L-Sign: Large-Vocabulary Sign Gestures Recognition System

Zhiwen Zheng, Qingshan Wang, *Member, IEEE*, Dejun Yang, *Senior Member, IEEE*, Qi Wang, Wei Huang, and Yinlong Xu

***Abstract*—Understanding sign gestures is an essential step to helping individuals with hearing impaired. The existing works can only identify a small set of gestures accurately and the accuracy rate drops sharply with an increasing number of gestures. Because there are two challenges—a large number of similar gestures in sign language and the various signing speed of different people. Based on commercial smart bracelets, this article proposes a large-vocabulary sign language recognition system (which we call L-sign). First, we propose an entropy-based forward and backward matching algorithm to segment each gesture signal. Second, we design a gesture recognizer including a candidate gesture generator and semantic-based voter. The candidate gesture generator is aimed at providing candidate gesture designs based on a 3-branch convolutional neural network. The purpose of a semantic-based voter is to select the target gesture from candidate gestures by scoring, where the semantic distances between the last gesture in the current sentence and any candidate gestures is calculated, and a multilayer k-means algorithm is proposed to obtain a multilayer sign word structure to complete the scores of candidate gestures. Lastly, we deployed L-sign on the MYO bracelet. For 200 commonly used Chinese sign gestures, the experimental results show that the average accuracy rate was greater than 90%.**

***Index Terms*—Gesture, multilayer word structure, recognition, semantic distance, sign.**

I. Introduction

According to the latest sampling survey by the World Health Organization, more than 466 million people currently suffer from hearing loss as a result of a disability. The WHO projects this number could increase to more than 630 million by 2030 [1]. These hearing-impaired individuals suffer from communication barriers daily, which can cause their inability to study, enjoy their lives, and seek medical treatment in the same way as those with normal hearing [2]–[4]. It is urgent to help those with hearing impaired to communicate with people with normal hearing.

Individuals with hearing impaired communicate with each other using sign gestures, which are difficult for people with normal hearing to understand. (In this article, "gesture" means hand motions corresponding to a sign language word). Thus, we propose a large-vocabulary Chinese universal sign gesture recognition system (which we call L-sign) to help individuals with hearing impaired.

The current research on sign language recognition can be divided into two groups, according to the means of signal acquisition: visual-based group and sensors-based group. The former was first utilized in the works of sign language recognition. This method recognizes sign language by processing pictures or video recordings of signers [5], [6]. The signer stands in front of a camera and performs sign gestures. At first, this method utilized a front view camera and a head-mounted camera to record and collect the signer's sign gesture data [7]. Later on, this method evolved to deploy multifunction cameras. For example in the study [5] and [8], the depth camera is utilized to capture the 3-D information of the signer's hand. These visual-based methods can recognize about 40 sign gestures. In recent studies, some commercial devices such as Kinect [6], [9], [10] and LeapMotion Controller [11], [12] are utilized in sign language recognition. This method has characteristics of high average accuracy and low deployment cost. However, the utilization of camera equipment often violates the privacy of those with hearing impaired.

In sensor-based approaches, various sensors are fixed to the signer's finger, wrist, arm, and other parts for capturing information when the signer performs gestures. These sensors are portable, rich in data, and stable in signal transmission, including surface electromyography (sEMG) signal sensors [13]–[15], acceleration sensors [16], [17], and gyroscopes [18]. In [14], the sensor is placed on the signer's wrist. When the signer signs, the sensor detects the surface muscle and the current signal of the signer's skin to identify the sign language. For 40 most commonly utilized gestures, this system can achieve a 95.94% recognition rate. Recently, smartwatches [19] are also being utilized for sign language recognition. For example, SignSpeaker [19] is deployed on smartwatches together with smartphones to recognize 103 commonly used American sign gestures. In these approaches, the sensors can collect data stably because of the little impact caused by environmental factors,

Manuscript received July 14, 2021; revised November 4, 2021 and January 19, 2022; accepted January 20, 2022. Date of publication February 15, 2022; date of current version March 14, 2022. This work was supported by the National Natural Science Foundation of China under Grant 61571179 and Grant 91538112. This article was recommended by Associate Editor Vittorio Fuccella. *(Corresponding author: Qingshan Wang.)*

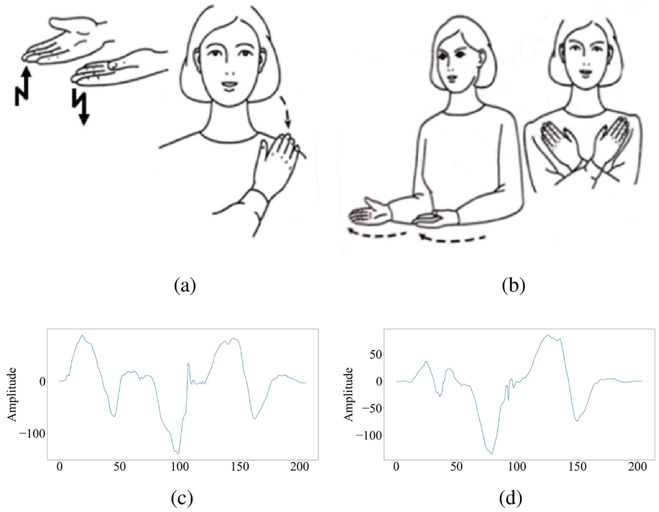
Zhiwen Zheng, Qingshan Wang, Qi Wang, and Wei Huang are with the School of Mathematics, Hefei University of Technology, Hefei 230601, China (e-mail: 2018111283@mail.hfut.edu.cn; qswang@hfut.edu.cn; wangq@hfut.edu.cn; whuang@hfut.edu.cn).

Dejun Yang is with the Department of Computer Science, Colorado School of Mines, Golden, CO 80401 USA (e-mail: djyang@mines.edu).

Yinlong Xu is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: ylxu@ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3146787.

Digital Object Identifier 10.1109/THMS.2022.3146787



(a)

(b)

(c)

(d)

Fig. 1. Similarity between the sign gestures for "Judge" and "Request." (a) Judge. (b) Request. (c) IMU signal of gesture "Judge". (d) IMU signal of gesture "Request".

which makes it suitable for large-scale sign language recognition. However, most of the existing sensor-based sign language recognition studies are based on small-scale gestures. When the scale is larger, the number of gestures with similar movements increases, and recognition difficulties increase sharply.

Hence, in this article, the main focus of our study is accurately recognizing large-scale gestures. We propose the L-sign, which deploys a bracelet with built-in sensors (which is the MYO bracelet in this article). The Chinese universal sign language is a sign language widely used by Chinese people with hearing impaired in daily communication [10]. The latest version of the Chinese universal sign language has 8223 unique sign gestures comprised of different shapes, the directions of movement, and orientation. The signer wears a bracelet to perform sign gestures on the dominant hand (usually the right hand), the bracelet collects sign gesture signals to identify the sign gestures. To achieve accurate recognition of the large-scale Chinese universal sign language, the following two challenges must be addressed. 1) There are similarities between gestures, and the number of similar gestures increases with the scale of gestures. For example, as shown in Fig. 1(a) and (b), the sign gestures of "Judge" and "Request" have similar hand movements, and both of them can be divided into two parts. Their first part is both hands outstretched and palms up, and the difference is that the former's hand movements include shaking the palm up and down, while the latter does not. And their second part is hands back. The movements of latter include the left hand resting on the right shoulder and the right hand resting on the left shoulder, while the movements of former only includes the right hand resting on the left shoulder. Fig. 1(c) and (d) shows the inertial measurement unit (which we call IMU) signals corresponding to these two gestures. It can be seen that the IMU signals corresponding to these two gestures are similar in shape because of their similar actions. It is challenging to construct an appropriate model to recognize gestures. 2) Different people signing different gestures at various speeds results in high levels of variation, making it difficult to accurately extract the gestural signals from the collected signals.

To distinguish similar gestures, we designed a gesture recognizer that combines the candidate-gesture generator and the semantic-based voter to obtain the final recognition results gesture-by-gesture. The candidate gesture generator is based on a 3-branch convolutional neural network aimed at providing candidate gestures. The semantic-based voter is aimed at selecting the target gesture by scoring the candidate gestures.

To solve the challenge of different individuals signing speeds, we conducted a lot of experiments. When the user makes gestures, we find that the sEMG signal peaks and the IMU signal fluctuates frequently. In the nongesture period, the peak value of the sEMG signal is low, and the fluctuation of the IMU signal tends to be flat. On this basis, a signal entropy-based forward and backward matching algorithm was proposed and accurately separates the gesture signal from the nongesture signal.

The main contributions of this article are summarized as follows.

1) We propose the signal entropy-based forward and backward matching algorithm to adaptively find the corresponding signal of each gesture.
2) We design a gesture recognizer that comprises a candidate gesture generator and semantic-based voter to select the target gesture. The candidate gesture generator provides candidate gestures, and the semantic-based voter scores the candidate gestures by using the semantic relationship of sign gestures.
3) We utilize an MYO bracelet for acquiring sign gesture data and evaluate the performance of the L-sign. We expanded L-sign's recognition database on 200 common Chinese sign gestures. Experimental results show that the accuracy of L-sign is above 90% across different users in different experimental environments.

The rest of this article is organized as follows. Section II introduces the existing related work. Section III is an overview of the L-sign. Section IV introduces the gesture signal acquisition and the signal entropy-based forward and backward matching algorithm in detail. Section V proposes the L-sign system and Section VI presents the performance evaluation. Finally, Section VII concludes this article.

II. Related Works

The existing human motion recognition studies can be mainly categorized into two classes: *visual-based* and *sensors-based*.

A. *Visual-Based Works*

Computer vision [5]–[8], [10], [12], [20]–[24] is a convenient way to recognize human movements. These approaches require the utilization of cameras or other noninvasive sensors to record images of user actions. Early visual-based motion recognition research utilized cameras. Starner *et al.* [7] utilized a front view camera and a head-mounted camera to track the users' hand movements, and can accurately recognize continuous American sign language composed of 40 sign gestures. Also using the camera for motion recognition, Feriset *et al.* [20] utilized a multiflash camera to take flash photos at multiple different locations to capture the depth information generated during the action to identify fingerspelling. Norooziet *et al.* [21] proposed human pose detection and dynamic human pose estimation methods based on RGB and 3-D images. Following the development of camera equipment, some people gradually utilized advanced camera equipment for research. Shotton *et al.* [8] proposed a method that can quickly and accurately utilize single-depth images to predict the 3-D position of human joints, and design an intermediate model of human body composition to transform difficult posture statistic problems into simpler pixel classification problems. Azari *et al.* [25] created a 2-D linear and generalized additive model (GAM) of video-recorded hand movements to predict expert-evaluated performance on a series of surgical movement scales. Recent visual-based works also employed commercial devices. For example, Kinect [9] and LeapMotion Controller [11] are the most commonly employed devices in visual-based works. For example, Hamda *et al.* [22] proposed a comparative study to recognize six hand gestures in real-time using the Kinect sensor. Potter *et al.* [12] presented an early exploration of the suitability of the leap motion controller for Australian sign language (ASL) recognition. With the reliability of its camera array, Kinect is also utilized to recognize Chinese universal sign language. Zhang *et al.* [10] recognized Chinese universal sign language using an adaptive hidden Markov model with self-built datasets based on Kinect. Gaglio *et al.* [26] presented a method for recognizing human activities using Kinect, and combined three different machine learning techniques, namely K-means clustering, support vector machines, and hidden Markov models, to detect and classify the postures involved in performing activities. Although the above computer-vision methods achieved high accuracy in human motion recognition, the use of devices may result in the user's privacy suffering from invasion.

B. Sensors-Based Works

Sensors-based studies utilize sensors to obtain the characteristics of individuals' activities. In this kind of work, common sensor types are sEMG sensor, acceleration sensor, and gyroscope, for example, [13]–[15], [27]–[35], which can be utilized to identify human movements. Kang *et al.* [13] proposed a new gesture recognition system on the premise of limiting the number of electromyogram signal (EMG) sensors, in which three signal channels could classify nine simple gestures. Wu *et al.* [14] combined the sEMG sensor and the acceleration sensor to design the American sign language recognition (ASLR) system to realize smooth communication between the hearing impaired and the able-bodied, and the recognition rate of the ASLR system is 95.94% for 40 common gestures. He *et al.* [15] compared the performance of single-channel ultrasonic wave and sEMG signals, and demonstrated their apparent complementary advantages. Research studies [18], [36], and[37] utilized acceleration sensors and multilayer perceptron classifiers to identify human activities. Kuroda *et al.* [38] utilized 24 inductors and 9 contact sensors to manufacture their device StrinGlove. Some recent works [18], [36], [37] can differentiate finger-level gestures using inertial sensors on wearables. Wang *et al.* [39] proposed a digital glove based on the ASL recognition system developed by the multidimensional hidden Markov model. All of the above studies utilize special data gloves or wrist-worn sensors. In recent years, there are studies [19], [40], [41] that utilize existing smart devices, such as smartwatches to recognize sign gestures. SignSpeaker [19] is an ASLR system that is deployed on smartwatches together with smartphones. Generally speaking, the sensors-based method for gesture recognition has the characteristics of portability, low cost, and stable signal transmission, which is suitable for large-scale Chinese common sign gesture recognition. In this article, an MYO bracelet integrated with sEMG sensors and IMU sensors is adopted for gesture acquisition and recognition.

III. System Overview

A. Data Collection.

For data collection, we collected sEMG and IMU signals from the MYO bracelet as shown in the blue box in Fig. 2. The L-sign utilizes a sensor bracelet as its sign language signal acquisition device. The bracelet includes sensors, batteries, and Bluetooth modules. The sensor system consists of an accelerometer, a gyroscope, and eight sEMG signal sensors.

The signal collected by the MYO bracelet is an 18-D vector: an 8-D sEMG signal, a 4-D gyroscope signal, and a 6-D acceleration signal. According to the characteristics of the collected signals, we divide the 18-D signals into two categories. The first type is an 8-D sEMG signal. The sEMG signal is the comprehensive effect of the superficial muscle and nerve trunk electrical activity on the skin surface, which can reflect the neuromuscular activity to a certain extent. This kind of signal has certain typical characteristics. When the muscle tested by the sEMG signal sensor is under mild load, there are isolated single low-amplitude movement unit potentials at certain intervals and frequencies. This is a pure phase. On the contrary, when the muscle is under heavy load, there is high amplitude potential with different frequencies and amplitude, and it is difficult to distinguish the difference and overlap. This is the interference phase. Therefore, the system can look for sEMG signal frequency, amplitude, and other characteristics to describe the gesture changes. The second category is IMU signal [17] which includes: 4-D gyroscope signals, 6-D X/Y/Z axis acceleration signals, and angular velocity signals (where X/Y/Z axis are static relative to the MYO bracelet). In the MYO bracelet, the sampling frequency of the sEMG signal is 200 Hz, and the sampling frequency of the IMU signal is 50 Hz.

We consulted sign language experts from special education schools and concluded that most people can complete any Chinese common sign gesture within 4 s. Thus, we set the acquisition time of a gesture to 4 s. A gesture data file contains 800 lines of sEMG data and 200 lines of IMU data.

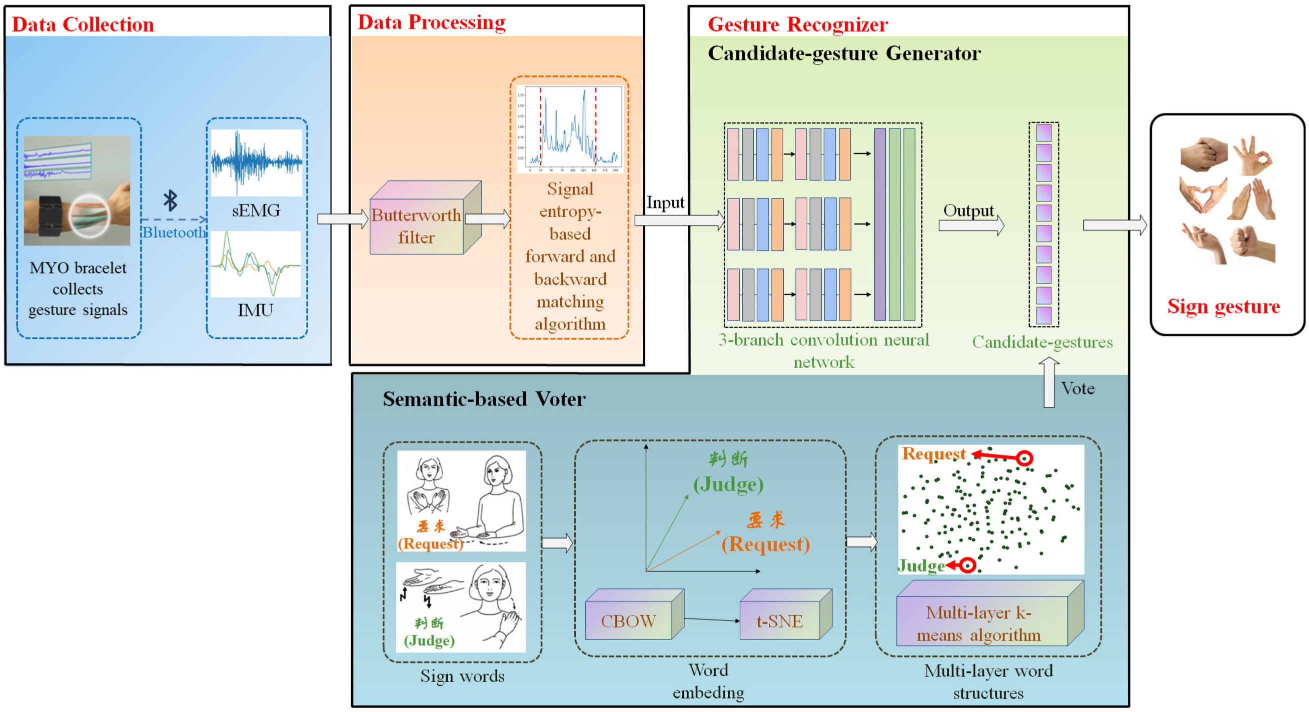

Data Collection
MYO bracelet collects gesture signals
Bluetooth
sEMG
IMU
Data Processing
Butterworth filter
Signal entropy-based forward and backward matching algorithm
Input
Gesture Recognizer
Candidate-gesture Generator
3-branch convolution neural network
Output
Candidate-gestures
Vote
Sign gesture
Semantic-based Voter
要求
(Request)
判断
(Judge)
Sign words
CBOW
t-SNE
Word embeding
Request
Judge
Multi-layer k-means algorithm
Multi-layer word structures

Fig. 2. L-sign system overview.

B. Data Processing

First, we apply Butterworth filter processing on the collected gesture signals to remove the high-frequency noise generated by the equipment and wireless transmission. Then, we propose the signal entropy-based forward and backward matching algorithm to segment the processed gesture signal and extract the effective gesture signal. The data processing is in Section IV.

C. Gesture Recognition

At this stage, we designed the gesture recognizer including the semantic-based voter and the candidate-gesture generator to recognize the target gesture. In Section V, we present the gesture recognizer.

IV. Data Processing

This section finds that the signal entropy of the existing collected data significantly changes when the hearing impaired performs sign gestures or not. Based on these changes, we propose a signal entropy-based forward and backward matching algorithm to find the corresponding signal of each sign gesture adaptively. The processing of the sEMG signal and IMU signal collected by the bracelet mainly includes denoising and determining the start and end positions of the gesture signal.

First, we utilize a Butterworth filter [42] to filter the collected signal. The signal collection may be affected by environmental factors, such as thermal noise generated by the electronic components of the acquisition device. Therefore, the collected signal needs to be filtered. Given that the frequency of background noise is much higher than that of human gestures, we perform a Butterworth filtering operation on the collected signal to remove the signal's high-frequency part.

Then, the execution time for the same gesture varies among different signing individuals. Moreover, the same hearing-impaired user has different execution times for different gestures. Therefore, we need to extract the signal part corresponding to the sign gesture from the overall signal of a sign gesture collected. At present, most of the cutting algorithms [43]–[46] adopt the sliding window method to judge the segmentation point, which have a better recognition effect for the signal with less severe fluctuation. In this article, aiming at the scene where the signal contains singular values, we extracted the signal entropy feature of the signal to reduce the influence of singular value on the selection of segmentation points and improve the accuracy of the cutting algorithm. We find that the arm and finger muscles of signers were tense when signers perform gestures, which is reflected in the peak of sEMG signal and the frequent fluctuation of IMU signal. On the contrary, when no sign gesture is performed, both the sEMG signal peak value and IMU signal fluctuation tend to be low. Then let us take the IMU signal as an example. The signal we collected is $A = \{a_{k,t} | 1 \le k \le 10, 1 \le t \le 200\}$. In this equation, t represents the packet number, k represents the dimension of gesture data, $a_{k,t}$ represents the amplitude of the tth packet in the kth dimension. We calculate the contribution $P(a_{k,t})$ of signal amplitude $a_{k,t}$

$$P(a_{k,t}) = \frac{a_{k,t}}{\sum_{t=1}^{200} a_{k,t}}. \tag{1}$$

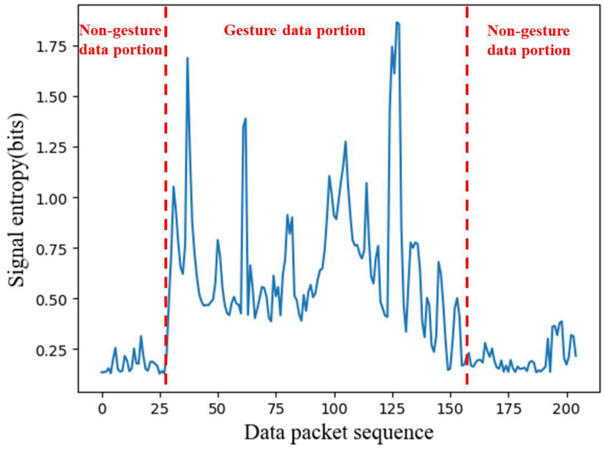


Fig. 3. Signal entropy of gesture "Handle."

We define the signal entropy of $a_{k,t}$ as

$$H(a_{k,t}) = \log_2\left(\frac{1}{P(a_{k,t})}\right). \tag{2}$$

We calculate the average of the signal entropy of $a_{k,t}$, $H(a_{k,t})$ in (1), for dimension k, and obtain the average entropy value of the signal in time series of tth packet

$$h(t) = \mathbf{E}[H(a_{k,t})] = \frac{1}{10}\sum_{k=1}^{10} H(a_{k,t}). \tag{3}$$

Fig. 3 shows the IMU signal entropy sequence of the gesture "Handle". It can be observed that the signal entropy corresponding to the gesture "Handle" is high, while the signal entropy corresponding to no gesture is low.

The signal entropy-based forward and backward matching algorithm is shown in Algorithm 1. The terms "forward" and "backward" in the algorithm refer to traversing the generated signal entropy sequence, which is described in (3). They determine the starting point and ending point of the gesture signal according to the variance's difference of the sequence, respectively. Specifically, the algorithm flow of Algorithm 1 is as follows. The input of Algorithm 1 is sensor signal s and threshold δ for signal clipping, and the output from Algorithm 1 is the sensor signal s after clipping and stretching.

First, we calculate the signal entropy described in (3), which is called h. And we calculate the standard deviation σ between two adjacent variables of the signal entropy sequence h. For example, for two adjacent variables $h[t+1]$ and $h[t]$, the variance $\sigma[t]$ is expressed as follows:

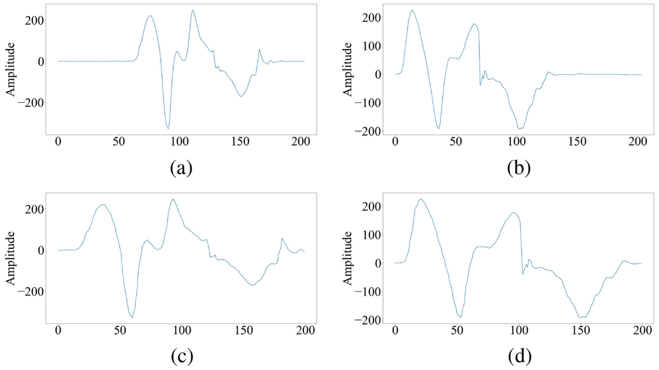
$$\sigma[t] = \sqrt{h[t+1]^2 - h[t]^2} \tag{4}$$

where $0 \leq t \leq \text{len}(h) - 1$, and $\text{len}(h)$ represents the length of the h. Among that, $\sigma[0]$ is initialized to $h[0]$. We assign 0 to θ, which is utilized to control the number of cycles. These can be seen in lines 1–3.

Then, in lines 4–14, we cut the signal by controlling two cycles. The first cycle is performed to determine the starting point of the signal when $\theta = 0$. Specifically, we clip the signal sequence h according to the difference of σ sequence. For two adjacent variables $\sigma[t+1]$ and $\sigma[t]$, if $\sigma[t+1] - \sigma[t]$ greater than δ, we set this location as the starting point of the data, and cut the data in this point. After that, we inverse the data, operate the cycle again, obtain the data endpoint, and cut the data at that point.

Algorithm 1: Signal Entropy-Based Forward and Backward Matching Algorithm.

```
 Input: $s$, $\delta$
 Output: $s$
 1:  Calculate the signal entropy of $s$, and obtain the signal
     entropy called $h$;
 2:  Calculate the standard deviation $\sigma$ of $h$;
 3:  $\theta \leftarrow 0$;
 4:  while $\theta \leq 1$ do
 5:    $t \leftarrow 1$;
 6:    while $\sigma[t+1] - \sigma[t] \leq \delta$ do
 7:      $t \leftarrow t + 1$;
 8:    if $i \leq \text{len}(h)$ then
 9:      $s \leftarrow s[t : \text{len}(s)]$;
10:    if $\theta = 0$ then
11:      $s \leftarrow s$.reverse();
12:    $\theta \leftarrow \theta + 1$;
13:  Pull the cut data up to the original length;
14:  return $s$
```



(a)

(b)

(c)

(d)

Fig. 4. Effect of Algorithm 1 on the IMU signal of gesture "Handle." (a) The raw IMU signal of Volunteer 1. (b) The raw IMU signal of Volunteer 2. (c) The IMU signal processed by Alg. 1 of Volunteer 1. (d) The IMU signal processed by Alg. 1 of Volunteer 2.

Last, in line 15, we stretch the cut data to the original length. We interpolate the updated s with the ratio of difference $\frac{\text{len}(h)-\text{len}(s)}{\text{len}(h)}$, where $\text{len}(h)$ represent the length of raw data.

Through this algorithm, we can cut and stretch the data, and solve the challenge that different people signing different gestures at various speeds results in high levels of variation. Fig. 4 shows the effect of Algorithm 1 on the IMU signal of the gesture "Handle." Fig. 4(a) and (b) shows raw IMU signals corresponding to sign gesture "Handle" performed by two different volunteers. It can be seen that there are some nongesture signals at both ends of the signal, and the IMU signals corresponding to the two volunteers are different due to their speeds. Fig. 4(c) and (d) shows IMU signals after processing by Algorithm 1 corresponding to Fig. 4(a) and (b). It can be seen that the IMU signals are similar. Thus, Fig. 4 illustrates that gesture signals can be extracted from the collected signals by Algorithm 1.

V. Gesture Recognizer

This section designs a gesture recognizer including a candidate-gesture generator and a semantic-based voter, which are utilized to recognize a large volume of sign gestures. We design a gesture recognizer including a candidate gesture generator and semantic-based voter.

A. *Word Vector Representation of Sign Language*

In this section, the t-SNE (t-distributed stochastic neighbor embedding) algorithm is utilized to convert sign words into vectors according to their semantics. Based on the frequency of sign language used by those with hearing impaired in China, we chose 200 commonly used sign words from the Chinese universal sign gesture corpus.

First, we apply the CBOW model to embed these sign words. The CBOW (Continuous Bag-of-Words) model [47]–[49] utilizes the context of a word to predict the word. The noise contrastive estimation (which we call NCE) loss function [50] can improve the training efficiency of the model by transforming the classification problem into a general logistic regression task. And the NCE loss function is often used in the training of the CBOW model. In the training process of the CBOW model, a noise dataset is constructed, and we train the CBOW model to learn the difference between real data and noise data by utilizing the NCE loss function. Thus, the semantic features of each word in the corpus will be learned by the CBOW model. By training the CBOW model, we can store the semantic information of every word in the whole corpus in the form of word-embedded representations (called word vectors). In our work, we train the CBOW model to obtain the word vectors of these 200 sign words.

The CBOW model has three layers in total, namely, the input layer, the hidden layer, and the output layer. Since the CBOW model is used to predict a word from its context, the final output of the CBOW model is the probability of the current word. After training, the values of the hidden layer in the model are regarded as the word embedding representations of these words. Let V represents these 200 word vectors, which is given by

$$V = \{v_i | 1 \leq i \leq 200\} \tag{5}$$

where $v_i = (v_{i1}, v_{i2}, \ldots, v_{i200})$ is a 200-D vector corresponding to a sign word.

We then utilize the t-SNE algorithm to reduce the dimension of the sign words. Compared with other dimensionality reduction algorithms, the t-SNE algorithm [51] can effectively solve the problem of crowding of data points after dimensionality reduction according to the characteristics of t-distribution [52].

In this article, these 200 sign word vectors are arranged from high to low according to their usage frequency. We add a weight ϑ_i to each word vector v_i and define it as follows:

$$\vartheta_i = \frac{p_i}{\sum_{i=1}^{200} p_i} \tag{6}$$

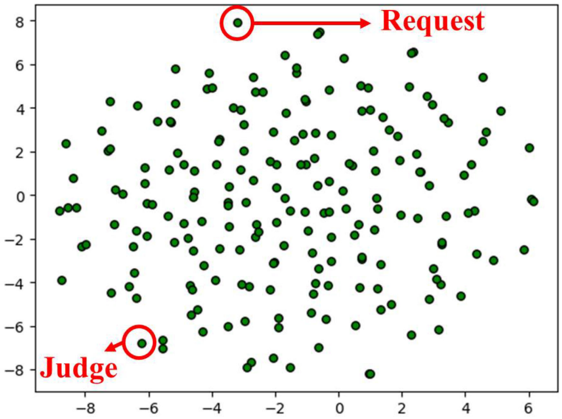


Fig. 5. Sign words vector distribution.

where p_i is the usage frequency of each sign word in the corpus [53], [54]. We then utilize Kullback–Leibler divergence (which we call KL distance) as our loss function for optimization and utilize the stochastic gradient descent algorithm to train the t-SNE model. The word vector v_i can be reduced dimension as $v'_i = (v'_{i1}, v'_{i2})$.

As shown in Fig. 5, it is the result of the t-SNE reduction to 2 dimensions for the word vectors of 200 commonly used sign gestures.

The words after dimension reduction constitute a semantic space, where each point represents a sign gesture. As can be seen from Fig. 5, the two sign gestures "judge" and "request," which originally had very similar gestures, are now located in two distant locations in Fig. 5. Thus, we achieve the effect of distinction.

B. *Clustering Analysis With Multilayer K-Means Algorithm*

To obtain a clustering with multilayer word structure, we propose a multilayer k-means algorithm to perform clustering analysis on the word vectors after dimensionality reduction.

We first define the semantic distance between sign words to express the semantic relationship between sign words. For any two word vectors v'_i and v'_j after dimensionality reduction in Section V-B, their semantic distance $d(v'_i, v'_j)$ is defined as follows:

$$d(v'_i, v'_j) = \sqrt{(v'_{i1} - v'_{j1})^2 + (v'_{i2} - v'_{j2})^2} \tag{7}$$

where $v'_i = (v'_{i1}, v'_{i2})$ and $v'_j = (v'_{j1}, v'_{j2})$ $(1 \leq i, j \leq 200)$. The closer the two points are, the smaller their semantic distance, which indicates that their semantic relationship is closer and vice versa. As shown in Fig. 5, the semantic distance between the two similar gestures "judge" and "request" is 25.346, which means their semantics are greatly different from one another.

Moreover, we present the multilayer k-means algorithm, whose main idea is that running d times k-means algorithm which is utilized to each class obtained from the last times on the word vectors. Thus, a new class number for each word vector in each time is appended. And Fig. 6 shows the flow chart of the multilayer k-means algorithm. The specific steps as follows: In

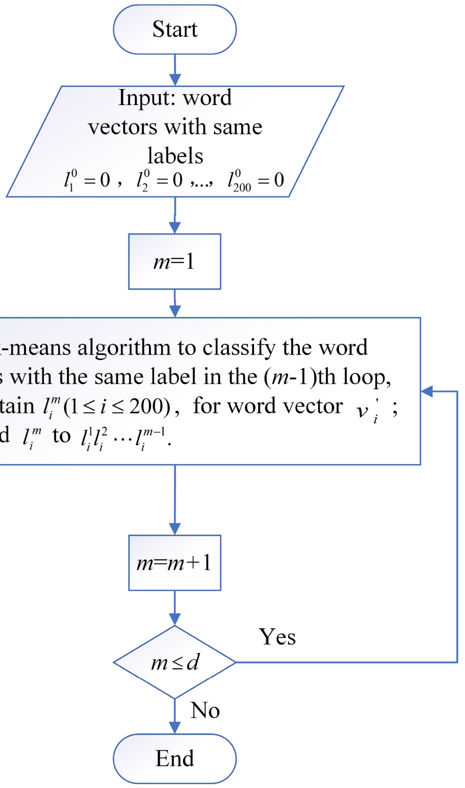

Start
Input: word vectors with same labels $l_1^0=0$, $l_2^0=0$,…, $l_{200}^0=0$
m=1
Used k-means algorithm to classify the word vectors with the same label in the (m-1)th loop, and obtain $l_i^m(1\le i\le 200)$, for word vector v_i' ; Append l_i^m to $l_i^1 l_i^2\cdots l_i^{m-1}$.
m=m+1
$m\le d$
Yes
No
End

Fig. 6. Flowchart of multilayer k-means algorithm.

step (1), we initialize the label of word vector v_i' $(1 \leq i \leq 200)$ as $l_i^0 = 0$. In step (2), we utilize k-means algorithm to classify these word vectors with the same label in the $(m-1)$th loop, and obtain class number l_i^m for word vector v_i'. At same time, we append l_i^m to $l_i^1 l_i^2 \ldots l_i^{m-1}$, thus v_i''s new label is $l_i^1 l_i^2 \ldots l_i^m$. In step (3), we compare the current number of cycles m and d to determine whether to terminate the cycle.

After the multilayer k-means algorithm, any sign word vector v_i' is attached with label $l_i^1 \ldots l_i^d$, and the d-layer sign word structure can be obtained. We determine the value of d according to the number of sign words. d is set to be 1 when the number of sign words is small, for example, 100; d is set to be 2 when the number of sign words is large, for example, the number is more than 100 and less than 300. We select $d = 2$, $k = 4$ in this article, and Fig. 7 shows the 2-layer word structure generated by multilayer k-means algorithm. As shown in Fig. 7, 200 sign words are divided into 4 classes by the blue curve in the first cycle, and each classes is divided into 4 smaller classes by the red dotted line in the second cycle.

C. Gesture Recognizer Including a Candidate-Gesture Generator and Semantic-Based Voter

This section designs a gesture recognizer including a candidate-gesture generator and semantic-based voter to recognize sign gestures.

1) Candidate-Gesture Generator: We construct a 3-branches 1-D convolution neural network (CNN) to construct the framework of the candidate-gesture generator, as shown in

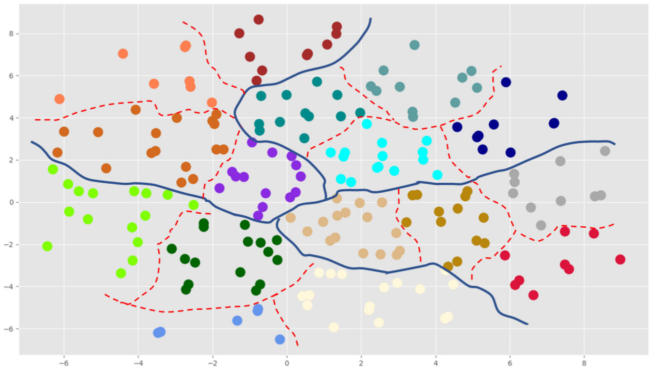


Fig. 7. Classification with two-layer word structure.

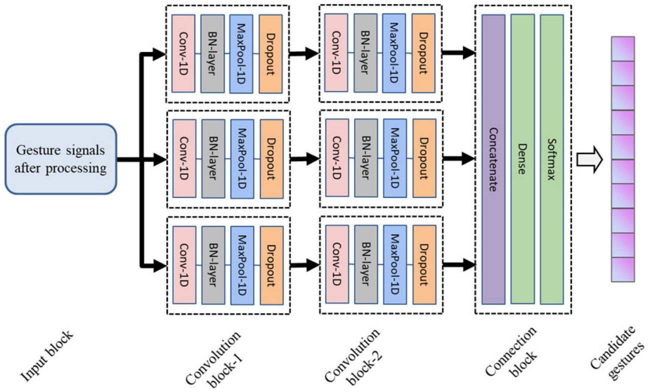


Fig. 8. Candidate-gesture generator.

Fig. 8. Compared with the LSTM neural network, the 1-D CNN can accurately extract the local information of the signals using fewer parameters. The inputs of the three branches are sEMG signal, gyroscope quaternion, and X, Y, Z three-axis velocity related signals (X, Y, Z three-axis acceleration and X, Y, Z three-axis angular velocity) after filtered and cut. For each sign gesture, the sizes of them are 800×8, 200×3, and 200×7, respectively. For these three branches, the 1-D convolution kernel sizes are 1×8, 1×3, and 1×7, respectively. Through 1-D convolution operation, we can reshape signals with different sizes to the same size. Lastly, the candidate-gesture generator provides the M candidate gestures $y_1, \ldots, y_M$ obtained from the softmax layer, and their probabilities are $p_1, \ldots, p_M$.

2) Semantic-Based Voter: Let x be the last gesture of the current sentence, which has been recognized by the gesture recognizer. First, it calculate the semantic distance $d(y_j, x)(1 \leq j \leq M)$ between candidate gesture y_j with previous gesture x based on (7). Then, it scores gesture y_j based on previous gesture x's label $l_x^1 l_x^2$ and candidate gesture y_j's label $l_{y_j}^1 l_{y_j}^2$ as follows:

$$\text{score}(y_j) = \begin{cases} p_j + \lambda d(y_j, x) & l_{y_j}^1 = l_x^1, l_{y_j}^1 = l_x^2 \\ p_j + \lambda(d(y_j, x) + \alpha) & l_{y_j}^1 = l_x^1, l_{y_j}^1 \neq l_x^2 \\ p_j + \lambda(d(y_j, x) + \alpha + \beta) & l_{y_j}^1 \neq l_x^1 \end{cases} \tag{8}$$

where $\lambda(\lambda < 0)$, $\alpha(\alpha > 0)$ and $\beta(\beta > 0)$ are all empirical values, and we obtain their values by grid search in the experiment.

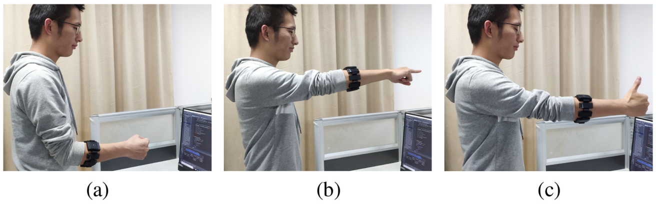


Fig. 9. Performance of the sign gesture for "Hello." (a) Initial state. (b) Action 1. (c) Action 2.

Lastly, it selects the target gesture corresponding to the highest score among score$(y_j)(1 \leq j \leq M)$. And it should be noted that we need to retrain the semantic-based voter for these newly selected sign words when other sign words are selected.

VI. Performance Evaluation

This section evaluates the performance of L-sign system proposed in this article. This experiment is realized using an MYO bracelet and a PC. As a signal acquisition device, the bracelet collects gesture signals at the speed of 100 data packets/second and transmits them to a PC. The PC works as the receiver end of the signal, and carries out signal processing and sign gesture recognition processes. The PC is equipped with an Intel Core i9-10900 K processor, 64 GB of memory, an Nvidia GeForce RTX 3090 graphics card, 24 GB graphics memory, and a preinstalled Ubuntu 18.04 operating system. The three parameters λ, α, and β are empirical values obtained by the grid search method in the experiment. Here, $\lambda = -0.001$, $\alpha = 0.2$ and $\beta = 0.1$. In the training process, batch size of the candidate-gesture generator is set as 512. Adam optimizer is used to optimize the network parameters. The learning rate is 0.001 and 1000 epochs are used to train.

In this experiment, we have 12 volunteers between 18 and 40 years old, which include 6 university students (3 males and 3 females), 2 hearing-impaired people (1 male and 1 female), and 4 Chinese sign language experts (2 males and 2 females). Before the experiments, we require them to learn all the selected 200 Chinese universal sign gestures used in this experiment, and they can accurately execute them through 8 hours of gesture training. Each volunteer performs each sign gesture 20 times and a total of 48 000 samples are collected.

In this article, the volunteers are right-handed. And we stipulated the correct wearing posture of the bracelet. The bracelet was worn at the uppermost third of the right forearm, and the logo of the bracelet was facing the middle finger. In Fig. 9, a volunteer performs the sign gesture for "Hello." Fig. 9(a) is the initial state, and Fig. 9(b) and (c) shows the two continuous actions including in the sign gesture "Hello".

A. Comparison With Existing Methods

1) Compare With Baseline: In this section, we compare the recognition performance of our system with an LSTM neural network [38], [55], [56] and a convolutional 3-D model (called C3D) [57]–[60], and test their recognition performance with different sign vocabulary sizes. We perform leave-one-volunteer-out cross-validation to validate their capacity. The training is done on the gesture samples of 11 volunteers from the 12 volunteers and the rest one is left out for testing. For different vocabulary sizes of sign gesture, the above is repeated 12 times for calculating the average accuracy while changing the test part one-by-one until testing has been done on all the datasets. The datasets are processed by Butterworth filter, then cut and stretched in the proposed Algorithm 1. Specifically, we remove semantic-based voter in the L-sign and replaced the 3-branches 1-D convolutional neural network with an LSTM neural network and C3D in the candidate-gesture generator, respectively. The network structures and the inputs of LSTM neural network and C3D are as follows.

In LSTM neural network [56], the number of neural network layers and neurons in each layer of LSTM are 3 and 512, respectively. For the requirement in LSTM neural network, the sEMG signals and IMU signals need to be fused and reshaped to the same size as follows. Since the sampling frequency of the sEMG signals is 4 times that of the IMU signals, we perform interpolation processing on the latter and scale it to the same length as the former. For every two IMU signals x_t and x_{t+1} $(1 \leq t \leq 200)$, we insert 3 frames of IMU signals a_{tk}, a_{tl}, a_{tm} which are given by $a_{tk} = 0.75x_t + 0.25x_{t+1}$, $a_{tl} = 0.5x_t + 0.5x_{t+1}$, and $a_{tm} = 0.25x_t + 0.75x_{t+1}$, respectively. Thus, the interpolated IMU signals are obtained as $\ldots, x_t, a_{tk}, a_{tl}, a_{tm}, x_{t+1}, \ldots$. We then splice the sEMG signals and the IMU signals to the size of 800×18 according to their minimum dimension.

With reference to works [58], [60], we design the structure of the C3D. There are five layers which include 3-D convolutional layer, 3-D pooling layer, fatten layer, and dense layer in C3D. The sizes of 3-D convolutional kernels and 3-D pooling are $2 \times 2 \times 1$ and $2 \times 1 \times 1$, respectively. For sEMG signals, we combine 32 data packets into one and add the dimension of height. Thus, we can obtain the sEMG signals with size of $200 \times 1 \times 1 \times 32$. For IMU signals, in order to align them with the dimension of sEMG signals, we add the dimension of height to them and obtain the IMU signals with size of $200 \times 1 \times 1 \times 10$. Moreover, we fuse the sEMG signals and the IMU signals together and obtain the sign gesture signals with size of $200 \times 1 \times 1 \times 42$ as the inputting signals of C3D.

The experimental results shown in Fig. 10 illustrate that L-sign has the best recognition performance among the three methods. The average accuracies of the LSTM neural network and the C3D decline obviously when the number of sign gestures exceeds 80, and the average accuracy of LSTM neural network decreases faster than that of C3D. When the number of sign gestures reaches 200, the LSTM neural network and the C3D are unable to perform gesture recognition while the average accuracy of L-sign still remains at 90.05%. The reason is that the L-sign considers both the sign gesture's signal and semantics, which is effective for recognizing sign gestures with similar actions in large vocabulary sign gestures.

2) Compare With State of Art: In this section, a comparative experiment with the SignSpeaker [19] is conducted to evaluate

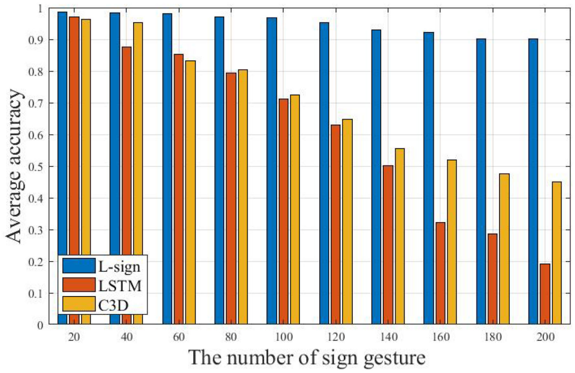


Fig. 10. Comparison with baseline.

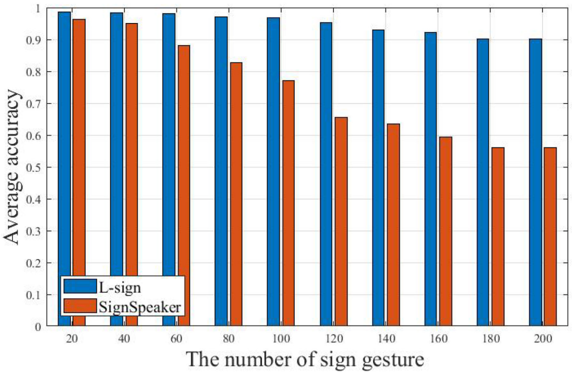


Fig. 11. Comparison with SOTA.

the recognition performance of L-sign. The leave-one-volunteer-out cross-validation method mentioned above is adopted in this experiment.

The experimental results shown in Fig. 11 illustrates that the SignSpeaker and the L-sign show good recognition performance and the recognition rate of SignSpeaker is slightly lower than that of L-sign when the vocabulary of sign gestures is small. This experimental result indicates that the L-sign is more suitable for Chinese sign gestures recognition. When the vocabulary of sign gestures increases, the recognition performance of SignSpeaker decreases while that of L-sign maintains a high level. The L-sign improves the average accuracy by 34% over the SignSpeaker. The experimental results show that L-sign is very effective for large-vocabulary Chinese sign gestures.

3) Compare With the Signal Cutting Algorithm in State of Art: To verify the effect of the signal entropy-based forward and backward matching algorithm, we conduct a comparative experiment with the signal segmentation algorithm in the state of art. Zhang *et al.* [43] proposed the signal segmentation algorithm based on moving averaged energy stream (called EMA) to segment signals in their works. We replace the signal entropy-based forward and backward matching algorithm in L-sign with Zhang's algorithm and compare their results to verify the effectiveness of our algorithm. This experiment is implemented on the collected dataset of 200 sign gestures. The average accuracy is obtained with leave-one-volunteer-out cross-validation method.

TABLE I
COMPARE WITH SIGNAL CUTTING ALGORITHM IN STATE OF ART

Alg. 1 in L-sign	Zhang's algorithm
90.05%	82.54%

TABLE II
RECOGNITION RESULTS OF SIMILAR GESTURES

Sign gestures	L-sign	LSTM
Tourism	90.01%	65.21%
Judge	90.32%	21.45%
Everybody	91.04%	20.36%
Doctor	90.15%	18.65%
Morning	90.00%	23.64%
Afternoon	90.14%	24.01%
Hello	91.43%	20.12%

The experimental results are shown in Table I. As can be seen from Table I, our algorithm is more conducive than Zhang's algorithm to improve the recognition performance of L-sign.

The reason has two points. On the one hand, the signal entropy can reduce the influence of singular value on the cutting points selection. On the other hand, our algorithm correctly determines the starting point and ending point of the signal from the forward direction and the reverse direction, respectively.

B. Recognition Reliability

To verify the reliability of the L-sign system, we randomly select the sample of 7 sign gestures from the collected 200 sign gesture dataset, which were poorly recognized by the LSTM neural network, and then utilize the L-sign for recognition purposes.

The experimental results are shown in Table II, and we can see that the average accuracy of L-sign remains over 90%, while the accuracy of the LSTM neural network at most 65.21%.

The reason is that the proposed neural network of L-sign is based on sign gestures themselves, as well as their semantic structures. Therefore, the experimental results show that our model can greatly improve the low average accuracy of the existing methods utilized in similar sign gestures.

C. Ablation Comparison Experiment

To verify the effectiveness of signal segmented of Algorithm 1 and semantic-based voter of L-sign, we conduct the ablation comparison experiment based on the collected dataset of 200 sign gestures. The L-sign system is trained with the leave-one-volunteer-out cross-validation method mentioned in VI-A1. Thus, there are 12 tests for 12 volunteers.

1) Without Signal Segmented: To verify the effectiveness of Algorithm 1, we input the raw signal without segment and the segmented signal by Algorithm 1 to the L-sign for training and recognition, respectively. Algorithm 1 can accurately remove the signals corresponding to nongesture actions at both ends of the data, then shrink the gesture signals to the same length.

TABLE III
EXPERIMENTS ON THE EFFECTIVENESS OF SEGMENTATION

Signal after segmentation	raw signal
90.05%	50.23%

TABLE IV
EXPERIMENTS ON THE EFFECTIVENESS OF SEMANTIC-BASED VOTER

L-sign	Without semantic-based voter
90.05%	85.26%

The recognition results shown in Table III illustrate that the accuracy of using the raw signal is lower apparently than that of using the segmented signal in Section IV, which illustrates that the proposed Algorithm 1 is useful for improving the recognition performance of gesture recognizer. The reason is that the sign gesture signals extracted from the collected signals by Algorithm 1 are of great help to improve the recognition performance of L-sign.

2) Without Semantic-Based Voter: To explore the influence of the semantic-based voter module on gesture recognizer, we dismantle the semantic-based voter module and only utilize the candidate-gesture generator model. In test step, we choose the sign gesture with the highest corresponding probability as the output of L-sign after the predicted probability of each sign gesture is generated by candidate-gesture generator.

The experimental results are shown in Table IV. From Table IV, we can find that the semantic-based voter can improve the recognition performance of L-sign. We believe that the semantic information of sign gestures can be introduced by semantic-based voter to assist L-sign for target judging.

D. Independency Judgment

In this section, we mainly evaluate the independence of L-sign in terms of the different individuals and experimental scenarios.

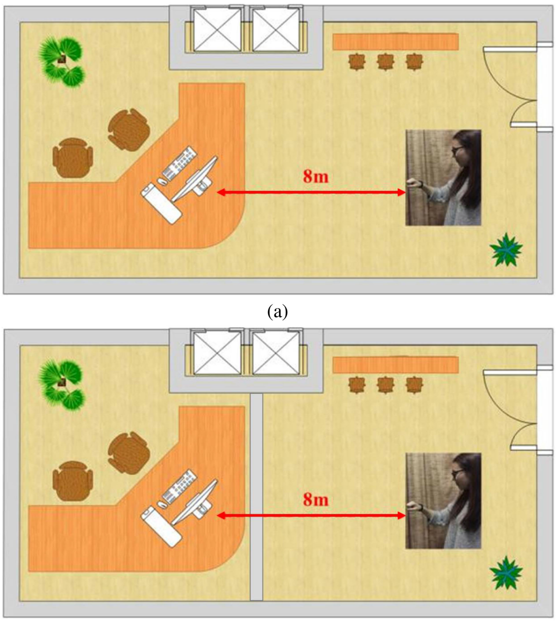
1) Individual Independence: To explore the individual generalization of L-sign for hearing-impaired individuals whose gesture data are not collected, we conduct the individual independence experiment. We select other 5 volunteers who trained for 8 hours to collect the 200 sign gestures mentioned above in the same experimental scenarios. Each volunteer performs each sign gesture 10 times and 10 000 samples are collected. We evaluate the performance of L-sign for the newly collected sign gesture samples and calculate the average accuracy of each volunteer's sign gestures without training L-sign.

Table V shows the experimental results that the L-sign has a high recognition rate for data collected from these five volunteers. It illustrates that the L-sign has user independence and can help hearing-impaired people's daily communication.

2) Experimental Scenarios Independence: To evaluate the recognition effect of L-sign in different scenarios, we conduct experiments in line-of-sight (LOS) and no-line-of-sight (NLOS) scenarios. Fig. 12(a) shows the LOS scenario, in which the volunteer carries out sign gestures with the bracelet about 8 m away from the PC, without any obstacle between the subject and the PC. Fig. 12(b) shows the NLOS scenario, which is the same as LOS except that there is a wall made of bricks between the volunteer and the PC. In this experiment, the 5 volunteers are randomly selected from those 12 volunteers to collect the 200 sign gestures in NLOS scenarios. Each volunteer performs each sign gesture 10 times and the 10 000 samples are collected. We evaluate the performance of L-sign in NLOS scenarios by inputting these samples into L-sign in the test step.

TABLE V
EXPERIMENTAL RESULTS OF USER INDEPENDENCE

Voulnteers	Accuracy
No. 1	90.20%
No. 2	90.05%
No. 3	90.30%
No. 4	91.00%
No. 5	90.40%
Grand mean	90.05%



(b)

Fig. 12. Different experimental scenarios. (a) Line-of-Sight. (b) Non-Line-of-Sight.

TABLE VI
EXPERIMENTAL RESULT ON LOS AND NLOS

Experimental scenarios	Average accuracy
LOS	90.05%
NLOS	89.81%

The experimental results shown in Table VI indicate that the L-sign can achieve high recognition performance in both scenarios.

We believe there are two reasons. The first is that the bracelet transmits the sign signal to the PC via Bluetooth, which offers a greater anti-interference function. The second is that the L-sign gains robustness by simultaneously considering the sign gesture's signal and semantics through the gesture recognizer.

TABLE VII
EXPERIMENTAL RESULTS OF SIGN GESTURES EXTENSION

Number of sign gestures	Accuracy
200	90.05%
210	89.01%
220	88.95%
230	88.34%
240	88.15%
250	88.01%

E. Sign Gestures Extension Experiment

To investigate the impact of the number of sign gestures on the recognition performance of L-sign, we increase the number of sign gestures from 200 to 250. According to the use frequency of gesture, we select 50 commonly used gestures that are not included in the above 200 gestures and add them to the vocabulary of experimental gestures. The newly selected 50 common Chinese sign gestures are collected by the same volunteers mentioned in Section VI. Each sign gesture is collected 20 times. Thus, 12 000 samples are collected. By now, the 250 sign gesture dataset are obtained. In this experiment, the leave-one-volunteer-out cross-validation method is performed to validate the capacity of L-sign. The test is repeated 12 times for calculating the average accuracy when the number of sign gestures increases from 200 to 250.

Table VII shows the average accuracy of L-sign when the sign gesture vocabulary is from 200 to 250. The experimental result indicates that the average accuracy of L-sign decreases slightly with the increase of sign gesture vocabulary. The reason is that the number of similar sign gestures increases with the scale of sign gestures. However, L-sign still maintains a high recognition performance even if the sign gesture vocabulary increases to 250.

VII. CONCLUSION

In this article, we present the L-sign system to recognize the large vocabulary of the Chinese universal sign gestures for the hearing impaired with a bracelet. In the data processing stage, this article proposes a signal entropy-based forward and backward matching algorithm to extract sign gesture signals. A gesture recognizer including a semantic-based voter and candidate-gesture generator is designed to recognize gestures. The candidate-gesture generator is designed on a 3-branch CNN to select the M candidate sign gestures. The semantic-based voter aims to score the candidate gestures based on the proposed semantic distance between the last gesture and the candidate gesture. Among them, the labels are obtained from the proposed multilayer k-means algorithm, and the target gesture is chosen based on the scores. The experimental results of 200 commonly used Chinese universal sign gestures show that the average accuracy of L-sign is over 90% even for sign gestures recognized difficultly by other methods.

In future research, we extend the L-sign system to all Chinese sign gestures and recognize the continuous sign gestures in real-time. The challenge is how to segment a series of continuous sign gesture signals into some discrete gesture signals. By observing the sEMG signals of a series of continuous gestures, we find that there are a lot of "zeros" generated in sEMG signals when signers switch from one gesture to the next. Fig. 13 shows the spectrum of continuous sign gesture signals: There is a flower shop next to a vegetable market. The dark part of Fig. 11 shows more "zeros" in the signal, while the bright part shows fewer "zeros" in the signal. Thus, the discrete sign gesture signals can be extracted from it. Based on this phenomenon, we analyze the distribution of zero in gesture signals and utilize them as the basis for classifying sign gestures.

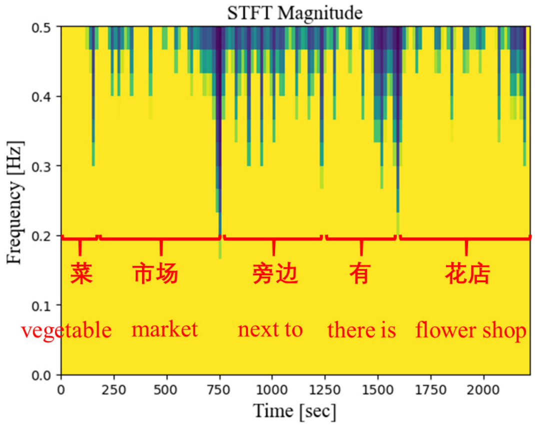


Fig. 13. Spectrum of continuous sign gesture signals.

REFERENCES

[1] Organization, "Avoiding linguistic neglect of deaf children," 2018. [Online]. Available: https:www.who.intdeafnessworld-hearing-dayWorld-Hearing-Day-2018-activity-reportrev1.pdf?ua=1

[2] T. Humphries *et al.*, "Avoiding linguistic neglect of deaf children," *Social Serv. Rev.*, vol. 90, no. 4, pp. 589–619, 2016.

[3] P. P. Wilcox, "My mother made me deaf: Discourse and identity in a deaf community," *J. Anthropological Res.*, vol. 75, no. 3, pp. 425–426, 2019.

[4] G. Davis, "Made to hear: Cochlear implants and raising deaf children by Laura Mauldin," *Amer. J. Sociol.*, vol. 122, no. 6, pp. 2026–2028, May 2017.

[5] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4913–4921.

[6] Y. Yan, Z. Li, Q. Tao, C. Liu, and R. Zhang, "Research on dynamic sign language algorithm based on sign language trajectory and key frame extraction," in *Proc. IEEE 2nd Int. Conf. Electron. Technol.*, 2019, pp. 509–514.

[7] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[8] J. Shotton, T. Sharp, and A. Kipman, "Real-time human pose recognition in parts from single depth images," *Commun. ACM.*, vol. 56, no. 1, pp. 116–124, Jan. 2013.

[9] Microsoft, "Kinect for xbox," 2017. [Online]. Available: http://www.xbox.com/en-US/xbox-one/accessories/kinect

[10] J. Zhang, W. Zhou, X. Chao, J. Pu, and H. Li, "Chinese sign language recognition with adaptive HMM," in *Proc. Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.

[11] L. Motion, "Leap motion," 2017. [Online]. Available: http://leapmotion.com

[12] L. E. Potter, J. Araullo, and L. Carter, "The leap motion controller: A view on sign language," in *Proc. Comput. Hum. Interact. Conf.*, 2013, pp. 175–178.

[13] K. Kang and H. C. Shin, "EMG based gesture recognition using feature calibration," in *Proc. Int. Conf. Inf. Netw.*, 2018, pp. 10–12.

[14] J. Wu, Z. Tian, L. Sun, L. Estevez, and R. Jafari, "Real-time American sign language recognition using wrist-worn motion and surface EMG sensors," in *Proc. IEEE 12th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2015, pp. 1–6.

[15] J. He, H. Luo, J. Jia, J. T. W. Yeow, and N. Jiang, “Wrist and finger gesture recognition with single-element ultrasound signals: A comparison with single-channel surface electromyogram,” *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1277–1284, May 2019.

[16] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. Ailisto, “Identifying users of portable devices from gait pattern with accelerometers,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 2, pp. 973–976.

[17] T. Y. Pan, C. H. Kuo, H. T. Liu, and M. C. Hu, “Handwriting trajectory reconstruction using low-cost IMU,” *IEEE Trans. Comput. Intell.*, vol. 3, no. 3, pp. 261–270, Jun. 2019.

[18] H. Koch, A. Konig, A. W. Seitz, K. Kleinmann, and J. Suchy, “Multisensor contour following with vision, force, and acceleration sensors for an industrial robot,” *IEEE Trans Instrum. Meas.*, vol. 62, no. 2, pp. 268–280, Feb. 2013.

[19] J. H. Hou, X. Y. Li, P. D. Zhu, and P. L. Yang, “Signspeaker: A real-time, high-precision smartwatch-based sign language translator,” in *Proc. MobiCom.*, 2019, vol. 2, pp. 1–15.

[20] R. Feris, M. Turk, R. Raskar, K. Tan, and G. Ohashi, “Exploiting depth discontinuities for vision-based fingerspelling recognition,” in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 155–161.

[21] F. Noroozi, D. Kaminska, and C. Corneanu, “Survey on emotional body gesture recognition,” *IEEE Trans. Affect. Comput*, vol. 12, no. 2, pp. 505–523, 1 Apr.–Jun. 2021.

[22] M. Hamda and A. Mahmoudi, “Hand gesture recognition using Kinect’s geometric and hog features,” in *Proc. Int. Conf. Big Data, Cloud Appl.*, 2017, pp. 1–5.

[23] M. S. Alam, K.-C. Kwon, and N. Kim, “Implementation of a character recognition system based on finger-joint tracking using a depth camera,” *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 3, pp. 229–241, Jun. 2021.

[24] Y. Liu, M. Peng, M. R. Swash, T. Chen, R. Qin, and H. Meng, “Holoscopic 3D microgesture recognition by deep neural network model based on viewpoint images and decision fusion,” *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 2, pp. 162–171, Apr. 2021.

[25] D. P. Azari *et al.*, “A comparison of expert ratings and marker-less hand tracking along Osats-derived motion scales,” *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 1, pp. 22–31, Feb. 2021.

[26] S. Gaglio, G. L. Re, and M. Morana, “Human activity recognition process using 3-D posture data,” *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 5, pp. 586–597, Oct. 2015.

[27] C. Pacchierotti, S. Sinclair, M. Solazzi, A. Frisoli, V. Hayward, and D. Prattichizzo, “Wearable haptic systems for the fingertip and the hand: Taxonomy, review, and perspectives,” *IEEE Trans. Haptics*, vol. 10, no. 4, pp. 580–600, Oct.–Dec. 2017.

[28] G. A. Casula, A. Michel, P. Nepa, G. Montisci, and G. Mazzarella, “Robustness of wearable UHF-band PIFAs to human-body proximity,” *IEEE Trans. Antennas. Propag.*, vol. 64, no. 5, pp. 2050–2055, May 2016.

[29] K. Van Volkinburg and G. Washington, “Development of a wearable controller for gesture-recognition-based applications using polyvinylidene fluoride,” *IEEE Trans. Biomed. Circuits. Syst.*, vol. 11, no. 4, pp. 900–909, Aug. 2017.

[30] T. Ogasawara, H. Fukamachi, K. Aoyagi, S. Kumano, H. Togo, and K. Oka, “Archery skill assessment using an acceleration sensor,” *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 3, pp. 221–228, Jun. 2021.

[31] W. Wang, R. Li, Z. M. Diekel, Y. Chen, Z. Zhang, and Y. Jia, “Controlling object hand-over in human–robot collaboration via natural wearable sensing,” *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 1, pp. 59–71, Feb. 2019.

[32] T. Zhao, J. Liu, Y. Wang, H. Liu, and Y. Chen, “PPG-based finger-level gesture recognition leveraging wearables,” in *Proc. Conf. Comput. Commun.*, 2018, pp. 1457–1465.

[33] F. Adib, Z. Kabelac, and D. Katabi, “3D tracking via body radio reflections,” in *Proc. 11th USENIX Symp. Netw. Syst. Des. Implementation*, 2014, pp. 317–329.

[34] T.-H. Chen, S.-I. Sou, and Y. Lee, “Witrack: Human-to-human mobility relationship tracking in indoor environments based on spatio-temporal wireless signal strength,” in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput.*, 2019, pp. 788–795.

[35] Z. Wang, T. Zhao, J. Ma, H. Chen, and J. Ren, “Hear sign language: A real-time end-to-end sign language recognition system,” 2020. [Online]. Available: https://doi.org/10.1109/TMC.2020.3038303

[36] E. G. Bakhoum, M. H. M. Cheng, and R. A. Kyle, “3-axis, ultrahigh-sensitivity, miniature acceleration sensor,” *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 8, no. 2, pp. 244–250, Feb. 2018.

[37] H. Ahmed and M. Tahir, “Improving the accuracy of human body orientation estimation with wearable IMU sensors,” *IEEE Trans. Instrum. Meas.*, vol. 66, no. 3, pp. 535–542, Mar. 2017.

[38] T. Kuroda, Y. Tabata, A. Goto, H. Ikuta, and M. Murakami, “Consumer price data-glove for sign language recognition,” in *Proc. Int. Conf. Disabil. Virtual Reality Assoc. Tech.*, 2004, pp. 253–258.

[39] H. Wang, C. L. Ming, and C. Oz, “American sign language recognition using multi-dimensional hidden Markov models,” *J. Inf. Sci. Eng.*, vol. 22, no. 5, pp. 1109–1123, 2006.

[40] D. Ekiz, “Sign sentence recognition with smart watches,” in *Proc. Signal Proces. Commun. Appl. Conf.*, 2017, pp. 1–4.

[41] H. Y. Wen, J. R. Rojas, A. K. Wang, and Y. Dey, “Serendipity: Finger gesture recognition using an off-the-shelf smartwatch,” in *Proc. Conf. Hum. Factors Comput. Syst.*, 2016, pp. 3847–3851.

[42] A. V. Astafiev, A. L. Zhiznyakov, and A. A. Demidov, “The use of butterworth filter to compensate for noise in signals from bluetooth low energy beacons in autonomous navigation systems,” in *Proc. Int. Russian Autom. Conf.*, 2020, pp. 1117–1121.

[43] 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 Trans. Syst., Man, Cybern. - Part A: Syst. Humans*, vol. 41, no. 6, pp. 1064–1076, Nov. 2011.

[44] J. Wu, B. Yin, and W. Qi, “Video motion segmentation based on double sliding window,” in *Proc. Int. Symp. Comput. Intell. Des.*, 2011, pp. 232–235.

[45] Y. Fan, M. Xu, Z. Wu, and L. Cai, “Automatic emotion variation detection using multi-scaled sliding window,” in *Proc. Int. Conf. Orange Technol.*, 2014, pp. 232–236.

[46] R. Gupta and N. Jha, “Real-time continuous sign language classification using ensemble of windows,” in *Proc. Int. Conf. Adv. Comput. Commun. Syst.*, 2020, pp. 73–78.

[47] T. Mikolov, K. Chen, J. Corrado, and G. Dean, “Efficient estimation of word representations in vector space,” in *Proc. Int. Conf. Comput. Learn. Representations*, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[48] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proc. Int. Conf. Mach. Learn.*, vol. 32, no. 2, 2014, pp. 1188–1196.

[49] P. Bojanowski, E. Grave, and A. Joulin, “Enriching word vectors with subword information,” *IEEE Trans. Assoc. Comput. Linguistics*, vol. 1, no. 5, pp. 135–146, Jun. 2017.

[50] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.

[51] J. Cheng, H. Liu, F. Wang, H. Li, and C. Zhu, “Silhouette analysis for human action recognition based on supervised temporal t-SNE and incremental learning,” *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3203–3217, Oct. 2015.

[52] L. Yin, S. Liu, W. K. Ho, and K. V. Ling, “Indoor tracking with the generalized t-distribution noise model,” *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 3, pp. 915–926, May 2018.

[53] S. Said, H. Hajri, L. Bombrun, and B. C. Vemuri, “Gaussian distributions on Riemannian symmetric spaces: Statistical learning with structured covariance matrices,” *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 752–772, Feb. 2018.

[54] M. Ling and X. Geng, “Indoor crowd counting by mixture of Gaussians label distribution learning,” *IEEE Trans Inf. Theory*, vol. 28, no. 11, pp. 5691–5701, Nov. 2019.

[55] C. Wang, X. N. Guo, Y. Wang, and Y. Y. Chen, “Friend or foe?: Your wearable devices reveal your personal pin,” in *Proc. Conf. Comput. Commun. Secur.*, 2016, pp. 189–200.

[56] X. Jiang, K. Xu, and T. Sun, “Action recognition scheme based on skeleton representation with DS-LSTM network,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2129–2140, Jul. 2020.

[57] H. Xu, S. A. Da, and K. Saenko, “R-C3D: Region convolutional 3D network for temporal activity detection,” in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5794–5803.

[58] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, “Using convolutional 3D neural networks for user-independent continuous gesture recognition,” in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 49–54.

[59] K. Liu, W. Liu, C. Gan, M. K. Tan, and H. D. Ma, “T-C3D: Temporal convolutional 3D network for real-time action recognition,” in *Proc. Conf. Artif. Intell.*, 2018, pp. 7138–7145.

[60] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, “Hand gesture recognition with 3D convolutional neural networks,” in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 1–7.