Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Real-Time Chinese Sign Language Recognition Based on Artificial Neural Networks*

Zhen Zhang, Ziyi Su, Ge Yang
School of Mechatronic Engineering and Automation
Shanghai University
Shanghai, China
{zhangzhen_ta & randy_sw & 322yg}@shu.edu.cn

***Abstract* - Sign language is the main method for the deaf-mute to communicate with each other. However, normal people cannot understand the meaning of specific gestures without training. In this study, a real-time Chinese Sign Language (CSL) recognition model based on surface electromyographic (sEMG) signals and Artificial Neural Networks (ANN) is proposed. The model achieves an average accuracy at 88.7% on 15 CSL gestures with average response time around 300ms (timing at movement begins). In our approach, the MYO armband is used for acquiring raw sEMG signals. Signal preprocessing includes rectification, filter and extracting the muscle activation section. We apply a sliding window and a muscle detection approach to segment the signals and extract features. A test method is used to recognize the gesture when the totality of same label, which returned by ANN classifier, reach the activation times threshold.**

Index Terms* - *Sign language recognition; Surface electromyography; Artificial neural networks; Real time

I. INTRODUCTION

Sign language (SL) is commonly used among the people who are deaf-mute or blind. As a kind of structured language, SL is comprised of abundant abstract hand and arm gestures. Therefore, normal people cannot communicate with the deaf-mute without receiving special training. Sign language recognition (SLR) provides an efficient approach for translating sign language to speech or text [1]. Accordingly, it is important to develop a system that can translate sign language instantly and automatically.

According to the diversity of sensors, traditional approaches for SLR can be divided into two main categories: the vision-based approach and the data glove-based approach. For vison-based approaches, Cooper et al. [2] used Hidden Markov Models (HMM) on 40 German sign words dataset achieving 85.1% classification accuracy. J. Zhang et al.[3] proposed a computer vision-based framework with adaptive HMM, which fuses the spatio-temporal information, obtained the accuracy of 92% over 100 signs words and 86% over 500 signs. However, the precision of visual system can be significantly influenced by the external factors, such as the illumination and occlusion, and the real-time performance still cannot meet the requirement. The other approach is based on the data gloves, which can track the posture of hands and depicting the trajectory. K. Li et al. [4] used a pair of digital gloves to collect real-world data to develop a new SLR framework and achieved accuracy rate of 87.4%. X. Zhang et al.[5] presented a system for fusing information from accelerometer and electromyography sensors, obtaining average accuracy about 90.6% for 72 Chinese sign language words. Although, wearing a pair of data gloves to collect the movement data from hands and fingers can really acquire a high recognition precision, potentially run counter to the intention of efficient and convenient.

In recent years, the increasing development of the surface electromyography (sEMG) application has given a great push to the development of hand gesture recognition, and provide a more convenient and stable approach of human-computer interaction compared with methods which based on vision or data gloves. Some wearable sEMG sensors, such as MYO armband [6] issued by the Thalmic, provide a better user experience than many traditional technologies. Machine learning is a state-of-the-art method for hand gesture recognition based on sEMG. The practical models for classification including decision tree [1], support vector machines [7], k-nearest neighbors, artificial neural networks [6, 8]. The conventional features for gesture recognition can be defined in in three main parts: the time-domain features, such as mean absolute value (MAV) and root mean square (RMS); frequency-domain features, such as wavelets; and frequency-domain features, such as frequency histograms. Consequently, the exploration of sign language recognition is still an open area for more and more new researchers.

In this paper, we present a SLR model based on sEMG signals and artificial neural networks (ANN), which achieve a decent average accuracy with real-time response. As an attempt for real-time sign language recognition system based on sEMG and ANN, the whole model is composed of following blocks, as shown in Fig.1: for data acquisition, we use MYO armband to collect sEMG signals of 15 Chinese sign language gestures. For preprocessing, we use low-pass filter to smooth the signal and remove the noise. For feature extraction, we extract several time-domain features with preprocessed sEMG signals in a sliding window to form the feature vectors. For classification, we develop an ANN classifier, dividing the dataset stochastically into two groups for training and testing, respectively.

* This work is jointly supported by Shanghai science and technology commission under grant No 18JC1410402.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

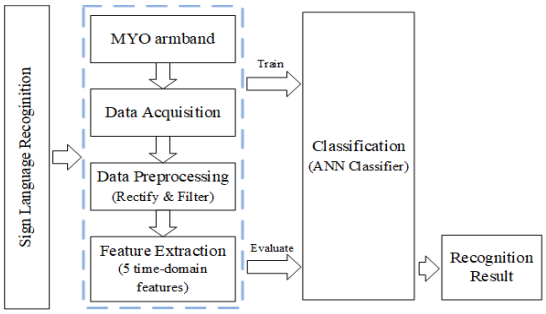

Fig.1 Main methodologies used in our approach

The rest of the paper is organized as follows. In the Section 2, we introduce the methodology and the ANN model we present in this work. The experimental results and analysis are discussed in Section 3. In section 4, we conclude the paper.

II. METHODOLOGY

A. Data Acquisition

In our approach, we use the MYO armband, which is illustrated in Fig.2, to collect sEMG signals. It integrates 8 channels of sEMG sensors which measure at 200Hz. Each channel returns the digital signal to the host computer via Bluetooth in real-time during acquisition. In addition, the armband also can be adjusted from 19cm to 34cm while only 93g weight, which really ensure the comfortable user experience.

In this study, we aim to recognize 15 Chinese sign language gestures, including "lin", "yi", "er", "san", "si", "wu", "liu", "qi", "ba", "jiu", "shi", "a", "b", "c", "d" (Fig.3). The set of the other gestures that our model cannot classify, including the relax posture, are referred as "no-gesture". It is worth noting that these pictures illustrate gestures at their final position. For convenience, we denote these 15 sign language gestures by English letters from "a" to "o" in the following text.

10 healthy subjects, including 6 males and 4 females, participated in this study. All the volunteers were trained before the experiment. In the beginning, each subject was asked to perform the 15 kind of sign language gestures in a random order. For the training dataset, every subject performed 5 repetitions of the 15 gestures and each of the sample was recorded during 2 seconds. For the test dataset, every subject performed 30 repetitions of each gesture during 2 seconds. As for the data of the "no-gesture", volunteers were asked to preform hand in the relaxed position while recording. For every repetition we carried out in following steps: relax, preform to the final position, relax. The reason why we set the ratio of training set to test set at 1:6 is that we can ensure the recognition performance of our model with limited training samples on account of users will not operate so many repetitions to train the system when applications.

Fig.2 MYO Armband

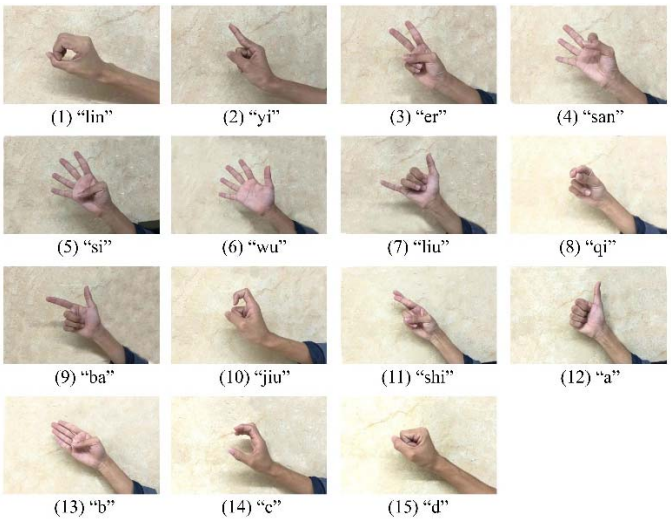

Fig.3 15 kind of Chinese Sign Language gestures

After this period, we can obtain the training dataset $D_{train}=\{(P_1,Y_1),\ldots,(P_N,Y_N)\}$ for each subject. The letter N denotes the number of training samples, here N = 75. Where $P_i \in [-150,150]^{8\times400}$ denotes the i^{th} channel of sEMG signals recorded in 2 seconds. The label vector $Y_i \in \{a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,no\ gesture\}$ stands for the category of the signal P_i. As for the test dataset, we obtain $D_{test}=\{(Q_1,Y_1),\ldots,(Q_M,Y_M)\}$ of each subject and the totality of instances M = 450. Each feature matrix Q_i is an instance of the range $[-150,150]^{8\times400}$. Unlike the training set, we remove the category "no-gesture" because the aim of our study is the recognition of 15 specific SL gestures.

B. Signal Preprocessing

The raw sEMG signals we obtained in last period are instable and contain additional noises. In order to make it easier for feature extraction, we normalize the original signal at first, which make each signal being in the range of [-1,1]. Then, an absolute value function is applied for rectifying the signal. We use the 4^{th} Butterworth low-pass filter to eliminate noises and smooth signals. We set a reasonable cut-off frequency at 5Hz, which is obtained by Fourier transform.

For extracting the muscle activation section, we apply a muscle detection function to remove the relaxed state in the head and the tail in each sEMG instance of training set. In the meantime, an energy spectrum approach is used to determine the muscle activity region, which is introduced in [9]. Firstly, we calculate the sum of sEMG envelopes, shown in Fig.4. Then, we use the short-term Fourier transform to obtain the energy spectrum and detect muscle activation area.

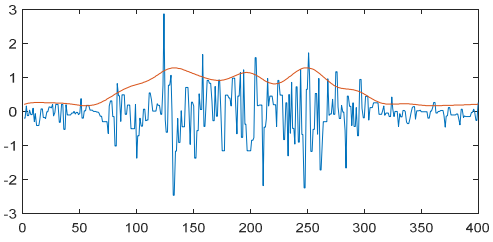

Fig.4 The raw sEMG signal (blue line) and its envelope (red line)

C. Feature Extraction

For the extraction of the features, we applied a sliding window approach in our model, which can divide the time series into data segments as illustrated in Fig.5. The length of the sliding window we denote as l. To obtain a high real-time performance, the step length of the window is set at 5ms (one sampling point), which is denoted as s. We use the data in the sliding window for feature extraction.

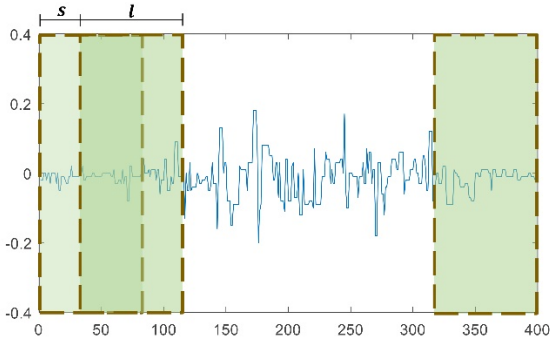

Fig.5 Using the sliding window to segment the sEMG signals

The features of our study are mainly composed by two categories: the values of preprocessed sEMG signals and 5 time-domain features including RMS, MAV, WL (wave length), SSC (slope sign change) and HP (Hjorth parameter). We set the sEMG signal as $\mathrm{a}(i)$, these 5 time-domain feature algorithms are listed in TABLE I:

TABLE I
THE FEATURES AND ALGORITHMS USED IN MODEL

Feature	Algorithm
RMS	$\mathrm{R}=\sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathrm{a}(i)^2}$
MAV	$\mathrm{M}=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\lvert\mathrm{a}(i)\rvert$
WL	$\mathrm{W}=\sum_{i=2}^{\mathrm{N}}\lvert\mathrm{a}(i)-\mathrm{a}(i-1)\rvert$
SSC	$\mathrm{S}=\sum_{i=2}^{\mathrm{N-1}}\lvert(\mathrm{a}(i)-\mathrm{a}(i-1)\times(\mathrm{a}(i)-\mathrm{a}(i+1))\rvert$
Activity Parameter (HP)	$\mathrm{A_{hp}}=\mathrm{VAR}(\mathrm{a}(i))=\frac{1}{\mathrm{N-1}}\sum_{\mathrm{k-1}}^{\mathrm{N}}\mathrm{a}(i)^2$
Mobility Parameter (HP)	$\mathrm{M_{hp}}=\sqrt{\frac{\mathrm{VAR}\left(\frac{\mathrm{da}(i)}{\mathrm{d}i}\right)}{\mathrm{VAR}(\mathrm{a}(i))}}$
Complexity Parameter (HP)	$\mathrm{C_{hp}}=\frac{\mathrm{Mob}\left(\frac{\mathrm{da}(i)}{\mathrm{d}i}\right)}{\mathrm{Mob}(\mathrm{a}(i))}$

To determine the length of the sliding window, we apply the T-Distributed Stochastic Neighbor Embedding (T-SNE) to reduce the dimension and visualize the feature vectors' cluster in feature space. The clustering results are shown in Fig.6, we can see that with the increase of the window size, the feature vectors of each class of SL gesture get closer to each other. Owing to the totality of feature vectors tend to reduce as the increasing of the window size, the changes become no obvious effect after the size is bigger than 80 points (400ms).

To obtain a high recognition performance and avoid over fitting, we determine the sliding window size at 40 points (200ms) in our model.

After determining the window size, the intact feature vector X_i of each sample can be obtained. The feature vector is mainly composed by two parts: preprocessed sEMG signals and the 5 time-domain features, which are denoted as R_i and F_i respectively. The length of R_i is equal to $\text{Window Size} * 8$, so there is $|R_i| = 40 * 8 = 320$ features. After we calculating these 5 time-domain features, we can acquire the length of vector F_i is $7 * 8 = 56$ features. Therefore, we obtain the final feature vector $X_i = [R_i, F_i]$ in the length of $|R_i| + |X_i| = 320 + 56 = 376$ features.

D. Classification

In this section, we develop ANN for classification, which has three layers: the input layer, the hidden layer and the output layer. The number of the input layer is corresponding to the element of feature vectors. The number of the nodes in hidden layer which is set about half of the input elements and the output layer contains 17 elements corresponding to the different prediction target gestures. We set the sigmoid function as the transfer function. The cross-entropy cost function and full batch gradient descent algorithm is applied to train the network. We use the train dataset to train the classifier and use the test set to evaluate it.

III. DISCUISSON

In order to evaluate the performance of the proposed model, we assess, in terms of recognition accuracy. The accuracy is defined as the number of correct classified samples over the totality of classified samples. In this section, we represent our results and compared them with the other researchers' work. Meanwhile, we will illustrate the real-time performance of our method as well.

A. Recognition Accuracy

The confusion matrix of test set is shown in Fig.7. It shows the overall recognition accuracy and each accuracy to the corresponding sign language gesture samples of all experimenters. As we can see from the chart, the overall recognition accuracy is about 88.7%. The gesture of the meaning "liu" obtains the highest sensitivity which at 95.3%, owing to the most distinguishable muscle motion among all gestures. By contrast, sign language gestures of "lin" and "ba" both get the lowest of 79.7%. As for precision, the gestures of "d" ranks at the first (96.1%) and the "wu" is the lowest (78.6%). Therefore, we conclude that the best recognition sign language gesture of our model is the gesture of the meaning "d". Additionally, to some

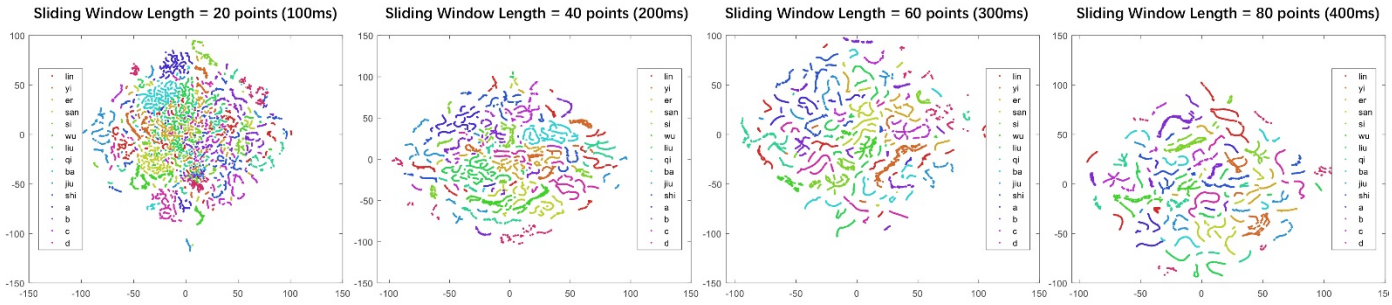

Fig.6 The T-SNE result of a single subject

		Targets																%PRECISION
		NG	lin	yi	er	san	si	wu	liu	qi	ba	jiu	shi	a	b	c	d	
Predictions	NO-GESTURE	0	0	1	2	3	1	1	0	0	1	2	0	0	0	0	0	NaN
	lin	0	239	0	6	1	1	1	0	5	0	1	1	0	2	1	0	92.6%
	yi	0	11	247	3	0	0	3	0	2	5	11	2	1	0	0	0	86.7%
	er	0	3	15	270	3	0	1	0	0	3	8	8	0	1	0	0	86.5%
	san	0	2	0	0	246	5	6	0	0	1	1	1	0	4	0	0	92.5%
	si	0	2	1	3	4	266	3	0	0	2	0	5	0	20	2	0	86.4%
	wu	0	3	2	0	33	9	279	2	0	15	4	1	3	3	1	0	78.6%
	liu	0	2	2	0	1	0	0	286	0	14	0	0	10	0	0	0	90.8%
	qi	0	6	3	0	0	0	0	0	284	1	9	0	0	0	0	1	93.4%
	ba	0	1	6	0	0	0	0	2	0	239	0	0	3	1	3	0	93.7%
	jiu	0	8	11	4	0	1	0	0	0	3	259	4	0	0	2	0	88.7%
	shi	0	2	7	8	5	1	0	0	1	2	3	266	0	2	1	0	89.3%
	a	0	3	0	0	0	0	0	10	0	7	2	0	279	0	3	7	89.7%
	ba	0	3	1	4	4	16	6	0	0	0	0	9	0	264	10	0	83.3%
	c	0	13	1	0	0	0	0	0	7	5	0	3	0	3	277	0	89.6%
	d	0	2	3	0	0	0	0	0	1	2	0	0	4	0	0	292	96.1%
	%SENSITIVITY	NaN	79.7%	82.3%	90.0%	82.0%	88.7%	93.0%	95.3%	94.7%	79.7%	86.3%	88.7%	93.0%	88.0%	92.3%	93.7%	88.7%

Fig.7 The confusion matrix of the result

extent our model tends to mistake the gesture of "san" as the gesture of "wu", which may cause by analogous skeletal muscle movements.

Besides, the gesture of "wu" is more likely to be incorrected classified in our model. It is worth noting that owing to some samples' strength are too low to pass the thresholds of the preprocessing and postprocessing, they tend to be recognized as "NO-GESTURE" as shown in the figure.

B. Real-time Recognition performance

In real practice, SLR requires not only high recognition accuracy but real-time response. For a gesture recognition system to operate in real time, the response time need to be limited in no more than 300ms after the movement begins [9]. However, in current practical researches, most pattern recognition approaches start timing after the action is completed, which cannot achieve an actual real-time recognition performance. F. Wang et.al[10] proposed a SLR model based on CNN and achieve average accuracy over 90%. Their model can response in less than 50ms after the movement finished. Unlike others methods, our approach starts timing at the beginning of the gesture instead of the end. An overflow algorithm is applied to count the label that classifier return to the model through the sliding window. If the number of the label for the same gesture over the activation threshold we set, then our model returns this gesture as the predicted gesture. The lower activation times threshold we set, the shorter time will the classifier return the predicted label. However, if the threshold is set too low, the incorrect classification may increase. We set an appropriate activation threshold at 30.

Fig.8 compares the average time for sEMG signal activation with model response time for each gesture instances of all experimenters. Most sEMG signals of each gesture active over 1500ms, while the average response time is only around 300ms. Apparently, the response time of our model is much shorter than the sEMG signal activation time, which means our model can recognize the sign language gesture much earlier than the gesture ends.

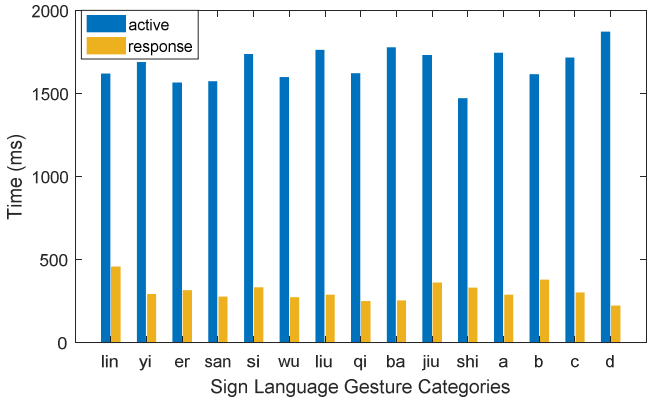

Fig.8 The average active and response time for each gesture of all subjects

IV. CONCLUSION

In this paper, we present a real-time sign language gesture recognition model using sEMG signals and ANN. We use the MYO armband to acquire raw sEMG signals. A sliding window and the energy spectrogram approach are applied to extract features. We divide the dataset into two parts, and use the train set to train the classifier and evaluate it through the test set. An overflow approach is used that the predicted label will be returned if the sum of same label reaches the activation threshold. The average accuracy of our model to 15 specific Chinese sign language gestures is 88.7%, while the average respond time is only around 300ms, which shows a decent real-time performance.

REFERENCES

[1] X. Yang, X. Chen, X. Cao, S. Wei, and X. Zhang, "Chinese Sign Language Recognition Based on an Optimized Tree-Structure Framework," *Ieee Journal of Biomedical and Health Informatics,* vol. 21, no. 4, pp. 994-1004, Jul 2017.

[2] H. Cooper, E. Ong, N. Pugeault, and R. Bowden, "Sign Language Recognition using Sub-Units, " *Journal of Machine Learning Research*, vol. 13, pp. 2205-2231, Jul 2012.

[3] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese Sign Language Recognition with Adaptive Hmm, " *2016 Ieee International Conference on Multimedia & Expo (Icme),* 2016.

[4] K. Li, Z. Zhou, and C. Lee, "Sign Transition Modeling and a Scalable Solution to Continuous Sign Language Recognition for Real-World Applications, " *Acm Transactions on Accessible Computing*, vol. 8, no. 2, Jan 2016.

[5] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors, " *Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans,* vol. 41, no. 6, pp. 1064-1076, Nov 2011.

[6] Z. Zhang, K. Yang, J. Qian, and L. Zhang, "Real-Time Surface EMG Pattern Recognition for Hand Gestures Based on an Artificial Neural Network, " *Sensors,* vol. 19, no. 14, Jul 18 2019.

[7] N. Dardas and N. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques, " *Ieee Transactions on Instrumentation and Measurement,* vol. 60, no. 11, pp. 3592-3607, Nov 2011.

[8] C. Motoche and M. Benalcazar, "Real-Time Hand Gesture Recognition Based on Electromyographic Signals and Artificial Neural Networks, " *Artificial Neural Networks and Machine Learning - Icann 2018, Pt I,* vol. 11139, pp. 352-361, 2018.

[9] M. Benalcazar *et al.*, "Real-Time Hand Gesture Recognition Using the Myo Armband and Muscle Activity Detection, " *2017 Ieee Second Ecuador Technical Chapters Meeting (Etcm),* 2017.

[10] F. Wang, S. Zhao, X. Zhou, C. Li, M. Li, and Z. Zeng, "An Recognition-Verification Mechanism for Real-Time Chinese Sign Language Recognition Based on Multi-Information Fusion, " *Sensors,* vol. 19, no. 11, Jun 1 2019.