Sign Language Recognition and Translation Method based on VTN

Wuyang Qin
Nanjing Tech University
College of Electrical Engineering and Control Science
Nanjing, China
e-mail: 1169024572@qq.com

Xue Mei *
Nanjing Tech University
College of Electrical Engineering and Control Science
Nanjing, China
* Corresponding author: mx@njtech.edu.cn

Yuming Chen
Nanjing Tech University
College of Electrical Engineering and Control Science
Nanjing, China
e-mail: cym@njtech.edu.cn

Qihang Zhang
Nanjing Tech University
College of Electrical Engineering and Control Science
Nanjing, China
e-mail: 546017449@qq.com

Yanyin Yao
Nanjing Tech University
College of Electrical Engineering and Control Science
Nanjing, China
e-mail: 935136577@qq.com

Shi Hu
Chizhou Vocational And Technical College
Department of Mechatronics and Automotive
Chizhou, China
e-mail: 503181147@qq.com

2021 International Conference on Digital Society and Intelligent Systems (DSInS) | 978-1-6654-0630-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/DSInS54396.2021.9670588

***Abstract*—Sign language recognition plays an important role in real-time sign language translation, communication for deaf people, education and human-computer interaction. However, vision-based sign language recognition faces difficulties such as insufficient data, huge network models and poor timeliness. We use VTN (Video Transformer Net) to construct a lightweight sign language translation network. We construct the dataset called CSL_BS (Chinese Sign Language-Bank and Station) and two-way VTN to train isolated sign language and compares it with I3D (Inflated three Dimension). Then I3D and VTN are respectively used as feature extraction modules to extract the features of continuous sign language sequences, which are used as the input of the continuous sign language translation decoding network (seq2seq). Based on CSL-BS, two-way VTN achieves 87.9% accuracy while two-way I3D is 84.2%. And the recognition speed is increased by 46.8%. In respect of continuous sign language translation, the accuracy of VTN_seq2seq is 73.5% while I3D_seq2seq is 71.2%, the recognition speed is 13.91s and 26.54s respectively.**

Keywords-component; continuous sign language recognition; Isolated sign language recognition; sign language feature extraction; encoding-decoding network; sign language dataset

I. INTRODUCTION

China has the largest number of hearing disabilities in the world. At present, there are few hearing people proficient in sign language. The automatic recognition of sign language can not only promote the communication between deaf-mute people and hearing people, but also has great application prospects in real-time sign language translation and education [1].

Sign language expresses language meaning by gesture actions. Currently, most of the dataset in the field of Chinese sign language research are collected by depth camera, such as the DEVISIGN dataset [12] and CSL dataset [4]. In terms of sign language recognition, 3D convolution method can complete the task but the recognition speed is slow. In order to solve the above problems, we carry out the recognition framework based on VTN. The main work includes: 1) Exploring the correlation between gestures and realizing the recognition of isolated sign language through VTN. 2) Proposing a continuous sign language translation framework which can be divided into three modules: isolated sign language training, continuous sign language feature extraction and continuous sign language recognition. 3) According to the "National Common Sign Language Common Vocabulary List", we construct the dataset called CSL-BS.

Sign language recognition consists of two subtasks, isolated sign language recognition and continuous sign language recognition [2]. Early isolated sign language recognition relies on artificially designed features, such as SIFT-based feature extraction [3, 5], HOG-based feature [6-8], and frequency domain Features [9-10], use Hidden Markov Model or Dynamic Time Warping [13] to establish the temporal relationship in the sign language action sequence on time correlation, and finally use classification algorithms such as SVM [14] to achieve isolated sign language classification. Convolutional neural network based such as Li et al. [15], Hamid Reza et al. [16] achieved better performance. There are also 3D networks that directly obtain spatiotemporal features of actions [17]. 3D convolution can complete the recognition task, but the amount of calculation is large and cannot meet the realtime requirements.

The sequence modeling of continuous sign language recognition is usually realized through the encoder-decoding network or CTC (Connectionist Temporal Classification) classification. The CTC method proposed by Graves [18] et al. is an ideal method to solve the problem of weakly labeled vision. The encoding-decoding model allowing input and output sequences to be of unequal length. Kalchbrenner et al. [19] first

978-1-6654-0630-7/21/$31.00 ©2021 IEEE

proposed using a single RNN for encoding and decoding tasks. On this basis, Cho [20] et al. used two independent RNNs for encoding and decoding. In this paper, we use the VTN recognizes isolated sign language and extracts continuous sign language features which be saved as the input of seq2seq model. Accoding to the model, we achieve continuous sign language translation with fast recognition speed and high accuracy.

II. SIGN LANGUAGE RECOGNITION AND TRANSLATION BASED ON VTN-SEQ2SEQ

The sign language recognition system is shown in Figure 1, which is divided into three parts: isolated sign language, continuous sign language feature extraction and continuous sign language translation module. The isolated sign language module takes isolated word data as input, trains and saves the model. The continuous sign language feature extraction module is an isolated sign language module that loads the model. The difference is that the layer of softmax is removed, the feature vector of the continuous sign language is retained. The continuous sign language translation module takes continuous action features as input. Learn the mapping relationship between continuous sign language feature vectors and labels.

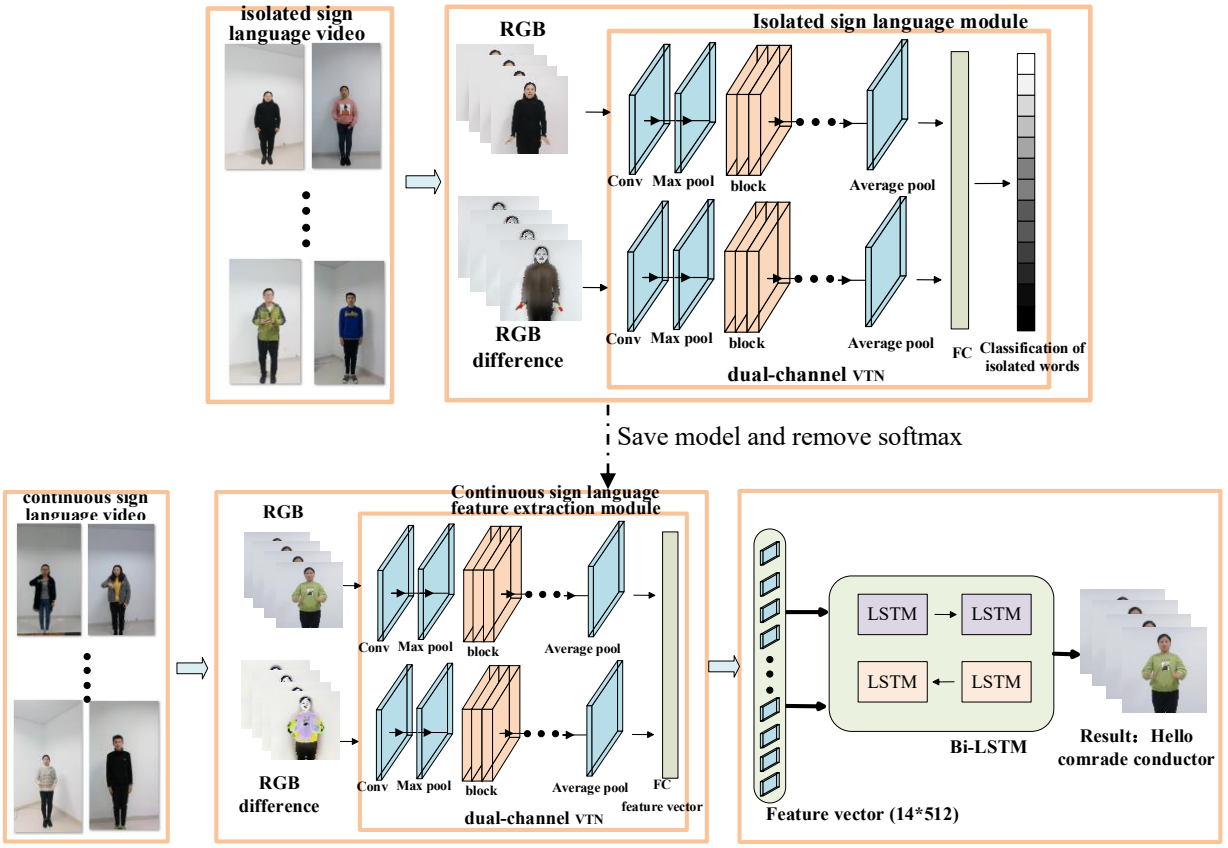

Fig.1 The flow chart of sign language recognition

A. Structure of VTN

VTN is a lightweight network structure for real-time behavior recognition [11]. It uses Transformer network as a method of sequence modeling and consists of two parts: encoding and decoding. In this paper, ResNet-34 is used as the encoder, and each frame of the input sequence is processed independently to obtain the frame embedding of each frame. The decoder is composed of a multi-head self-attention module and a feedforward network. Timing relationship. Figure 2 shows the structure of VTN:

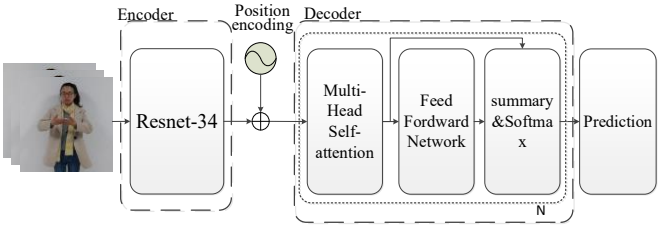

Fig.2 The Structure of VTN

B. Data preprocessing

The focus of recognition is on the change and movement of hand. Therefore, we locate the face position of the video, so as to focus on the changes of hand shape and hand motion trajectory around the upper body. We use dlib to detect the face of each frame and locate the position of the face. the height of

each frame of picture is selected at a ratio of 1:5. If the position of the presenter's face in a certain frame cannot be detected, we adopt the face position of the previous frame to extract the half-length region. We select different sampling intervals based on the video length dynamically.

C. Video-based dual-channel self-encoding network

We propose a dual-channel convolutional network based on VTN. The RGB sequence frame and the RGB difference sequence frame are respectively used as the input of ResNet-34, and then the frames of each channel are embedded for feature fusion, sent to the decoder, and finally isolated the classification prediction of word sign language, Figure 3 is a dual-channel network structure.

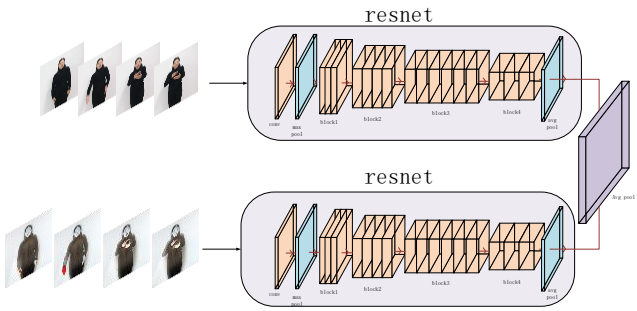

Fig.3 Structure of dual-channel VTN

The input of the dual-channel network is the RGB sequence frame and the RGB difference sequence frame, where the RGB difference is obtained by subtracting the average picture from each frame in the input sequence. Figure 4 Figure 5 Figure 6 are the average image, the original image and the corresponding RGB difference picture. The average picture is the average value obtained from all the frames in the dataset. By subtracting the picture, the RGB difference frame obtained is the part of the relative motion of the current frame. For sign language actions, the result obtained is reflected in the changes of hand movement. Compared with optical flow, the calculation amount of RGB difference is greatly reduced.

Fig.4 mean_image

Fig.5 RGB_image

Fig.6 RGB difference

D. Continuous Sign Language Translation and Decoding Network

For continuous sign language recognition, we use a two-way VTN network to extract video features, merges the features of RGB and RGB difference into a 512-dimensional feature vector, and use the feature vector as the input of the continuous sign language translation module to learn the mapping relationship between sequence and label. The basic framework of continuous sign language translation is Bi-LSTM (Bi-directional Long Short-Term Memory).

Bi-LSTM [21] processes fixed-length data. First we determine the input data according to the length of label word. Assuming that the longest continuous sign language text label in the data set does not exceed N words, the video sequence is converted into an $N * M$ feature vector. If there are less than N clips, the last clip is repeated multiple times to complement N segments, where M is the data dimension of the fully connected layer of the feature extraction network. Bi-LSTM overcomes the problem that LSTM can only learn one-way and ignore future information. The output result at time t depends not only on the previous frame in the action sequence, but also on the upcoming next frame. The model structure is shown in Figure 7.

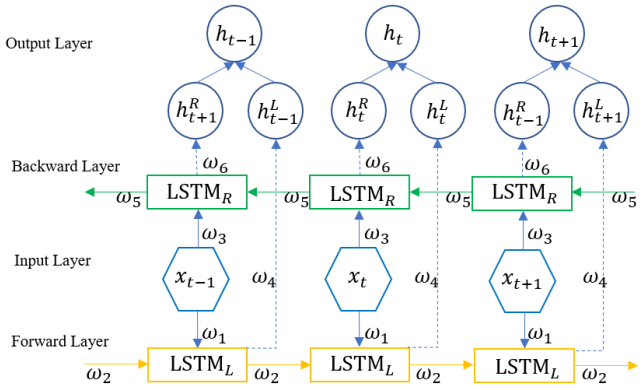

Fig.7 The model of Bi-LSTM

Where X_t is the feature vector obtained by the continuous sign language feature extraction module for the sign language sequence, ω_i represents the weight from one unit to another, and LSTM_L represents the forward LSTM unit. The hidden state obtained is h_t^L, and the expression is shown in (1), LSTM_R represents the reverse LSTM unit, and the obtained hidden state is h_t^R. The expression is shown in (2). The equation (3) indicates output vectors of the two LSTM layers are averaged to obtain h_t.

$$\mathrm{LSTM}_L: \{\mathrm{h}_0^L, \mathrm{h}_1^L \ldots \mathrm{h}_n^L\} \tag{1}$$

$$\mathrm{LSTM}_R: \{\mathrm{h}_0^R, \mathrm{h}_1^R \ldots \mathrm{h}_n^R\} \tag{2}$$

$$h_t = \frac{\mathrm{h}_t^L + \mathrm{h}_t^R}{2} \tag{3}$$

$$\mathrm{Bi-LSTM}: \{\mathrm{h}_0, \mathrm{h}_1 \ldots \mathrm{h}_n\} \tag{4}$$

III. RESULTS AND DISCUSSION

In this section, we mainly introduce the dataset and various experimental settings used in the experiment, and conduct two sets of experiments to verify the advantages of our method:

(1) isolated sign language recognition;

(2) continuous sign language recognition.

A. Dataset

According to the "National Common Sign Language Common Vocabulary", we cooperate with a special education school to construct CSL-BS. The dataset contains two kinds of data: isolated and continuous sentences. The information of our dataset is shown in Table 1. Figure 8 is a sequence of video frames. In addition to the dataset CSL-BS, we integrate the public dataset CSL [12] into the CSL-BS dataset to constructed for experiments, and the fusion dataset is Mixed-CSL.

Table 1 Dataset of CSL-BS

type	category	samples	Video length	people
Station words	230	12880	2-5	7
Bank words	133	7448	2-5	7
Station sentences	55	3080	6-10	7
Bank sentences	44	2464	6-10	7

Fig.8 Frame of CSL-BS dataset

B. Experimental setup

The model was trained using an Intel(R)Core i7,3.8GHZ with 32GB of RAM. The network was created using PyTorch framework and the GPU model is RTX2080ti with 11GB memory. The Adam optimizer was used to train the weight and set for 100 epochs with initial learning rate 0.001. The input clip length was set as 16, with the size 224*224 and the batch_size was set as 8 with the dropout 0.1.

C. Evaluation index

We use WER (Word Error Rate), BLEU (Bilingual Evaluation Understudy) and model response speed as the performance indicators of the recognition model.

WER represents the word error rate. S means substitution, I means insertion and D means deletion, N means total number of words. The larger the WER, the worse the translation effect. The calculation formula is:

$$WER = \frac{S+D+I}{N} \cdot 100\% \quad (5)$$

BLEU is the frequency of cooccurring words in two sentences. The calculation formula is:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \omega_n log P_n\right) \quad (6)$$

Where BP is the penalty factor, l_c is length of the translation sentence, and l_s is length of the standard sentence. When $l_c > l_s$, $BP = 1$. When $l_c \leq l_s$, $BP = e^{1-l_s/l_c}$. P_n represents the degree of matching between the translated text and the standard text, and the calculation formula for P_n is as follows:

$$P_n = \frac{\sum_{C \in candidates} \sum_{n-gram \in C} count_{clip}(n-gram)}{\sum_{C' \in candidates} \sum_{n-gram' \in C'} count(n-gram')} \quad (7)$$

$n-gram$ represents a sequence of n consecutive words, $candidates$ are the set of all sentences that need to be translated, and $count_{clip}(n-gram)$ refers to the minimum number of occurrences of n-gram. ω_n represents the weight value, we use BLEU-4 in experiment. Among them $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 0.25$.

D. Experimental results and analysis

1) Isolated sign language recognition

We carry out experiments based on two-way VTN and I3D respectively, and 15% of isolated sign language samples were tested respectively. The recognition accuracy and average time based on different dataset are shown in Table 2.

Table 2 Comparison of different model

	CSL_BS	Mixed_CSL	Time_cost
Dual VTN	87.9%	89.5%	0.99
I3D	84.2%	86.4%	1.86

Compared with the I3D model, the model in this paper has increased by 3.7% and 3.1% on the CSL-BS dataset and Mixed_CSL dataset respectively, and the recognition speed has increased by 0.87s.

We visualize the results of feature extraction and draw it into a heat map. It can be seen from Figure 9 that the class activation heat map is distributed in a circular shape with the hand as the center. The color of the area where the hand is located is red. For different actions, the network always pays the highest attention to the position of the hand, followed by other areas of the human body, and hardly pays attention to the background image. This is consistent with the expected focus of the sign language action when expressing the hand movement.

Fig.9 Heat charts of feature regions

2) Continuous sign language recognition:

On the basis of isolated sign language recognition, the sign language feature extraction module reads continuous sign language video sequences in the form of a sliding window with a sliding window size of 16 and a step length of 8. The continuous sign language frame is processed into the feature vector form that can be directly processed by the Bi-LSTM layer, and the VTN network and I3D network are respectively used as feature extraction modules. Table 3 shows the performance index comparison of the two models.

Table 3 Comparison of continuous sign language recognition

Method	Accuracy	WER	BLEU-4	Time cost
I3D+seq2seq	71.2%	26.6%	0.709	26.54
VTN+seq2seq	73.5%	24.9%	0.736	13.91

Among the table, accuracy means the recognition result and the label are exactly the same. As is seen from the table, compared with the I3D-seq2seq model, the accuracy of the continuous sign language recognition model proposed is improved by 2.3%, WER is reduced by 1.7%, and BLEU-4 is improved by 0.027. When recognizing a continuous sign

language sentence, the average time used by I3D-seq2seq is 26.54s, which cannot meet the real-time requirements of sign language recognition. The average time of the continuous sign language recognition model constructed is 13.91s, which greatly improves the speed of recognition and obtain good continuous sign language translation performance.

IV. CONCLUSION

This paper proposes a continuous sign language recognition model based on VTN-seq2seq. The two-way network combines the RGB sequence and the RGB difference sequence, which can effectively extract the change information between the action sequences; the application of the self-attention mechanism calculates the weight scores between current frame and all other frames in entire sequence. The weight scores between all frames are used to capture the dependencies between long-distance frames. Multiple independent self-attention modules can learn relevant information in different subspaces, avoid overfitting and achieve parallel computing to reduce running time. In the continuous sign language decoding network, seq2seq model encodes both forward and backward actions at the same time, which better captures the connection between the forward and backward movements of the gesture. Compared with I3D-seq2seq using I3D to extract features, this paper uses two-way VTN to extract continuous sign language features. Both in recognition accuracy and speed be significantly improved. At the same time, this paper constructs the sign language dataset CSL-BS, which includes Sign language is commonly used in station and bank scenes to enrich the data in the field of sign language research. However, the model constructed in this paper has shortcomings in the recognition of similar sign language actions, and the extraction of long-sequence key frames needs to be improved. In future work, this problem will be further studied.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No.61973334), Research Center of Security Video and Image Processing Engineering Technology of Guizhou (China) under Grant SRC-Open Project ([2020]001]),2020 Anhui Province University Outstanding Young Backbone Teacher Visit Study Project(gxgnfx2020161), Rudong Yifu Special Education School. We are also grateful to the High Performance Computing Center of Nanjing Tech University for supporting the computational resources.

REFERENCES

[1] Duarte A, Palaskar S,Ventura L, Ghadiyaram D, DeHaan K, Metze F, Torres J, Giro-i-Nieto X. (2021) How2Sign: A Large-scale Multimodal D-ataset for Continuous American Sign Language[EB/OL]. https://arxiv.org/abs/2008.08143.

[2] Zhou H, Zhou W G, Qi W Z, PU J F and Li H Q.(2021) Improving Sign Language Translation with Monolingual Data by Sign Back-Translation[EB/OL].https://arxiv.org/abs/2105.12397.

[3] Yang Q. (2010) Chinese sign language recognition based on video sequence appearance modeling. In: 2010 5th IEEE Conference on Industrial Electronics and Applications. Taichung, Taiwan. pp: 1537–1542.

[4] Huang J, Zhou W G, Li H Q and Li W P. (2019) Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 29(9): 2822-2838.

[5] Yasir F, Prasad P. W. C, Alsadoon A and Elchouemi A. (2015) Sift based approach on Bangla sign language recognition. In: 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA). Hiroshima, Japan. pp. 35–39.

[6] Liwicki S and Everingham M. (2009) Automatic recognition of fingerspelled words in British Sign Language. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Miami, FL, USA. pp. 50–57.

[7] Buehler P, Zisserman A and Everingham M. (2009) Learning sign language by watching TV (using weakly aligned subtitles). In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. pp. 2961–2968.

[8] Cooper H, Ong E J, Pugeault N and Bowden R. (2012) Sign Language Recognition Using Sub-units. The Journal of Machine Learning Research, 13(1): 2205-2231.

[9] Al-Rousan M, Assaleh K and Talaa A. (2009) Video-based signer-independent Arabic sign language recognition using hidden Markov models. Applied Soft Computing, 9(3): 990–999.

[10] Badhe P C and Kulkarni V. (2015) Indian sign language translator using gesture recognition algorithm. In: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS). Bhubaneswar, India. pp. 195–200.

[11] Kozlov A, Andronov V and Gritsenko Y. (2020) Lightweight Network Architecture for Real-Time Action Recognition. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC '20). New York, USA. pp. 2074-2080.

[12] Wang H J, Chai X J, Hong X P, Zhao G Y and Chen X L. (2016) Isolated Sign Language Recognition with Grassmann Covariance Matrices. Acm Transactions on Accessible Computing, 8(4): 1-21.

[13] Lichtenauer J F, Hendriks E A and Reinders M J. (2008) Sign language recognition by combining statistical DTW and independent classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11): 2040–2046.

[14] Nagarajan S and Subashini T S. (2013) Static hand gesture recognition for sign language alphabets using Edge Oriented histogram and multi class SVM. International Journal of Computer Applications, 82(4): 28-35.

[15] Li D X, Opazo C R, Yu X and Li H D. (2020) Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Snowmass, CO, USA. pp. 1448–1458.

[16] Joze H R V and Koller O. (2019) MS-ASL: A largescale data set and benchmark for understanding American sign language[EB/OL]. https://arxiv.org/abs/1812.01053v1.

[17] Ye Y C, Tian Y L, Huenerfauth M and Liu J Y. (2018) Recognizing American sign language gestures from within continuous videos. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognitio-n Workshops (CVPRW). Salt Lake City, UT, USA. pp. 2145–214509.

[18] Graves A, Fernández S, Gomez F, and Schmidhuber J. (2006) Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In: Proceedings of the 23rd international conference on Machine learning. New York, USA. pp. 369–376.

[19] Kalchbrenner N and Blunsom P. (2013) Recurrent Continuous Translation Models. In: Empirical Methods in Natural Language Processing (EMNLP). Seattle, USA. pp. 1700-1709.

[20] Cho K, Merriënboer B V, Bahdanau D and Bengio Y S. (2014) On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar. pp. 103-111.

[21] FU Z R, WU S X, WU X Y, GU X S. (2021) Human Action Recognition Using BI-LSTM Network Based on Spatial Features. Journal of East China University of Science and Technology, 47(2): 225-232.