Paper:

Dynamic Sign Language Recognition Based on Improved Residual-LSTM network

Da Ma[⋆1], Kaoru Hirota[⋆2], Yaping Dai[⋆3], Zhiyang Jia[⋆4]

[⋆1⋆2⋆3⋆4]School of Automation, Beijing Institute of Technology, No. 5, Zhongguancun South Street, Haidian District, Beijing, China
E-mail: d.ma@bit.edu.cn

Sign language recognition aims to recognize meaningful hand movements. In order to solve the problem of low recognition accuracy during the process of sign language recognition, a dynamic sign language recognition method based on improved Residual-LSTM network is proposed. There are two main innovations in this paper. Firstly, the residual learning mechanism and channel attention mechanism are introduced to improve the convolutional neural network. The improved network can effectively extract enough image information and enhance the ability of spatial feature extraction. Secondly, the long-term and short-term memory network is introduced to extract the temporal features of image sequences while reducing the amount of calculation. In addition, YOLO network is used for hand detection, which greatly reduces the amount of network calculation. The experiment is carried out on SLR data set and the recognition accuracy is 91.2%. The results show that this method can recognize the dynamic sign language accurately.

Keywords: dynamic sign language recognition, computer vision, image processing, deep neural network

1. Introduction

Millions of hearing-impaired people around the world use sign language to communicate. However, due to the lack of systematic learning of sign language, it is difficult for many people to communicate with deaf mutes. In order to help the deaf and dumb people communicate normally in their daily life, it is very important and meaningful to study the dynamic sign language recognition algorithm.

In recent years, with the development of human-computer interaction, many researchers turn their attention to sign language recognition, which makes sign language recognition get some applications[1]-[3]. According to the different input methods, the existing sign language recognition methods are mainly divided into wearable sensor based system and vision based system. Among the two systems, vision-based system has the characteristics of flexibility, scalability and low cost, which is currently popular in the research of gesture interaction technology. Because sign language is constructed by a series of actions and includes fast actions with similar characteristics, static sign language recognition is difficult to deal with the complexity and large changes of vocabulary set in manual actions. Therefore, the research of dynamic sign language recognition is a more effective method to solve related problems. This paper is aimed to study dynamic sign language recognition based on vision system.

According to different feature extraction methods, dynamic sign language recognition can be divided into non deep learning method and deep learning method. Many researchers use non deep learning methods for sign language recognition. For example: Yang et al. [4] use Kinect to collect multi-modal sign language data and used the Hidden Markov Model(HMM) for time series modeling; Yang et al. [5] propose a hierarchical construction algorithm, combined with HMM, to dynamically divide and recognize sign language words; Jangyodsuk et al. [6] first extract the Histogram of Oriented Gradient(HOG) features and trajectory features from sign language data, and then use the Dynamic Time Warping(DTW) modeling to analyze features; Jiang et al. [7] propose the application of dictionary learning and sparse representation in sign language recognition. Besides, many researchers use deep learning methods for sign language recognition in the recent years. For example, Barros et al. [8] construct a multi-channel convolutional neural network(CNN) network; Pigou et al. [9] also achieve sign language feature extraction through CNN network; Huang et al. [10] design a multi-channel 3D-CNN network; Wu et al. [11] use Deep Belief Network (DBN) to process trajectory information, and 3D-CNN to process color images and depth information; Pigou et al. [12] use the improved 3D-CNN to extract sample features and construct bidirectional Long Short-Term Memory(LSTM) network to classify gesture; Huang et al. [13] build a multi-layer attention decoding network. However, due to the complexity of hand shapes and motion changes, the time sequence modeling process of sign language recognition task is complicate, and the amount of data of sign language recognition task is large. The current dynamic sign language recognition methods still have the problem of low recognition accuracy.

A progressive method of dynamic sign language recognition based on improved residual-LSTM network is proposed in this paper. This method is divided into three main parts. The first part is the hand positioning module based on the framework of YOLO, the hand is lo-

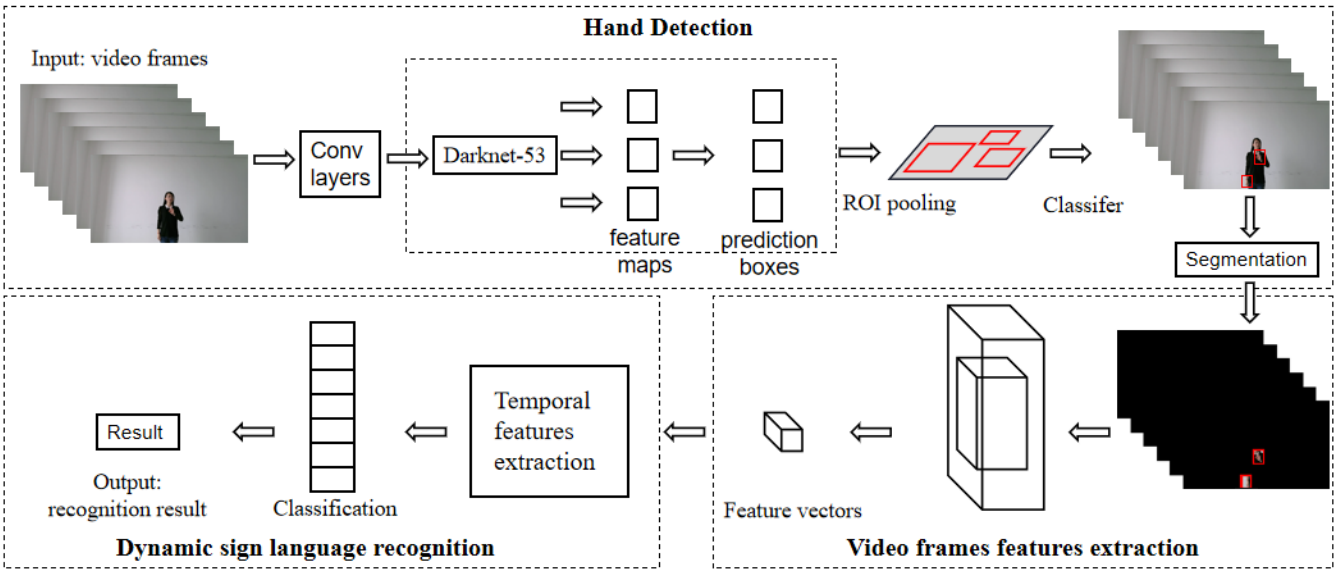

Fig. 1. Framework of our proposed method.

cated in the video frame to reduce the time and space complexity of network computing. The second part is the video sequence features extraction module, the improved Residual-LSTM network extracts spatiotemporal features from video sequences. The third part is dynamic sign language recognition module. Through analyzing each video feature vectors, the dynamic sign language could be recognized effectively.

The rest of the paper is organized as follows. Section 2 introduces our proposed method in detail. Experiments and results analysis are in Section 3. The conclusion and future work is summarized in Section 4.

2. Methodology

This section introduces the details of our proposed method.

2.1. Method Overview

In order to recognize sign language gestures, a method of dynamic sign language recognition based on video sequence is proposed, and its framework is shown in the **Fig. 1**. This method is divided into three main parts. The first part is the hand positioning module based on the framework of YOLO, which is used to capture the position information of the hand. Video frames are trained by convolution layer for feature extraction. The ROI area of the hand is obtained through the target detection network. With the accurate information of hand position, the hand region could be segmented from the background by coding algorithm. After that, the segmented video frames are inputted into the next part for spatial feature extraction. The second part is the video sequence features extraction module, which performs the task of spatial feature extraction with inputting segmented video frames. Each video feature vector will be provided to the third part for analyzing dynamic information of sign language. The third part is dynamic sign language recognition module, which can analyze long-term temporal dynamics and predict the hand gesture label. Through analyzing each video feature vectors, the frames label could be predicted. According to the top label prediction scores, this label will be regarded as the label of video sequence and be outputted as the recognition result. Therefore, the dynamic sign language could be recognized effectively.

2.2. Hand Detection Using YOLO

Hand detection is very important for image segmentation and subsequent recognition module. In order to obtain the accurate information of hand position in frame image, its essential to choose an excellent object detection algorithm. At present, deep learning target detection algorithms can be divided into two categories: one is the regression based target detection algorithm represented by YOLOv3[14], such as SSD, YOLOv4[15], etc., the other is the region recommendation based target detection algorithm represented by Faster R-CNN, such as SSP, etc. Among them, the first kind of target detection algorithm directly uses the detection network to generate the location information and category of the target, and does not generate candidate regions, so it has higher detection speed and meets the real-time requirements. The second algorithm needs to generate candidate regions, which makes the detection speed slow and does not meet the real-time requirements. In gesture recognition, the accuracy and speed of gesture recognition are required to be as high as possible. Therefore, this paper selects YOLOv4 for hand recognition, which can balance the recognition accuracy and speed well. The framework of YOLO algorithm is shown in **Fig. 2**.

As can be seen from **Fig. 2**, the YOLOv4 detection

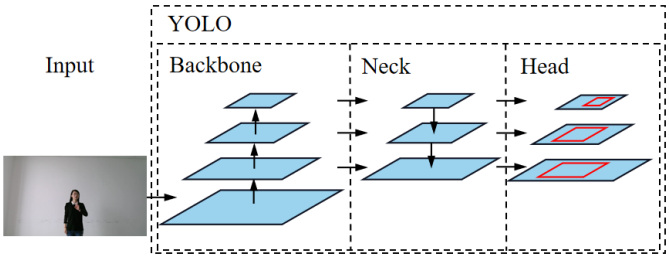

Fig. 2. YOLOv4 detection framework.

framework can be divided into three parts, including the backbone network, neck and head. In order to detect the hand position accurately and quickly, CSPDarknet53 is used as the backbone network to extract the hand features of the input image. CSPDarknet53 has five more CSP modules than darknet53, the backbone network of YOLOv3. The CSP module divides the feature mapping of the basic layer into two parts, and then merges the cross stage hierarchical structure to reduce the amount of calculation and ensure the accuracy. In order to make the prediction frame get larger receptive field and effectively separate the salient features from the data, YOLOv4 adds the SPP module after the CSPDarknet53 structure. SPP structure can effectively increase the feature acceptance range of the backbone network and significantly separate the important feature information. YOLOv4 uses FPN + PAN module as neck layer. Different from single FPN in YOLOv3, YOLOv4 adds a bottom-up feature pyramid after FPN, which contains two PAN structures. FPN layer conveys strong semantic features from top to bottom, while feature pyramid conveys strong positioning features from bottom to top. The improved structure can be used for parameter aggregation of different levels of detectors, which further improves the ability of feature extraction, thus enhancing the ability of hand detection. Finally, YOLOv3 is used as the head of YOLOv4 architecture.

After obtaining the precise hand position, the hand can be segmented from the background through the precise coordinate points of the hand position in the video frame. Therefore, using the YOLO framework to detect the hand, we can get the accurate information of the segmentation position from the background of the frame image.

2.3. Improved Residual-LSTM network

A residual-LSTM model for dynamic sign language recognition based on video sequences is presented in this paper. This method can accomplish video sequence features extraction and learn spatiotemporal features. For the dynamic sign language recognition, different sign language gestures correspond to different videos with different labels. Therefore the hand gesture could be recognized by classifying the labels. Through extracting spatiotemporal feature of different videos and classifying videos with different characteristics, this model can perfectly realize the various dynamic sign language including complex hand gestures. To improve the recognition accuracy of dynamic sign language, the feature sequences are analyzed by the long short-term memory units. After that, the final features are classified by softmax function. The network structure is shown in the **Fig. 3**.

2.3.1. Spatial Feature Extraction Unit

As shown in **Fig. 3**, the input image sequence first passes through the spatial feature extraction network. The spatial feature extraction module is composed of multi-layer convolution layer. In order to avoid the gradient disappearing in the training process and better extract the spatial features of the image, the residual network structure is adopted. The image is used as the input of the model through the two-dimensional channel of $H \times W$, where H and W are the height and width respectively. Then, 2D convolutions are applied with a kernel size of 7×7 on each channels separately. The 2×2 down-sampling is applied on each of the feature maps in the convolution layer, which leads to the same number of features maps with a reduced spatial resolution. After that, the residual network consists of four bottle blocks, and each bottle block consists of three residual blocks. The shortcut connections are inserted to the networks which could be turned into counterpart residual version. The identity shortcuts can be directly used when the input and output are with the same dimensions. In the residual block, the main branch uses three convolution layers. The first is the convolution layer with convolution kernel size of 1×1, which can be used to compress the channel dimension. The second is the convolution layer with convolution kernel size of 3×3. The third is convolution layer with convolution kernel size of 1×1, which can be used to restore channel dimension.

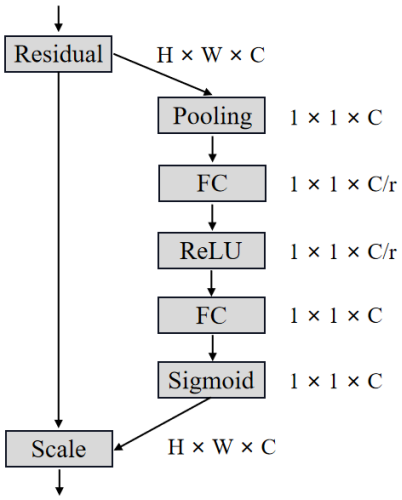

Fig. 4. Structure of channel attention used in this paper.

Each bottle block contains three residual blocks, so there are 37 convolution layers in the spatial feature extraction unit. However, the output of convolution layer does not consider the dependence on each channel. In order to selectively enhance the maximum amount of in-

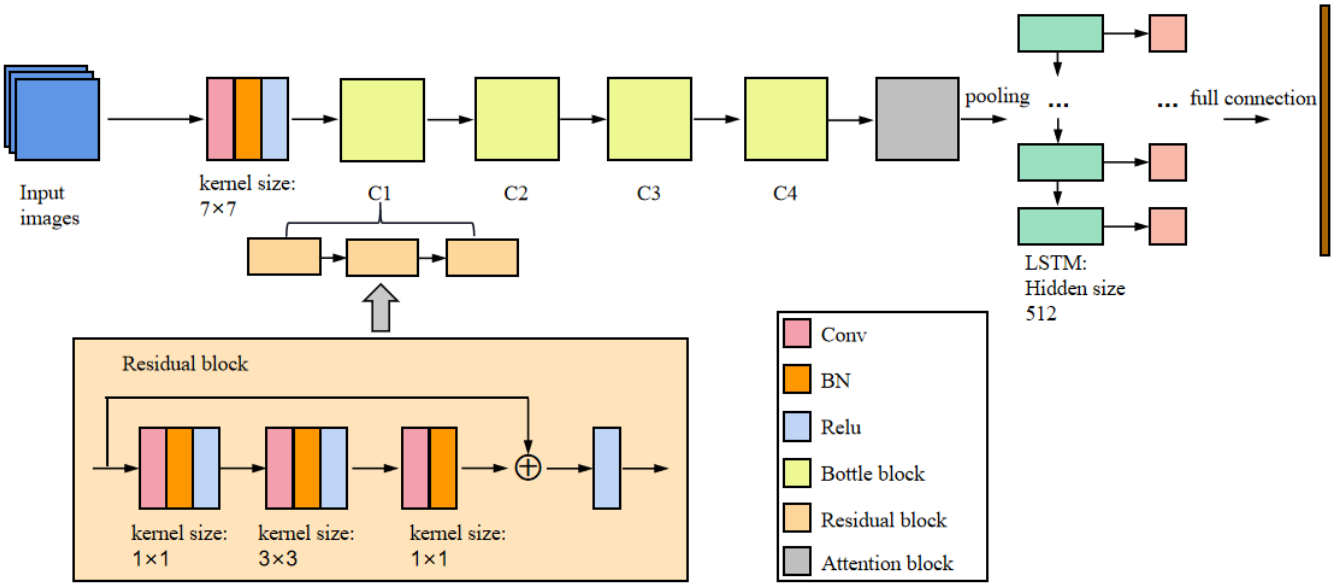

Fig. 3. Structure of improved residual-LSTM network.

formation in the network, the channel attention mechanism is added after the residual block to improve the network. The structure of channel attention block is shown in **Fig. 4**. In the **Fig. 4**, H is the height of input features, W is the width of input features, C is the channel of input features. In the attention block, global average pooling is performed on the input features to get $1 \times 1 \times C$ of features. Then the feature dimension is reduced through the full connection layer and the activation function layer. The channel number is compressed first, and then reconstructed back to the channel number. Finally, the obtained features are input into sigmoid layer to generate attention weights of 0 - 1 between channels. Finally, the features are changed into input feature size through scale layer.

By introducing attention mechanism, we can make full use of the most informative features and suppress useless features. After the multiple layers of convolution and channel attention layer, the input frame has been converted into a 512D feature vector capturing the spatial information in the input frame. After 16 frames of input image are processed, the feature vectors are then fed into an LSTM network for temporal feature extraction.

2.3.2. Temporal Feature Extraction Unit

With the above steps, the spatial features of image sequence are extracted. Then LSTM is used to extract the temporal features. LSTMs are an important part of deep learning models to analyze long-term temporal dynamics for human gesture recognition. Conceptually, the memory cell stores the past contexts, the input and output gates allow the cell to store contexts for a long period of time. Meanwhile, the memory in the cell can be cleared by the forget gate. The structure of LSTM cell used in this paper is shown in **Fig. 5**.

The following formulas are given, including an input sequence $x = \{x_1, x_2, \ldots, x_t\}$, the cell states $c = \{c_1, c_2, \ldots, c_t\}$, and the hidden states $h = \{h_1, h_2, \ldots, h_t\}$. The gates i_t, f_t, o_t and c_t are the input gate, forget gate,output gate, and memory cell activation vectors respectively. The equations for a LSTM cell are as follows:

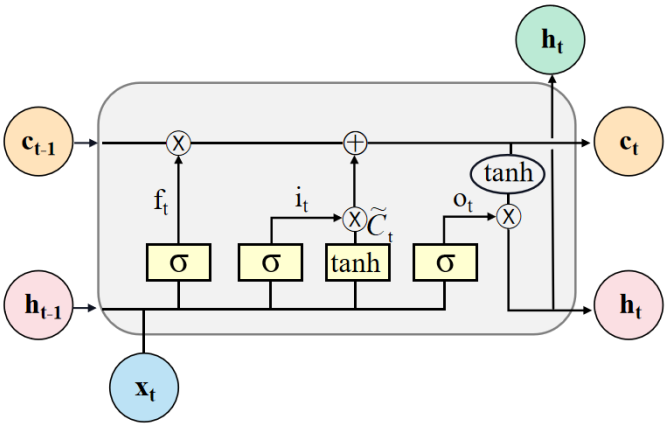

Fig. 5. Structure of LSTM cell used in this paper.

$$i_t = \sigma(w_{xi}x_t + w_{hi}h_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma(w_{xf}x_t + w_{hf}h_{t-1} + b_f), \tag{2}$$

$$\tilde{C}_t = tanh(w_C x_t + w_C h_{t-1} + b_C), \tag{3}$$

$$o_t = \sigma(w_{xo}x_t + w_{ho}h_{t-1} + b_o), \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tilde{C}_t, \tag{5}$$

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

where σ is sigmoid function, and tanh is the hyperbolic tangent function. The forget gate f_t decides when information should be cleared from the memory cell c_t. The input gate i_t decides when new formation should be incorporated into the memory. The tanh layer $\tilde{C}_t$ generates a candidate set of values which will be added to the memory cell if the input gate allows it. Referred to **(5)**, based on the output of the forget gate f_t, input gate i_t and the new candidate values $\tilde{C}_t$, the memory cell c_t is updated. In **(6)**, the output gate o_t controls the status and memory information of the hidden state. Finally, the hidden state is represented as a product between a function of the memory cell state and the output gate. After that, the final features are classified by soft-max. Therefore, the Residual-LSTM network could obtain the full features of the input video.

3. Experiment Result and Analysis

3.1. The Database

In order to prove that this method can recognize dynamic sign language effectively, we choose two datasets for experiment. Oxford hand is used for hand detection experiment, and SLR data set is used for sign language recognition experiment. More details about these two datasets are as follows.

1) Oxford hand dataset

Oxford hand dataset[16] is collected and produced by Oxford University. The dataset contains about 4170 high-quality hand instances, and a total of 13050 hand instances are annotated. When collecting data, there were no restrictions on the posture or visibility of people, nor on the environment. In each image, all hands that can be clearly perceived by humans are annotated by rectangular boxes. Some examples of the Oxford hand dataset are shown in the **Fig. 6**.

Fig. 6. Examples of the Oxford hand dataset

2) SLR dataset

The Chinese sign language recognition data set[17] was collected and produced by the University of science and technology of China. The isolated SLR dataset contains 500 Chinese sign words. Each sign video is performed by 50 signers with 5 times. Since, there are 250 instances for each sign word. The isolated SLR dataset contains 125K labeled video instances and every video instance is annotated by a professional Chinese sign language teacher. Some examples of the SLR dataset are shown in the **Fig. 7**.

Fig. 7. Examples of the SLR dataset

3.2. Network Training

Firstly, yolov4 network is pre-trained with marked Oxford hand dataset. After that, opencv is used to divide the video into several frames, and some useless frames are removed. Then, the pre-trained yolov4 network is used to detect the hand of the selected frames. The hand regions in the video frames are preserved by image clipping. Next, the frames with hand areas are input into the improved residual-LSTM model for training. When training the model, the batch size is set to 16. The stochastic gradient descent (SGD) algorithm is employed for weights update with the initial learning rate $lr = 1e-4$, momentum $m = 0.9$ and weight decay $wd = 1e-5$. The improved residual-LSTM model is implemented based on the deep learning platform Pytorch, the GPU is NVIDIA GTX1080Ti with 11 GB of memory.

After the experimental parameters are set, the improved residual-LSTM model is training for the dynamic sign language recognition, which mainly can extract spatiotemporal features from the input videos for analyzing temporal dynamics to predict the hand gesture label. To evaluate the performance of the model for dynamic sign language recognition, the recognition accuracy is employed as the criterion. The calculation formula of accuracy is shown in the **(7)**.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{7}$$

where TP(True Positive) is the number of positive cases that are correctly classified, TN(True Negative) is the number of negative cases that are correctly classified, FP(False Positive) is the number of positive cases that are not correctly classified, FN(False Negative) is the number of negative cases that are not correctly classified.

3.3. Result and Analysis

After training, the training accuracy of the network on 500 classes of the SLR dataset is 0.998, and the test accuracy is 0.912. Here we compared the proposed method with some other models for sign language recognition.

All the compared models are trained on the SLR datasets. The comparison of dynamic sign language recognition results are shown in **Table. 1**.

Table 1. A Comparison of our method with the other methods on the SLR database.

Methods	Accuracy
BLSTM[18]	0.566
HMM-DTC[19]	0.652
DNN[20]	0.658
C3D[21]	0.735
CNN-LSTM	0.832
3DResNet	0.921
Ours	**0.912**

From the **Table 1**, we can see that our method can achieve better performance (0.912) than only using BLSTM network (0.566), HMM-DTC (0.652), DNN network(0.658), C3D network(0.735) and CNN-LSTM(0.832). Our method achieves a competitive accuracy compared with 3DResNet(0.921). The comparison results show that the proposed method can recognize the dynamic sign language accurately.

4. Conclusions

In this paper, we propose a model for the task of dynamic sign language recognition based on improved Residual-LSTM network. The image sequences obtained after hand detection can reduce the amount of calculation and time complexity of the network. By analyzing the image sequences, the proposed model could effectively recognize different hand gestures with extracting video spatiotemporal features and analyzing features sequence. With our model, it could get a good performance on sign language recognition. In the experimental result on SLR dataset, different sign language could be accurately distinguished. The experimental results demonstrate that our proposed method improves accuracy in dynamic sign language recognition.

In future work, the sign sentence recognition for real world video in complex environment will be considered for further research, and the information of bone points will be introduced and combined with the existing work to carry out the task of dynamic sign language recognition.

Acknowledgements

This work is supported by the National Talents Foundation under Grant No.WQ20141100198.

References:

[1] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 873-891, 2005.

[2] Z. Zafrulla, H. Brashear, P. Yin, et al, "American sign language phrase verification in an educational game for deaf children," in 2010 20th International Conference on Pattern Recognition, 2010.

[3] Z. Ma, H. Yu, W. Chen, et al, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," IEEE Trans. Veh. Technol., vol. 68, no. 1, pp. 121-128, Jan. 2019.

[4] H.-D. Yang, "Sign language recognition using Kinect," Journal of Advanced Engineering and Technology, vol. 8, no. 4, pp. 299-303, 2015.

[5] W. Yang, J. Tao, and Z. Ye, "Continuous sign language recognition using level building based on fast hidden Markov model," Pattern Recognit. Lett., vol. 78, pp. 28C35, 2016.

[6] P. Jangyodsuk, C. Conly, and V. Athitsos, "Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features," in Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments, 2014.

[7] Y. Jiang, J. Tao, W. Ye, et al, "An isolated sign language recognition system using RGB-D sensor with sparse coding," in 2014 IEEE 17th International Conference on Computational Science and Engineering, 2014.

[8] P. Barros, S. Magg, C. Weber, et al, "A Multichannel Convolutional Neural Network for Hand Posture Recognition," in Artificial Neural Networks and Machine Learning C ICANN 2014, Cham: Springer International Publishing, 2014, pp. 403-410.

[9] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017.

[10] J. Huang, W. Zhou, H. Li, et al, "Sign Language Recognition using 3D convolutional neural networks," in 2015 IEEE International Conference on Multimedia and Expo (ICME), 2015.

[11] D. Wu , "Deep Dynamic Neural Networks for multimodal gesture segmentation and recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 8, pp. 1583-1597, 2016.

[12] L. Pigou, A. van den Oord, S. Dieleman, et al, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," Int. J. Comput. Vis., vol. 126, no. 2-4, pp. 430-439, 2018.

[13] J. Huang, W. Zhou, Q. Zhang, et al, "Video-based Sign Language Recognition without temporal segmentation," 32nd AAAI Conference on Artificial Intelligence (AAAI-18), 2018.

[14] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," IEEE Conference on Computer Vision and Pattern Recogbition, 2018.

[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," Computer Vision and Pattern Recognition, 2020.

[16] A. Mittal, A. Zisserman, and P. Torr, "Hand detection using multiple proposals," in Procedings of the British Machine Vision Conference 2011, 2011.

[17] J. Zhang, W. Zhou, C. Xie, et al, "Chinese sign language recognition with adaptive HMM," in 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016.

[18] P. Kumar, H. Gauba, P. Pratim Roy, et al, "A multimodal framework for sensor based sign language recognition," Neurocomputing, vol. 259, pp. 21C38, 2017.

[19] H. Wang, X. Chai, and X. Chen, "Sparse observation (SO) alignment for sign language recognition," Neurocomputing, vol. 175, pp. 674C685, 2016.

[20] T. Kim, "Lexicon-free fingerspelling recognition from video: Data, models, and signer adaptation," Comput. Speech Lang., vol. 46, pp. 209C232, 2017.

[21] T. Liu, W. Zhou, and H. Li, "Sign language recognition with long short-term memory," in 2016 IEEE International Conference on Image Processing (ICIP), 2016.