Research on Sign Language Recognition Algorithm Based on Improved R(2+1)D

Yueqin Sheng, Qunpo Liu, Ruxin Gao
School of Electrical Engineering and Automation, Henan Polytechnic University, 2001 Century Avenue
Jiaozuo, Henan 454003, China

Hanajima Naohiko
College of Information and Systems, Muroran Institute of Technology, 27-1 Mizumoto-cho
Hokkaido, Hokkaido 050-8585, Japan
E-mail: 212007010029@home.hpu.edu.cn, lqpny@hpu.edu.cn, gaoruxin@hpu.edu.cn
hana@mondo.mech.muroran-it.ac.jp
www.hpu.edu.cn, www.muroran-it.ac.jp

Abstract

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 R(2+1)D 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.

Keywords: Sign Language Recognition; R(2+1)D Convolution; 3D Convolution; CELU Activation Function

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. 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

Fig. 1. Example diagram of sign language

Traditional SLR methods mainly include Hidden Markov 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. However, 3D convolution is difficult to optimize, slow and requires high hardware.

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

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. Each residual block consists of two convolution layers followed by a ReLU activation function. If *x* represents the input data size of 3×*L*×*H*×*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 *i*th residual block is:

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

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.

R(2+1)D introduces hyperparameter M_i and uses M_i two-dimensional space convolution kernels with size of and N_{i-1}×1×d×d and N_i one-dimensional time convolution kernels with size of M_i×*t*×1×1 to replace N_i three-dimensional convolution kernels with size of N_{i-1}×*t*×*d*×*d*, so as to maintain approximately the same number of parameters as the three-dimensional residual network. The following relation can be obtained:

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

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

When the input is single channel, the 3D convolution kernel and (2+1)D convolution kernel are shown in Fig. 2. The left is the 3D convolution kernel with the size of *t*×*d*×*d*, where *t* represents time depth and d represents the height or width of the space. The right is the (2+1)D convolution kernel formed by decomposing 3D convolution kernel. The number of 2D convolution kernel after decomposition is M_i.

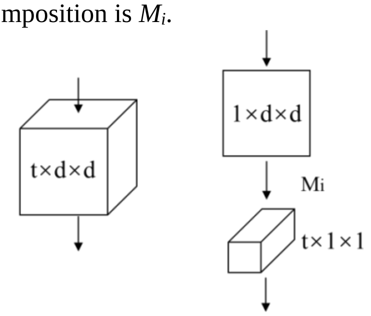

Fig. 2. 3D convolution kernel and (2+1)D convolution kernel

3.2. *Optimization of activation function*

The R(2+1)D model proposed by Tran et al.[5] 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. (4). The calculation formula of CELU activation function is shown in Eq. (5). In this paper, the value of *α* of CELU activation function is 0.05. The output comparison between ReLU and CELU is shown in Fig. 3.

$$\mathrm{ReLU}\left(x\right) = max\left\{0, x\right\} \tag{4}$$

$$\mathrm{CELU}\left(x,\alpha\right) = max\left\{\alpha\left(\exp\left(\frac{x}{\alpha}\right) - 1\right), x\right\} \tag{5}$$

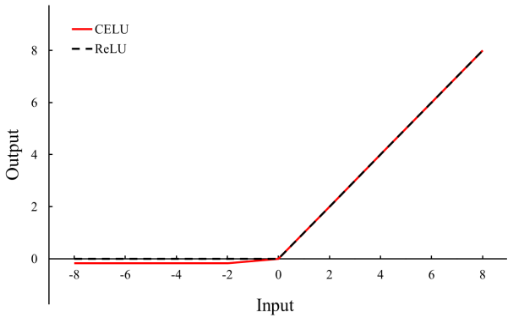

Fig. 3. Activation function curves of CELU and ReLU

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

The structure of the improved (2+1)D-ResNet18 model proposed in this paper is shown in Fig. 4. The video frame sequence first enters the fully connected layer and the max pooling layer, then enters four improved (2+1)D residual convolution blocks. After that, the average pooling layer and the fully connected layer are entered successively. Finally, the classifier outputs the classification results.

4. Experimental Results and Analysis

The data set we used is CSL isolated word sign Language dataset from University of Science and Technology of China, which contains 500 commonly used sign language words. Considering the training time and hardware requirement, we selected 100 of these words to conduct experiments on 3D-ResNet18, (2+1)D-ResNet18 and improved (2+1)D-ResNet18 in this paper. The data set was divided into training set, validation set and test set in a ratio of 7:2:1. Each continuous video was extracted into discrete video frames. 16 frames were sampled from each video frame set as the input of the model using uniform sampling method. The experiments were carried out in the same experimental environment. Fig. 5 and Fig. 6 show the validation result curves of the three models. Table 1 shows the accuracy of the three models on the test dataset.

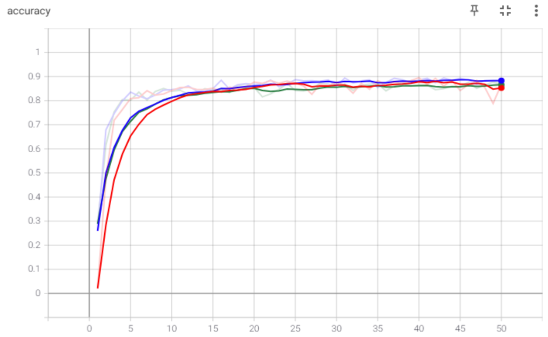

Fig. 5. Validation accuracy curves of 3D-ResNet18(green), (2+1)D-ResNet18(red) and improved (2+1)D-ResNet18(blue)

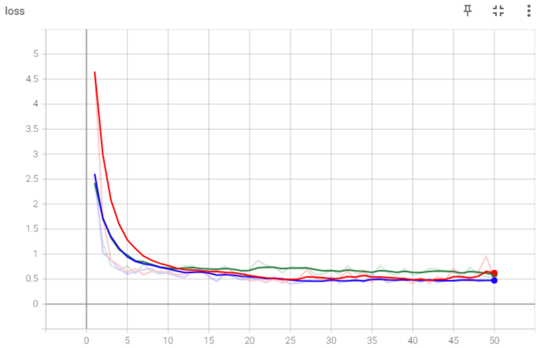

Fig. 6. Validation loss curves of 3D-ResNet18(green), (2+1)D-ResNet18(red) and improved (2+1)D-ResNet18(blue)

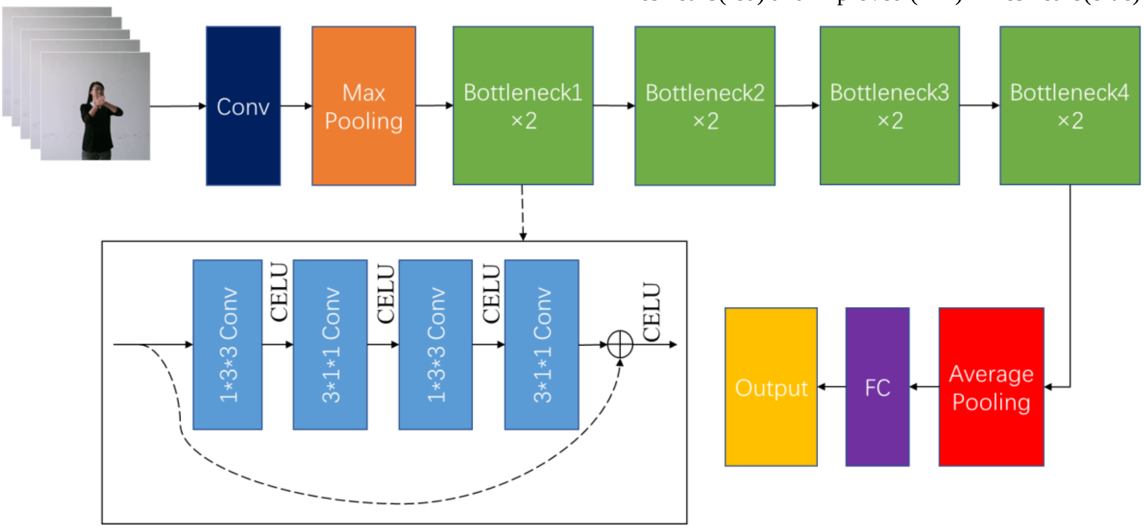

Fig. 4. Structure diagram of improved (2+1)D-ResNet18 model

As can be seen from the validation curves, the improved (2+1)D-ResNet18 has the fastest speed of accuracy increase and loss reduction. It is the first one to reach the minimum value of loss, and its curve is the smoothest.

Table 1. Test results of models

Model	Test Accuracy
3D-ResNet18	86.94%
(2+1)D-ResNet18	87.76%
Improved (2+1)D-ResNet18	88.92%

It can be seen from Table 1 that 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

This paper proposes an improved R(2+1)D model for isolated word recognition. The model separates spatial convolution and time convolution and use CELU as the activation function, reaching the accuracy of 88.92%. In the future, sign language in complex environments will be further studied.

References

1. Hanjie Wang, Xiujuan Chai, Xilin Chen. Sparse Observation (SO) Alignment for Sign Language Recognition. Neurocomputing, Volume 175, Part A, 2016, Pages 674-685.
2. Y. Yan, Z. Li, Q. Tao, C. Liu and 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 and H. Li, "Sign language recognition with long short-term memory," 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 2871-2875.
4. Pu J., Zhou W., Li H. (2016) Sign Language Recognition with Multi-modal Features. In: Chen E., Gong Y., Tie Y. (eds) Advances in Multimedia Information Processing - PCM 2016. PCM 2016. Lecture Notes in Computer Science, vol 9917. Springer, Cham. https://doi.org/10.1007/978-3-319-48896-7_25.
5. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M., "A Closer Look at Spatiotemporal Convolutions for Action Recognition", <i>arXiv e-prints</i>, 2017.
6. Barron, J. T., "Continuously Differentiable Exponential Linear Units", <i>arXiv e-prints</i>, 2017.

Authors Introduction

Ms. Yueqin Sheng

She received her Bachelor's degree from the Department of Automation, Henan Polytechnic University (China) in 2020. She is currently a Master's course student in Henan Polytechnic University (China).

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. Hanajima Naohiko

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.