SUGISAKA MASANORI

Journal of Robotics, Networking and Artificial Life
Vol. 9(3); December (2022), pp. 268–274
ISSN (Online):2352-6386; ISSN (Print):2405-9021
https://alife-robotics.org/jrnal.html

Research Article

Sign Language Recognition Based on Deep Learning with Improved (2+1)D-ResNet

Yueqin Sheng[1], Qunpo Liu[1], Ruxin Gao[1], Naohiko Hanajima[2]
[1]*School of Electrical Engineering and Automation, Henan Polytechnic University, 2001 Century Avenue, Jiaozuo, Henan 454003, China*
[2]*College of Information and Systems, Muroran Institute of Technology, 27-1 Mizumoto-cho, Hokkaido, Hokkaido 050-8585, Japan*

ARTICLE INFO

Article History
Received 24 November, 2021
Accepted 18 September 2022

Keywords
Sign language recognition
(2+1)D convolution
3D convolution
CELU activation function

ABSTRACT

Sign language is an important communication tool for deaf and hearing-impaired people. The study of sign language recognition can not only promote the communication between deaf-mutes and normal people, but also push the development of intelligent human-computer interaction. Sign language recognition based on deep learning has advantages in processing large scale dataset. Most of them use 3D convolution, which is not conducive to optimization. In this paper, an improved (2+1)D-ResNet model is proposed for isolated word recognition. The model convolves the video frame sequence in space and time dimensions and optimizes the parameters respectively. Based on CELU activation function, the accuracy of sign language recognition is improved effectively. The validity of proposed algorithm is verified on CSL dataset..

1. Introduction

Sign language is an important tool for deaf-mutes to communicate, but most normal people have not learned it, which makes it difficult for deaf-mutes to communicate with others.

Different countries and regions use different sign language. Even under the same standard, there are great difference in action made by different signers because of left-handed or right-hander and speed of motion. Besides, part of sign language motion is obscured by hands, so sign language recognition (SLR) is a very challenging task. According to the type of sign language motion, the study of SLR can be divided into isolated word recognition and sentence recognition. Isolated word recognition corresponds to the sign language action of each word. Sentence recognition corresponds to the sign language action of a sentence, which involves not only the sequential connection between words but also the grammar of sign language. This paper studies SLR based on isolated words. Fig. 1 shows a partial frame of the sign language "situation".

2. Study on Sign Language Recognition

The research on SLR can be traced back to the 1980s. Traditional SLR methods mainly include Hidden Markov

Fig. 1. Example diagram of sign language

Corresponding author's E-mail: 212007010029@home.hpu.edu.cn, lqpny@hpu.edu.cn, gaoruxin@hpu.edu.cn, hana@mondo.mech.muroran-it.ac.jp
URL: www.hpu.edu.cn, www.muroran-it.ac.jp

Model, Dynamic Time Warping (DTW) and Conditional Random Field. Wang et al [1]. achieved 91% recognition accuracy in a data set containing 370 words based on hidden Markov model and gaussian mixture model. Yan et al [2]. improved the traditional DTW by combining dynamic trajectory with type information of key sign language. It is better than traditional DTW in speed and accuracy.

Traditional methods can only solve the problem of SLR in a certain scale dataset. In the current era of big data, SLR based on deep learning is mainstream research trend.

Liu et al [3]. proposed a SLR model based on long short-term memory, which took the motion trajectories of four joints as input. Using skeleton data alone may ignore facial features. Pu et al [4]. obtained the gesture changes of the video through 3D-Convolutional Neural Network (CNN) and used the shape context to describe the trajectory characteristics of the joint to construct a SLR system with two-channel data, which achieved good results in their self-made data set. However, 3D convolution is difficult to optimize, slow and requires high hardware.

3. Sign Language Recognition Model Based on Improved (2+1)D-ResNet

3.1. *(2+1)D convolution*

In static SLR, 2D-CNN plays an irreplaceable role. 3D-CNN that introduces space-time dimension promotes the progress of dynamic SLR. However, both of them have shortcomings. 2D-CNN cannot process the information of time series. 3D-CNN has many parameters, large computation, slow speed and high requirements for hardware.

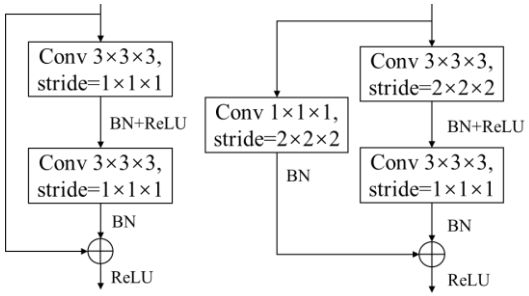
Based on the above problems, Tran et al [5]. proposed a spatio-temporal feature extraction method that optimizes the 3D convolution kernel into (2+1)D convolution kernel under the situation that 3D convolution has been applied to ResNet. 2D network is limited in its ability to process video tasks, while 3D network has a large number of parameters. Mixed convolution can achieve performance equivalent to 3D network with fewer parameters. (2+1)D performs structural decomposition of the spatio-temporal expression so that additional nonlinearity can be obtained.

Fig. 2 shows two connection modes of 3D residual blocks. Each residual block is composed of two convolution layers. If x represents the input data size of $3\times L\times H\times W$, where L represents the number of frames, H and W represent the height and width of video frames respectively, and 3 is the RGB channel of image, the output of ith residual block can be obtained by Eq. (1)

$$z_i = z_{i-1} + F\left(z_{i-1};\theta_i\right), \tag{1}$$

where z_{i-1} is the output of (i-1)th residual block; $F\left(z_{i-1};\theta_i\right)$ is the output obtained through two convolution layers and two activation functions.

When the 3D convolution kernel is split into (2+1)D



(a) Residual connection mode 1 (b) Residual connection mode 2

Fig. 2. 3D residual block structure

convolution kernel, the hyperparameter M_i is introduced. (2+1)D-ResNet uses M_i two-dimensional space convolution kernels with size of and $N_{i-1}\times 1\times d\times d$ and N_i one-dimensional time convolution kernels with size of $M_i\times t\times 1\times 1$ to replace N_i three-dimensional convolution kernels with size of $N_{i-1}\times t\times d\times d$, so as to maintain approximately the same number of parameters as the three-dimensional residual network. Eq. (2) is the relation of parameter quantity. M_i can be obtained from Eq. (3).

$$N_{i-1}\times t\times d^2\times N_i = N_{i-1}\times d^2\times M_i + M_i\times t\times N_i, \tag{2}$$

$$M_i = \frac{td^2N_{i-1}N_i}{d^2N_{i-1}+tN_i}. \tag{3}$$

When the input is single channel, the 3D convolution kernel and (2+1)D convolution kernel are shown in Fig. 3. The left is the 3D convolution kernel with the size of $t\times d\times d$, where t represents time depth and d represents the height or width of the space. The convolution is performed in both spatial and temporal dimensions. The right is the (2+1)D convolution kernel formed by decomposing 3D convolution kernel. The convolution is performed firstly in the spatial dimension and then in the time dimension. The number of 2D convolution kernel after decomposition is M_i.

The two connection modes in Fig. 2 correspond to the two connection modes (a) and (b) in Fig. 4 respectively

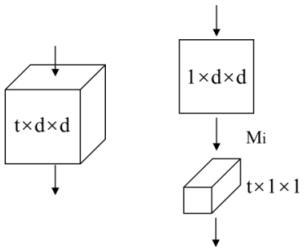
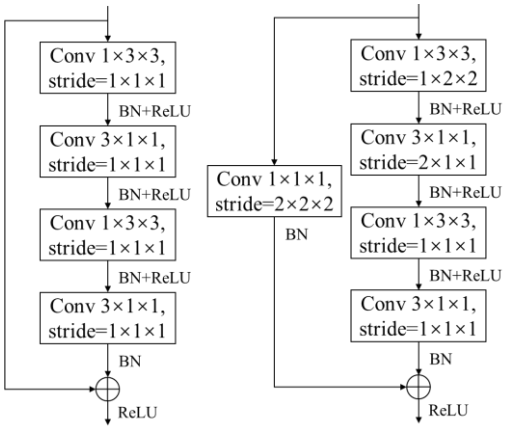


Fig. 3. 3D convolution kernel and (2+1)D convolution kernel

after convolution kernel decomposition. The first basic block of Conv3_x, Conv4_x and Conv5_x of the model in this paper uses the connection mode 2 to change the size of the feature map. The other basic blocks all use the connection mode 1 as shown in Fig. 4(a).

3.2. *Bach Normalization Layer*



(a) Residual connection mode 1 (b) Residual connection mode 2

Fig. 4. (2+1)D residual block structure

In the training process of neural network, the change of parameters of each layer will affect the input of the next layer and the data distribution of each batch will also change. As a result, the neural network needs to learn different data in each iteration, which increases the difficulty of network learning and the risk of network overfitting. In order to solve the above problems, Ioffe et al [6]. proposed a data processing method named Batch Normalization (BN) in 2015. Network training generally adopts mini-bach training method. The whole data set is divided into several batches. Each batch contains multiple groups of data. During training, data is input in batches and optimized once. The advantage of this is that the data set can be optimized multiple times instead of once for each iteration, which speeds up the training of the network. During each batch operation, two additional parameters are introduced to realize the BN operation. Its calculation formula is:

$$\mu_\beta = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{4}$$

$$\sigma_\beta^{\ 2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_\beta\right)^2, \tag{5}$$

$$\hat{x}_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^{\ 2} + \varepsilon}}, \tag{6}$$

$$y_i = \gamma \hat{x}_i + \beta, \tag{7}$$

where Eq. (4) calculates the mean value of data. Eq. (5) calculates the variance of input data. Eq. (6) carries out data normalization. ε is a constant, which is to prevent the calculation of Eq. (6) from being invalid when the variance is 0. The value is 0.00005. Eq. (7) carries out data offset. γ and β are the parameters for extension and translation. These two parameters are learnable parameters introduced into the network by BN. After the normalization of BN layer, its distribution will be closer to the origin, so that the activation function can obtain a large gradient. The distribution of data will also become denser. Dense data is easier to fit and less likely to overfit. In addition, the distribution of all kinds of data tends to be unified, which improves the generalization performance of the network.

3.3. *Optimization of activation function*

The original (2+1)D-ResNet model uses ReLU activation function. ReLU is an activation function commonly used in neural networks, characterized by fast computing speed and good performance. However, when input $x<0$, the function output is 0. The loss gradient disappears during back propagation, resulting in the failure of parameter updating. To solve this problem, the improved R(2+1)D model in this paper selects CELU [7] as the activation function. CELU is a continuous and differentiable exponential smoothing function with nonlinear turning point which is beneficial to the convergence and generalization of neural networks. The calculation formula of ReLU activation function is shown in Eq. (8). The calculation formula of CELU activation function is shown in Eq. (9). α is a constant that avoids vanishing gradient. In this paper, the value of α of CELU activation function

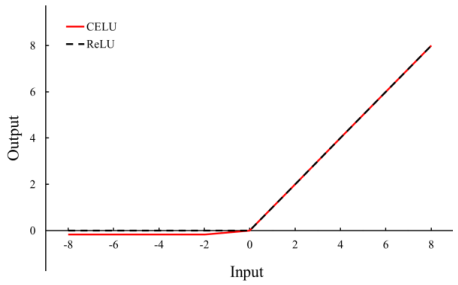


Fig. 5. Activation function curves of CELU and ReLU

is 0.05. The output comparison between ReLU and CELU is shown in Fig. 5.

$$\mathrm{ReLU}(x)=max\{0,x\} \tag{8}$$

$$\mathrm{CELU}(x,\alpha)=max\left\{\alpha\left(\exp\left(\frac{x}{\alpha}\right)-1\right),x\right\} \tag{9}$$

3.4. *Improved (2+1)D-ResNet18 model*

There are 18, 34, 50, 101 and 152 layers of networks in the ResNet family. We select 18 layers of network to build our model. The overall structure of the improved (2+1)D-ResNet18 model proposed in this paper is shown in Fig. 6. The image sequence firstly enters the fully connected layer and max pooling layer to extract input features and reduce the size of the feature map. Then, features are fed into four improved (2+1)D residual blocks successively to extract higher-level features. After that, high-level features are sent to softmax classifier through average pooling layer and fully connected layer for outputting class.

4. Experimental Results and Analysis

The data set we used is CSL [8] isolated word sign Language dataset from University of Science and Technology of China, which contains 500 commonly used sign language words. Due to the large amount of original data, long training time and high requirements on hardware, we selected 100 words for the experiment. Each category is recorded by 50 signers for 5 times of sign language video, so there are 250 video samples for each category. Each sample is composed of 16 frames sampled evenly from each video, so there are 100×250=25000 samples in total. The details of data are shown in Table 1.

Table 1. The number of data used in the experiment

Class	Total	Training	Validation	Test
1	250	175	50	25
2	250	175	50	25
3	250	175	50	25
……	……	……	……	……
100	250	175	50	25
Total	25000	17500	5000	2500

The data set was divided into training set, validation set and test set according to the ratio of 7:2:1. There are 17500 samples for the training set, 5000 samples for the validation set, and 2500 samples for the test set. We employed uniform sampling method to extract 16 frames from every video. The size of original image is 1280×720. Since the original image contains a large amount of background redundant information, we extracted a 600×600 area of each image centered on the signer and adjust the image size to 224×224 as the network input by using

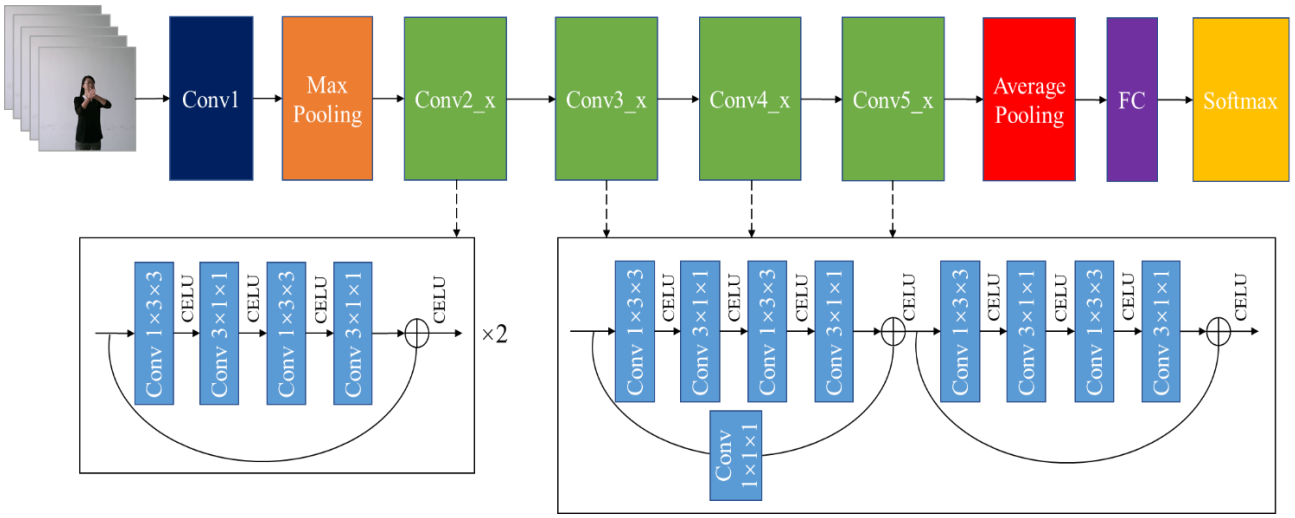


Fig. 6. Structure diagram of improved (2+1)D-ResNet18 model

resize function, so as to reduce the amount of network input data without losing important information. The original images and the preprocessed images are shown in Fig. 7 and Fig. 8. The epoch of the experiment was set to 50.

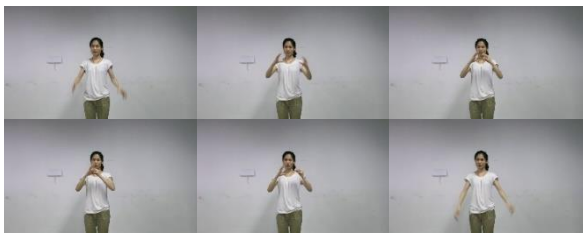


Fig. 7. Original images

Fig. 8. Processed images

In order to verify the effectiveness of the proposed algorithm, experiments were carried out on 3D-ResNet18, (2+1)D-ResNet18 and the improved (2+1)D-ResNet18 in this paper. The experiments were carried out in the same experimental environment. Fig. 9 and Fig. 10 show the validation result curves of the three models. Fig. 11 is the scatter plot of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Table 2 shows the accuracy of the three models on the test dataset.

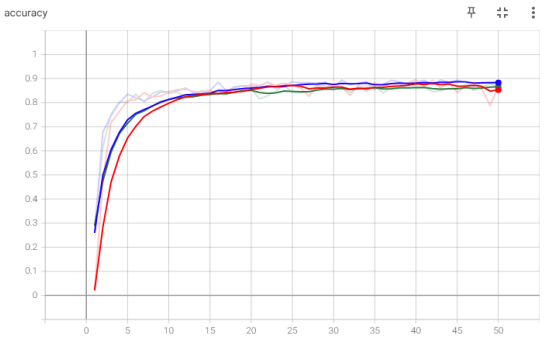


Fig. 9. Validation accuracy curves of 3D-ResNet18(green), (2+1)D-ResNet18(red) and improved (2+1)D-ResNet18(blue)

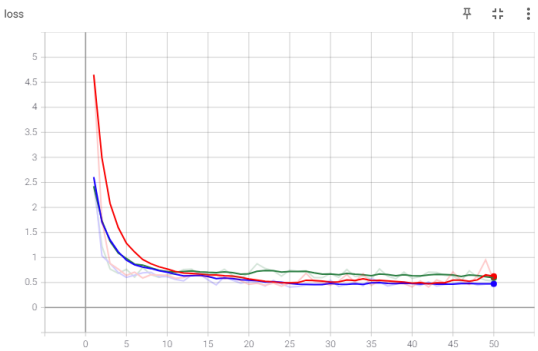


Fig. 10. Validation loss curves of 3D-ResNet18(green), (2+1)D-ResNet18(red) and improved (2+1)D-ResNet18(blue)

As can be seen from the curve of validation accuracy, 3D-ResNet18 has the lowest accuracy, followed by the original (2+1)D-ResNet18. Due to the decomposition of the convolution kernel that is more conducive to parameter optimization, the accuracy of original (2+1)D-ResNet18 increases significantly faster than that of 3D-ResNet18. Our model uses the CELU activation function to achieve the highest validation accuracy. And its overall curve is the smoothest. In Fig. 10, improved (2+1)D-ResNet18 has the fastest speed of loss decreasing and reaches the stable value first. Its final value is smaller than the other two models. The original (2+1)D-ResNet18 achieves the second smallest validation loss. However, 3D-ResNet18 has the largest loss and the most volatile curve.

As can be seen from the scatter plots of TP and TN, the number of correctly classified samples of (2+1)D-ResNet18 and improved (2+1)D-ResNet18 is more than that of 3D-ResNet18. It is difficult to compare who has the most. However, from the average TP (ATP) and average TN (ATN) in Table 2, the number of correctly classified samples with Improved (2+1) D-ResNet18 is the largest, followed by (2+1)D-ResNet18 and 3D-ResNet18 is the least. As can be seen from the scatter plots of FP and FN, 3D-ResNet18 has the largest number of misclassified samples. According to the average FP (AFP) and average FN (AFN) in Table 2, improved (2+1)D-ResNet18 has the smallest number of misclassified samples overall.

Table 2. Test results of models

Model	Test Accuracy	ATP	ATN	AFP	AFN	Time/ms
3D-ResNet18	86.94%	21.69	2471.69	3.31	3.31	33.31
(2+1)D-ResNet18	87.76%	21.94	2471.94	3.06	3.06	34.29
Improved (2+1)D-ResNet18	88.92%	22.41	2472.41	2.59	2.59	33.18

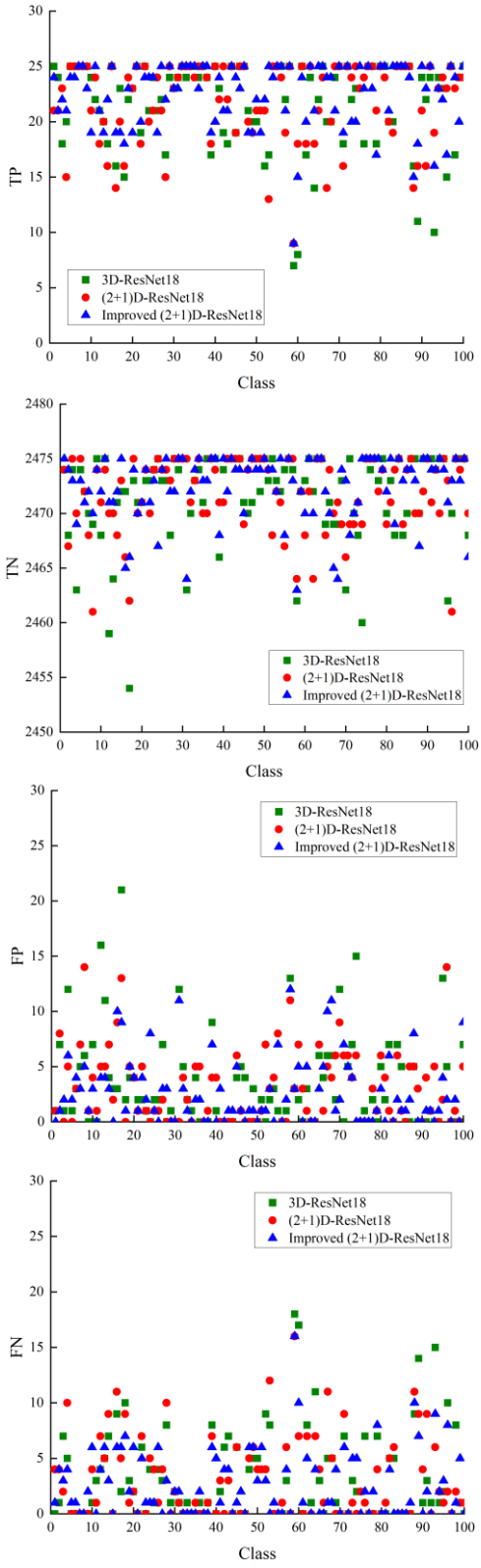


Fig. 11. Scatter plots of TP, TN, FP and FN

It can be seen from Table 2 that the test experiment obtains similar results to the validation experiment. The improved (2+1)D-ResNet18 proposed in this paper has the best performance, followed by (2+1)D-ResNet18. The 3D-ResNet18 has the worst performance. The test accuracy of 3D-ResNet18 is 86.94%. The accuracy of (2+1)D-ResNet18 obtained by separating spatial dimension and time dimension is 87.76%. The improved (2+1)D-ResNet18 has the highest accuracy of 88.92%. The CELU activation function in the improved model solves the problem of gradient disappearance during back propagation, thus further improving the accuracy.

5. Conclusion

In order to solve the problem of insufficient ability of 2D convolution to process sign language video and large amount parameters of 3D convolution, this paper proposes an improved (2+1)D-ResNet model to recognition isolated sign language words. The model performs convolution and feature extraction step by step in spatial dimension and time dimension. Parameter optimization can be carried out in space dimension and time dimension respectively during back propagation. In addition, the original ReLU activation function is prone to the problem of gradient disappearance during the back propagation. We employ CELU activation function to introduce nonlinear factors into the neural network. When the input is zero, parameters can still be updated. The model in this paper achieves an accuracy of 88.92% on the CSL dataset, which can recognize sign language well.

In the future work, we will further design and optimize the network structure, so that it can't only perform well on the existing data set, but also show good performance in the real complex environment.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China (No. U1804147) and Science and Technology Innovation Talents in Universities of Henan Province (20IRTSTHN019), Henan Provincial Science and Technology Research Project (No. 212102210508). The data supporting the study is available from https://ustc-slr.github.io/openresources/cslr-dataset-2015/index.html but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. However, data are

available from the authors upon reasonable request and with permission of https://ustc-slr.github.io/openresources/cslr-dataset-2015/index.html.

References

1. H. Wang, X. Chai, X. Chen, Sparse Observation (SO) Alignment for Sign Language Recognition, Neurocomputing, Volume. 175, Part A, pp. 674-685, 2016. January.
2. Y. Yan, Z. Li, Q. Tao, C. Liu, R. Zhang, Research on Dynamic Sign Language Algorithm Based on Sign Language Trajectory and Key Frame Extraction, 2019 IEEE 2nd International Conference on Electronics Technology (ICET), 2019, pp. 509-514.
3. T. Liu, W. Zhou, H. Li, Sign language recognition with long short-term memory, 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 2871-2875.
4. J. Pu, W. Zhou, H. Li, Sign Language Recognition with Multi-modal Features, In: E. Chen, Y. Gong, Y. Tie (eds), Advances in Multimedia Information Processing-PCM 2016, Lecture Notes in Computer Science, vol. 9917, Springer. Cham, 2016, pp. 252-261.
5. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A Closer Look at Spatiotemporal Convolutions for Action Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6450-6459.
6. S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 448-456.
7. J. Barron, Continuously Differentiable Exponential Linear Units, arXiv e-prints, 2017.
8. J. Huang, W. Zhou, Q. Zhang, H. Li, W. Li, Video-Based Sign Language Recognition Without Temporal Segmentation, Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2018.

Authors Introduction

Ms. Yueqin Sheng

She graduated from Henan Polytechnic University (China) in 2020 with a bachelor's degree in automation. She is currently studying for a master's degree in control science and engineering at Henan Polytechnic University. She is mainly engaged in research on image processing and sign language recognition.

Dr. Qunpo Liu

He graduated from the Muroran Institute of Technology (Japan) with a Ph.D. in Production Information Systems. He is an associate professor and master tutor at the School of Electrical Engineering and Automation, Henan Polytechnic University (China). He is mainly engaged in teaching and research work in robotics, intelligent instruments and industrial automation equipment.

Dr. Ruxin Gao

He graduated from Huazhong University of Science and Technology (China) with a ph. D. in Pattern Recognition and Intelligent Systems. He is an associate professor and master tutor at the School of Electrical Engineering and Automation, Henan Polytechnic University (China). He is mainly engaged in teaching work of computer application and research work of image processing and visual related.

Dr. Naohiko Hanajima

He graduated from the Hokkaido University (Japan) of Technology in Japan with a Ph.D. He is a professor at the College of Information and Systems at Muroran Institute of Technology (Japan). He is mainly engaged in the research work of robotics and intelligent equipment.