Research on Gesture Recognition Based on Convolutional Neural Network and SVM

Shoujin Wang[a], and Dongxu Li[b]

Information and Control Engineering Faculty, Shenyang Jianzhu University, Shenyang, China

[a]23917240@qq.com; [b]429708299@qq.com

Keywords: Convolutional Neural Network; Support Vector Machine; Gesture Recognition

Abstract: In the new generation of human-computer interaction technology, gesture recognition is a very important research direction. Different from traditional gesture recognition methods, convolutional neural network can automatically perform feature extraction without manual intervention. In this paper, the method of combining a convolutional neural network and support vector machine is employed to gesture recognition, using the American Sign Language Data Set to train and test. Experiments show that the method has lower complexity and better recognition effect.

1. Introduction

With the development of science and technology, human-computer [1] interaction has occupied an increasingly important position in people's daily life. Keyboard, mouse and touch screen-based interactions are still not efficient enough to meet people's current needs.[2] Compared with other methods, using gestures to communicate with a computer is more direct, concise, and natural. Gesture recognition [3] has become a hot research topic in the fields of computer vision and pattern recognition.

Traditional gesture recognition algorithms need to acquire gesture data in complex backgrounds, then perform a series of complex operations for feature extraction of achievability segmentation, finally, identify it. Different from the traditional method [4], convolutional neural networks do not require manual extraction of features, automatically complete feature learning through input images, it has achieved great success in the fields of face recognition and verification, and semantic segmentation of the scene.

Convolutional neural network is a kind of deep learning technology [6]. Based on image data, it extracts abstract information automatically and extracts the feature information from the layer by layer. It can solve the problem of image classification well. Therefore, this paper studies the method of gesture image classification based on convolutional neural network, and uses convolutional neural network to complete feature learning. Due to the limited data of gesture samples, the data provided to the convolutional neural network training is not sufficient enough to lead to over-fitting, while the support vector machine still has better generalization ability and classification accuracy in the case of relatively small samples. This paper proposes a new network structure based on LeNet-5 model, which uses support vector machine [7] to replace the Soft-Max layer in convolutional neural network to train features. Thereby, a classifier of the gesture image is obtained, which further improves the classification effect of the gesture image.

2. Convolutional Neural Network

Inspired by visual nerve cells, the convolutional neural network is a multilayer perceptron that recognizes two-dimensional shapes. Compared with the traditional neural network, its particularity is reflected in two aspects: local sensing, Partial connectivity is the primary means of connection between neurons, and convolution operations are implemented through local sensing, weight sharing, the connection weights in the perceptual range of the neuron and the sample image can be shared with

DOI: 10.25236/iciss.2019.017

the connection weights in other feature maps, which can greatly reduce the number of weight parameters.

Based on these two specialities, a deeper network structure similar to a biological neural network can be constructed, which enhances the fitting ability and reduces the complexity of the network model.

The structure of a convolutional neural network generally consists of an input layer, a plurality of alternating convolution and pooling layers, a fully connected layer, and an output layer. [8] The input layer is usually a matrix, such as an image in this experiment. The convolutional layer and the pooled layer are special hidden layers for feature extraction and dimensionality reduction. Fully connected layer is a common hidden layer. In addition to this, in order to introduce nonlinear features into the network, there is also an activation function layer.

The convolution layer performs a convolution operation on the input image. In essence, the convolution kernel [9] (characteristic matrix) moves in a certain way on the image matrix and multiplies it with the corresponding position on the image. Finally, the result is added to obtain a value. When the convolution kernel stops moving, these values form a new image matrix, the essence of which is the feature map extracted from the previous image. Its mathematical expression is shown in Eq.1:

$$x_j^l = f\left(\sum_{i \in m_j} x_i^{l-1} \times K_{ij}^l + b_j^l\right) \tag{1}$$

This article uses the ReLu activation function, whose mathematical expression is shown in Eq.2:

$$f(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \tag{2}$$

The pooling layer uses the principle of local correlation to subsample the image data. Retain useful information while reducing the dimensionality of the data so that the features are invariant to scaling and displacement. Its mathematical expression is shown in Eq.3:

$$x_j^l = f\left(\beta_j^l down\left(x_j^{l-1}\right) + b_j^l\right) \tag{3}$$

In this paper, the max-pooling method is used to maximize the pixels in each image block, thus reducing the output dimension of the image.

3. Support Vector Machine

Support Vector Machine (SVM) is a generalized linear classifier that maps low-dimensional nonlinear problems to high-dimensional spaces, making them linearly separable. At the same time, finding the optimal classification hyperplane in the new space can not only correctly classify the data, but also maximize the data classification interval. For the nonlinear case, the mathematical expression of the optimal classification hyperplane solution problem of support vector machine is as shown in Eq.4:

$$min\left\{\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \varepsilon_i\right\} \tag{4}$$

This paper uses the RBF radial basis kernel function, and its expression is as in Eq.5:

$$K\left(x_i, x_j\right) = exp\left(-\left\|x_i - x_j\right\|^2 / \sigma^2\right) \tag{5}$$

4. Algorithm Structure Design Based on Convolutional Neural Network and SVM

By increasing the number of convolutional and pooling layers, the LeNet-5 structure is improved, and SVM is used instead of the traditional classifier to form the network structure used for gesture recognition. The image data is input into the convolutional neural network, and the obtained multi-scale feature vector is input to the SVM classifier, and the CNN+SVM model is established through secondary training. There are input layer, 3 convolutional layers, 3 pooling layers, and 1 fully connected layer. The specific structure is shown in Fig.1:

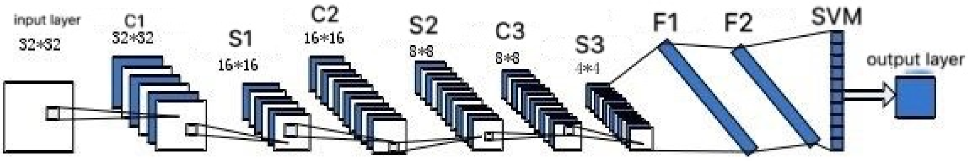

Figure 1. The Structure of Convolutional Neural Network

First, the neural network begins to propagate forward., Including: (1) input layer: input 32*32 image into network; (2) the convolution operation of the input layer and the convolution kernel with the receptive field size of 5*5 is performed to obtain a feature map of 32*32; (3) S1 layer: pooling the feature map of the C1 layer to obtain a feature map of 16*16; (4) C2 layer: Convolution operation of S1 layer and convolution kernel with 5*5 receptive field size to obtain 16*16 feature map; (5) S2 layer: pooling the feature map of the C2 layer to obtain a feature map of 8*8; (6) C3 layer: Convolution operation of S2 layer and convolution kernel with 5*5 receptive field size to obtain 8*8 feature map; (7) S3 layer: pooling the feature map of the C3 layer to obtain a 4*4 feature map; (8) The F1 and F2 layers perform dimensional transformation on the S3 layer to obtain a one-dimensional vector, which is input into the SVM to obtain the corresponding probability value. Then the neural network starts back propagation and uses the gradient descent method to update the network weights. Next, perform forward propagation, backpropagation, and update weight operations until the end of the iteration, The CNN+SVM model training is completed. The parameters are as shown in the following Table 1:

Table 1 Parameters of Convolutional Neural Network

Layers	Parameters
C1	5*5, convolution kernel,32
S1	2*2, max-pooling
C2	5*5, convolution kernel,64
S2	2*2, max-pooling
C3	5*5, convolution kernel,64
S3	2*2, max-pooling
F1	1024 dimensions
F2	24 dimensions

This article sets the network learning rate to 0.005, the number of batch samples to 50, and the number of iterations to 10,000.

5. Experimental Process and Results Analysis

The gesture recognition algorithm in this paper is implemented by Python language and Tensorflow deep learning framework. Tensorflow is a symbolic mathematics system based on data flow programming. It is mainly used in the programming implementation of various machine learning algorithms. Its predecessor is Google's neural network algorithm library DistBelief. Its network model is built in the form of data streams, with great flexibility, and a large number of interfaces can also simplify the operation of building neural networks. This article uses the American

Sign Language data set, which contains 24 types of gesture pictures. The training set has 50,000 images and the test set has 10,000 images.

5.1. Image Preprocessing.

Since the convolutional neural network is highly invariant to image scaling, translation, tilt, or other forms of deformation, this paper randomly performs the above operations on the pictures in the data set. The network parameters are trained by data expansion to prevent over-fitting and improve the accuracy and robustness of the algorithm.

5.2. Analysis of Results.

The performance of the above two models was compared by experimenting with the accuracy of the original CNN model and the CNN+SVM model. Fig.2 is a graph showing the accuracy of the pre-processed data set in the CNN model and the CNN+SVM model.

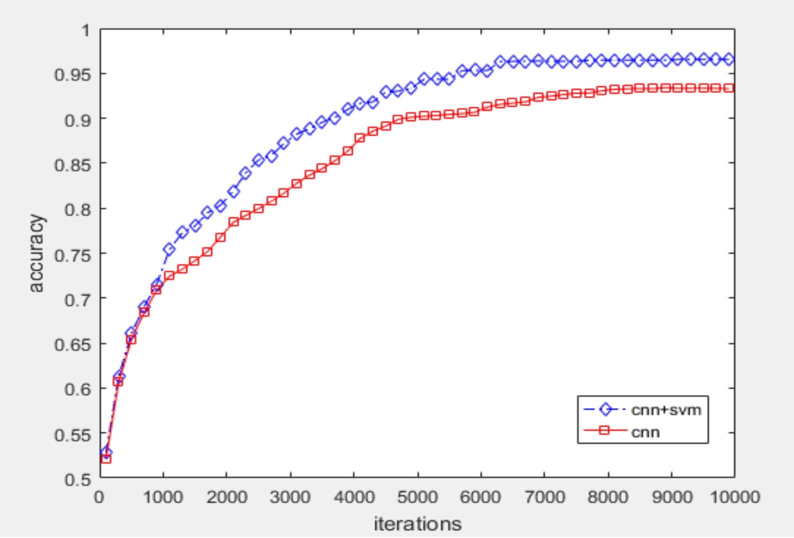

Figure 2. Comparison of the accuracy of the network model

The graph shows that the CNN+SVM model converges approximately 6500 iterations, while the CNN model converges approximately 8,000 times. Regardless of convergence speed or accuracy, the CNN+SVM model is significantly higher than the CNN model. This shows that the CNN+SVM model is more efficient and faster than the original CNN model.

In order to compare the superiority, the method of this paper is compared with some traditional methods to analyze the accuracy of ASL data sets on different gesture recognition algorithms.

Table 2 Comparison of different gesture recognition algorithms

recognition methods	accuracy
HSF+RDF[10]	75%
SP-EMD[11]	75.8%
SAE+PCA[12]	99.05%
CNN	90.24%
CNN+SVM	93.76%

It is shown in the above table that the gesture recognition algorithm based on CNN+SVM obtained in this paper has an accuracy of 93.76%, which is superior to the original CNN model and the identification methods of references [10, 11].

6. Conclusion

Compared with the traditional method, the gesture recognition algorithm of this paper is realized by convolutional neural network, which avoids complicated image preprocessing and manual extraction of features. The characteristics of local sensing and weight sharing reduce the number of parameters and reduce the complexity of the algorithm. The pooling technique avoids recognition errors caused by image distortion, and greatly reduces the amount of calculation, resulting in a more compact network structure and a stronger function fitting ability. In this paper, a new model based on convolutional neural network is proposed and tested. It is verified that the gesture recognition algorithm implemented by this model has better accuracy and robustness. In the follow-up work, the network model is continuously trained by constructing a more complex gesture dataset, which further improves the accuracy of the algorithm in gesture recognition.

References

[1] J.WU: *Research on Gesture Recognition Based on Deep Learning*(MS.,University of Electronic Science and Technology of China,China 2015),p.36.

[2] W.WU: *Research on ASL Letter Recognition Method Based on SAE-PCA Model*(MS.,Xiamen University,China 2014),p.105.

[3] Y.Y.ZHANG: Computing Technology and Automation, Vol. 34 (2015) No.1, p.131

[4] X.C.DU: *Research and Implementation of Vision-based Dynamic Gesture Recognition Related Technology* (MS.,University of Electronic Science and Technology of China, China 2012),p.15.

[5] M.Simon, E.Rodner: The IEEE International Conference on Computer Vision (ICCV)(Santiago, Chile, December 13-16, 2015).Vol.1.p.1143

[6] Vapnik V, Cortes: Machine learning, Vol.20 (1995) No.3, p.273

[7] Y. Le Cun, L. Bottou: Proceedings of the IEEE, Vol. 86 (1998) No.11, p.2278

[8] J.CAI:Computer system application, Vol.24 (2015) No.4, p.113

[9] Pugeault N, Bowden N: IEEE International Conference on Computer Vision Workshops (Barcelona, Spain, November 6-13, 2011). Vol. 1, p.1114-1119.

[10] Wang C, Liu Z: IEEE Transactions on Multimedia, Vol.17 (2014) No.1, p.29

[11] Li Z,Yu B,Wu: Neurocomputing, Vol.151 (2015) No.2,p.565