Multimedia Tools and Applications (2020) 79:15697–15715
https://doi.org/10.1007/s11042-019-08345-y

An eight-layer convolutional neural network with stochastic pooling, batch normalization and dropout for fingerspelling recognition of Chinese sign language

Xianwei Jiang[1] · Mingzhou Lu[2] · Shui-Hua Wang[3,4]

Received: 4 June 2019 / Revised: 8 August 2019 / Accepted: 1 October 2019
Published online: 19 December 2019

Abstract
Fingerspelling recognition of Chinese sign language rendered an opportunity to smooth the communication barriers of hearing-impaired people and health people, which occupies an important position in sign language recognition. This study proposed an eight-layer convolutional neural network, combined with three advanced techniques: batch normalization, dropout, and stochastic pooling. The output of the stochastic pooling was obtained via sampling from a multinomial distribution formed from the activations of each pooling region. In addition, we used data augmentation method to enhance the training set. In total 10 runs were implemented with the hold-out randomly set for each run. Our method achieved the highest accuracy of 90.91% and overall accuracy of 89.32 ± 1.07%, which was superior to three state-of-the-art approaches compared.

Keywords Convolutional neural network · Hyperparameter optimization · Deep learning · Stochastic pooling · Batch normalization · Dropout

1 Introduction

According to the survey, there are more than 360 million hearing-impaired people in the world, and there are nearly 27.9 million deaf people in China, which is a huge group [15]. Sign language is the main way for hearing-impaired people to communicate directly, but most healthy people do not understand or are not familiar with sign language, which makes the huge

✉ Shui-Hua Wang
shuihuawang@ieee.org

Xianwei Jiang
jxw@njts.edu.cn

Mingzhou Lu
lmz@njau.edu.cn

Extended author information available on the last page of the article

communication gap between these two groups. Thus, the hearing-impaired people have encountered great challenges in employment, learning, even living areas such as medical treatment and legal counseling. Communication barriers also bring about a loss of resources such as social labor and special group intelligence. The real-time translation of sign language and spoken language through sign language interpreters is a solution to the communication problem, but it requires advance scheduling, which is costly and often unrealistic. Therefore, researchers consider introducing artificial intelligence and machine learning to develop and implement automatic translation of sign language recognition. The research of sign language recognition technology is profound, which can not only smooth the communication barriers of hearing-impaired people and healthy people, promote the integration of aphasias people into society, but also promote the development of more friendly and intelligent human-computer interaction interfaces.

Sign Language (SL) is a complete communication system consisting of a series of elements such as hand shape, movement, expression and posture. Chinese Sign Language (CLS) can be divided into two categories: finger sign language and gesture sign language (See Fig. 1). 30 finger letters (including 26 single letters a-z, 4 double letters zh, ch, sh, ng) and some numbers constitute the basic unit of the finger sign language, and each Chinese pinyin letter is represented by the shape of the finger. It is easy to learn and has a small number of gestures, which can easily express professional terminology and abstract concepts and occupies an important position in sign language recognition [3]. Gesture sign language simulates the meaning to be expressed through the image and movement of the gesture, which is relatively difficult to use and identify.

Sign Language Recognition (SLR) refers to the use of computer technology to translate or convert sign language information into text, language and other information to facilitate the understanding and communication of others [1]. At the beginning of sign language recognition technology, the focus of research was on designing dedicated hardware devices to input data, and then was on the study of marker gestures and human palms. After that, the identification of natural hand recognition became a popular trend. At present, based on data input levels, sign language recognition technology can be divided into contact and non-contact. Commonly used sensors mainly include data gloves, EMG signal arm rings and depth cameras. The accuracy of data glove-based sign language recognition is high, but the data glove equipment is expensive and inconvenient to carry, which makes it difficult to promote and popularize. Including carrying problems, the sign language recognition using the EMG signal armband also has

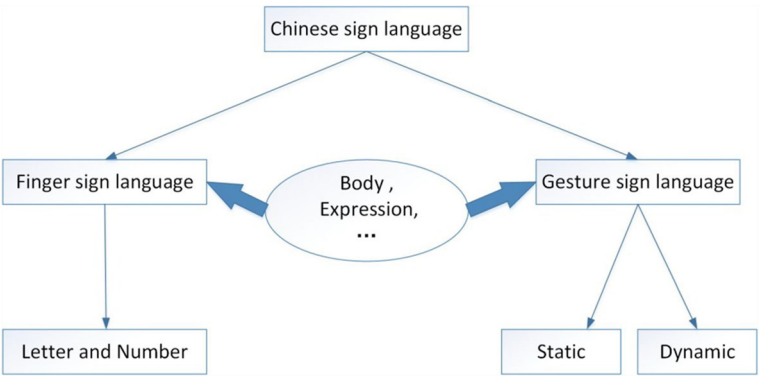


Fig. 1 Chinese sign language type

the situation that when the EMG signal is a weak electric signal or the sign language is changed more quickly, the arm muscle movement cannot be clearly captured and cannot be accurately recognized. In addition, neither of the above methods can effectively identify when the same gesture points to different locations. However, this problem was not found in the depth camera-based sign language data set. The computer vision-based solution is to obtain information from video images and complete the recognition by means of image processing technology [8]. It is free from the constraints of hardware devices, flexible and convenient to operate. It has no affection to users, and is generally welcomed by the market. Currently, this field has attracted a large number of researchers to participate in.

In order to improve recognition accuracy and enhance practical effects, some classical and effective image processing [32–34] and recognition algorithms [7, 10, 26, 29, 37, 39] are widely mentioned, such as hidden Markov model (H-MM), support vector machine (SVM), k-nearest neighbor (k-NN), artificial neural network (ANN), dynamic time warping (DTW), long short-term memory network (LSTM), skin color modeling, random forest, extreme learning machine (ELM), recurrent neural network (RNN), convolutional neural network (CNN), including their various variants. Kumar, et al. [11] employed hidden Markov model (HMM) and developed a position and rotation invariant framework sign language recognition model. Lee, et al. [13] combined support vector machine (SVM) and hidden Markov model (HMM) to develop Taiwanese sign-language recognition. Yang and Lee [38] proposed a new method called hierarchical conditional random fields (HCRF). In combination with dynamic time warping and secondary classification, Lichtenauer, et al. [17] obtained an average recognition rate of 92.3%. Li, et al. [16] proposed combining HMM, K-means, ant colony algorithm to Taiwan sign language recognition, and the average recognition rate reached 91.3%. Pariwat and Seresangtakul [24] presented a finger-spelling sign language system in SVM kernel with an average accuracy of 91.2%. ANN classifier was trained by P. V. V. Kishore to get an average word matching score over 90% [27].

The adoption of these technologies has achieved favorable results, but they still have different disadvantages. For example, HMM is a statistical model and needs to be based on successful detection, which is difficult to use for real-time identification. DTW also needs to create templates in advance, which brings a huge amount of work. ELM often lacks superior generalization performance and robustness in gesture recognition. Classification based on SVM and k-NN requires higher feature extraction, and consumes a large amount of system resources in the classification process. Some combination algorithms have relatively high recognition rates, but their data sets are insufficient. The emergence of neural network technology provides a new idea of solution. It has strong self-learning ability and organizational capability. The distribution characteristics are obvious. It also can effectively resist noise. Deep learning is derived from artificial neural networks, which combine the features of lower layers to form more abstract high-level representation attributes or features to discover the distribution characteristics of the data. As the most classical deep neural network, convolutional neural network (CNN) is very suitable for image classification and recognition [12, 21]. In particular, it can perform network training on multi-dimensional image samples, avoiding complex manually feature extraction operations in traditional recognition algorithms [4].

In this paper, based on image processing techniques [25, 31, 35, 36], eight-layer convolutional neural network with stochastic pooling, batch normalization and dropout for Chinese finger sign language recognition was proposed. This CNN is fully optimized on each layer. Besides, stochastic pooling and data augmentation were introduced to achieve excellent

performance. In the experiments, we compared stochastic pooling against average pooling and maximum pooling method. Finally, our method is found to be superior to state-of-the-art approaches.

The contributions of this paper are listed below: (i) we utilized some advanced technologies to overcome common issues in traditional CNN, for instance, stochastic pooling and dropout were employed to avoid overfitting, batch normalization was applied to speed up learning convergence, data augmentation was adopted to enhanced train set; (ii) our study rendered an opportunity to smooth the communication barriers of hearing-impaired people and health people and elevate the integration of hearing-impaired people into society; (iii) vision-based sign language recognition was free from the constraints of hardware and with no affection to patients, which was flexible and convenient.

2 Dataset

2.1 Data collection

According to the Chinese finger sign language standard, there are 30 categories, including 26 basic monosyllabic letters and 4 double syllable letters. More than 44 volunteers are selected from different departments to help create the self-built sign language image database using a camera. Figure 2 demonstrates part samples of the main hand shapes in 30 categories intercepted from those sample images. A total of 1320 images of Chinese finger sign language are acquired and normalized to 256×256 background-optimized samples. Our experiment was accomplished with this pre-processed 1320 Chinese finger sign language samples.

2.2 Data augmentation

Hold-out validation method was used. 80% of the total 1320 images, i.e., 1056 images were used for training, and the rest 264 images were used for test. Data augmentation was used on the 1056 training images.

Fig. 2 Part of samples of Chinese finger sign language

(i) Scaling. Images were scaled with scaling factor s from 0.7 to 1.3 with increase of 0.02.
(ii) Noise injection. The zero-mean 0.01-variance Gaussian noise was embedded to the sign language images in original dataset to generate 30 new noised images.
(iii) Random translation. The hand gesture image was randomly shifted by 30 times. The value of the random shift at both horizontal and vertical directions $t = [t_x, t_y]$ lies in the scope of [−15, 15] pixels, and obeys uniform distribution.
(iv) Gamma correction. The gamma factor R differed in the range of [0.4, 1.6] with increase of 0.04.
(v) Affine transform. It exerted deformation to the images, while preserved straight lines.
(vi) PCA color augmentation. It shifted the color values which were the most present in original images.

Thus, one original image will generate 180 new images. The augmented training set now has 1056 * 181 = 191,136 images, as shown in Table 1. The experiment repeated ten times. Each time the data split was reset randomly.

3 Methodology

3.1 Convolutional layer

The convolutional neural network is a typical feedforward neural network that contains multiple layers of deep structures and combines two functions of feature extraction and classification recognition. It generally contains input, convolutional layer, pooling layer, fully connected layer, output, and so on. The convolution layer extracts the input data by convolution operation, and the pooling layer realizes data dimensionality reduction and controls the calculation burden to prevent over-fitting. The fully connected layer mainly performs the classification function. As needed, we can also add function functions in the middle of these layers, such as batch normalization, dropout.

The convolutional layer is composed of various convolutional units with learning capabilities. The subsequent convolution layer extracts more complex features based on the previous low-level features, and finally achieves feature extraction of the target object by adding a larger number of convolution layers [28]. Figure 3 shows the entire process of convolutional layer starting from the input and finally outputting as a feature map through a series of filters (Table 2).

The two-dimensional convolution of the convolutional layer is done between the three-dimensional input and the learned filters, with the directions of width and height [20]. Suppose that the input, filter and output size are shown in Table 2. The feature map size is calculated as follows:

Table 1 Partition of dataset

Partition	Number of Images
Training	1056
Augmented training	191,136
Test	264

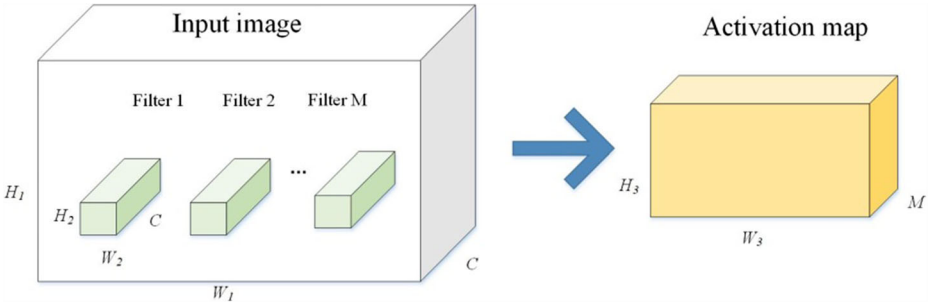


Fig. 3 Illustration of convolution operation

$$W_3 = 1 + \frac{(W_1 - W_2 + 2P)}{S} \tag{1}$$

$$H_3 = 1 + \frac{(H_1 - H_2 + 2P)}{S} \tag{2}$$

Here, the input size is $W_1 \times H_1 \times C$, output size is $W_3 \times H_3 \times M$. In the specified hyperparameter, M indicates the number of filters, S denotes the stride size, and P is padding size.

3.2 Pooling layer

Pooling is also called subsampling. To avoid overfitting and reduce computational burdens, the pooling layer is often used to achieve dimensionality reduction by using a neuron value to represent an area until all neurons are represented [22, 23, 30]. This achieves compression of the convolutional layer output size. Moreover, pooling can help to maintain translation invariance. There are two common pooling methods: max pooling and average pooling.

Max pooling is achieved by selecting the maximum value of the pooling region while the average pooling obtains a condensed feature map by calculating the average of the elements in each pooling region. An example is shown in Fig. 4, where the filter size equals 2 and stride is 2.

Suppose the pooling region is R, we can define the activation set X included in R as

$$X = [x_i | i \in R] \tag{3}$$

The max pooling P_M is expressed as:

$$P_M = \max(X_R) \tag{4}$$

Table 2 Size of Input, Filter and Output

Operator	Size
Input	$W_1 \times H_1 \times C$
Filter	$W_2 \times H_2 \times C \times M$
Output	$W_3 \times H_3 \times M$

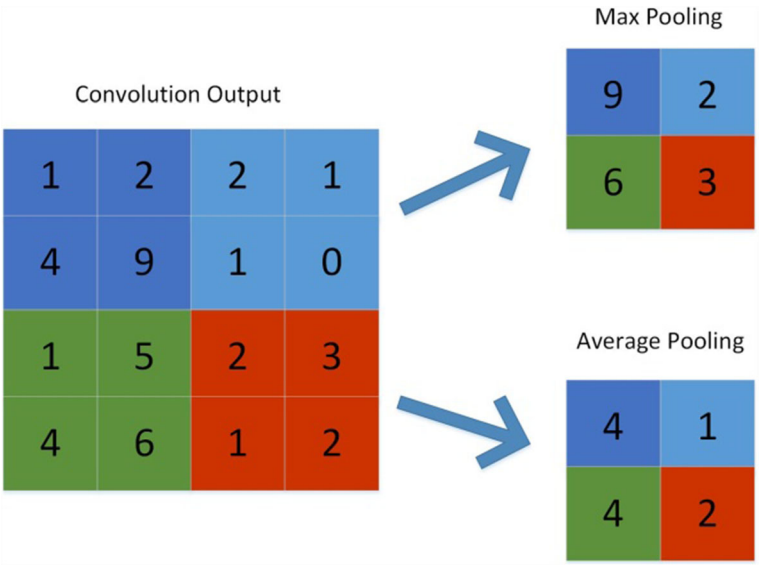


Fig. 4 Example of Max Pooling and Average Pooling

While the following equation gives the definition of average pooling P_A

$$P_A = \frac{\sum X_R}{|X_R|} \tag{5}$$

here $|X_R|$ is the number of the elements in the set X.

Although both methods are popular, they have their own shortcomings. In general, average pooling can only reduce the error of the estimated variance caused by the limited size of the neighborhood, and it retains more background information of the image. Max pooling can only reduce the offset of the estimated mean due to convolutional layer parameter errors, it retains more texture information. In addition, the max pooling usually overfits training data [5].

In order to bridge these gaps, the researchers turn to the probabilistic pooling method. Stochastic pooling (SP) was proposed, which is somewhere in between. By giving the probability of the pixel points according to the numerical value, and then subsampling according to the probability, in the average sense, it is similar to average pooling, and in the local sense, it obeys max pooling guidelines. This process can be expressed as the following two steps:

(1) Calculate the probability map p_i via original activation map x_i.

$$p_i = \frac{x_i}{\sum_X x_i} \tag{6}$$

(2) Pick a location k within the activation region X according to the probability p. Therefore, stochastic pooling P_S can be defined as follow

$$P_S = x_k, where\ k \sim P(p_1, \ldots, p_i, \ldots) \tag{7}$$

Figure 5 presents a stochastic pooling example. It generates a probability map firstly and then randomly chooses the location k as 4, which has corresponding position at (2, 1) and value of 0.4. Finally, the output of P_S is 4 of the original activation map.

3.3 Batch normalization

In the process of deep network training, the change of the parameters of the previous layer often influences the distribution of the data in the latter layer, and also affects the speed of training [14]. With function of unified decentralized data and optimized neural network, Batch Normalization (BN) algorithm can settle this problem well. By inserting a normalization layer and performing a normalization operation after each layer, BN forces the input values of any neurons in each layer of the neural network to be distributed back to the standard normal distribution, that is, the mean is 0, and the variance is 1. This prevents the issue of gradient disappearance and accelerates learning convergence [6].

The formula for the forward conduction process of BN network layer is as follows:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{8}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \mu)^2 \tag{9}$$

$$z_i^{'} = \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{10}$$

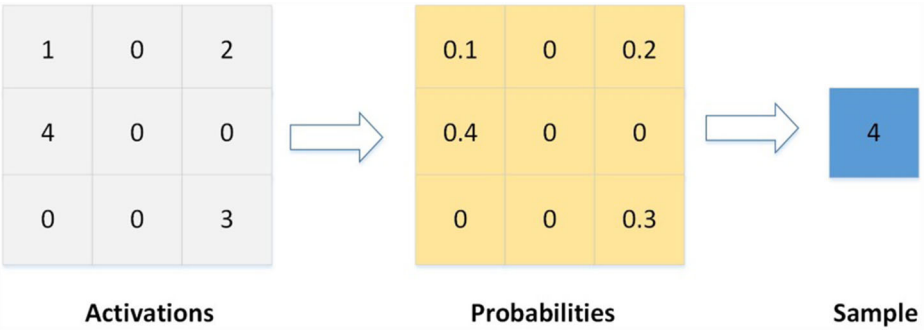


Fig. 5 Example of Stochastic Pooling

$$o_i = \gamma z_i^{'} + \beta \equiv ZN_{\gamma,\beta}(z_i) \tag{11}$$

where $[z_i]$ indicates input set, $Z = [z_{1\ldots n}]$, and $[o_i]$ is a mini-batch output. In this study, we define the mini-batch size as 256.

As shown in Fig. 6, the output from convolutional layer or fully connected layer supplies the source for the input of batch normalization, and then the output of BN turns into the input of other layers.

3.4 Dropout

Overfitting and time consuming are two major embarrassments in training deep neural networks. The dropout can validly resolve the occurrence of overfitting and achieve regularization to some extent. The realization of dropout can be divided into two steps, first training the entire neural network, and then averaging the results of the whole collection. Dropout traverses layer by layer, dropping out several neurons randomly with probability P, and keeps other neurons with probability $Q = (1\text{-}P)$, where the value of P is commonly set as 0.5. The output of all discarded neurons is set to zero, which ultimately results in a network with fewer nodes and smaller scales, reducing the links and making the neural network easier to train [9].

An example of dropout neural network is shown in Fig. 7, where the blue solid circle represents a normal neuron and the dotted circle denotes a dropout neuron. It can be seen that each layer drops out some neural units at a certain dropout rate while preserving the remaining neural units. Taking the second layer in Fig. 7 as an example, three neural units are discarded, and the other two are retained, which refines the original network layer. Obviously, network after the application of dropout has fewer nodes and been shrunk.

3.5 Experiment setting

This experiment was in-house developed and run on the platform of a personal computer with 2.5 GHz Core i7 CPU, and 16 GB memory, under the operating system of Windows 10. The maximum epoch was set to 30. The mini batch size was set to 256. The global learning rate was set to 0.01, and decreased to one-tenth of its previous value every 10 epochs. In total, the setting is listed in Table 3.

We evaluate the experiment results using “overall accuracy”, which is defined as proportion of samples that are correctly classified in all samples. It is computed by dividing the number of correctly predicted items by the total of item to predict [2].

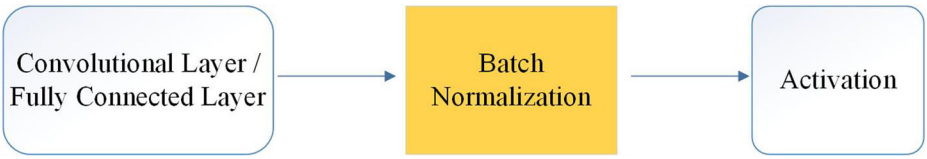


Fig. 6 Illustration of batch normalization

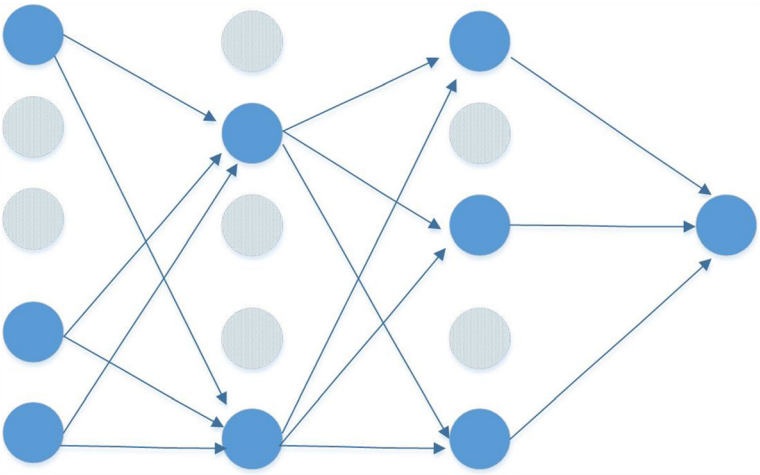


Fig. 7 An example of dropout neural network

4 Experiment results

4.1 Data augmentation results

We use the ch image as an example, which was shown in the top-left corner of Fig. 2. The data augmentation results are shown below in Fig. 8. From Fig. 8a–f, six enhancement methods such as gamma correction, PCA color augmentation, affine transform, noise injection, scaling and random shift are listed, respectively. A total of 180 new augmented ch images were generated, which created new training sets 181 times as larger as before. As we all know, sufficient image data sets are benefit for deep learning. Data augmentation can expand the data set, and it also helps to overcome over-fitting and improve classification accuracy.

4.2 Structure of proposed CNN

After tuning, we finally determine an eight-layer CNN with 6 convolutional layers and 2 fully-connected layers. Their details are listed in Table 4. For instance, in Block_2, the hyperparameters represent that the number of filters is 64 and its width is 3, the height is 3, the channel is 32, respectively. Here, batch normalization (BN),

Table 3 Setting of neural network model

Hyperparameter	Value
Maximum epoch	30
Mini batch size	256
Initial learning rate	0.01
Drop factor of learning rate	0.1
Drop period of learning rate	10

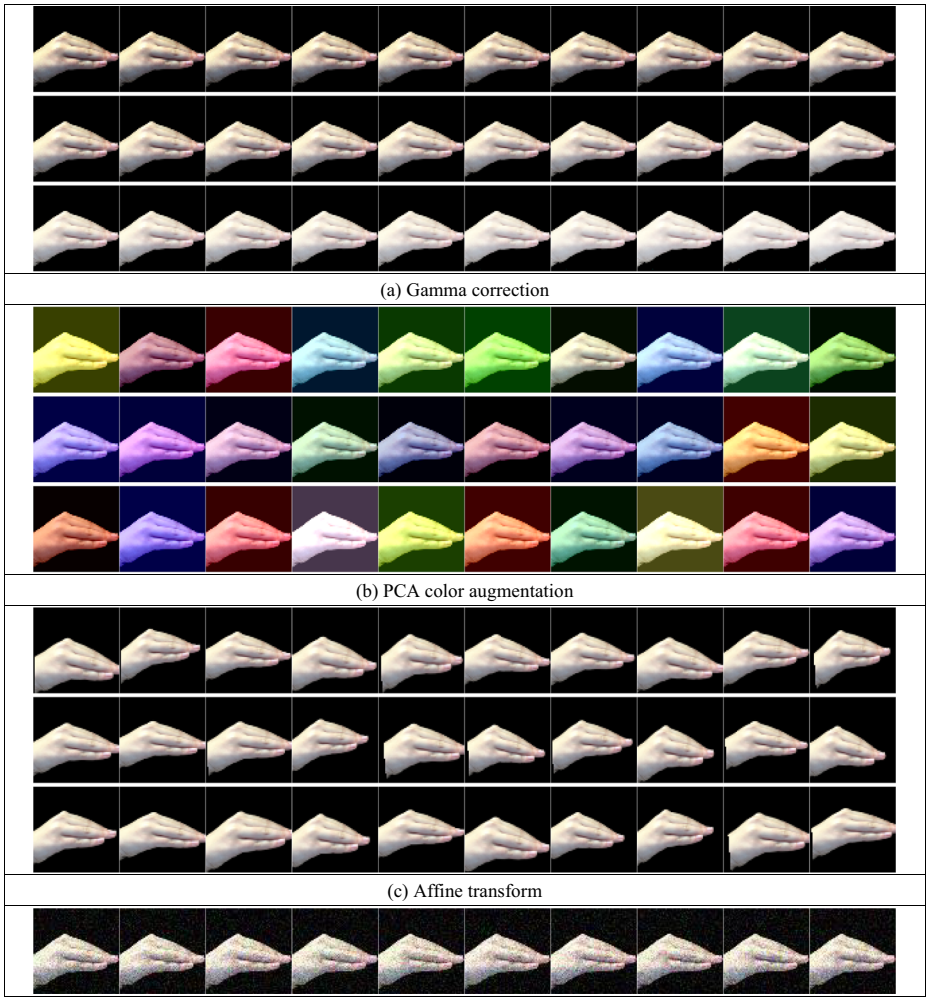
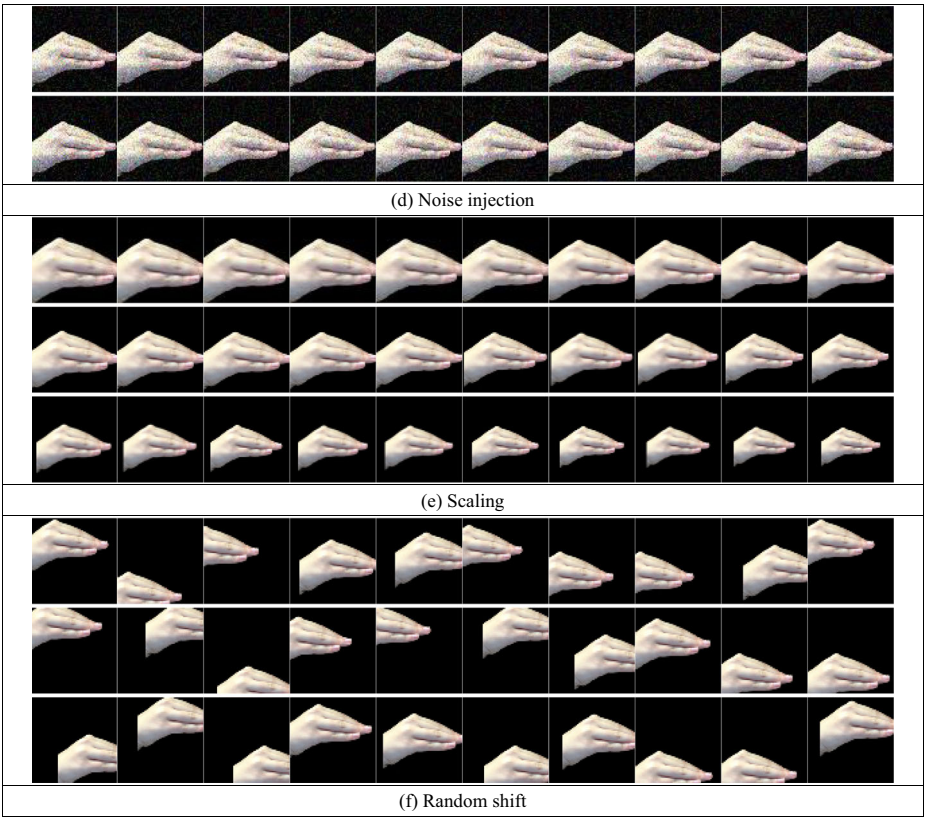


Fig. 8 Data augmentation of ch sample image

rectified linear units (ReLU) and stochastic pooling (SP) components are employed with the convolutional layer. Meanwhile, the value of stride is set to 2. The other blocks are similar in parameters setting. In first fully-connected layer which includes dropout, the dropout rate is decided as 0.4 by seeking in experiments.

4.3 Statistical analysis

We used this proposed eight-layer CNN, in which we employed batch normalization, dropout, and stochastic pooling components. The results of 10 runs are shown in Table 5. It can be seen that the highest accuracy is 90.91% which has been marked in bold and the minimum accuracy is 87.12%. In addition, the accuracy of eight runs exceeds 89%, and the overall accuracy reaches $89.32 \pm 1.07\%$, which is relatively efficient and stable.



(d) Noise injection

(e) Scaling

(f) Random shift

Fig. 8 (continued)

4.4 Pooling method comparison

In this experiment, we compared the stochastic pooling (SP) with other two orthodox pooling methods: average pooling (AP) and maximum pooling (MP). The comparison is shown in Table 6. The average accuracies of SP, AP and MP are $89.32 \pm 1.07\%$, $86.67 \pm 1.01\%$, and $88.86 \pm 1.42\%$, respectively. It demonstrates that SP is considerably better than AP and MP in measure of accuracy. We also can see that SP and MP

Table 4 Details of each layer in proposed CNN

Layer	Activations	Weights	Bias
Input	256x256x3		
Block_1 (32 7x7x3 /2 Conv, BN, ReLU, and 3 × 3 /2 SP)	64x64x32	7x7x3x32	1x1x32
Block_2 (64 3x3x32 Conv, BN, ReLU, and 3 × 3 /2 SP)	32x32 × 64	3 × 3 × 32x64	1x1x64
Block_3 (128 3x3x64 Conv, BN, ReLU, and 3 × 3 /2 SP)	16x16x128	3x3x64x128	1x1x128
Block_4 (128 3x3x128 Conv, BN, ReLU, and 3 × 3 /2 SP))	8x8x128	3x3x128x128	1x1x128
Block_5 (256 3x3x128 Conv, BN, ReLU, and 3 × 3 /2 SP)	4x4x256	3x3x128x256	1x1x256
Block_6 (256 3x3x256 Conv, BN, ReLU, and 3 × 3 /2 SP)	2x2x256	3x3x256x256	1x1x256
Dropout (40) and FCL_1 (100-d)	1x1x100	100 × 1024	100 × 1
FCL_2 (30-d)	1x1x30	30 × 100	30 × 1

Table 5 Ten runs of our method

Run	Our Method
1	89.77
2	87.12
3	89.77
4	87.88
5	89.77
6	89.39
7	89.39
8	89.39
9	89.77
10	**90.91**
Average	**89.32 ± 1.07**

both achieve the highest accuracy of 90.91% while AP doesn't touch this line. Furthermore, the minimum accuracy of AP is 84.85%, which is much lower than 87.12% in SP and MP.

4.5 Dropout rate

We varied the dropout rate from 0% to 90%, and recorded the 10-run results in Table 7. The error bar was shown in Fig. 9. As can be seen, when dropout rate is 40%, the highest overall accuracy reaches 89.32 ± 1.07%, which gives the best performance. In general, with the increase of the dropout rate, the overall accuracy is rising and reaches its peak with dropout rate of 40%. Then it begins to decline, the second highest overall accuracy 88.98 ± 1.29% appears with dropout rate of 70%. After that, the overall accuracy continuously keeps dropping. Therefore, the optimal dropout rate was sought at 40%.

4.6 Comparison to state-of-the-art approaches

In this experiment, we compared our method with state-of-the-art approaches: HMM [11], SVM-HMM [13], HCRF [38]. The comparison results are listed in Table 8. We can observe that our method is better than HMM of 83.77%, SVM-HMM of 85.14% and HCRF of 78.00%. Our leading edge derives from deep learning that combines

Table 6 Comparison of average pooling, maximum pooling, and stochastic pooling

Run	Average Pooling	Maximum Pooling	Stochastic Pooling (Ours)
1	87.12	87.50	89.77
2	88.64	**90.91**	87.12
3	86.36	89.02	89.77
4	86.74	88.26	87.88
5	85.61	87.12	89.77
6	86.36	88.26	89.39
7	86.74	90.53	89.39
8	84.85	87.88	89.39
9	87.12	88.26	89.77
10	87.12	**90.91**	**90.91**
Average	86.67 ± 1.01	88.86 ± 1.42	89.32 ± 1.07

Table 7 10-run results against different dropout rate

Run	Dropout Rate									
	0	10	20	30	40	50	60	70	80	90
1	85.98	88.64	89.77	88.64	89.77	89.39	87.88	89.02	89.39	87.50
2	86.36	88.26	89.02	90.15	87.12	88.26	88.64	89.39	87.88	87.50
3	87.50	86.74	89.39	88.64	89.77	87.12	87.88	88.64	89.77	88.64
4	88.64	87.12	88.26	88.26	87.88	89.02	88.26	86.36	87.88	88.26
5	87.88	85.98	89.02	88.64	89.77	90.15	88.64	91.29	87.88	90.15
6	87.12	89.39	87.50	88.64	89.39	90.53	87.12	89.39	89.77	89.02
7	86.36	90.15	88.26	88.26	89.39	89.39	88.26	88.26	87.50	86.36
8	87.12	86.36	87.12	87.50	89.39	87.50	87.12	89.02	89.39	87.50
9	87.88	88.26	85.98	87.50	89.77	85.98	89.39	90.15	87.50	88.64
10	87.88	85.98	87.88	90.53	90.91	86.74	88.64	88.26	88.64	89.39
Avr ± SD	87.27 ± 0.84	87.69 ± 1.46	88.22 ± 1.15	88.67 ± 0.99	89.32 ± 1.07	88.41 ± 1.53	88.18 ± 0.71	88.98 ± 1.29	88.56 ± 0.94	88.30 ± 1.11

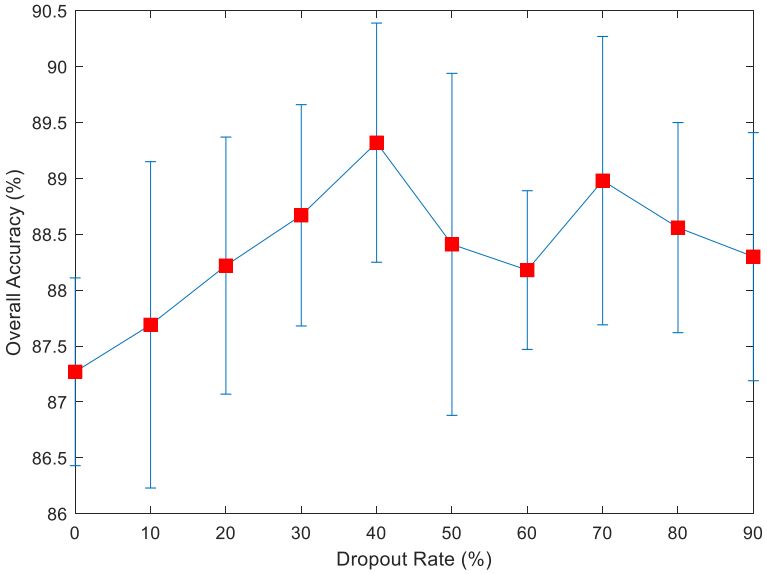


Fig. 9 Error bar of overall accuracy against dropout rate

multiple technologies. Firstly, stochastic pooling can resolve the overfitting and down-weight issue. Secondly, batch normalization can help to accelerate learning convergence and prevent the issue of gradient disappearance. Thirdly, dropout can effectively solve the occurrence of overfitting and achieve regularization. Finally, data augmentation was applied to enhance the generality of deep neural network. Thus, our method obviously has superiority to other state-of-the-art approaches compared.

5 Conclusion

This study proposed an optimized eight-layer convolutional neural network with stochastic pooling, batch normalization and dropout for fingerspelling recognition of Chinese sign language. The result demonstrated that our method was superior to three state-of-the-art approaches, even better than the second best method SVM-HMM by 4% in terms of overall accuracy. We compared stochastic pooling against average pooling and maximum pooling method. The experiment outcomes indicated the excellence of stochastic pooling, which reduced overfitting effectively. Besides, batch normalization, dropout and data augmentation were employed to achieve superior performance. All these advanced technologies could

Table 8 Comparison with state-of-the-art approaches

Approach	# of Images	Overall Accuracy
HMM [11]	2700	83.77%
SVM-HMM [13]	300	85.14%
HCRF [38]	12,960	78.00%
eight-Layer CNN (Ours)	1320	89.32% ± 1.07%

overcome common issues in traditional CNN, which offered a big opportunity to elevate the integration of hearing-impaired people into society.

Nevertheless, there are some shortcomings to deal with. To improve accuracy, current data size is insufficient and more data need to be collected. To achieve excellent experiment results, the hyperparameters obtained by experience need to be optimized.

In the future, we will try to verify and filter a deep neural network of the appropriate depth and take more advanced technology to improve accuracy. The data set also will be further enlarged. We will try to shift the profits of this study to other fields, such as biomedical imaging, clinical oncology, blind fever screening, which will greatly help those in need. Besides, transfer learning [18, 19] is an alternative way to solve our task.

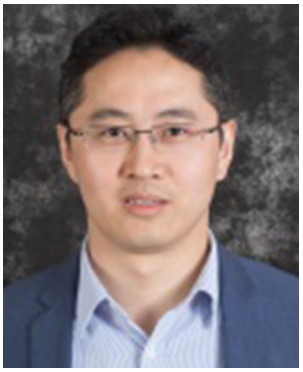
Acknowledgements This work was supported from Jiangsu Overseas Visiting Scholar Program for University Prominent Young & Middle-aged Teachers and Presidents of China, Henan Key Research and Development Project (182102310629), Natural Science Foundation of China (61602250).

References

1. Cheok ZOMJ, Jaward MH (2019) A review of hand gesture and sign language recognition techniques. Int J Mach Learn Cybern 10:131–153
2. Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 37(1):35–46
3. Dingqian SXG, Yuanyuan Y (2005) The analysis of Chinese sign language's basic words (basic movements). Chin J Spec Educ 2:65–72
4. Du T, Ren X, Li H (2018) Gesture recognition method based on deep learning. In: 33rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Nanjing, China, pp 782–787. IEEE.
5. Wang S-H, Tang C, Sun J, Yang J, Huang C, Phillips P and Zhang Y-D (2018) Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling. Front. Neurosci. 12:818. https://doi.org/10.3389/fnins.2018.00818
6. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML), vol 37, pp 448–456. ACM.
7. Jiang Y (2018) Exploring a smart pathological brain detection method on pseudo Zernike moment. Multimed Tools Appl 77(17):22589–22604
8. Huang J, Zhou W, Zhang Q, Li H, Li W (2018) Video-based sign language recognition without temporal segmentation. Thirty-Second AAAI Conference on Artificial Intelligence: 2257–2264
9. Khan SH, Hayat M, Porikli F (2019) Regularization of deep neural networks with spectral dropout (in English). Neural Netw 110:82–90
10. Kong FQ (2018) Ridge-based curvilinear structure detection for identifying road in remote sensing image and backbone in neuron dendrite image (in English). Multimed Tools Appl 77(17):22857–22873
11. Kumar P, Saini R, Roy PP (2017) A position and rotation invariant framework for sign language recognition (SLR) using Kinect. Multimed Tools Appl 77:8823–8846
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436
13. Lee GC, Yeh F, Hsiao Y (2016) Kinect-based Taiwanese sign-language recognition system. Multimed Tools Appl 75:261–279
14. Leopold H A, Orchard J, Zelek J S, Lakshminarayanan V (2019) PixelBNN: Augmenting the pixelCNN with batch normalization and the presentation of a fast architecture for retinal vessel segmentation. Journal of Imaging 5(2): 26
15. Li X (2017) Research on Chinese Sign Language Recognition for Middle and Small Vocabulary based on Neural Network. University of Science and Technology of China, pp 1–2
16. Li T H S, Kao M C, Kuo P H (2016) Recognition system for Home-Service-related Sign Language Using Entropy-Based$K$-Means Algorithm and ABC-Based HMM. IEEE transactions on systems, man, and Cybernetics: systems 46(1):150–162
17. Lichtenauer JF, Hendriks EA, Reinders MJT (2008) Sign language recognition by combining statistical DTW and independent classification. IEEE Trans Pattern Anal Mach Intell 30(11):2040–2046

18. Liu J. Detecting cerebral microbleeds with transfer learning. Mach Vis Appl. Accessed on 22 April. Available https://doi.org/10.1007/s00138-019-01029-5
19. Lu S (2019) Pathological brain detection based on AlexNet and transfer learning. J Comput Sci 30:41–47
20. Muhammad K (2019) Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. Multimed Tools Appl 78:3613–3632
21. Oscar Koller SZ, Ney H, Bowden R (2018) Deep sign: enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. Int J Comput Vis 126:1311–1325
22. Pan C (2018) Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. J Comput Sci 27:57–68
23. Pan C (2018) Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU. J Comput Sci 28:1–10
24. Pariwat T, Seresangtakul P (2017) Thai finger-spelling sign language recognition using global and local features with SVM. 9th International conference on knowledge and smart technology (KST), IEEE: 116–120
25. Zhang Y, Wu L, Peterson B, Dong Z (2011) A two-level iterative reconstruction method for compressed sensing MRI. Journal of Electromagnetic Waves and Applications 25(8-9):1081–1091
26. Qian P (2018) Cat swarm optimization applied to alcohol use disorder identification. Multimed Tools Appl 77(17):22875–22896
27. Rao GA, Kishore PVV, Kumar DA, Sastry ASCS (2017) Neural network classifier for continuous sign language recognition with selfie video. Far East Journal of Electronics and Communications 17(1):49
28. Sellami A, Hwang H (2019) A robust deep convolutional neural network with batch-weighted loss for heartbeat classification (in English). Expert Syst Appl 122:75–84
29. Sun J (2018) Preliminary study on angiosperm genus classification by weight decay and combination of most abundant color index with fractional Fourier entropy. Multimed Tools Appl 77(17):22671–22688
30. Tang C (2018) Twelve-layer deep convolutional neural network with stochastic pooling for tea category classification on GPU platform. Multimed Tools Appl 77(17):22821–22839
31. Wei G (2010) Color image enhancement based on HVS and PCNN. SCIENCE CHINA Inf Sci 53(10): 1963–1976
32. Zhang Y, Wu L (2008) Improved image filter based on SPCNN. Science in China Series F-Information Sciences 51(12):2115–2125
33. Wu LN (2008) Pattern recognition via PCNN and Tsallis entropy (in English). Sensors 8(11):7518–7529
34. Zhang Y, Wu L (2009) Segment-based coding of color images. Science in China Series F-Information Sciences 52(6):914–925
35. Wu L (2011) Optimal multi-level Thresholding based on maximum Tsallis entropy via an artificial bee Colony approach. Entropy 13(4):841–859
36. Yan J (2010) Find multi-objective paths in stochastic networks via chaotic immune PSO. Expert Syst Appl 37(3):1911–1919
37. Yang J (2019) An adaptive encoding learning for artificial bee colony algorithms. J Comput Sci 30:11–27
38. Yang H-D, Lee S-W (2010) Robust sign language recognition with hierarchical conditional random fields. In: 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp 2202–2205. IEEE
39. Zhao G (2018) Smart pathological brain detection by synthetic minority oversampling technique, extreme learning machine, and Jaya algorithm. Multimed Tools Appl 77(17):22629–22648

Mr. Xianwei Jiang received his B.S. degree from Nanjing Normal University (1998-2002) and M.S. degree from Nanjing University of Aeronautics and Astronautics (2007-2011). He visited the Department of Informatics, University of Leicester, UK, from September 2018 to September 2019. Now, he is an associate professor in the College of Mathematics and Information Science, Nanjing Normal University of Special Education. His current research interests include deep learning, computer vision and special education informatization.



Dr. Mingzhou Lu received a B.S. from Nanjing Normal University (1998-2002) and a M.S. from Nanjing University of Aeronautics and Astronautics (2007-2010). He received his Ph.D. from Nanjing Agricultural University (2010-2014). He worked as an associate professor in Nanjing Agricultural University since 2013. He worked in KU Leuven as a visiting scholar (2015-2016). He is a member of Chinese Computer Federation (CCF), member of the Committee of Computer Vision of CCF, member of Chinese Society of Agricultural Engineering. He is working on the precision Agriculture technology based on computer vision. In the past three years, he worked as a principle investigator in 2 research projects funded by Chinese government. He has published more than 20 publications. He also holds nearly 10 patents which were issued by China.

Dr. Shui-Hua Wang received a B.S. from Southeast University (2005-2008) and a M.S. from The City University of New York (2010-2012). She worked as a Research Assistant in Columbia University (2012-2014). She received her Ph.D. from Nanjing University (2014-2017). She worked as a Research Associate in Loughborough University (2018-2019). She is now working as a Research Fellow in University of Leicester. She published over 100 papers in SCI-indexed journals. She served as the editor of IEEE Access and Journal of Alzheimer's disease from 2018, and the managing guest editor of Multimedia Tools and Applications (2017-2018).

Affiliations

Xianwei Jiang [1] · Mingzhou Lu [2] · Shui-Hua Wang [3,4]

[1] Nanjing Normal University of Special Education, Nanjing 210038, China

[2] College of Engineering/Jiangsu Province Engineering Lab for Modern Facility Agriculture Technology & Equipment, Nanjing Agricultural University, Nanjing 210031, China

[3] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, People's Republic of China

[4] School of Architecture Building and Civil engineering, Loughborough University, Loughborough LE11 3TU, UK