Volume: 14 Issue: 06 Pages: 5666 - 5671 (2023) ISSN: 0975-0290

Chinese Sign Language Recognition Based on Two-stream CNN and LSTM Network

School of Computer and Communication, Hunan Institute of Engineering,Xiangtan 411104, China
LUOYin
Email: 571816459@qq.com
HU Ying*
Email: huying1983@hnie.edu.cn
LIUDi-kun
Email:2678984@qq.com
LIRui
Email:1254200540@qq.com
YANGMeng-hao
Email:1169352664@qq.com

----------**ABSTRACT**----------
Sign language recognition is the use of computer technology to convert sign language into text or speech to facilitate the communication between deaf-mute people and normal people. This paper takes Chinese sign language words as the research object, and proposes a new method of sign language recognition based on two-stream 3D-CNN and LSTM network. First, the key frame extraction algorithm is used to remove redundant data frames in the original data, and then a two-stream3D-CNN is used to learn local hand change features and global trajectory features at the same time, and aggregated as the feature input of the video clip to the LSTM codec network. In order to focus on the video frames that express the meaning of sign language, a time attention mechanism is introduced in the LSTM encoding and decoding network. On the DEVISIGN-D sign language data set, an experiment was compared with three sign language recognition algorithms, the experimental results show that the method can identify Chinese isolated words sign language very well, with an accuracy rate of 98.4%.

Keywords **:3D Convolutional Neural Network, Attention mechanism,Long and Short-Term Memory Network, Sign language recognition,Key frame**

Date of Submission: April 10,2023 Date of Acceptance: April 27,2023

I. INTRODUCTION

Sign language recognition is the use of computer technology to convert sign language into text or speech to facilitate the communication between deaf-mute and normal people. Sign language as the main communication channel between deaf-mute and normal people, plays a very important role in daily life. In recent years, deep learning has achieved very good results in image classification, target detection and other fields. The Convolutional Neural Network (CNN) model can automatically extract deeper and more comprehensive features of the image, thereby greatly improving the accuracy of image recognition[1]. Sign language words are composed of video sequences, using 2D-CNN network to extract features will lose time information. 3D-CNN network can extract the spatiotemporal features of the video, a major breakthrough has been made in behavior recognition [2], which provides a new way for sign language recognition.

Sign language recognition research is divided into isolated word sign language recognition and continuous sentence sign language recognition. The object of isolated word recognition is a single vocabulary,continuous sentences are meaningful and complete sentences produced by a series of sign languages and gestures. In this paper, we use Chinese isolated word sign language as research object,proposed asign language recognition method based on two-way 3D-CNNand Long and Short-Term Memory (LSTM).The contributions are mainly as follows:

(1)Extract key frames as the input of the network, filter out transitional frames and invalid frames, and reduce redundant information in the data;

(2)Introducing the attention mechanism, focusing on video frames that express the meaning of sign language.

(3)Using 3D-CNN to extract the spatio-temporal features in the sign language video can capture the movement information of the sign language. A dual-stream CNN framework is proposed, which extracts useful features from global motion trajectory information and local palm motion information, and uses feature fusion to classify sign language. Multi-modal data features can increase the accuracy of sign language recognition.

II. RELATED WORK

Traditional sign language recognition methods build time series models by manually extracting features, using time series models such as hidden Markov, conditional random field, and dynamic time warping . Manual feature extraction relies on the designer's experience, and the timing modeling process is cumbersome, and no breakthrough has been made for many years. In recent years, deep learning has achieved remarkable results in the field of computer vision.

Volume: 14 Issue: 06 Pages: 5666 - 5671 (2023) ISSN: 0975-0290

Convolutional neural networks can extract high-level semantic features in images and have more powerful learning and generalization capabilities. The use of deep learning to achieve sign language recognition has attracted researcher's attention concern. Since 2013, some research teams have conducted a series of isolated word sign language recognition studies based on CNN, mainly incorporating multimodal data [3] [4][5] (including depth, skeleton, human key points, etc.), focusing on hand posture features [6], feature fusion [7], and other related optimization strategies, achieving good recognition results.Compared with CNN, the 3D-CNN model adds time-dimensional information and can be better used in recognition tasks with timing relationships. Huang et al. [8] designed a multi-channel 3D-CNN network, through automatic learning and network parameter adjustment, an accuracy rate of 94.2% was achieved on the 25 Chinese sign language word data collected by the Kinect device.Li et al. [9] proposed a gesture recognition algorithm based on 3D-CNN network and large-scale RGB D data, which sent RGB and depth video into the C3D model to extract spatiotemporal features, and finally used SVM for classification.Liang et al. [10] proposed a sign language recognition algorithm based on multimodal data and 3D-CNN networks, and performed convolutional fusion on multiple data sets to verify its effectiveness on large-scale datasets. Zhao et al. [11] proposed a 3D convolutional neural network method combining optical flow processing to improve recognition accuracy. Overall, existing research on 3D-CNN networks mainly uses multimodal data forms such as color, depth, and bone for sign language recognition, and has low generalization on sign language datasets mostly composed of RGB data.

In addition to the convolutional neural network, the recurrent neural network(RNN) cyclically uses the information of the previous moment and the information of the current moment for modeling, can handle time-series tasks well, and has powerful time-series modeling capabilities.Reference [12] incorporated Long Short Term Memory (LSTM) networks, which represent contextual information, into the study of sign language recognition in response to the uncertainty of manually designed features. The motion trajectories of four skeleton joint points were used as network inputs, achieving good recognition results on the Chinese isolated word sign language dataset. Huang et al. [13] proposed a sign language recognition algorithm based on keyframe video sequences to address the issue of redundant information affecting recognition accuracy. The keyframe algorithm was embedded into the RNN network, allowing for different attention to input data and achieving significant recognition results. Lin et al. [14] proposed a method that combines Res-C3D network with mask and LSTM network for skeleton data modeling, and simultaneously processes RGB-D video data, achieving 68.42% recognition accuracy on Challean dataset. Liao et al. [15] proposed a sign language recognition framework based on Bi Long term temporal sequence modeling (BLSTM). Firstly, a detection network was used to segment the hand, and then the segmented hand features were fed into LSTM along with the original RGB data to achieve dynamic long-term feature modeling. Finally, the classification results were output, achieving accurate recognition on two large Chinese sign language isolated word datasets.

III. PROPOSED WORK

3.1 System architecture

Isolated words contain 50-200 frames of images, most of these video frames are excessive frames, and there are not many key frames that can express semantic features.The network structure designed in this paper first extracts the key frames of the sign language video, filters out the transition frames and invalid frames, and reduces the redundant information in the data.Then the sign language features are extracted in two ways, one way is to directly input the video data into the 3DCNN and LSTM models to extract the arm's movement trajectory information to obtain global features,the other way first passes the video data through the YOLOv4 model to detect and separate the hands region, and then input the hand region data into the 3DCNN and LSTM model to extract the local features of the gesture region.Finally, the features extracted from the two paths are merged to complete the recognition of sign language actions.Fig.1 shows the sign language recognition framework.

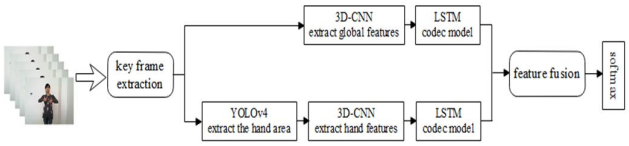

Figure 1.Sign language recognition framework

3.2 Extract key frames

At a recording speed of 30 frames per second, the number of frames for a sign language isolated word is in the range of 50-200. Most of these video frames are excessive frames, and there are not many key frames that can express semantic features. When the image content of two adjacent frames in the video changes greatly, it is considered as a key frame.

Use openpose[16] pose extraction library to locate the joint point coordinates of the human body in the video data, the coordinate positions of the 15 joint points of the human body can be determined, Fig.2 shows the schematic diagram of human joint coordinate extraction.The inter-frame similarity algorithm based on joint point coordinates finds the distance between adjacent frames, and finally takes the first 30 frames of images as the key frame according to the size of the inter-frame distance. Fig.2 shows theschematic diagram of human joint coordinates.

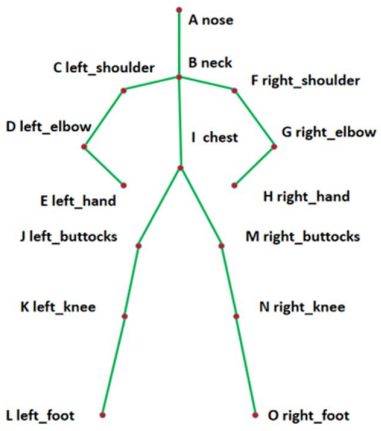

Figure 2.Schematic diagram of human joint coordinates

Volume: 14 Issue: 06 Pages: 5666 - 5671 (2023) ISSN: 0975-0290

Because sign language communication mainly occurs in the upper body, only 9 joint points of the upper body are selected for calculation. Select the coordinates of the neck B(x, y) point as the reference coordinates of the origin of the picture, and calculate the relative coordinates of the 9 joint points to B. Then the coordinates of the 9new joint points can be expressed as:

$$\begin{aligned} A' &= A(x,y) - B(x,y) \\ B' &= B(x,y) - B(x,y) \\ &\cdot \\ &\cdot \\ O' &= O(x,y) - B(x,y) \end{aligned} \quad .(1)$$

Use Euclidean distance to calculate the distance between adjacent frames,Euclidean distanceis defined as:

$$dis(i+1,i) = \sqrt{(A'_{i+1} - A'_i)^2 + (B'_{i+1} - B'_i)^2 + \cdots (O'_{i+1} - O'_i)^2} \quad (2)$$

3.3YOLOv4 algorithm

YOLOv4 [17] model is optimized on the basis of the YOLOv3 target detection network architecture. Fig.3 shows the YOLOv4 network structure.The backbone network draws on the network structure of CSPNet and builds the basic network model CSPDarknet53. The introduction of Mish activation function, SPP, PANet and other structures enhances the learning ability of the network, and at the same time optimizes the data enhancement, loss function, etc., reduces the amount ofcalculation, and achieves the best balance between detection accuracy and detection speed.

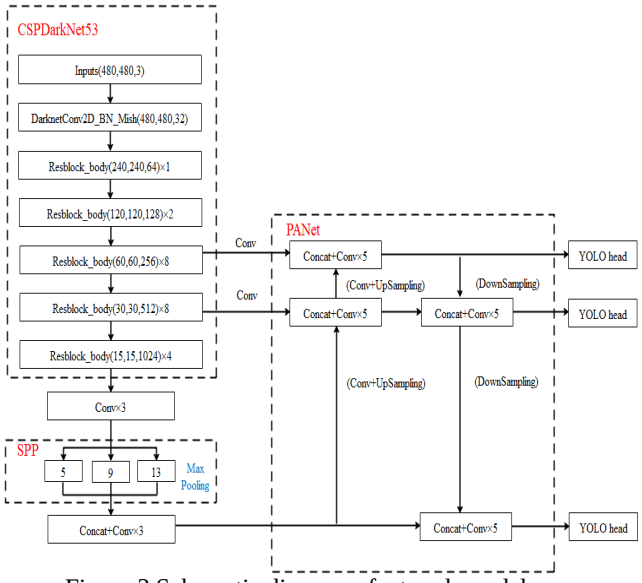

Figure 3.Schematic diagram ofnetwork model

The 9 priori boxes used by the YOLOv4 model are calculated on the VOC data set, while the hand area of the sign language data set is dominated by small targets,the size of the original priori box and the size of the target in the data set are quite different. As a result, the accuracy of the detection model is not high, so the K-means algorithm is used to perform clustering analysis on the data set, and 9 new priori boxes are obtained respectively (8, 14), (10, 17), (13, 21), (15, 27), (20, 34), (27, 45), (39, 67), (59, 108), (116, 203).

3.43D Convolutional Network

In this paper, referring to the literature [18], a 3D-CNN network architecture is designed, and the 30-frame key frames input into the 3D-CNN to extract information in the two dimensions of time and space. For sign language recognition tasks, the 3D-CNN network architecture has made some modifications on the C3D[19] network structure. The network parameters are shown in Table 1. The structure consists of 5 convolutional layers, 4 maximum pooling layers and 2 full Connection layer composition.

Using the center cropping method, each picture is processed to a unified size of 160×120, and then the pictures are binarized.The dimension of the input data of the 3D-CNN model is 160×120×30×1, and the length and width of each frame of image are 160 and 120 respectively. the number of frames of the video sequence is 30, the image is a single-channel binarization image. The 3D-CNN network architecture has a large amount of parameters and requires sufficient sample data for training. In this paper ,we uses the transfer learning method to solve the problem of a small number of training samples. The UCF-50 behavior recognition database collects 50 action categories, and uses the UCF-50 data set to pre-train the 3D-CNN network model to learn the underlying features of the image as the initialization parameters of the sign language recognition network model.

Table 1:3D Convolutional Networkstructure

network layer	convolution kernel	step size	output channels	output size
Conv1	3×3×3	1×1×1	32	158×118×28
Pool1	2×2×2	2×2×2	32	79×59×14
Conv2	3×3×3	1×1×1	64	77×57×12
Pool2	2×2×1	2×2×1	64	38×28×12
Conv3	3×3×3	1×1×1	128	36×26×10
Pool3	2×2×1	2×2×1	128	18×13×10
Conv4	3×3×3	1×1×1	256	16×11×8
Pool4	2×2×1	2×2×1	256	8×5×8
Conv5	3×3×3	1×1×1	512	6×3×6
Fc6	—	—	—	1024
Fc7	—	—	—	1024
Softmax				50

3.5 LSTM-Attention network

Sign language words are composed of video sequences and have temporal and spatial features,the CNN network only extracts spatial features,using LSTM's temporal expression capabilities for temporal coding, the temporal features of sign language words can be obtained. At the encoding end, the features extracted by the CNN network are input frame by frame in chronological order. At the decoding end, the temporal and spatial sign language features obtained from the encoding end are used as initialization. Each decoding moment judges the optimal output based on the output of the previous decoding moment, so that the context within the sign language words can be constructed, and finally the corresponding expression of the sign language sample can be obtained.

The attention mechanism [20] learns the human visual system, can learn the weight distribution of each moment, and screen out more important information. Each sign language word obtains 30 key frames through preprocessing, but the contribution of each frame to sign language semantics is different, and different frames should be given different weights. In this paper, an attention mechanism is introduced in the LSTM network to achieve different weights for

Volume: 14 Issue: 06 Pages: 5666 - 5671 (2023) ISSN: 0975-0290

different video frames. Figure 4 showsthe LSTM network based on the attention mechanism.

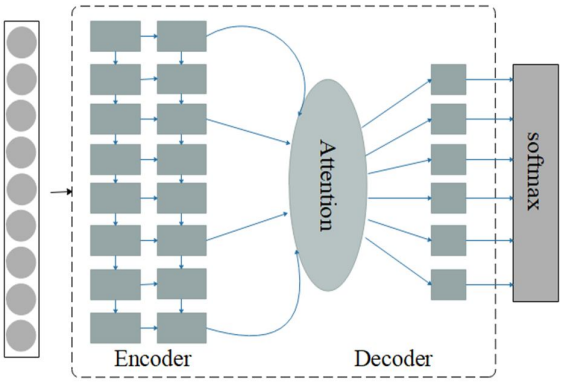

Figure 4.LSTM-Attention networkstructure diagram

This paper adopts the global attention mechanism[21],the global attention mechanism considers all hidden states when calculating the semantic vector. Semantic vector $c_t$is defined as:

$$c_t = \sum_s \alpha_{ts}\, \bar{h}_s \qquad (3)$$

Where$s$ represents the sequence number of the hidden layer vector at the encoding end, $\bar{h}_s$represents the hidden layer state at the encoding end, and α_{ts} represents the attention weight of a source state at$t$ time. $\alpha_{ts}$is defined as:

$$\alpha_t(s) = align(h_t, \bar{h}_s) = \frac{exp\left(score\left(h_t, \bar{h}_s\right)\right)}{\sum_{s'=1}^{S} exp\left(score\left(h_t, \bar{h}_{s'}\right)\right)} \quad (4)$$

$$score\left(h_t, \bar{h}_s\right) = h_t \bar{h}_s \quad (5)$$

Combine the hidden layer state h_t and the semantic vector c_t to complete the combination of the two vector information,$\tilde{h}_t$ is defined as:

$$\tilde{h}_t = tanh(W_c[c_t; h_t]) \quad (6)$$

Where W_c is the weight of the fully connected layer.Figure 5 shows theprinciple of global attention mechanism.

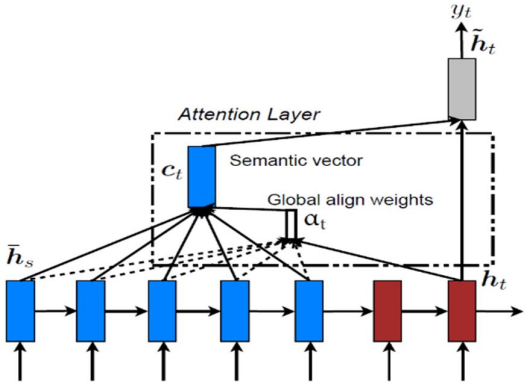

Figure 5.Schematic diagram of global attention mechanism

3.5 Multimodal fusion

The global features of sign language and the local features of the hand region are respectively input to the encoding and decoding network of LSTM, and at each decoding time t, the respective output probability vectors $P_{global}(w)$ and $P_{local}(w)$ are obtained respectively, then the outputs of thetwo models are fused with fixed weights[22]. $P(w)$is defined as:

$$P(w) = \alpha * P_{global}(w) + (1-\alpha) * P_{local}(w) (7)$$

w represents the word element,the weight α is continuously adjusted according to the test set to obtain the optimal value.$P(w)$ represents the probability distribution of each dictionary element at the current decoding moment.Calculate the cross entropy between p and the decoded label at the current moment to obtain the loss function value, and select the dictionary element corresponding to the maximum probability as the output at the current decoding moment.The output $y(t)$ at time t is defined as:

$$y(t) = argmax(P(w)) (8)$$

IV. Experiment Results and Discussion

4.1 Operating environment and data set

The hardware configuration of the server: CPU is Intel Xeon E5-2680, 128GB memory, 4 NVIDIA TITAN XPGPUs. The running software environment is the Ubuntu16.04 operating system, which builds the tensorflow2.0 deep learning framework.

This article usesthe public sign language data set DEVISIGN-D[23] , the data set consists of 500 vocabulary,each sign language vocabulary records 12 sign language videos, recorded by 8 people, a total of 6000 videos, the frame rate is 30fps.Choose 50 common words in DEVISIGN-D as experimental samples, Randomly divide it into training set and test set according to 4:1.

4.2 Training parameters

Each iteration randomly selects 32 video samples, the loss function uses cross entropy, the Adam optimizer is used, and the momentum is 0.9. The learning rate is set to 0.001, the convolutional layer uses the ReLU activation function, and the network parameters use batch normalization.encoding length is 512,decoding length is 256.

Figure 6 shows loss value change curve. The loss value of the test set decreases as the number of iterations increases, indicating that the model does not have an overfitting phenomenon.

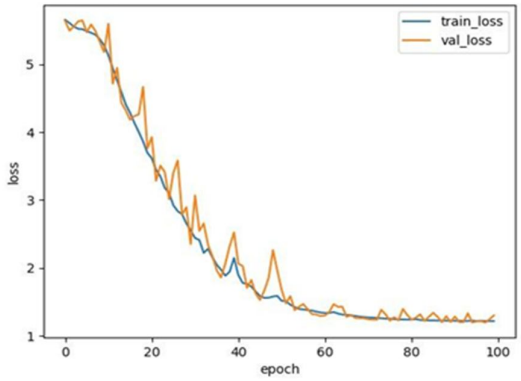

Figure 6.Loss value graph

Volume: 14 Issue: 06 Pages: 5666 - 5671 (2023) ISSN: 0975-0290

4.3 Experiment analysis

In order to verify the feasibility of the model proposed in this article, comparative experiments were carried out with GMM-HMM [24], 3D-CNN [25] and CNN-LSTM [26] three sign language recognition methods under the same parameter settings and data sets. The experimental results are shown in Table 2, indicating that the use of deep learning methods has more advantages than traditional methods. The experimental results verify the effectiveness of the model proposed in this article.

Table 2: Comparison of different training algorithms

Model	top-1accuracy	top-5 accuracy
GMM-HMM	0.645	0.726
3D-CNN	0.852	0.872
CNN-LSTM	0.885	0.905
Proposed work	0.936	0.988

V. CONCLUSION AND FUTURE WORK

This paper proposes a method for isolated word sign language recognition based on two-way CNN and LSTM,relying on the YOLOv4 model to obtain the gesture area, the dual-channel CNN-LSTM is used to fuse the global motion information and the local hand shape change information, and the model fusion method is used for sign language recognition.Use the inter-frame difference method to extract the key frames as the input of the network to reduce the redundant information in the data. An attention mechanism is introduced in the LSTM network, focusing on video frames with rich voice information.The experimental results verify that the method proposed in this article has a high accuracy rate for isolated word sign language recognition, but the method proposed in this article still has some shortcomings, The future work mainly focuses on the following two aspects:(1) Sign language words are composed of video sequences, and using 2D-CNN network to extract features will lose time information. Introduce the 3D-CNN network to extract the spatiotemporal features of sign language videos.(2) Regarding sign language recognition as a task similar to video description, refine the label granularity to better perceive the difference between two sign language words with similar actions.

ACKNOWLEDGEMENTS

This study was supported by 2021 Hunan Provincial Science and Technology Innovation Talents Plan College StudentScience and Technology Innovation and Entrepreneurship Project ”Hunan Institute of Engineering New Engineering Talents Science and Technology Innovation and Entrepreneurship Ability Training Base”(XKJ [2021] No. 40, Project No.: 2021RC1011), 2022 National College Student innovation and entrepreneurship training program project”Research and implementation of traffic sign recognition in complex urban scenes”,2022 Scientific research project of Hunan Provincial Department of Education(22B0735)”Research on Deep Learning Method of Sign Language Recognition Based on Attention Mechanism”.

REFERENCES

[1]. Alex K, Ilya S, et al, Image Net Classification with Deep Convolutional Neural Networks, Neural Information Processing Systems (NIPS), Lake Tahoe, USA, 2012,1097-1105.

[2]. Ji S W, Xu W, et al, 3D convolutional neural networks for human action recognition, Pattern Analysis and Machine Intelligence,35(1),2013,221-231.

[3]. Tang A,Lu K,et al, A real-time handposture recognition system using deep neural networks, ACM Transactions on Intelligent Systems and Technology, 6(2),2015: 1–23.

[4]. Hossen M A,Sultana S,et al, Bengali sign language recognition using deep Convolutional Neural Network, International Conference on Informatics, Informatics,Elec-tronics& Vision, Kitak yushu, Japan, 2018,369–373.

[5]. Zhang H W,Hu Y,Zou Y J,et al, Fingerspelling Identification for American Sign Language Based on Resnet-18,Int. J. Advanced Networking and Applications,1(13),2021,4816-4820.

[6]. Kim S, Lee KB,Ji Y H,An effective sign language learning with object detection based ROI segmentation, IEEE International Conference on Robotic Computing (IRC), Laguna Hills, USA,2018,330-333.

[7]. Hu H z, Zhou W G,Li H Q, Hand-model-aware sign language recognition, Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, MenloPark, 2021, 1558-1566.

[8]. Huang J,Zhou W,Li H,et al,Sign language recognition using 3d convolutional neural networks, Proceedings of Multimedia and Expo(ICME),Turin,Italy, 2015,1-6.

[9]. Li Y, Miao Q, et al, Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model, IEEE Transactions on Circuits and Systems for Video Technology, 28(10) ,2018,2956–2964.

[10]. Liang Z J, Liao S B, et al, 3D convolutional neural networks for dynamic signl anguage recognition, The Computer Journal,61(11), 2018,1724–1736.

[11]. Kai Z, Zhang K J, et al, Real-time sign language recognition based on video stream, International Journal of Systems, Control and Communications,12(2),2021,158-174.

[12]. Liu T, Zhou W G, Li H Q, Sign language recogntion with long short-term memory, IEEE International Conference: on Image Processing (ICI P),Phoenix,USA,2016,2871–2875.

[13]. Huang S L, Mao C S, et al, A novel chinese sign language recognition method based on key frame-centered clips, IEEE Signal Processing Letters, 25(3),2018, 442–446.

[14]. Lin C, Wan J, Liang Y Y, etal. Large-scale isolated gesture recognition using a refined fused model

Volume: 14 Issue: 06 Pages: 5666 - 5671 (2023) ISSN: 0975-0290

based on maskedResC3D network and skeleton LST M, The 13th IEEE International Conference on Automatic Face & Gesture Recognition , Xi'an, China, 2018,52–58.
[15]. Liao Y Q, Xiong P W, et al,Dynamic sign language recognition based on videosequence with BLSTM-3D residual networks, IEEE Access, 7, 2019,38044–38054.
[16]. Jangyodsuk P, Conly C, Athitsos V, Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features, Proceedings of Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments, Petra ,2014,1-6.
[17]. Bochkovskiy A,Wang C ,YLiao H,YOLOv4:Optimal speed and accuracy of object detection, Computer Vision and Pattern Recognition, 17(9),2020, 198-215.
[18]. Huang J, Zhou W, et al, Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition,IEEE Transactions on Circuits and Systems for Video Technology, 29(9),2018, 2822-2832.
[19]. Du T,Lubomir B, Rob F, et al,Learning Spatiotemporal Features with 3D Convolutional Networks,2015 International Conference on Computer Vision, Santiago, Chile, 2015,4489-4497.
[20]. Huang J, Zhou W, et al, Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition, IEEE Transactions on Circuits and Systems for Video Technology, 29(9),2018,2822-2832.
[21]. Du T,Lubomir B, Rob F, et al, Learning Spatiotemporal Features with 3D Convolutional Networks,2015 International Conference on Computer Vision, Santiago, Chile, 2015,4489-4497.
[22]. Mao C S, Research on Chinese Sign Language Word Recognition Method Based on Convolutional Networks and Long Short Term Memory Networks, master diss., University of Science and Technology of China,Hefei,2018.
[23]. Wang H J, Chai X J, Hong X P, et al, Isolated Sign Language Recognition With Grassmann Covariance Matrices,ACM Transactions on Accessible Computing, 8(4),2016, 14-21.
[24]. Wang H,Chai X,Zhou Y,et al,Fast sign language recognition benefited from low rank approximation,The11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition,ljubljana, Slovenia,2015,1-6.
[25]. Huang J, Zhou W,LiH,et al, Sign Language Recognition using 3D convolutional neural networks, IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA,2018:1-6.
[26]. Han N J,Research on Sign Language Recognition Method Based on Deep Learning ,master diss.,Jilin University,Changchun,2021.