Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-7193-4

A hand gesture recognition algorithm based on DC-CNN

Xiao Yan Wu[1]

Received: 3 November 2018 / Revised: 27 December 2018 / Accepted: 9 January 2019 /
Published online: 24 January 2019

Abstract
In the process of hand gesture recognition, the diversity and complexity of gesture will greatly influence the recognition rate and reliability. In the task of hand gesture recognition, the traditional method based on manual feature extraction is time-consuming, and the recognition rate is low. In order to improve the recognition rate, a novel recognition algorithm based on double channel convolutional neural network (DC-CNN) is proposed. Firstly, the preprocessing, denoising and edge detection of original gesture images are performed to obtain the hand edge images. Secondly, the hand gesture images and the hand edge images are respectively selected as two input channels of the CNN. Each channel contains the same number of convolutional layers and the same parameters, but each has a separate weight. Finally, the feature fusion is performed at the full connection layer and the output result is classified by softmax classifier. Experiments on the Jochen Triesch Database (JTD) and the NAO Camera hand posture Database (NCD) show that the proposed algorithm has improved the rate of hand gesture recognition and has enhanced the generalization ability of the CNN.

Keywords CNN · Double channel · DC-CNN · Gesture recognition · Deep learning

1 Introduction

With the development of computer performance, the interaction between human and computer is becoming more and more frequent. Gestures are a way of human-computer interaction, and gestures recognition is to analyze the specific meaning of each gesture by computer. By recognizing gestures, we can obtain the expression of people, which can be used to achieve the purpose of intuitive and intelligent man-machine interaction.

There are many common gesture recognition methods. The recognition method based on geometric features is used to recognize the gesture structure, edge, contour and other features [6, 19, 26]. Although the recognition method based on geometric features has good stability, the improvement of the recognition rate cannot be achieved by improving the sam-

✉ Xiao Yan Wu
37678324@qq.com

[1] Sichuan University of Arts and Science, DaZhou City, SiChuan Province, 635000, China

ple size. The recognition method based on hidden markov model has the ability to describe the spatial and temporal changes of gestures, however the recognition speed of this method is not satisfactory [20].

Deep learning is a new research field of machine learning. and its essence is a nonlinear network model with multiple hidden layers. The feature, expressed the original data, can be extracted from the network model. The feature can be used to predict or classify samples by training for large-scale raw data. In the field of image recognition and computer vision, convolutional neural network (CNN) has achieved the most remarkable results [14, 15, 24]. In addition, deep learning is widely used in pedestrian detection [16], gesture recognition [3], natural language processing [9], data mining and speech recognition. CNN can deal with the two-dimensional image directly, compared with other deep neural networks such as the depth confidence network [13], S layer automatic coding [23]. When the two-dimensional image is converted into one, the spatial structure characteristics of the input data are lost.The problem of insufficient feature extraction caused by artificial extraction features, has been solved by CNN through training and learning the local and global features of input images. CNN has strong feature extraction and classification ability, and has achieved remarkable achievements in image classification [18], face recognition [8] and speech recognition [1], and so far, it has been widely applied in many fields of pattern recognition.

In the field of image processing, the main application direction of CNN is image classification, target recognition and image segmentation. CNN has the characteristics of local connection, weight sharing, depth stratification and automatic feature extraction, which brings new ideas to the task of gesture recognition [5, 27]. There are many scholars have explored the field of gesture recognition. Yamashita [29] discussed the fusion of the pretreatment process and gesture recognition process before the input convolution network, then the end-to-end gesture recognition is realized, and the recognition rate is improved. Pavlo et al. [29] successfully solved the problem of the dynamic gesture recognition input rules. They change the first layer of convolutional layer of the CNN to a three-dimensional convolution, so that the dynamic gesture can be entered into a model in the form of a stereoscopic view. Pablo et al. [2] creatively used the stereo convolution kernels for gesture recognition, and a better recognition rate of gesture recognition was obtained.

Multi-channel convolution neural network is a novel structure of convolution neural network. It can accept different features of image as input and perform convolution processing respectively. Then these features are combined for image classification. In order to improve the rate of hand gesture recognition, a hand gesture recognition algorithm based on double channel convolutional neural network (DC-CNN) is proposed. The original gesture images are preprocessed, denoised and edge detected. Then the edge images and hand gesture images are respectively used as two input channels of the CNN. Each channel contains the same number of convolutional layers and the same parameters, but each has a separate weight. The feature fusion is performed at the full connection layer and the recognition result classified by softmax classifier finally.

The rest of the paper is organized as follows. In Section 2, the SAR image model, the wavelet filter and fast guided filter are introduced, then the proposed algorithm is presented. Section 3 gives the experimental results. At last, Section 4 draws a conclusion of this paper.

2 Methodology

In this section, we will introduce the convolution neural network firstly. Secondly, we present the edge detetion method. The proposed algorithm will be introduced at last.

2.1 Convolution neural network

The understanding of CNN images is from local to global, because the local pixel image is closer and far away from the local correlation. At first CNN perceives local features and then combines these local features at a higher level. In this way, it can obtain the global characteristics of the image and its topological structure, then the properties and categories of the images can be determined. Therefore, CNN is highly invariant to the translation, scaling, inclination or other forms of deformation.

There are two typical features of CNN. The first one is the locally connection between two layers of neurons by convolution kernels rather than the fully-connection. Therefore, the convolution layer connected to the input image is a local link built for the pixel block, rather than the traditional pixel point-based full connection. Second, the weight parameter of convolution kernel is shared in each same layer. These two characteristics have greatly reduced the number of the parameters of the deep web, the complexity of the model is reduced and the training speed is accelerated. It makes CNN have a big advantage in the pixel value of processing units. The main components of CNN include the convolutional layer, the pooling layer, the activation function, the full connection layer and the classifier, shown as Fig. 1. The first layer that is directly connected to the input image is the convolution layer, which carries the task of connecting images directly. By processing pixel values, the input is transformed into a form understandable by the convolution network. Then it's propagated in the convolution layer. The pooling layer and the activation function are usually attached at the back of the convolutional layer, alternating with the convolutional layer.

2.1.1 Convolution layer

The convolution layer is a core component of CNN. And its main function is to extract local features of the input and move through the fixed step length of the convolution kernel.

The output of each element on the feature map is the output of a neuron. The input of the neuron connection is a local area of the previous output feature graph. The input through the feel field is calculated by a set of synaptic weights and the neuron output is obtained by activating the function. In the process of generating the feature graph, the number of parameters can be greatly reduced by sharing this set of synaptic weights. For a network, the size of the convolution kernel is fixed, but the weight parameter of the convolution kernel is obtained through training sample training.

The convolution kernel is the core of convolution layer and it is a mapping relation of local receptive field extracting image features. The convolution kernel can also be considered as an eigenmatrix. When the convolution is operated, the convolution kernel moves

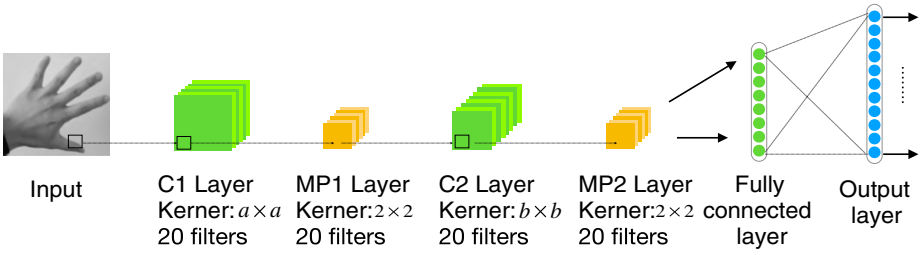

Fig. 1 The structure of CNN

continuously at the input end. The convolution value of the sensory field is obtained by integrating the corresponding elements in the sensory field. After moving, the input characteristic matrix is obtained as well as the characteristic diagram. The mathematical expression of convolution operation is as follows.

$$\mathbf{a}_n^l = f(\sum_{\forall m}(\mathbf{a}_m^{l-1} * \mathbf{k}_{m,n}^l) + \mathbf{b}_n^l) \tag{1}$$

where a_n^l and a_m^{l-1} are the current layer's feature map and the previous feature map. $k_{m,n}^l$ represents the convolution kernel from the mth feature graph of the previous layer to the nth feature graph of the current layer. $f(x) = 1/[1 + exp(-x)]$ is neuron activation function. b_n^l represents the bias of neurons. It is the response of convolution kernel with the convolution kernel of the previous layer feature map, and the different convolution kernel can be used to extract different characteristics.

2.1.2 Pooling layer

In the pool layer, function transformation is performed on the non-overlapping rectangle region of the output feature graph of the previous layer. Then the invariant characteristics of higher layer can be obtianed. Its function is to transform the characteristic graph of convolution layer and reduce the dimension of feature graph. At the same time, the output is less sensitive to the deformations including tilting, displacement, and other forms. So the generalization ability of the model is improved. In the process of forward calculation, the input image and feature graph are gradually expanded and integrated into global characteristics through pooling operation. The commonly pooling methods are meanpooling and maxpooling. The pooling layer can reduce the feature dimension while preserving the original feature information. The pooling layer nodes output can be expressed as follows.

$$\mathbf{a}_n^l = f\left(\mathbf{k}_n^l \times \frac{1}{s^2}\sum_{s\times s}\mathbf{a}_n^{l-1} + \mathbf{b}_n^l\right) \tag{2}$$

where $s \times s$ is the lower sampling template dimension, k_n^l is the power of the template.

2.1.3 Fully connected layer

Since the multiple convolutional kernel templates are usually used by the convolution layer, the output is also the same size feature graph. One or more fully connected layers are required in order to fuse the feature maps for classification. The full connected layer is generally connected to the last layer of pooling layer and classifier, which is used to fuse the different characteristics expressed by multiple feature graphs. The powerful feature extraction capability of the convolutional neural network comes from the convolution operation of the multi-convolution kernel template. The output of each convolution kernel template represents a characteristic expression of a different angle. So it's an important process to merge these features together. Each neuron in the full connected layer is connected to all the neurons in previous layer of the output characteristics, which can be expressed as follows.

$$\mathbf{a}_n^{out} = f(\sum_{\forall m}(\mathbf{a}_m^{out-1} * \mathbf{k}_{m,n}^{out}) + \mathbf{b}_n^{out}) \tag{3}$$

The full connection layer combines all the features of previous feature map, then enters it into the softmax classifier. The common activation functions are Sigmoid, Tanh, and Relu (Rectifiedunit).

2.2 Edge detection

To get a smooth image without losing edge, the fast guided filter is adopted to denoise the orignial hand gesture image. Then the edge detection algorithm of Canny is performed to get the edge image.

2.2.1 Denoised by fast guided filter

The fast guided filter is an effective smoothing filter, and it has a good ability of edge-preserving smoothing [10–12]. The key assumption of the fast guided filter is a local linear model between the guidance image G and the filter output image F. The guidance image is guided for smoothing an input image. It is supposed that F is a linear transform of G in a window ω_k, centered at pixel k. The guided filter can be expressed as follows.

$$F_i = a_k G_i + b_k, \forall i \in \omega_k \tag{4}$$

where i is the index of a pixel. a_k, b_k are some linear coefficients assumed to be constant in w_k. This local linear model ensures that F has an edge only if G has an edge since $dtriF = adtriG$ [28]. Suppose that the filtering input image is I, then minimize the reconstruction error between F and G.

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i\in\omega_k} G_i I_i - \mu_k \mathrm{Imacr}_k}{\sigma_k^2 + \epsilon} \tag{5}$$

$$b_k = \mathrm{Imacr}_k - a_k \mu_k \tag{6}$$

where μ_k and σ_k are the mean and variance of G in the window ω_k. ϵ is a regularization parameter preventing a_k from being too large. Then the filtering output F can be computed.

$$F_i = \bar{a}_i G_i + \overline{b}_i \tag{7}$$

where $\overline{b}_i$ and $\overline{b}_i$ are the mean values in a_k and b_k. The main computation is a series of box filters, which does not need to be carried out at full resolution. The input image I and guidance image G can be subsampled by a ratio r. Then the two coefficient $\overline{b}_i$ and $\overline{b}_i$ are bilinearly upsampled to the original size. Finally, the output image F is still computed by $F_i = \bar{a}_i G_i + \overline{b}_i$.

In the process of denoising, the hand gesture image itself is selected as the reference image. As shown in Fig. 2, the hand gesture images ware filtered by fast guided filter, and the smooth image ware obtained.

2.2.2 Canny edge detection

John Canny proposed an edge detection algorithm, called the Canny operator [4]. Canny converts the edge detection problem into the maximum value problem of detection unit function. Canny believes that a good edge detection operator should have three characteristics, which are the good detection performance, good positioning performance, and low response times to the same edge.Three criteria for determining edge detection operators are proposed as follows.

(1) SNR (signal-to-noise ratio) criterion. The probability that the edge is not detected is minimized. And the probability of non-edge detection is minimized. In both cases, the probability decreases monotonously with the increase of SNR.

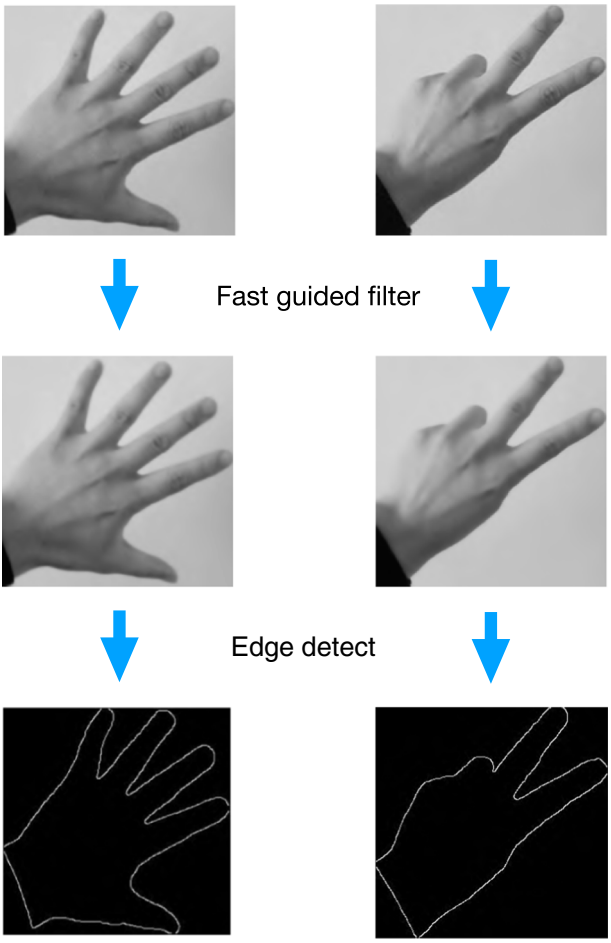

Fig. 2 The result of fast guided filter and Canny edge detection

(2) Positioning accuracy criteria. The detected edge is as far as possible in the center of the real edge.
(3) Single edge response criteria. The probability of multiple responses from a single edge is lower and the false edges are most suppressed.

Based on the above indexes and criteria, an expression composed of edge positioning precision and SNR product is derived by the method of functional derivation. This expression approximates the first derivative of gaussian function, which is the best approximation of the optimal function.

2.3 Proposed algorithm

The traditional Canny operator edge detection adopts the 1-order directional derivative in either direction of the two-dimensional gaussian function as the noise filter. Convolution filtering is performed on the image. Then, the filtered image is searched for the local extremum

to determine the edge of the image. Suppose that the two-dimensional gaussian function is as following.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} exp\left(\frac{1}{2\sigma^2}(x^2 + y^2)\right) \tag{8}$$

The first-order directional derivative of $G(x, y)$ in direction n can be expressed as following.

$$G_n = \frac{\partial G}{\partial n} = \mathbf{n} dtriG \tag{9}$$

$$w(x) = w(y) + w(n) \tag{10}$$

Equation 10 where $\partial n = \begin{bmatrix} cos\theta \\ sin\theta \end{bmatrix}^T$, $dtriG = \begin{bmatrix} \frac{\partial G}{\partial x} \\ \\ \frac{\partial G}{\partial y} \end{bmatrix}$, $\mathbf{n}$ is the direction vector, and $dtriG$ is the gradient vector.

Perform convolution transformation on images $f(x, y)$ with G_n, and change the direction of $\mathbf{n}$. When maximum $G_n f(x, y)$ is taken, the direction of the edge $\mathbf{n}$ can be obtianed.

Canny operator is used for the filtered gesture image, then the edge image can be obtained, which is shown in Fig. 2.

The multi-channel of CNN is used to access different views of the data (such as red, green, blue channel and stereo audio track of color images) [7]. It enables CNN to learn more abundant features and the classification effect of the model is improved by adding input information.

Therefore, in order to optimize the training process of gesture recognition, the double channel CNN (DC-CNN) is proposed in this paper, and the structure is shown in Fig. 3. The DC-CNN structure is composed of two relatively independent convolution neural network. The input of first CNN is the gesture image after preprocessing, and the input second CNN is gesture edge image. Each channel contains the same number of convolutional layers and parameters, but they have independent weights. After the pooling layer, the double channels are respectively connected to a full connection layer and a full connection map is performed.

The two channels are connected to a fully connected hidden layer, which produces the output of a logical regression classifier. The weight of each channel has its own update. But the final error is obtained through two output layers. So, two output layers are like a deviation from each other. This produces a specialized filter tuning based on the edge enhancement of the Canny filter. The training method of this model is forward calculation, then back propagation error update weight. In the process of forward calculation, in order to accelerate the training speed and reduce overfitting, the random gradient descent (SGD) method was adopted to calculate.

3 Experiment and analysis

The computer configuration of experiment is as follows, MacBook, macOS 10.12.6, 2.7GHz Intel Core i5 processor, 16GB 1867MHZ DDR, 256 GB hard disk. The software of experiment is MATLAB 2017a.

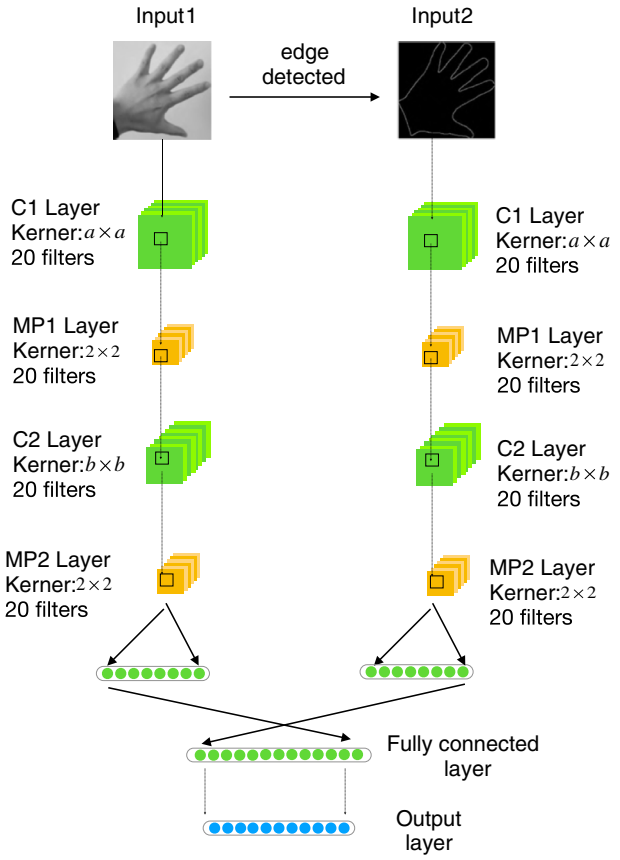

Fig. 3 The structure of DC-CNN

3.1 Experimental database

In order to evaluate the model, the experiment adopted the Jochen Triesch Database (JTD). This database contains 10 gestures, from 24 people in three different background, a light, a dark, completed before a complex background. All images have the size of 128×128, and centered in hand gesture. The examples of JTD database are shown in Fig. 4a.

To evaluate the model in a realistic man-machine interactive scene, a database called NAO Camera hand posture Database (NCD) was used. The examples of NCD database are shown in Fig. 4b. Each image has a resolution of 128×128 pixels, The database has significant execution differences, with a total of 400 500 examples per gesture. In each image, the hand is in a different position and is not always centered, sometimes obscured by some fingers. The CNN parameters, i.e. the number of convolutional layers, the dimension of the kernel and the max-pooling operation region were based on the research of [29]. The numbers of filters in each layer were found by evaluating the results for a range of numbers. Table 1 shows the parameters for each experimental setup.

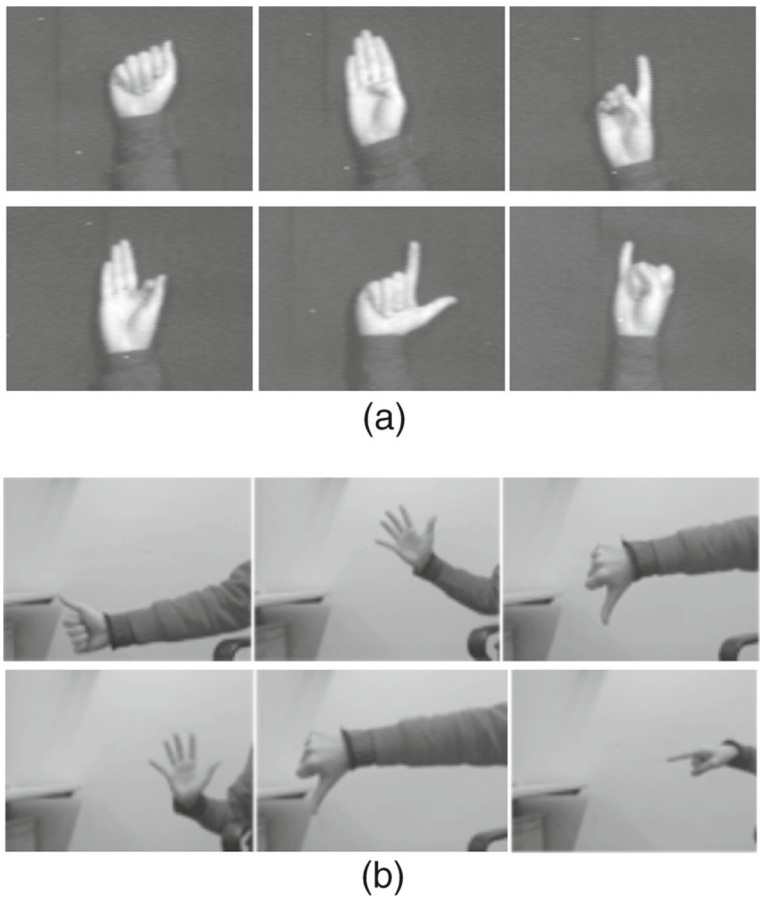

Fig. 4 **a** Examples of hand gestures with JTD. **b** Examples of hand gestures with NCD

3.2 Data preprocessing

The original gesture images in the database are 248×256 or 128×128 pixel size images, which contain a large amount of data. Moreover, the posture of the whole picture is not large, and the background is very redundant. If select the original image as the input to the convolution network directly, the amount of data that needs to be processed will be huge, and the classification results will be easily influenced by a sophisticated background. So

Table 1 The recognition rate of the double-channel and single-channel CNN

Model structure	Kernel size	JTD(%)	NCD(%)
Single-channel CNN	5×5	95.96	94.82
Single-channel CNN	7×7	96.21	95.05
Double-channel CNN	5×5 & 5×5	97.72	96.28
Double-channel CNN	7×7 & 7×7	98.02	97.29

the image is preprocessed first and then used as input to the model. Figure 4 is the preprocessing of gesture images, taking JTD database as an example. Preprocessing method is mainly to determine the center of the gesture, and then set a fixed scope, to pick up the full gesture and filter out most of the background. JTD database extracted the size of the gesture to 64×64 pixels, then reduced to 32×32 pixels by proportion. And the preprocessed image information is reduced to 1/64 of the original image.

In Fig. 5, the process of decreasing classification error and increasing iteration number in single-channel and double-channel CNNs training is compared. The experimental results show that the classification rate of the model tends to be stable after more than 20 iterations. The error rate of double-channel CNN is better than that of single-channel CNN, and it has a more stable convergence process.

3.3 Experiment and analysis

In order to verify the practical application of DC-CNN in gesture recognition, two groups of experiments are designed. In the first group of experiments, the neural network recognition effects of single channel and double channel convolution were compared. At the same time, the neural network identification accuracy of double-channel convolution kernel with different configuration is compared. In the second group of experiments, it compares the proposed DC-CNN algorithm with other gesture recognition algorithms.

3.3.1 The comparison of single-channel CNN and double-channel CNN

Because of different processing objects and different application scenarios, the size of the convolution kernel is not fixed. Therefore, only by matching the convolution kernel with the most appropriate convolution kernel can the optimal classification effect be obtained.

As a result, the experiment selected network including convolution kernel sizes are 5×5 & 7×7 single-channel convolution neural network. And the convolution kernel sizes

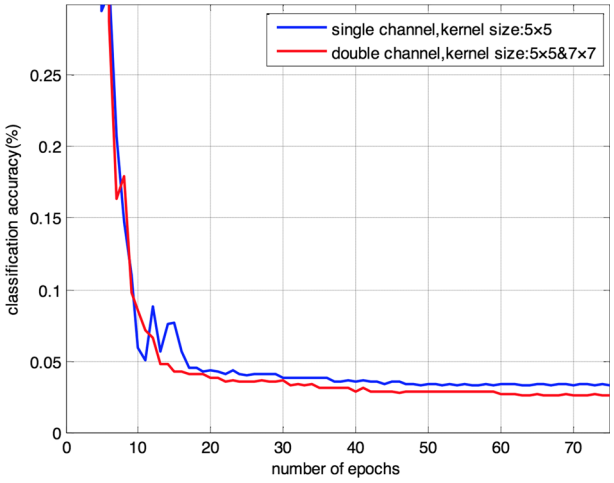

Fig. 5 The change curve of the recognition rate with the number of iterations

of double-channel convolution neural network are same with single-channel. The the experiments are performed respectively on JTD and NCD database.

Table 1 shows the recognition results of convolution kernels of different sizes on the database. After the double-channel convolution neural network was adopted, the rate of gesture recognition was significantly improved compared with single-channel convolution neural network. At the same time, it can be seen that the recognition rate is also different with different sizes of convolution kernels. The convolution kernel with dimensions of 7×7 and 7×7 has the best recognition effect. The experimental results show that, the proposed algorithm can obtain more abundant characteristic information by selecting the edge image as the second input channel. Therefore, the proposed algorithm can effectively improve the gesture recognition rate.

When training single-channel and double-channel convolution neural network, the classification error is reduced with the increase of the number of iterations, which is shown in Fig. 5. From Fig. 5, it can be seen that the classification accuracy of the model tends to be stable after more than 20 iterations. The error rate of double-channel convolution neural network is obviously better than that of single-channel convolution neural network, and it has a more stable convergence process.

3.3.2 The comparison of the proposed algorithm and other algorithms

The representative traditional gesture recognition algorithm and the recognition method based on CNN ware selected to compare with the proposed algorithm. Then the experiment was performed on JTD database. Based on the results of the previous experiment, the convolution kernel size of 7×7 & 7×7 is selected in the double-channel CNN model.

The results of comparison experiments are listed in Table 2. Among them, MPCNN [21] combines the maximum pooling with the CNN to form a deep CNN,a recognition of 97.38 percent is obtained. Bottom-up CNN [29] is an end-to-end deep CNN, and the rate of gesture recognition is 88.69 percent. The recognition rate of the proposed algorithm is 98.02 percent, which is higher than other models of CNN. Based on the above results, the following conclusions can be obtained.

(1) Double-channel convolution neural network can learn the more rich by using two input channels of hand gesture images and the edge images. Compared with the traditional single channel convolution neural network, the range of double-channel CNN feature extraction is wider. Therefore, the gesture recognition rate is higher.
(2) The proposed algorithm expands the traditional CNN, network training is also performed by use of supervision learning.The process of feature extraction requires no human involvement, it shows the excellent expansibility of the CNN. At the same time, it also shows the great potential of the structure extension of the convolution divinity network to improve the performance.

Table 2 The recognition rate of different gesture recognition algorithms

Algorithms	Reference	Recognition rate(%)
Spatial Pyramid	[17]	85.41
Bottom-up CNN	[29]	88.69
Tiled CNN	[22]	90.27
MPCNN	[21]	97.38
Proposed Method	–	98.02

4 Conclusion

In order to improve the hand gesture recognition rate, this paper presents a novel algorithm for gesture recognition, called DC-CNN. In the model of proposed algorithm, the hand gesture images and the edge mages are selected as two input channels. In order to obtain more abundant local information and overall topological structure of gestures, features are fused at the full connection layer after the pooling operation, and deeper classification information is extracted. The experiments are performed with two hand gesture databases, which databases are named JTD and NCD. Then the experiment results show that the proposed algorithm has improved the rate of hand gesture recognition. At the same time, it can adapt to simple and complex, bright and dark background forms and has strong generalization ability.

There is still a lot of space for research and development of DC-CNN, including the following three aspects. (1) Try to introduce more features of hierarchy and scale to further improve the model adaptability to complex background. (2) The rate of dynamic gesture recognition still has much space for improvement, and the model can be applied to the field of dynamic gesture recognition. (3) The convolution neural network model for gesture recognition requires a lot of image data with labels for training. In the future, training can be carried out through unsupervised learning or semi-supervised learning to reduce the dependence of the model on a large number of tag data.

References

1. Abdel-Hamid O, Deng L, Yu D (2013) Exploring convolutional neural network structures and optimization techniques for speech recognition[C]. Interspeech, pp 3366–3370
2. Barros P, Magg S, Weber C et al (2014) A multichannel convolutional neural network for hand posture recognition[C]. In: International conference on artificial neural networks. Springer, Cham, pp 403–410
3. Cai J, Cai JY, Liao XD et al (2015) Preliminary study on hand gesture recognition based on convolutional neural network[J]. Comput Syst Appl 24(4):113–117
4. Canny J (1986) Acomputationalapproachtoedgedetection[J]. IEEE Trans Pattern Anal Mach Intell 8(6):679–698
5. Cao X, Bo H (2016) Study on gesture recognition based on CNN [J]. Microcomputer Appl 35(9):55–57
6. Dong L, Ruan J, Ma Q, Wang L (2012) The gesture identification based on invariant moments and SVM[J]. Image Process Multimed Technol 31(6):32–35
7. Dumoulin V, Visin F (2016) A guide to convolution arithmetic for deep learning[J]
8. Farfade S S, Saberian MJ, Li L-J (2015) Multi-view face detection using deep convolutional neural networks[C]. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, 2015. ACM, pp 643–650
9. Goldberg Y (2017) Neural network methods for natural language processing[J]. Synthesis Lectures on Human Language Technologies 10(1):1–309
10. He K, Sun J (2015) Fast guided filter[J]. Computer Science
11. He K, Sun J, Tang X (2010) Guided image filtering. In: ECCV, pp 1–14
12. He K, Sun J, Tang X (2013) Guided image filtering. TPAMI 35(6):1397–1409
13. Hinton GE, Salakhutdinov RR (2006) Reducing the diensionality of data with neural networks. Science 313(5786):504–507
14. Hua-Fu LV (2018) Research on the static hand gesture recognition base on convolutional neural network[J]. Modern Computer
15. Jin LW, Zhong ZY, Yang Z (2016) Applications of deep learning for handwritten chinese character recognition: A review[J]. Acta Automatica Sinica 42(8):1125–1141

16. John V, Mita S, Liu Z et al (2015) Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks[C]. In: Proceedings of the 14th IAPR international conference on machine vision applications. IEEE, Tokyo, pp 246–249
17. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]. In: 2006 IEEE computer society conference on computer vision and pattern recognition. IEEE, vol 2, pp 2169–2178
18. Le QV (2013) Building high-level features using large scale unsupervised learning[C]. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 8595–8598
19. Liu Y, Yin Y, Zhang S (2012) Hand gesture recognition based on HU moments in interaction of virtual reality[C]. In: International conference on intelligent human-machine systems and cybernetics. IEEE, pp 145–148
20. Murthy GRS, Jadon RS (2010) Hand gesture recognition using neural networks[C]. In: Advance computing conference. IEEE, pp 134–138
21. Nagi J, Ducatelle F, Di Caro GA et al (2011) Max-pooling convolutional neural networks for vision-based hand gesture recognition[C]. In: 2011 IEEE international conference on signal and image processing applications (ICSIPA). IEEE, pp 342–347
22. Ngiam J, Chen Z, Chia D et al (2010) Tiled convolutional neural networks[C]. In: Advances in neural information processing systems, pp 1279–1287
23. Ranzato MA, Poultney C, Chopra S (2007) Efficient learning of sparse representations with an energy-based model. In: Proceedings of the 2007 advances in neural information processing systems. MIT Press, USA, pp 1137–1144
24. Razavian AS, Azizpour H, Sullivan J et al (2014) CNN features off-the-shelf: an astounding baseline for recognition[C]. In: 2014 IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 512–519
25. Scherer D, Muller A, Behnke S (2010) Evaluation of pooling operations in convolutional architectures for object recognition[J]. Artificial Neural Networks–ICANN 2010:92–101
26. Sui Y, Guo Y (2014) Hand gesture recognition based on combing Hu moments and BoF-SURF support vector machine[J]. Appl Res Comput 31(3):953–956
27. Wang L, Liu H, Wang B (2017) Gesture recognition method combining skin color models and convolution neural network [J]. Comput Eng Appl 53(6):209–214
28. Xie SJ, Lu Y, Yoon S et al (2015) Intensity variation normalization for finger vein recognition using guided filter based singe scale retinex[J]. Sensors 15(7):17089–17105
29. Yamashita T, Watasue T Hand posture recognition based on bottom-up structured deep convolutional neural network with curriculum learning[C]. In: 2014 IEEE international conference on image processing (ICIP). IEEE, pp 853–857, vol 2014

Xiao Yan Wu received the degrees at the School of Computer Science and Technology, China West Normal University, Nan Chong, China, in 2004.