A Modified LSTM Model for Chinese Sign Language Recognition Using Leap Motion

1st Bixiao Wu
College of Automation Science and Engineering
South China University of Technology
Guangzhou, China
wubixiao1997@163.com

2nd Zhenyu Lu
Bristol Robotics Laboratory
University of the West of England
Bristol, UK
zhenyu.lu@uwe.ac.uk

3rd Chenguang Yang*
Bristol Robotics Laboratory
University of the West of England
Bristol, UK
cyang@ieee.org

***Abstract*—At present, there are about 70 million deaf people using sign language in the world, but for most normal people, it is difficult to understand the meaning of the sign language expression. Therefore, it is of great importance to explore the ways of recognising the sign language. In this paper, we propose a dynamic sign language recognition method based on the modified long short-term memory (LSTM) model. Firstly, we use Leap Motion to collect the features of Chinese Sign Language (CSL). LSTM has a good effect in processing time series data, but the parameters of its hidden layer are shared, making it important information lost when dealing with long time series. The attention mechanism can give different attention weights to different features according to the correlation between the input data and output data, so as to enhance the model's attention to key information. Therefore, we combine LSTM with attention mechanism for dynamic sign language recognition. Experimental results show that the recognition accuracy of the modified LSTM model is 99.55%, which is higher than that of LSTM model. Finally, we developed a sign language human-computer interaction system, which verifies the real-time performance and effectiveness of the method proposed in this paper.**

***Index Terms*—Sign language recognition, LSTM, Attention mechanism, Leap Motion**

I. Introduction

Sign language recognition is a kind of gesture recognition, which is more practical than ordinary gestures. Although deaf-mutes can communicate barrier-free with each other through sign language without barriers, it is difficult to use sign language to communicate smoothly between deaf-mutes and ordinary people haven't learned sign language before. Therefore, the efficient sign language recognition has strong practical significance, which can help deaf-mutes to participate in normal social activities.

Sign language recognition can be divided into static and dynamic sign language recognition. Static sign language recognition refers to the recognition of the sign language expressed in static data. Kuznetsova *et al*. [1] proposed a high-precision method to recognize static sign language. They first obtained the depth image, then obtained the key features by rotating, translating and scaling of the image, and trained the multi-layer random forest to classify the feature vector for sign language recognition. In [2], Amaya *et al*. used principal component analysis and support vector machine to conduct a sign language recognition system. The method included color information detection, skin region separation, hand segmentation, morphological operation, gesture feature extraction and classification. And the results showed that the test accuracy of their method was higher than 80%, and the execution time of each frame was only 59 milliseconds. In [3], [4], the collected images containing gestures are used to train the convolutional neural network (CNN) recognition model for sign language recognition.

Static sign language recognition uses static data, which is easy be achieved in data acquisition, feature extraction and recognition. However, the actual sign language is often a series of continuous actions, rather than static, so the static sign language recognition has great limitations in practical. Different from static sign language recognition, dynamic sign language recognition uses continuous sequence data. It has high practical value, but at the same time it is more complex to implement. Huang *et al*. [5] proposed an sign language recognition method based on key frame center segment. They extracted the key information in sign language as features. Finally, this method was tested on the data set of isolated sign language and achieved a good effect.

Deep learning methods are also widely used for dynamic sign language recognition. In [6], Camgoz *et al*. used CNN to conduct a framework that could translate the video sign language into spoken language. Pu *et al*. [7] developed a method based on three-dimensional convolution residual network and the encoder-decoder network for sign language recognition. The former is used to learn the data features, while the latter is used to learn sequence models. Ariesta *et al*. [8] proposed a sentence-level sign language recognition method. They used CNN and bidirectional recursive neural network (RNN) to recognize sign language. They applied three-dimensional CNN to extract features from each video frame, while bidirectional RNN was used to obtain features from gesture actions in video sequences. Cui *et al*. [9] used CNN and RNN for feature extraction and sequence learning separately and then constructed an end-to-end sign language recognition framework.

However, RNN model has some problems, such as gradient disappearance and gradient explosion, which make RNN model perform poorly in long-time sequence data processing [10], [11]. In order to solve the problems of the RNN model, researchers proposed LSTM model [12], [13]. LSTM model can process sequence better, so it is often used as a method of dynamic sign language recognition [14]. Liu *et al*. [15] proposed an end-to-end sign language recognition method based on LSTM, considering that the LSTM can well learn the context information of sequence data. In [16], the 3D

This work was supported in part by the H2020 Marie Skłodowska-Curie Actions Individual Fellowship under Grant 101030691.

convolution residual network and bidirectional LSTM network were combined to realize dynamic sign language recognition. Kumar *et al*. [17] used Kinect and Leap Motion to get the feature of dynamic hand gesture, and proposed a multimodal recognition method with a Hidden Markov Model (HMM) and bidirectional LSTM. In [18], a human-robot interaction framework was constructed based on hand gesture recognition, Leap Motion was used to extract hand features and LSTM was adopted to predict the dynamic hand gestures. The method was applied in human-social robot interactions.

Extracting information from the natural data is important in dynamic sign language recognition and other cases [19]. There are two important steps to improve the efficiency and accuracy of sign language recognition: focusing on the critical information of the sequence data of dynamic sign language and reducing the attention to unimportant information. In neural network, attention mechanism is used to calculate the correlation between input data and output data to enable the model to focus on useful information [20]. Therefore, it has a excellent performance in the tasks such as natural language processing, image recognition and speech recognition [21]–[23].

LSTM model is an effective model to solve time series problems. It not only has the advantages of traditional neural network, but also can save the historical information of data. However, since the parameters of the hidden layer of the LSTM model are shared, important information is easily lost in long time series problems, resulting in a waste of information resources. Therefore, we introduce the attention mechanism into the model and combine it with LSTM network. Attention mechanism can solve the problems of LSTM model effectively. It can provide corresponding weights for each feature according to the correlation level between input data and output data, so as to strengthen the model's memory of important information and ignore irrelevant information.

Vision-based dynamic sign language recognition method extracts key features from videos for corresponding recognition. Although videos are easy to obtain and contain rich information, the computational cost of processing video is large, and it is easy to be affected by lighting conditions and other factors. In contrast, the skeleton data of hands can represent the information of the hand accurately, with less other interference factors. The Leap Motion is a kind of hand three-dimensional data acquisition equipment based on computer vision. The device can collect a variety of hand information such as the position and direction of hand joints and the speed of fingertip movement, among other things. In recent years, it has been widely utilized by researchers for dynamic gesture recognition [24]–[26].

Inspired by above works, we use Leap Motion to extract features of sign language and combine the LSTM and attention mechanism to achieve dynamic sign language recognition. The main contributions are as follows:

1. The LSTM is strengthened with the adoption of attention mechanism and then used for dynamic sign language recognition based on the features obtained by Leap Motion.
2. The proposed method is applied in a novel framework for human-computer interaction through sign language.

In this work, we use the key point data of hands detected by Leap Motion to recognize dynamic sign language, and use the sign language recognition results to realize human-computer interaction. The rest of the paper is organized as follows. In Section II, we mainly introduce the Leap Motion and the collected features in detail. The methodology is presented in Section III. In Section IV, we mainly compare the results of the different experiments and design the human-computer interaction system. In Section V, we summarize this work and discuss further work.

II. Data Collection

A. Leap Motion

Leap Motion is a somatosensory controller released in 2013. It has powerful algorithm supports, which can directly display the detection results of each joint of the hand on the user interface and realize the accurate grasping of objects in the virtual scene. Compared with other devices, Leap Motion has advantages of small volume, high accuracy of hand recognition and good real-time performance. Leap Motion uses the right-handed coordinate system, and each unit in the coordinate system represents a length of one millimeter. The center of the device is set as the origin of the coordinate. The vertical upward direction is the positive direction of the Y-axis, the horizontal plane is composed of X axis and Z axis. The coordinate system of Leap Motion is shown in Fig. 1.

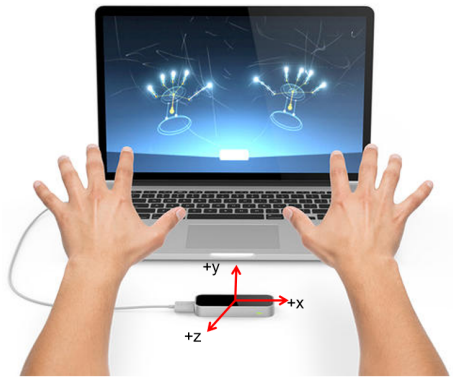


Fig. 1: The coordinate system of Leap Motion.

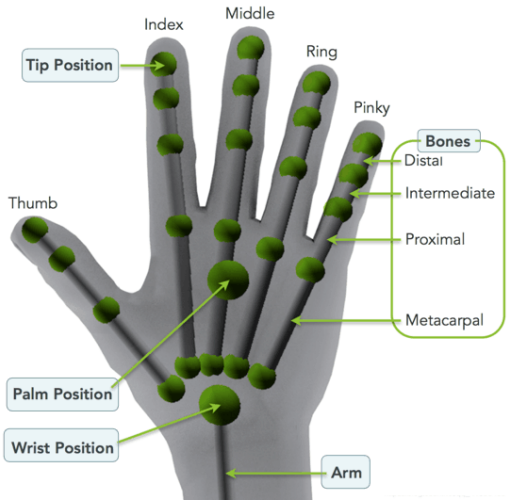


Fig. 2: Structure of the hand.

Leap Motion can accurately identify the hand and provide a series of gesture related information. Frame data is the core of Leap Motion data. Each frame dataset contains all the information of fingers, fingertips, gestures and their position, speed, direction, rotation angle and so on. For each hand, the corresponding hand model will be generated, including thumb, index finger, middle finger, ring finger, little thumb and wrist joint, as shown in Fig.2. The detectable area of Leap Motion is a spatial inverted pyramid, with a horizontal field angle of 140°, a vertical field angle of 120°, an interactive depth of 10cm-60cm, and a height of up to 80cm. If the hand appears in the recognition area of the Leap Motion, it will automatically track, and output a series of data frames in real time, refresh constantly. The Leap Motion can collect 200 frames of hand data per second with an accuracy of 0.01mm. In this paper, we use Leap Motion to collect the dynamic hand gestures of sign language.

B. Features collection

We used Leap Motion to collect the relevant features of the both hands, as shown in TABLE I. The pitch of the palm is represented the angle between the z-axis and the projection of the vector onto the y-z plane. The yaw and roll of the palm are represented by the angle between the z-axis and the projection of the vector onto the x-z plane and the angle between the y-axis and the projection of the vector onto the x-y plane, respectively. For a clearer description, we use Fig.3 to visualize the collected features.1

TABLE I: Sign language features collected by Leap Motion

Features	Dimensions
Position of the palm	6
Normal of the palm	6
Direction of the palm	6
Pitch of the palm	2
Yaw of the palm	2
Roll of the palm	2
Position of the fingers	30
Distance between the fingertips and the palm	10
Position of the joints	126

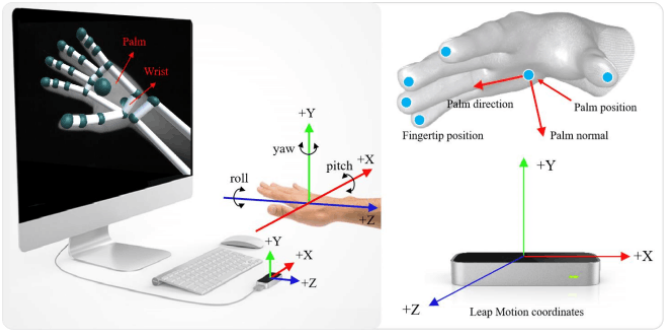


Fig. 3: Hand features collected by Leap Motion.

III. Methodology

A. LSTM

LSTM model is composed of a memory cell, a forget gate, an input gate, and an output gate, so it could save the historical information of the sequence and solve the timing problem effectively. Sign language can be regarded as a continuous series of dynamic gestures. Therefore, LSTM is very suitable for recognizing sign language actions. The structure of LSTM is illustrated in Fig. 4. x_t represents the input of LSTM model, h_t represents the output of LSTM model, f_t is the forget gate variable, i_t is the input gate variable and o_t is the output gate variable. t and $t-1$ indicate the current time and the previous time. c_t and $\widetilde{c}_t$ represent memory cell state and the memory gate respectively.

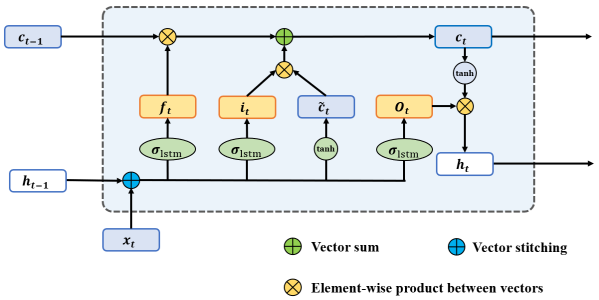


Fig. 4: Structure of LSTM.

The forget gate, input gate, and output gate control the flow of information, and their calculation methods are shown in (1) to (6).

$$f_t = \sigma_{lstm}(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma_{lstm}(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\widetilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t \tag{4}$$

$$o_t = \sigma_{lstm}(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

where W_f, W_i, W_c and W_o represent the weight matrices of corresponding subscripts. b_f, b_i, b_c and b_o are the biases of corresponding subscripts. σ_{lstm} and tanh represents a sigmoid function and a hyperbolic activation function respectively, and the expressions are shown in (7) and (8) .

$$\sigma_{lstm}(x) = \frac{1}{1+e^{-x}} \tag{7}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{8}$$

B. The modified LSTM model

Although the LSTM model has a good performance in dealing with timing problems, it still has some limitations. The hidden layer parameters of LSTM are shared, which may cause the model to lose some key information. For example, in the process of dynamic gesture recognition, LSTM may not distinguish the importance of continuous gesture features well, causing the model to lose important information. Attention mechanism can solve this problem effectively. Attention

mechanism [21] can get the correlation between the input data and output data, so that different information can be weighted differently. For the time period that has a greater impact on the results, the attention mechanism will give a larger weight to them, and for the time period that has a less impact on the output results, the attention mechanism will give a smaller weight to them. Therefore, the model can pay more attention to key data and weaken the influence of unimportant data, which greatly improves the effect and efficiency of the model.

The structure of attention mechanism is shown in Fig. 5. h_i is the input data of the model, s_i is used to indicate the correlation between input information and output information, the $softmax$ function is used to normalize the calculated correlation s_i, so that the attention weight coefficient a_i of input data can be obtained. Finally, the output sequence H can be obtained by weighted summation of the attention weight coefficients and the input data. Through the above steps, the network model can focus on the information with high weight coefficient and ignore the information with low weight coefficient, thereby improving the recognition accuracy of the model.

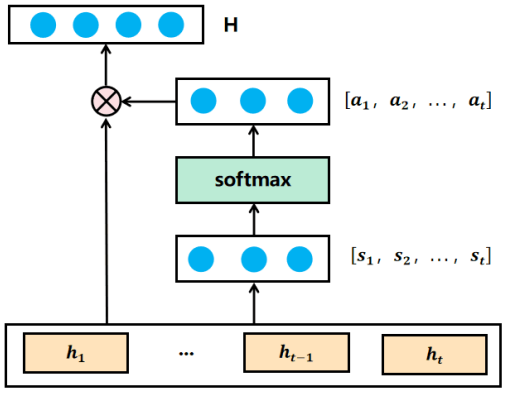


Fig. 5: Structure of attention mechanism.

Due to the multi-level and multi-dimensional characteristics of hand features, different hand features have different effects on the LSTM network. The traditional LSTM model cannot solve the problem of sign language recognition well, so we combine the LSTM with the attention mechanism in this paper. The model structure of the modified LSTM network is shown in Fig. 6. The model consists of five parts: input layer, LSTM network, attention layer, full connection layer and output layer. The specific implementation steps are as follows:

1. Firstly, the different features collected by Leap Motion are input into the LSTM model, and the features are encoded by the network to obtain the intermediate hidden state $[h_1, h_2, ..., h_t]$.
2. Take the hidden layer state $[h_1, h_2, ..., h_t]$ as the input of the attention layer, and then we use Multi-Layer Perception to calculate the correlation between vector $h_i(i = 1, 2...t)$ and output y_t, as shown in (9).

$$s_i = MLP(h_i, y) = v^T \tanh(W_1 h_i + W_2 y) \quad (9)$$

where $MLP(*)$ is the Multi-Layer Perception, tanh() is a hyperbolic activation function, and v, W_1 and W_2 are the parameters to be trained.

3. The obtained correlation s_i is normalized by $softmax$ function to obtain the attention weight coefficient a_i of each feature, as shown in (10).

$$a_i = softmax(s_i) = \frac{exp(s_i)}{\sum_{j=1}^{t} exp(s_j)} \quad (10)$$

4. Weighted summation of the attention weight coefficient a_i and the vector $h_i(i = 1, 2...t)$ to obtain the output sequence H, as shown in (11).

$$H = \sum_{i=1}^{t} a_i h_i \quad (11)$$

5. Finally, the output sequence H is adopted to the full connection layer to obtain the outputs.

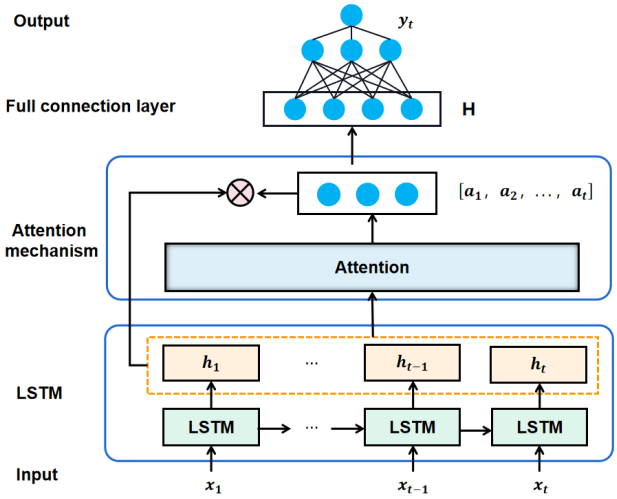


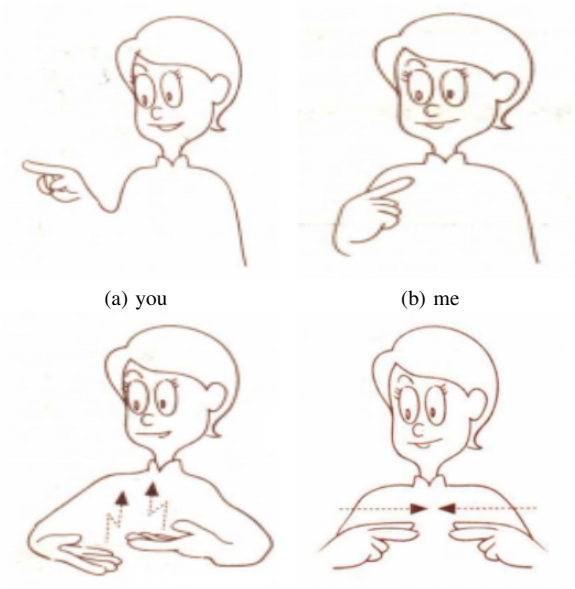
Fig. 6: Structure of the proposed method.

After adding the attention mechanism, the modified LSTM model can assign different weights to different features. The model can learn the importance of different features and extract important information from dynamic sign language, improving recognition accuracy.

IV. Experiment

A. Experimental setup

In this section, we will test the performance and efficiency of the proposed method. We conducted experiments on a laptop equipped with Intel Core i5-6200u CPU. Sign language generally includes one-handed sign language and two-handed sign language. Common one-handed sign languages include: "you", "me", "good", "thanks" and so on, as shown in Fig. 7a and Fig. 7b; Common two-handed sign language include: "what", "happy", "please" and so on, as shown in Fig. 7c and Fig. 7d. When Leap Motion collects one-handed sign language, we set the related data of the other hand to match the collected sign language data, thus ensuring the same dimension of the input data. We collected 20 dynamic sign languages to create a sign language dataset, as shown in TABLE II. These dynamic sign languages were collected from five individuals who repeated each gesture 100 times at different speeds.



(a) you (b) me

(c) happy (d) meet

Fig. 7: Several examples of sign language.

TABLE II: 20 common examples of sign language

you	me	can	good	please
and	go	want	here	home
where	thanks	what	give	look
how much	happy	no matter	meet	goodbye

B. Results

We use the features introduced in Section II to train and test neural networks. The experimental results are shown in Fig.8. It can be seen from the experimental results that after adding the attention mechanism, the modified LSTM model achieves a higher accuracy than the original LSTM model. The highest accuracy of the modified LSTM model is 99.55%, which is 0.46% higher than the result of the LSTM. And the combination of all features achieves the highest accuracy. It shows that different features include different characteristics of the hand, which can improve the recognition accuracy of the model. We use the trained model for recognizing sign language and part of experimental results are shown in Fig. 9 to Fig. 11. The experimental results show that these sign languages could be well recognized by the methods presented in this paper.

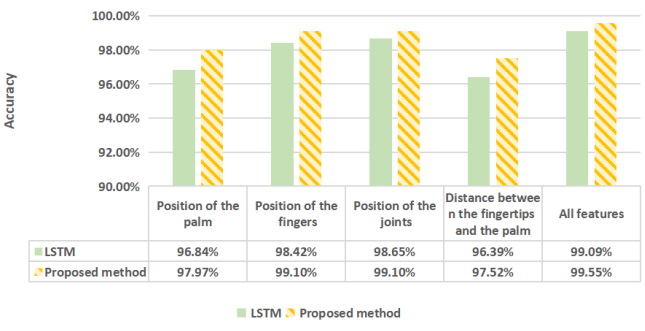


Fig. 8: Recognition accuracy of different features.

In addition, the second and third columns in Fig. 8 show that although the position of joints contain richer features than the position of fingertips, the accuracy obtained by the two experiments is the same. This shows that after adding the attention mechanism to LSTM, the model can pay more attention to the important information in the input data. In other words, among the position of all joints, the position of fingertips are the most critical feature.

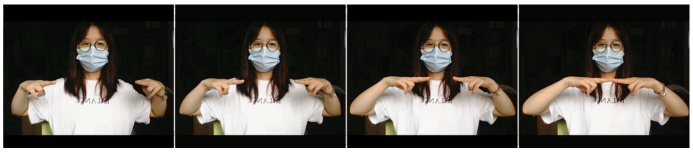


Fig. 9: Recognition results of ”meet”.

Fig. 10: Recognition results of ”happy”.

Fig. 11: Recognition results of ”please”.

C. Application

Based on our proposed method, we develop a sign language recognition interaction system that enables efficient interaction between deaf-mutes and computers. We use the CSL synthesis system that is developed by Institute of computing technology of the Chinese Academy of Sciences, as shown in Fig.12. This system can convert the sentences entered by the user into the corresponding sign language action which then displayed in the user interface. With this sign language synthesis system, we can give a real-time feedback to sign language recognition, to make whole sign language interaction system becomes more friendly and natural.

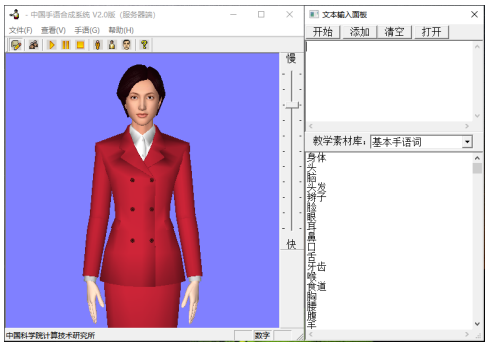


Fig. 12: Chinese sign language synthesis system.

The framework of sign language recognition interaction system is shown in Fig.13. Firstly, we input the dynamic sign language collected by Leap Motion into the trained neural network, and start the sign language synthesis software at the same time. Then, the content of the response is given based on the recognition results of the model. For example, if the recognition result is "thanks", the reply is "you are welcome". Finally, pywin32 library is called to obtain the window handle of the software, and sends the win system message to the specified software window through the window handle, so that the sign language action can be displayed on the screen. Through the above steps, we could realize the human-computer interaction system of CSL.

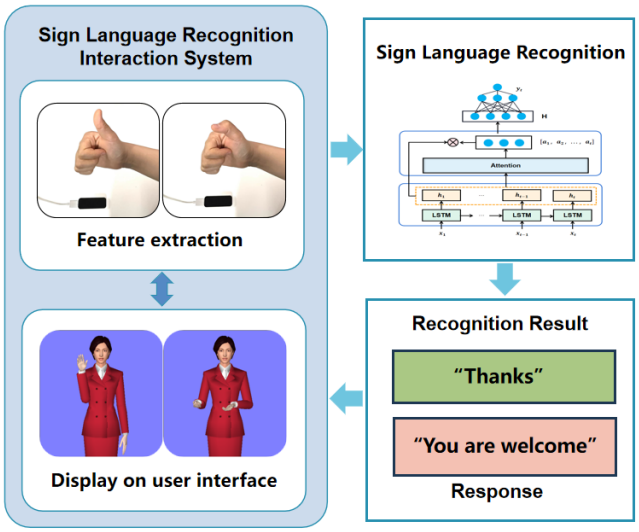


Fig. 13: Framework of sign language recognition interaction system.

V. Conclusion

This paper proposes a dynamic sign language recognition framework based on a modified LSTM. First, we use the Leap Motion to collect various features of sign language. Then, considering that the attention mechanism could make the neural network model pay more attention to the key information in sign language, we combine attention mechanism and LSTM to recognize dynamic sign language. The experimental results show that the modified LSTM model can effectively use the key features of the hand and improve the recognition accuracy of the model. Moreover, based on the proposed method, a human-computer interaction system for sign language has been developed. In the future, we will focus on developing a framework with stronger generalization ability and practicality, which can recognize more sign languages.

References

[1] A. Kuznetsova, L. Leal-Taixé, and B. Rosenhahn, "Real-time sign language recognition using a consumer depth camera," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 83–90.

[2] C. Amaya and V. Murray, "Real-time sign language recognition," in *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 2020, pp. 1–4.

[3] D. Aich, A. Al Zubair, K. M. Zubair Hasan, A. D. Nath, and Z. Hasan, "A deep learning approach for recognizing bengali character sign langauage," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–5.

[4] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural computing and applications*, vol. 32, no. 12, pp. 7957–7968, 2020.

[5] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel chinese sign language recognition method based on keyframe-centered clips," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 442–446, 2018.

[6] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7784–7793.

[7] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4165–4174.

[8] M. C. Ariesta, F. Wiryana, A. Zahra *et al.*, "Sentence level indonesian sign language recognition using 3d convolutional neural network and bidirectional recurrent neural network," in *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*. IEEE, 2018, pp. 16–22.

[9] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1880–1891, 2019.

[10] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen," *Diploma, Technische Universität München*, vol. 91, no. 1, 1991.

[11] J. Schmidhuber, "Learning complex, extended sequences using the principle of history compression," *Neural Computation*, vol. 4, no. 2, pp. 234–242, 1992.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[13] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International conference on artificial neural networks*. Springer, 2005, pp. 799–804.

[14] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified lstm model for continuous sign language recognition using leap motion," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7056–7063, 2019.

[15] T. Liu, W. Zhou, and H. Li, "Sign language recognition with long short-term memory," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 2871–2875.

[16] C. Wei, W. Zhou, J. Pu, and H. Li, "Deep grammatical multi-classifier for continuous sign language recognition," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2019, pp. 435–442.

[17] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, 2017.

[18] B. Wu, J. Zhong, and C. Yang, "A visual-based gesture prediction framework applied in social robots," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 510–519, 2021.

[19] Z. Lu, N. Wang, and C. Yang, "A novel iterative identification based on the optimised topology for common state monitoring in wireless sensor networks," *International Journal of Systems Science*, vol. 53, no. 1, pp. 25–39, 2022.

[20] Z. Da, J. Engelberg, and P. Gao, "In search of attention," *The journal of finance*, vol. 66, no. 5, pp. 1461–1499, 2011.

[21] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[22] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," *arXiv preprint arXiv:2102.05095*, vol. 2, no. 3, p. 4, 2021.

[23] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[24] T.-W. Chong and B.-G. Lee, "American sign language recognition using leap motion controller with machine learning approach," *Sensors*, vol. 18, no. 10, p. 3554, 2018.

[25] W. Zeng, C. Wang, and Q. Wang, "Hand gesture recognition using leap motion via deterministic learning," *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28 185–28 206, 2018.

[26] A. Vaitkevičius, M. Taroza, T. Blažauskas, R. Damaševičius, R. Maskeliūnas, and M. Woźniak, "Recognition of american sign language gestures in a virtual reality using leap motion," *Applied Sciences*, vol. 9, no. 3, p. 445, 2019.