Fingerspelling Identification for American Sign Language Based on Resnet-18

College of Computer and Communication, Hunan Institute of Engineering, Xiangtan 411104, China
ZHANG Han-wen
Email:2678984@qq.com
HU Ying*
Email: huying1983@hnie.edu.cn
ZOU Yong-jia
Email:1902850183@qq.com
WU Cheng-yu
Email:1914896168@qq.com

---ABSTRACT---

Sign language as the main communication channel for deaf and hearing people, plays a very important role in daily life. With the rapid development of the field of deep learning, the field of sign language recognition has ushered in new opportunities. Aiming at the small number of sign language samples and low detection accuracy, an American sign language detection method based on Resnet-18 and data augmentation is proposed. First, the sign language picture is adjusted to 64×64 size using the filling method, and then the data is augmented by methods such as chromaticity change, rotation, and noise addition to increase the diversity of data samples and improve the robustness of the detection method. The experimental results show that the accuracy of the sign language recognition method based on Resnet-18 can reach 99%, which provides a new method for sign language recognition.

Keywords **: Convolutional Neural Network, Data augmentation, Resnet-18, Sign language recognition**

Date of Submission: Aug 02, 2021 Date of Acceptance: Aug 16, 2021

I. INTRODUCTION

As the main communication channel for deaf-mute and hearing people, sign language plays a vital role in acquiring knowledge, communicating with hearing people, and improving the quality of life. Sign language recognition is the conversion of sign language into speech and text through computer processing, so that hearing people can quickly understand the meaning of deaf-mute people, and help deaf-mute people better integrate into society, which has important research value. Computer vision is the key technology of sign language recognition technology. In recent years, computer vision based on deep learning has developed rapidly, and the field of sign language recognition has also ushered in new opportunities.

In this paper, we proposed a Resnet-18 based transfer learning method equipped with data augmentation for fingerspelling recognition of American Sign Language (ASL) and gained a average accuracy of 99%.

II. RELATED WORK

The research object of finger language recognition is the appearance characteristics of the hands in the image. The difficulty lies in the fact that the sign language image will be affected by factors such as light and background environment.

Chen Hanxiong et al.[1]used Kinect to obtain skeletal coordinate information, constructed a spatial point set, and used K-curvature algorithm to recognize sign language with a recognition rate of 91%. Kong Y et al.[2] extracted the color, depth, and contour of the image through a camera, using multi-eye vision method to recognize finger language. Wang Wei et al.[3]proposed a static finger language recognition method based on SVM, using Leap Motion to obtain finger language, using Principal Component Analysis (PCA) dimensionality reduction, and finding the optimal parameters, the recognition rate of digits 0-9 can reach 99%. Mou Yu et al. [4] used HOG to extract image features and used LC-KSVD algorithm to sparsely represent finger language, effectively distinguishing finger language categories. Amaya et al. [5]proposed a static sign language recognition method based on PCA and Support Vector Machine (SVM) to recognize 5 vowels in English, PCA is used in the opponent’s part, Feature extraction is performed on regions, and SVM is used for classification. In recent years, Convolutional Neural Network (CNN) has achieved very good results in the field of image recognition. With the help of CNN's powerful ability to extract image features, hand shape features can be automatically extracted. Pias Paul et al. [6] proposed a custom convolutional neural network, which uses the migration learning method to pre-train the model,on the ASL dataset, the recognition rate can reach 92%. Aich et al. [7] constructed their own CNN is used on the data set to recognize 36 characters in Bengali Sign Language with an accuracy of 92.7%, but the image background is single and the scalability is not high. Jiang Xianwei et al. [8] proposed a Chinese finger language recognition method based on AlexNet, which uses data enhancement to expand the training data set, and the recognition rate can reach 91.5%. Hasan et al. [9] used CNN to recognize ASL letters, and the recognition rate reached 97.62%.

III. PROPOSED WORK

In this section, we introduced the techniques which make up the integral part of our approach. Sections 3.1 describes the dataset, sections 3.2 describes data preprocessing, the proposed models are discussed in sections 3.3 and 3.4.

3.1 Dataset

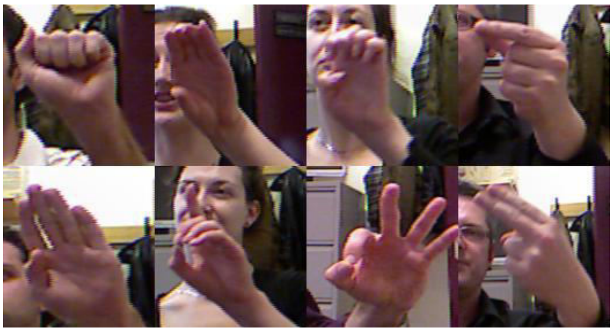

Figure 1. Samples from ASL dataset

Since different countries have different spelling methods, English finger language is used for static finger language recognition in this paper, and the data set for finger language recognition is the ASL Finger Spelling Datase [10] data set provided by Pugeault N. This data set was recorded by 5 people, each of whom recorded 24 finger language, and each person recorded about 550 finger language, totaling 65774 pictures. Since the letters 'j' and 'z' are dynamic sign languages, these two finger languages are not included in the ASL dataset. Some pictures of finger language are shown in Fig.1, which respectively represent the letters "a~h". It can be seen from the figure that the ASL dataset has a complex background.

3.2 Pre-processing

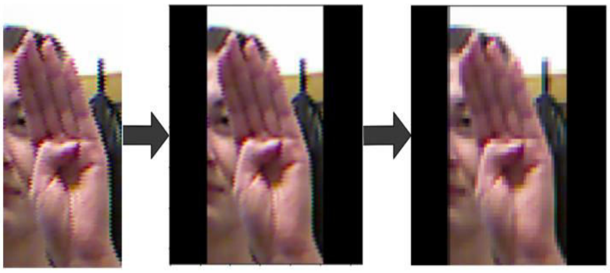

Figure 2.Image filling and scaling

(1) Image normalization. Since each sign language picture has a different size, it is necessary to normalize each picture to a size of 64×64×3. Use the filling method to adjust the sign language picture to be equal in length and width, the purpose is to better retain the original shape characteristics after the picture is scaled, and to ensure that the image will not be distorted due to the direct scaling of the picture. The steps of the filling method are to first obtain the length and width of the image, take the longer side l, create a full black image of all 0 pixels $l \times l \times 3$, and then place the sign language image in the full black image. After filling the image with the same length and width, the bilinear interpolation method is used to scale the image to 64×64 , The image processing process is shown in Fig.2.

(2) Augmented Dataset. Since the number of original data samples is relatively small, it will greatly affect the learning effect of deep learning, so it is necessary to use data augmentation [11] to increase the number of samples. The data augmentation method of chromaticity change, rotation and adding noise is adopted, the chromaticity enhancement factor is 0.4, the picture is rotated by 30 degrees and 60 degrees, and Gaussian noise with a signal-to-noise ratio of 0.8 is added to the picture. Through data augmentation, the data set was expanded to 120,000 pictures.

3.3 Model structure

Residual network [12] is a residual learning framework proposed on the basis of the existing training deep network, which has the advantages of easy optimization and small amount of calculation. The residual design is used to solve the degradation and gradient problems, so that the performance of the network can be improved as the depth increases, and the jump connection in the residual block uses the identity mapping, and no additional parameters are increased, Computational complexity is not increased.Fig.3 shows the structure diagram of the residual block.

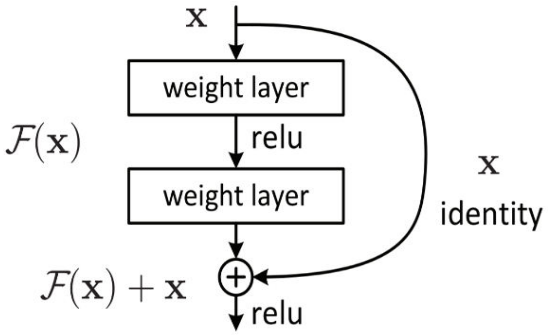
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 3.Residual Block

The Resnet-18 network model is used to train finger language, Fig.4 shows architecture of Resnet-18 network. The first layer of Resnet-18 is a convolutional layer, and then 8 residual blocks are connected, and average pooling is used. Each convolutional layer performs a Batch Normalization operation and a relu activation function. The last layer is a fully connected layer, which uses a softmax classifier. The network structure of Resnet-18 is shown in Table 1.

Table 1: Network structure of Resnet-18

Name	Nuclear size/step size	Output size
conv	$3 \times 3/1$	$64 \times 64 \times 64$
Resnet_block1	$3 \times 3/1$	$64 \times 64 \times 64$
Resnet_block2	$3 \times 3/2$	$32 \times 32 \times 128$
Resnet_block3	$3 \times 3/2$	$16 \times 16 \times 256$
Resnet_block4	$3 \times 3/2$	$8 \times 8 \times 512$
global avgpool		1×512
softmax		24

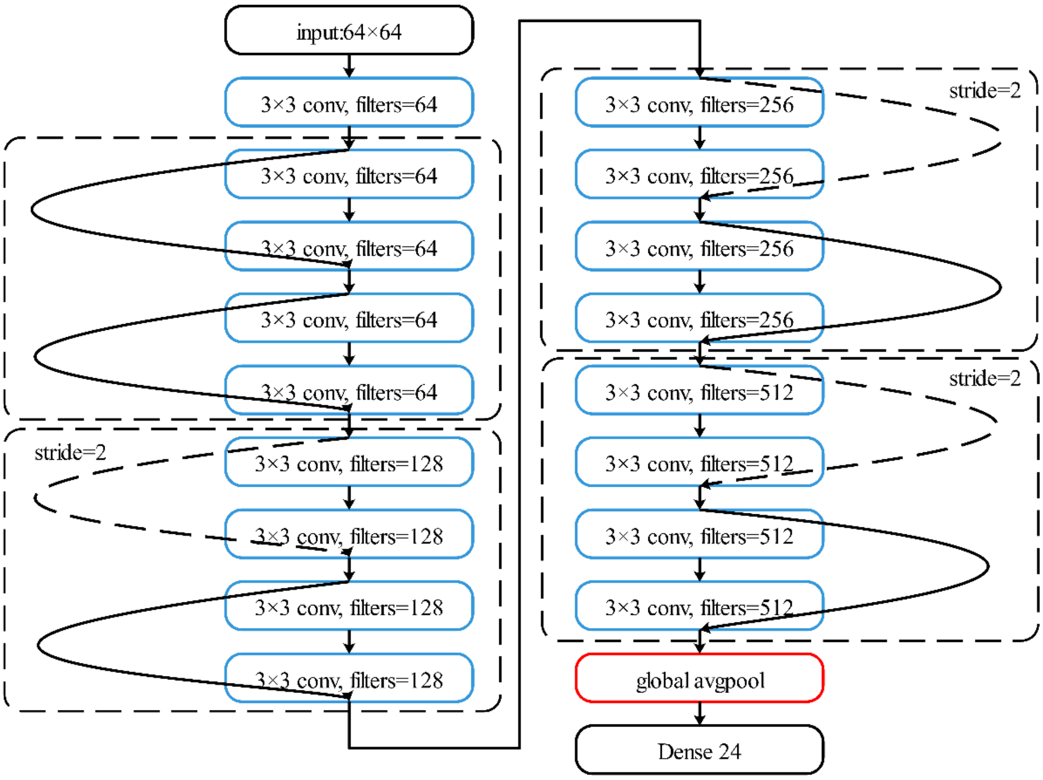

Figure 4. Resnet-18 model

3.4 Training Algorithms

The activation function of the convolutional layer adopts ReLU function, which basically solves the problem of "gradient disappearance", the ReLU function is defined as:

$$ReLU(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \quad (1)$$

The loss function is used to calculate the deviation between the input sample and the label value, in this paper, the cross entropy function is used as the loss function, the cross entropy function is defined as:

$$H(p,q) = -\sum_{i} p(x) log q(x) \quad (2)$$

Among them, x is a randomly distributed sample value, p is the expected probability distribution of sample x, and q is the actual probability distribution of sample x.

The optimizer is used to update the weight and bias values during back propagation. Adam optimizer has excellent performance in the training of neural networks, in this paper, Adam optimizer is used. The formula of Adam optimizer is defined as follows:

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g(w_t) \quad (3)$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g(w_t) \cdot g(w_t) \quad (4)$$

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t} \quad (5)$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \quad (6)$$

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t + \varepsilon}} \widehat{m}_t \quad (7)$$

In the formula: $g(w_t)$ is the loss gradient to the parameter w_t at time t. β_1 is the exponential decay coefficient, which is used to control the decay speed, and the default value is 0.9.β_2 is the exponential decay rate, which is used to control the deviation caused by the square of the gradient, and the default value is 0.999. α is the learning rate, and the default value is 0.001. In order to prevent the loss of nan when $\hat{v}_t$ is zero, ε is 10^{-7}.

IV. EXPERIMENT RESULTS AND DISCUSSION

The architecture was implemented using the Python library tensorflow2.0 for deep learning based on a platform of a computer whose main components include two NVIDIA TITAN XP GPU, 2.5GHZ Core i7 CPU, and 96 GB memory, under the operating system of Ubuntu16.04.Test evaluation of detection accuracy is defined as:

$$A = \frac{x}{y} \quad (8)$$

In the formula: A represents the accuracy rate (%), x represents the number of correct samples, y represents the total number of tested samples. Sign language samples are randomly divided into training set and validation set according to 4:1. The number of training samples in a single batch is 32, the initial learning rate is 0.001, and the learning rate attenuation is 0.0001. Using the early stopping method to monitor the loss value of the verification set, it is found that the loss value of the network no longer decreases in 34 rounds, and the training is ended early. The accuracy of the training set and the validation set have reached more than 99%. Fig.5 showns training results.

Randomly take 24 different sign language pictures in the test set and put them into the trained network model to predict. The result is shown in Fig.6, where the format of the picture title is "Original Finger Language: Predicted

Finger Language => Probability of Predicted Finger Language" . Every sign language can be well recognized.

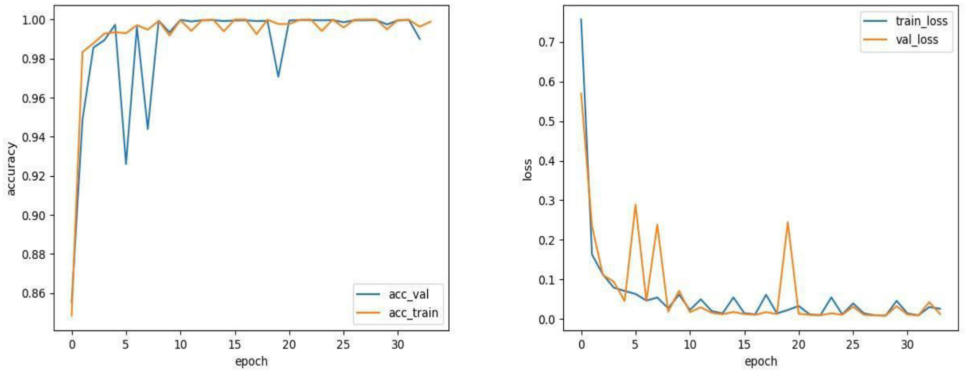

Figure 5. Curve of accuracy and loss value

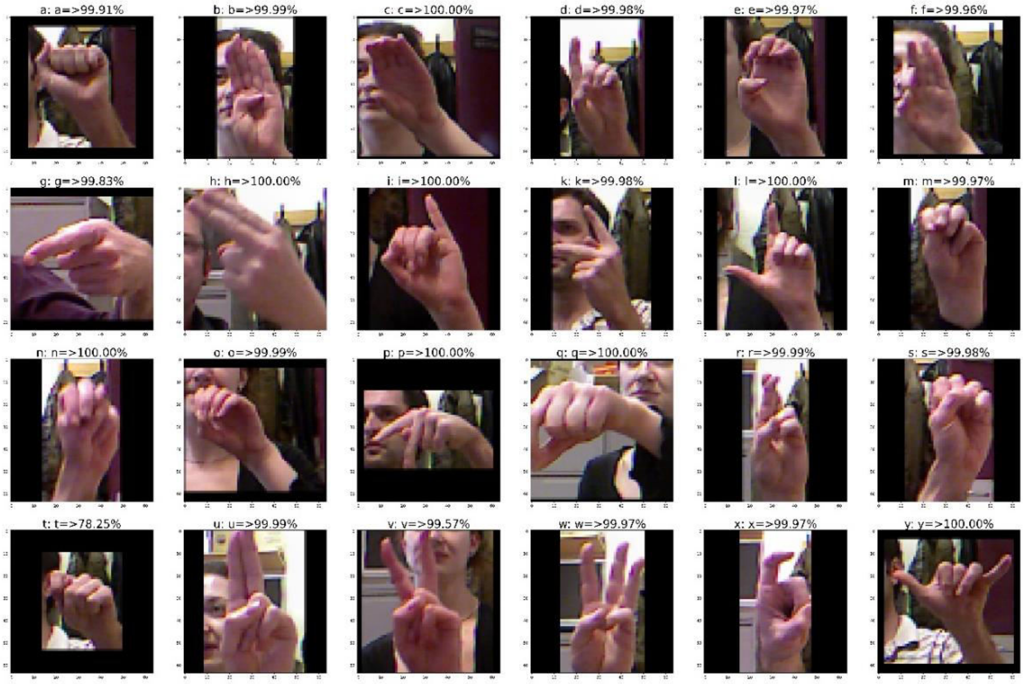

Figure 6. Sign language prediction results

Table 2 shows the comparison of average precision by previous works and our work for ASL dataset, the dataset is split into two parts: training (80%) and testing (20%) datasets.The experimental results show that the method we proposed outperformed other three approaches.

Table 2: Comparison of different training algorithms

Model	Training accuracy (%)	Validation accuracy (%)
KNN	87%	82%
SVM	94%	91%
CNN	96%	92%
Proposed work	99%	99%

V. CONCLUSION AND FUTURE WORK

In this work, we apply the Resnet-18 model to sign recognition of ASL, equipped with data augmentation. Experiment results demonstrated that our method achieved the average accuracy of 99%, which was excellent among the three state-of-the-art approaches. In the future, add the letters j and z to the dataset, we shall try to use Recurrent Neural Network to solve the problem of dynamic finger language recognition.

ACKNOWLEDGEMENTS

This study was supported by 2019 National College Student innovation and entrepreneurship training program project ”Research and implementation of sign language recognition system based on deep learning”, Scientific

Research Project of Hunan Provincial Department of Education(19C0478)" Research on Chinese Sign Language word recognition technology based on deep learning".

REFERENCES

[1]Chen H X, Huang Y Y, Liu Y, Research and implementation of air gesture tracking and recognition based on Kinect, *Video Engineering,39*(21), 2015,91-94.

[2]Kong Y, Satarboroujeni B, Fu Y, Learning hierarchical 3D kernel descriptors for RGB-D action recognition, *Computer Vision & Image Understanding, 144*(5), 2016, 14-23.

[3]Wang W, Zhang H J, Ren X Z, Single finger language recognition method based on SVM, *Computer Engineering and Design, 39*(10),2018,3234-3239.

[4]MU Y, *Research on finger language recognition method based on sparse coding*, master diss., Shenyang University of Technology, Shen Yang, 2020.

[5]Amaya C, Murray V, Real-time sign language recognition, *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing*, Lima, Peru, 2020, 1-4.

[6]Pias Paul,Moh.Anwar-Ul-Azim Bhuiya,et al, A modern approach for sign language interpretation using convolutional neural network, *PRICAI 2019: Trends in Artificial Intelligence 16th Pacific Rim International Conference on Artificial Intelligence*, LNAI, 2019, 431-444.

[7]D. AICH, A. AL ZUBAIR,et al, A deep learning approach for recognizing Bengali character sign language, *2020 11th International Conference on Computing, Communication and Networking Technologies*, Kharagpur, India, 2020, 1-5.

[8]JIANG X W, HU B,et al, Fingerspelling identification for chinese sign language via AlexNet-based transfer learning and adam optimizer, *Scientific Programming* , 2020, 1-12.

[9]HASAN M M, SRIZON A Y, SAYEED A, et al, Classification of sign language characters by applying a deep convolutional neural network, *2020 2nd International Conference on Advanced Information and Communication Technology,* Dhaka, Bangladesh, 2020,434-438.

[10]Pugeault N, Bowden R, Spelling it out: Real-time ASL fingerspelling recognition, *2011 IEEE International conference on computer vision workshops*, Barcelona, Spain, 2011,1114-1119.

[11]CUBUK E D,ZOPH B, et al, Autoaugment: Learning augmentation strategies from data,*Proceedings of the IEEE conference on computer vision and pattern recognition,* Long Beach,2019,113-123.

[12]HE K, ZHANG X, et al, Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition,* Las Vegas, 2016, 770-778.

[13]Dewinta Aryanie, Yaya Heryadi, American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier, *2015 3rd International Conference on Information and Communication Technology,* Nusa Dua, Bali, Indonesia ,2015, 533-536.

Biographies and Photographs

ZHANG Han-wen is currently pursuing his Bachelor's degree (Communication Engineering) from Hunan Institute of Engineering, Xiangtan, China. His current research areas are:Computer Vision, pattern recognition, machine learning and Deep learning.

HU Ying is Associate Professor in the College of Computer and Communication at Hunan Institute of Engineering. She has obtained his Master degree from Central South University, Changsha, China. Her research interests include Computer Vision, Artificial Intelligence, Machine Learning, Deep Learning, Video Processing.