order to comprehensively evaluate the robustness of the proposed method, on the basis of the standard gesture test sets, four different levels of robustness test sets are extended by introducing additional random angular bias, namely, four robustness test sets with random angular bias of ±(0°–9°], ±(9°–18°], ±(18°–27°], and ±(27°–36°] to verify the effect of slight to significant bending angle variations that may be encountered during practical applications on the recognition accuracy of the proposed method. As an example, a robustness test set introducing ±(0°–9°] random angular bias was constructed based on Equation (2), and a random integer bias of ±(0–50] was added to each data point in the standard gesture test set to simulate angular deviation in actual use. Random angular bias was added to the standard gesture test sets, but these data were not included in the original standard gesture test set. Instead, they were used to construct a separate robustness test set. Each robustness test set contains 100 gesture samples, totaling 3600 data records, the same size as the standard gesture test set. The remaining three robustness test sets were designed in the same way.

3. Method

3.1. Overview of the Proposed Method

The proposed method can be subdivided into four core parts: first, the preprocessed matrix F is input, the input matrix F is noisily injected to form a new matrix, and the formed new matrix is input to the self-attention enhancement module. In this phase,

the key features that determine the gesture classification are enhanced by augmenting the weight function of the features to highlight their importance. Subsequently, the self-attention mechanism is utilized to assign higher attention weights to these key features to ensure that the proposed method can focus on the most discriminative features. These features are further extracted and categorized in CNN. Based on the extracted features, gesture classification is performed to achieve accurate recognition of sign language data. This process not only improves the recognition accuracy of the proposed method but also enhances its robustness during use by different users. Figure 4 demonstrates the specific flow of the proposed method.

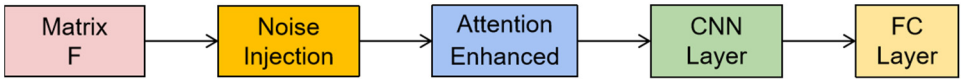


Figure 4. Gesture recognition flow diagram.

3.2. Data Diversity Enhancement Based on Noise Injection

Deep neural networks require a large number of cases for training. A number of data enhancement strategies have been proposed for image analysis, such as flipping, color space [16], panning, rotation, noise injection [17], image blending, random cropping, and generative adversarial networks [18]. In this paper, noise injection is used to add a small amount of random noise to the input data that is small and consistent with the distribution of the original data, as opposed to the data collected by the data glove.

In order to generate a noise matrix F_{noise} with the same shape as the input matrix F, where the dimension of matrix F is 1 row and 8 columns, with a single channel, the matrix F is regularized in order to maintain the consistency of the data and scale standardization, and after regularization, normally distributed biased noise obeying a mean of 0 and a standard deviation of 0.001 is added to each element, where the standard deviation is chosen to be 0.001 in order to make sure that the added noise is sufficiently small and will not change the data's original features [19], and at the same time an effective way to prevent overfitting [20] and increase sample diversity [21]. Then the matrix F' after noise injection can be expressed in the form of Equation (3).

$$F' = F + F_{noise} \tag{3}$$

The amount of data in the training set is expanded after noise injection for each static sign language data. The new training set consists of 10,800 data matrices from the original training set together with the added 10,800 data matrices after noise injection, and this process can be expressed by Equation (4).

$$F_{new} = \begin{pmatrix} F \\ F' \end{pmatrix} \tag{4}$$

As a consequence, the amount of data in the new training set is doubled from the original 10,800 data representing 36 gestures to 21,600 data covering the same 36 gestures. The increase in data volume enhances the diversity of the data and utilizes the limited data more effectively to simulate real-world variations, thus improving the robustness of the model.

3.3. Feature Enhancement Based on Self-Attention Mechanism

Self-attention mechanism (SAM) can demonstrate its excellent data processing ability in academic research, which is able to deeply mine and accurately parse the intricate contextual dependencies within the input sequence data [22]. By dynamically adjusting the allocation of attention, the mechanism can intelligently focus on the feature regions carrying key information in the sequence [23], realizing efficient screening and refining of information. And it is widely used in machine translation [24], machine vision [25], and gesture prediction [26].

This paper utilizes the unique advantage of the self-attention mechanism in extracting the core features of the data [27]. Given that the core of the gesture recognition task is to accurately identify the number of straightened fingers that constitute a particular gesture, a threshold range of finger bending angles in the angular interval of 0°–90° is set as the criterion for defining key features. According to Equation (2), the data interval of 2000 to 2500 corresponds to the angular interval of 0°–90°. In this paper, feature enhancement is performed for this feature data interval so as to increase the weight of key features. The feature enhancement process based on the self-attention mechanism can be specifically divided into the following four parts:

1. Constructing the weighting function

In order to deeply optimize the data processing and feature representation of the input matrix, a weight function is constructed in this paper. The core purpose of this function is to implement a differentiated weight assignment strategy for each element of the matrix based on the intrinsic importance of the data or its unique attribute characteristics. This process is realized by Equation (5), where $\alpha = 1.5$ and $\beta = 0.5$. The values of these parameters were determined through multiple experiments conducted within the range of 0 to 2, resulting in optimal performance.

$$f(x_i) = \begin{cases} \alpha x_i & x_i > 2000 \\ \beta x_i & x_i \leq 2000 \end{cases} \tag{5}$$

2. Feature transformation

After differentially assigning weights to each element in the input sequence by means of a weight function, a further linear transformation operation is performed to generate the three key vectors, namely query (Q'), key (K'), and value (V') [28]. This process is derived based on Equation (6) and aims to reveal the deep relationships and underlying patterns between data elements through linear mapping.

$$Q' = f(x_i)\mathrm{W}^{\mathrm{Q}}, \quad K' = f(x_i)W^k, \quad V' = f(x_i)W^v \tag{6}$$

where W^Q, W^K, and W^V are trainable weight matrices.

3. Allocating Attention to Feature Fingers

In order to assign higher attention to the feature fingers mentioned in the previous section during feature extraction to improve the robustness of static sign language recognition, the attention mechanism is applied. The attention mechanism generally relies on calculating the similarity between the input sequences and thus giving different weights to the individual elements, and a key step in calculating the similarity between the elements in the input sequences is to measure the similarity between them by the dot product between the query (Q') and the key (K') generated by the feature transformation in step 2. This mechanism allows the model to automatically focus on the most important parts of the input data, allowing the model to assign higher attention to the feature fingers, a process based on the complete content of the input matrix and aimed at revealing the intrinsic connections between the elements [29].

Specifically, for any two elements x_i and x_j (where i and j represent the position of the element in the sequence and $i \neq j$) in the sequence, the corresponding query vectors Q_i' and key vectors K_j' are first computed by predefined feature transformations, respectively. Subsequently, the dot product operation $Q_i' \cdot K_j'$ is performed, the result of which represents the element-to-element attention score. This score reflects the strength of similarity or correlation between two elements in the feature space, which is realized by Equation (7).

$$E' = \frac{Q' \cdot K'^T}{\sqrt{d_k}} \tag{7}$$

where d_k is the dimension of the key vector.

The scores are then transformed into probability distributions by the SoftMax function. This is performed by Equation (8).

$$\text{Attention}' = \text{softmax}(E') = \frac{\exp(E')}{\sum \exp(E')} \tag{8}$$

4. Weighted summation

The value vector V' is weighted and summed using the attention weights to generate the final output vector, which is implemented by Equation (9).

$$O' = \sum_{j} \text{Attention}' \cdot V' \tag{9}$$

The whole process of the self-attention enhancement module can be represented as the structure shown in Figure 5.

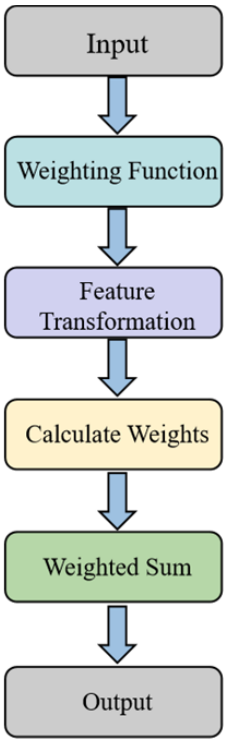


Figure 5. Self-attention enhancement module architecture.

3.4. Feature Extraction Based on Convolutional Neural Networks

After the introduction of a self-attention enhancement module to give higher attention to the feature fingers, the subsequent processing flow incorporates the advantages of convolutional neural networks and fully connected layers [30]. Among them, convolutional neural networks are designed to deeply mine and extract key features of finger data enhanced by self-attention mechanisms. This process not only enhances the model's ability to capture subtle feature differences but also promotes effective aggregation and representation of information. Subsequently, these finely refined features are passed to the fully connected layer (FC), which is responsible for learning and mapping out the complex relationships between the input features and the predefined categories, ultimately realizing the accurate classification of finger data. Figure 6 visualizes the overall architecture of the convolutional neural network.

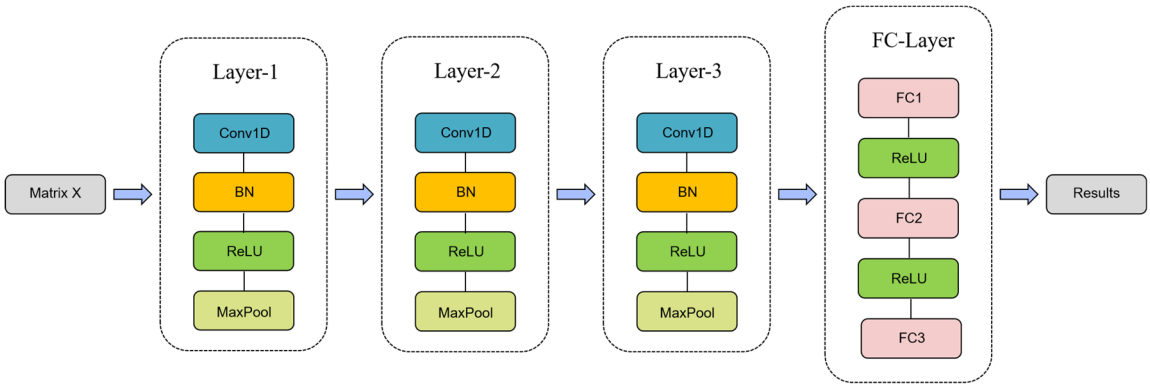


Figure 6. Overall Architecture of Convolutional Neural Networks.

Table 5 details the key parameters of each layer, including input and output dimensions, convolution kernels, etc., which provide a quantitative basis for the model architecture.

Table 5. Parameter Settings of each layer.

Network	Layer	Input Size	Output Size	Kernel Size	Padding
Conv-Layer-1	Conv1D BN	1 32	32	2 × 2	1
Conv-Layer-2	Conv1D BN	32 64	64	2 × 2	1
Conv-Layer-3	Conv1D BN	64 128	128	2 × 2	1
FC-layer	FC1 FC2 FC3	128 128 64	128 64 36		

In summary, the proposed method constructs an efficient gesture recognition model architecture by integrating a noise injection module, a self-attention enhancement module, and a convolutional neural network, which not only strengthens the feature extraction capability of finger bending data but also realizes the accurate recognition of sign language gestures through the fine parameter configuration and hierarchical design. In the next section, the proposed method will be systematically evaluated and validated in terms of recognition accuracy and robustness using a rigorous scientific method.

4. Results

In this section, the overall experimental results of the proposed method on static gesture recognition are elaborated. Through a series of experimental design and implementation, the performance of the proposed method in terms of feature extraction, recognition accuracy, etc., and also systematically evaluates its noise immunity and robustness in practical application scenarios.

4.1. Comprehensive Experimental Performance Evaluation

To ensure the reliability of the experimental results, this paper repeats five experiments independently executed on the training set and the standard gesture test set introduced in Section 2.2. In this paper, the average recognition accuracy is used as the evaluation metric. The average recognition accuracy (*Acc*) is defined as the ratio of the number of correctly classified gesture category samples to the total number of gesture samples [31], and Equation (10) demonstrates its specific calculation method.

$$Acc = \frac{\sum_{1}^{5} \frac{N_{True}}{N_{Total}}}{5} \tag{10}$$

where N_{True} denotes the number of correctly categorized gesture category samples in the test set and N_{Total} denotes the number of total gesture samples in the standard gesture test set. This metric can provide a quantitative evaluation of the performance of the proposed method on the gesture recognition task. As can be seen from the five experiments and calculating the mean value of the recognition accuracy, the proposed method achieves an average recognition accuracy of 99.52% in the standard gesture experiment.

Figure 7 visualizes the convergence and accuracy curves of both the training set and the standard gesture test set during the training process of the model built by the proposed method. Additionally, it displays the ROC curves of the standard gesture test set, which highlight the trade-off between true positive rate and false positive rate at various thresholds. These curves not only clearly reflect the rapid convergence of the learning process and the steady improvement in recognition accuracy, but also demonstrate the strong discriminative power of the proposed method, as evidenced by the ROC curves showing high true positive rates with low false positive rates. These results strongly confirm the efficient recognition capability of the proposed method on the standard gesture recognition task.

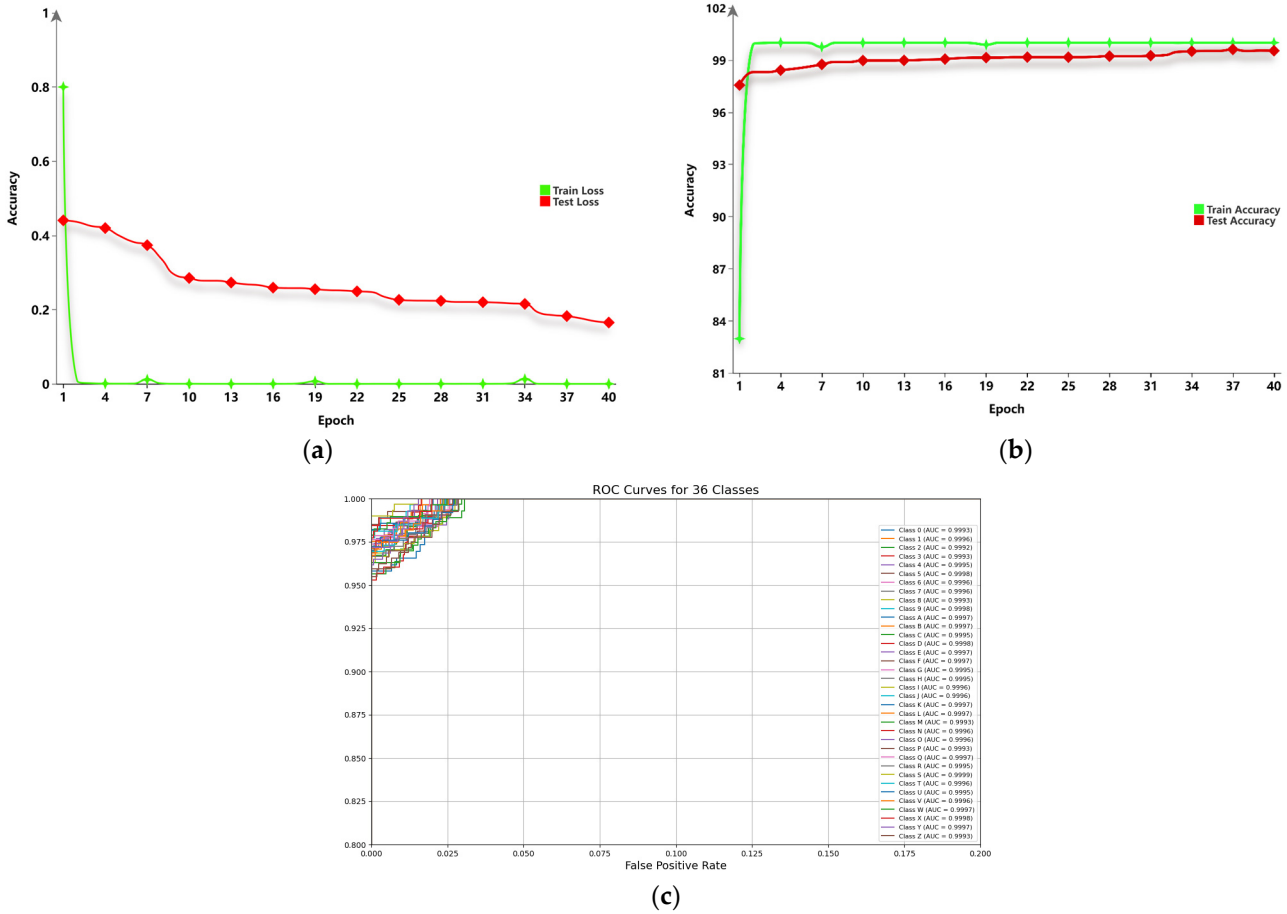


Figure 7. Training set and standard gesture test set: (**a**) Accuracy curves (**b**) Convergence curves. (**c**) ROC curves.

For the robustness experiments, five independent repetitions of each random angular bias test of the same range were conducted separately to calculate the average of the recognition accuracy to ensure the accuracy and consistency of the results. The experimental

results show that the average recognition accuracy of the proposed method shows a decreasing trend as the random angular bias range increases. The average recognition accuracy of each robustness test set is 98.63%, 86.33%, 66.72%, and 45.63%, respectively. As shown in Figure 8, the variation of accuracy with different levels of random angular bias applied is demonstrated.

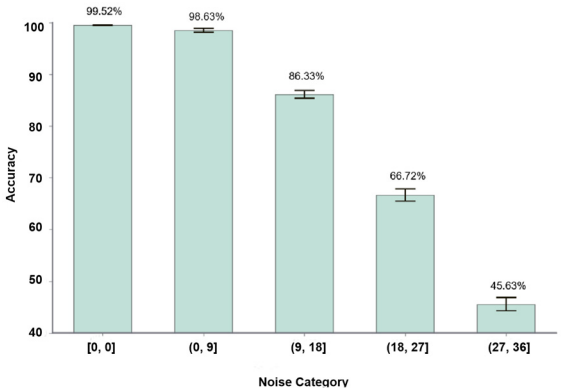


Figure 8. Average recognition accuracy under different levels of random angular bias.

From the analysis in Figure 8, it can be seen that the recognition accuracy decreases with the increase of random angular bias, and the accuracy fluctuation amplitude of multiple trials increases, but the proposed method still maintains the average recognition accuracy of 98.63% and 86.33% in the ranges that the finger bending bias angle is in the range of ±(0°–9°] and ±(9°–18°]. This result further consolidates the high noise immunity and strong robustness demonstrated by the proposed method in practical application scenarios.

4.2. Comparison of Identification Methods

In order to verify the effectiveness of the proposed method, the following paper compares the different state-of-the-art methods proposed by scholars based on wearable device gesture recognition, and Table 6 shows the specific comparison results.

Table 6. Comparison of existing advantageous methods.

Number	Method	Accuracy	Precision	Recall	F1-Score	Year
1	CNN + BilSTM [10]	98.21%	98.15%	98.10%	98.12%	2023
2	BP Network [11]	98.5%	98.45%	98.40%	98.42%	2024
3	SVM [12]	99.2%	99.15%	99.10%	99.12%	2024
4	Our Method	99.63%	99.53%	99.47%	99.48%	2024

By analyzing the data in Table 6, it can be observed that the proposed method achieves significant improvements across all evaluation metrics in the standard gesture recognition task compared to the current state-of-the-art techniques. This demonstrates not only the effectiveness of the proposed method but also highlights its superiority over other approaches. The observed improvements are mainly due to two factors. First, the incorporation of noise injection enhances data diversity, making the model more robust to input variations and improving generalization. Second, the self-attention mechanism allows for more precise extraction of key features by enabling the model to focus on the most relevant aspects of the input data. These factors lead to the enhanced performance of the proposed method.

In addition, in order to fully evaluate the noise immunity and robustness of the proposed methods in this paper, further robustness experiments are conducted for the methods compared above. Table 7 shows the comparative results of the robustness experiments. The experimental results show that the proposed method can still maintain an average recognition accuracy of 98.63% and 86.33% under the introduction of random angular bias

conditions of ±(0°–9°] and ±(9°–18°]. These experimental results not only demonstrate the stability of the proposed method in real application scenarios but also further prove the excellent noise immunity and robustness experiments of the proposed method.

Table 7. Comparison of noise resistance accuracy of different methods.

Angular Bias Range	Method	Accuracy	Precision	Recall	F1-Score
±(0, 9]	CNN + BiLSTM	96.77%	96.70%	96.65%	96.68%
	BP Network	95.17%	95.10%	95.05%	95.07%
	SVM	94.32%	94.25%	94.20%	94.22%
	Our Method	98.63%	98.55%	98.50%	98.52%
±(9, 18]	CNN + BiLSTM	81.65%	81.50%	81.45%	81.47%
	BP Network	80.46%	80.35%	80.30%	80.32%
	SVM	77.13%	77.00%	76.95%	76.97%
	Our Method	86.33%	86.20%	86.15%	86.17%

4.3. Comparison of Ablation Experiments

The core approach of the proposed method to improve robustness relies on two key modules: the noise injection module that increases data diversity through noise injection, and the self-attention enhancement module that applies self-attention mechanisms to strengthen key features. In order to fully assess the specific contribution of these modules to the performance of the proposed method in this paper, a series of ablation experiments will be conducted on the ASL dataset. These experiments will be conducted while keeping other parameters constant, aiming to quantitatively analyze the effects of the noise injection module and the self-attention enhancement module on the performance of the proposed method. After 40 epochs of iterative training on the base convolutional neural network based on adding the noise injection module as well as the self-attention enhancement module, respectively, with the learning rate set to 0.0001, Table 8 demonstrates the specific experimental results.

Table 8. Results of ablation experiments.

Angular Bias Range	Experiment	Method	Accuracy
0	Without Noise Injection	CNN	97.33%
		CNN + SAM	98.78%
	Noise Injection	CNN	98.44%
		CNN + SAM	99.52%
(0, 9]	Without Noise Injection	CNN	97.31%
		CNN + SAM	98.33%
	Noise Injection	CNN	97.50%
		CNN + SAM	98.63%
(9, 18]	Without Noise Injection	CNN	77.02%
		CNN + SAM	84.69%
	Noise Injection	CNN	82.33%
		CNN + SAM	86.33%
(18, 27]	Without Noise Injection	CNN	57.81%
		CNN + SAM	65.19%
	Noise Injection	CNN	64.66%
		CNN + SAM	66.72%
(27, 36]	Without Noise Injection	CNN	38.03%
		CNN + SAM	44.64%
	Noise Injection	CNN	42.52%
		CNN + SAM	45.63%

From Table 8, it can be analyzed that the proposed method improves the recognition rate of standard gestures from 97.33% to 99.52%, which is a 2.19% improvement in accuracy relative to the CNN network without the noise injection module and self-attention enhancement module. After adding the random angular bias of ±(0°–9°], ±(9°–18°], ±(18°–27°], ±(27°–36°], respectively, it can be seen from the comparison of the data in Table 8 that the recognition accuracy increased by 1.32%, 9.31%, 8.91%, and 7.60%, respectively. Figure 9 illustrates the changes in the growth rates.

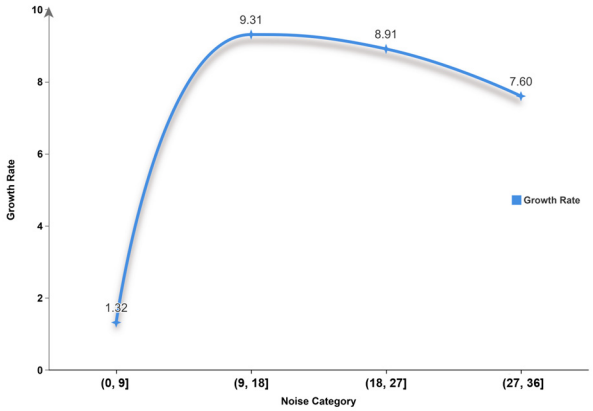


Figure 9. Change in recognition rate growth.

Finally, in order to further analyze the inference performance of the proposed method, the recognition performance of the proposed method for each gesture category is shown in detail on the standard gesture dataset using the confusion matrix. As shown in Figure 10, the confusion matrix of the gestures is represented. The values on the main diagonal of the confusion matrix indicate the percentage of correctly predicted samples in each gesture category, and the remaining positions indicate the cases where the model incorrectly predicts a given gesture as another category. From Figure 10, it can be observed that the model can recognize 32 gestures accurately, with the most confusion for the numeric gestures 1 and 6 and the alphabetic gestures E and O.

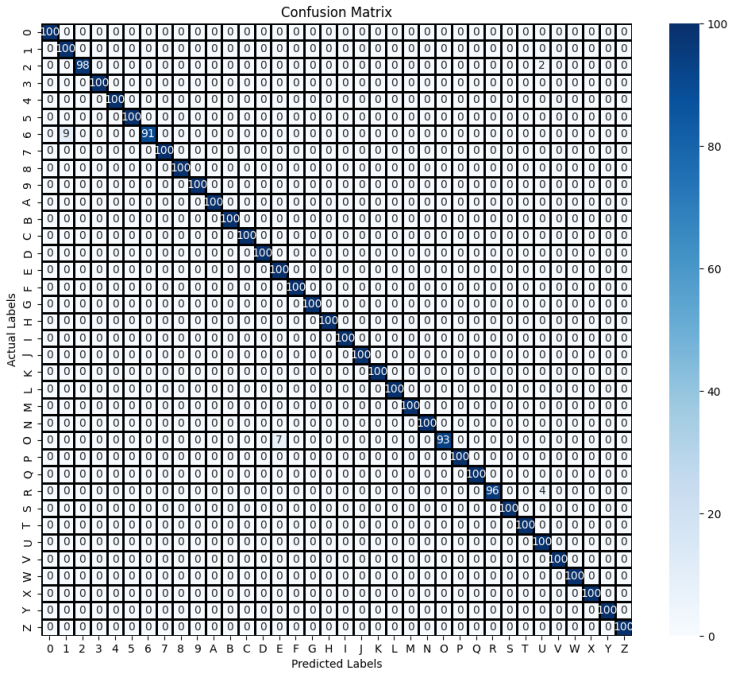


Figure 10. Confusion matrix of standard gesture test set.

5. Conclusions

In this paper, a static sign language recognition method based on self-attention mechanism enhancement is proposed to address the problems of poor noise immunity and low robustness of wearable devices under different usage objects. The method increases the sample diversity through a noise injection module and designs a self-attention enhancement module, which improves the weight of key features in the feature extraction process and enhances the robustness of gesture recognition.

The experimental results show that the proposed method achieves an accuracy of 99.52% under the standard gesture test experiment set of the new dataset customized with reference to the ASL dataset, and the accuracy rate has been improved to varying degrees in comparison with mainstream methods at home and abroad. In the robustness experiments, even under the random angular bias conditions of $\pm(0°–9°]$ and $\pm(9°–18°]$, the average recognition accuracies can still be maintained at 98.63% and 86.33%, respectively, and the accuracies are still improved compared with the latest mainstream methods, which show excellent noise resistance and robustness.

However, current research mainly focuses on static gesture recognition, while gesture interactions in real applications are usually dynamic and continuous. Therefore, future work will focus on dynamic gesture recognition and work on designing more efficient and accurate real-time interaction models.

Author Contributions: Conceptualization, Y.W. and H.J.; methodology, H.J.; software, L.X.; validation, H.J., Y.S. and L.X.; formal analysis, H.J.; investigation, Y.S.; resources, H.J.; data curation, H.J.; writing—original draft preparation, H.J.; writing—review and editing, Y.W. and H.J.; visualization, H.J.; supervision, Y.W.; project administration, H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available at https://github.com/jones12138/Dataset, and it was collected on 26 August 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASL	American Sign Language
AI	Artificial Intelligence
VR	Virtual Reality
sEMG	Surface Electromyography
Flex	Resistive Bending Sensor
CNN	Convolutional Neural Networks
FC	Fully Connected Layer
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
MLP	Multi-Layer Perceptron
IMU	Inertial Measurement Unit
KNN	K-Nearest Neighbors
HOG	Histogram of Oriented Gradients
SAM	Self-Attention Mechanism
Conv1D	1-Dimensional Convolution
BN	Batch Normalization
SVM	Support Vector Machine
BP Network	Backpropagation Network

Appendix A

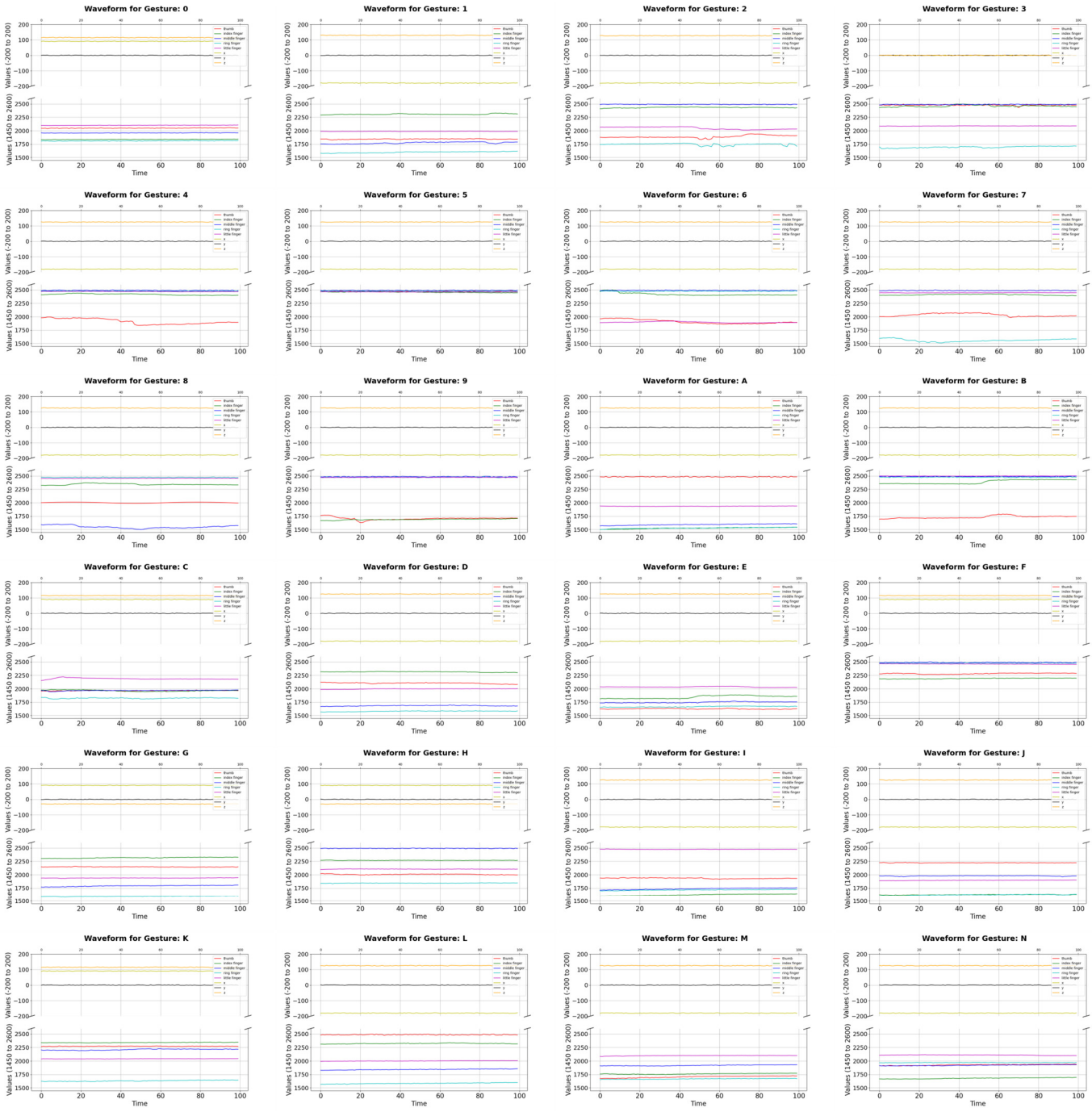


Figure A1. *Cont.*

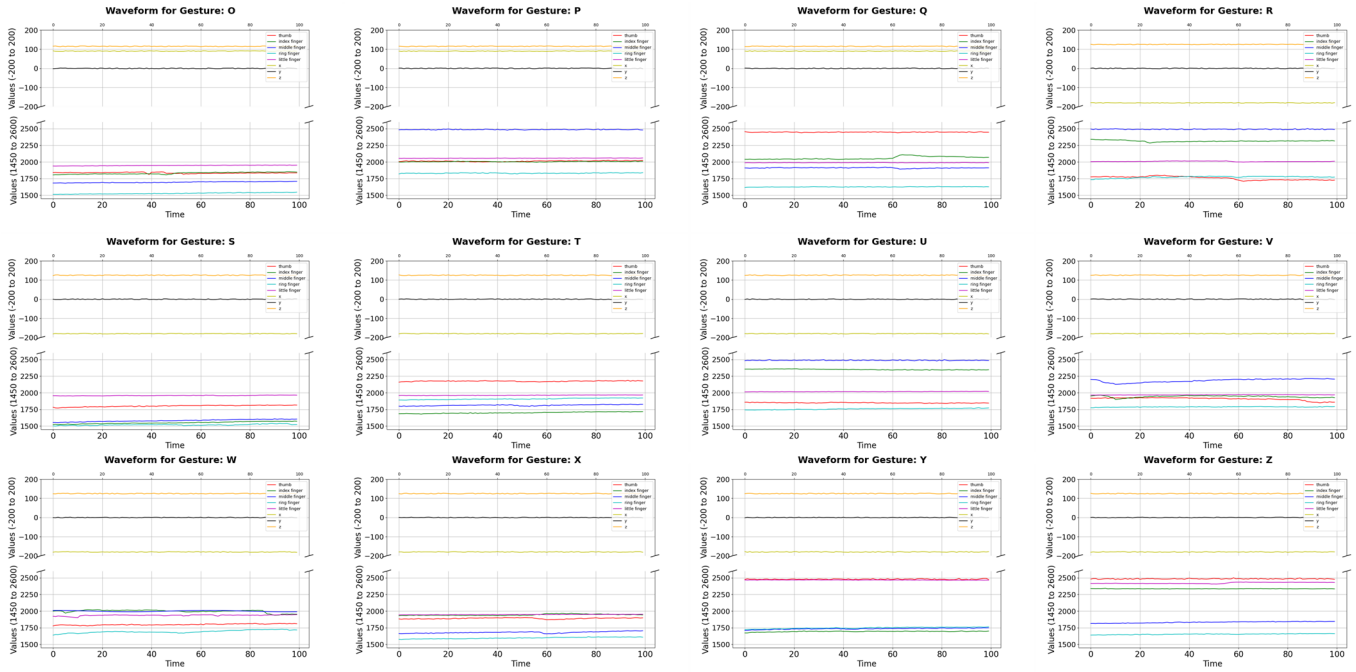


Figure A1. 36 static gesture waveform.

References

1. Sánchez-Vicinaiz, T.; Camacho-Perez, E.; Castillo Atoche, A.; García-Martínez, J.; Rodriguez, J. MediaPipe Frame and Convolutional Neural Networks-Based Fingerspelling Detection in Mexican Sign Language. *Technologies* **2024**, *12*, 124. [CrossRef]
2. Ali, A.R.; Moubarak, F. Exoskeleton Robotics Glove with A Dual Force Feedback Haptic Device for Immersive Virtual Reality Interaction. In Proceedings of the 2023 2nd International Conference on Smart Cities 4.0, Cairo, Egypt, 22–24 October 2023; pp. 304–307.
3. Majumder, S.; Mondal, T.; Deen, M.J. Wearable Sensors for Remote Health Monitoring. *Sensors* **2017**, *17*, 130. [CrossRef]
4. Xie, P.; Zhao, L.; Zhang, Y.; Xu, M.; Jiang, G.; Chen, J. EMG-Inertial Feature Fusion and Gesture Recognition Based on Multi-source Domain Transfer Learning. *J. Electron. Meas. Instrum.* **2024**, 1–9. [CrossRef]
5. Ferrara, E. Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling: A Survey of Early Trends, Datasets, and Challenges. *Sensors* **2024**, *24*, 5045. [CrossRef]
6. Zhang, Y.; He, X.; Wang, J.; Liu, Z.; Li, S.; Wu, L. Research on Gesture Recognition Techniques and Methods. *J. Northwest Minzu Univ.* **2024**, *45*, 21–36. [CrossRef]
7. Li, L.; Jiang, S.; Gu, G. SkinGest: Artificial skin for gesture recognition via filmy stretchable strain sensors. *Adv. Robot.* **2018**, *32*, 1112–1121. [CrossRef]
8. Mummadi, C.; Philips, F.; Verma, K.; Kasireddy, S.; Scholl, P.; Kempfle, J.; Van Laerhoven, K. Real-Time and Embedded Detection of Hand Gestures with an IMU-Based Glove. *Informatics* **2018**, *5*, 28. [CrossRef]
9. Sernani, P.; Pacifici, I.; Falcionelli, N.; Tomassini, S.; Dragoni, A.F. Italian sign language alphabet recognition from surface EMG and IMU sensors with a deep neural network. In Proceedings of the Recent Trends and Applications in Computer Science and Information Technology, Tirana, Albania, 26–27 April 2021.
10. Fan, L.; Zhang, Z.; Zhu, B.; Zuo, D.; Yu, X.; Wang, Y. Smart-Data-Glove-Based Gesture Recognition for Amphibious Communication. *Micromachines* **2023**, *14*, 2050. [CrossRef]
11. Zhu, Y.; Shen, H.; Wu, J.; Wang, X.; Liu, Y. Gesture Recognition Research Based on Flexible Strain Sensors for Data Gloves. *J. Agric. Mach.* **2024**, *55*, 451–458.
12. Deng, W.; Lu, C.; Song, C.; Xu, J.; Wu, H. Guiding Industrial Robot Motion through Body Movements. *Sens. Microsyst.* **2024**, *43*, 164–168. [CrossRef]
13. Shin, J.; Matsuoka, A.; Hasan, M.A.; Srizon, A. American Sign Language Alphabet Recognition by Extracting Feature from Hand Pose Estimation. *Sensors* **2021**, *21*, 5856. [CrossRef] [PubMed]
14. Lee, C.; Ng, K.K.H.; Chen, C.-H.; Lau, H.C.W.; Chung, S.Y.; Tsoi, T. American Sign Language Recognition and Training Method with Recurrent Neural Network. *Expert Syst. Appl.* **2021**, *167*, 114403. [CrossRef]
15. Gibson, H.; Byrne, A. ASL LITERACY AND CRITICAL THINKING DEVELOPMENT. *Rev. Bras. De Alf.* **2024**, *22*, 1–20.
16. Mei, X.; Ye, X.; Wang, J.; Wang, X.; Huang, H.; Yusong Liu, Y.; Jia, Y.; Zhao, S. UIEOGP: An underwater image enhancement method based on optical geometric properties. *Opt. Express* **2023**, *31*, 36638–36655. [CrossRef]
17. Zheng, H.; Wu, H.; Du, Z. Enhancing Gut Metagenome Images and Improving Classification Prediction Accuracy of Metabolic Diseases through Deep Learning. *Genetics* **2024**, 1–13. [CrossRef]

18. Zhao, J.; Lui, H.; Kalia, S.; Lee, T.; Zeng, H. Improving skin cancer detection by Raman spectroscopy using convolutional neural networks and data augmentation. *Front. Oncol.* **2024**, *14*, 1320220. [CrossRef] [PubMed]
19. Zeng, W. Image data augmentation techniques based on deep learning: A survey. *Math. Biosci. Eng.* **2024**, *21*, 6190–6224. [CrossRef]
20. Shorten, C.; Khoshgoftaar, T. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
21. Feng, R.; Chen, D.; Hua, B. A Review of Augmentation Methods for Text Data. *Data Anal. Knowl. Discov.* **2024**, 1–17. Available online: https://link.cnki.net/urlid/10.1478.G2.20240911.1803.004 (accessed on 19 September 2024).
22. Wu, J.; Ren, P.; Song, B.; Zhang, R.; Zhao, C.; Zhang, X. Data glove-based gesture recognition using CNN-BiLSTM model with attention mechanism. *PLoS ONE* **2023**, *18*, e0294174. [CrossRef]
23. Pratap, S.; Narayan, J.; Hattta, Y.; Ito, K.; Hazarika, S. Glove-Net: Enhancing Grasp Classification with Multisensory Data and Deep Learning Approach. *Sensors* **2024**, *24*, 4378. [CrossRef] [PubMed]
24. Zhang, Y.; Li, J.; Karas, Z.; Bansal, A.; Li, T.; McMillan, C.; Leach, K.; Huang, Y. EyeTrans: Merging Human and Machine Attention for Neural Code Summarization. *Proc. ACM Softw. Eng.* **2024**, *1*, 115–136. [CrossRef]
25. Cao, Z.; Qin, Y.; Jia, L.; Xie, Z.; Gao, Y.; Wang, Y.; Li, P.; Yu, Z. Railway Intrusion Detection Based on Machine Vision: A Survey, Challenges, and Perspectives. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 6427–6448. [CrossRef]
26. Wu, D.; Wang, J.; Zhao, W.; Geng, X.; Liu, G.; Qiu, S. Research on wearable sensor gesture recognition based on CNN_GRU and attention mechanism. *J. Intell. Fuzzy Syst.* **2024**, 1–12. [CrossRef]
27. Zhao, R.-X.; Shi, J.; Li, X. QKSAN: A Quantum Kernel Self-Attention Network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**. *early access*. [CrossRef]
28. Sun, Q.; Zhang, J.; Fang, Z.; Gao, Y. Self-Enhanced Attention for Image Captioning. *Neural Process. Lett.* **2024**, *56*, 131. [CrossRef]
29. Wang, Z.; Zhu, H.; Liu, F. SMSTracker: A Self-Calibration Multi-Head Self-Attention Transformer for Visual Object Tracking. *Comput. Mater. Contin.* **2024**, *80*, 605–623. [CrossRef]
30. Ali, M. Advancing Crowd Object Detection: A Review of YOLO, CNN and ViTs Hybrid Approach. *J. Intell. Learn. Syst. Appl.* **2024**, *16*, 175–221. [CrossRef]
31. Wu, J.; Tian, Q.; Yue, J. Static Gesture Recognition Based on Residual Dual Attention and Cross-Level Feature Fusion Modules. *Comput. Syst. Appl.* **2022**, *31*, 111–119. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.