Research on Gesture Recognition Method Based on Deep Learning

Yizhi Li*
College of Internet of Things, Jiangxi Teachers College, Yingtan City, Jiangxi Province, China, 335000
*Corresponding author’s e-mail: 2002018003@jxsfgz.com

Abstract. Gestures are necessary and indispensable in people's ordinary life communication. And gestures have become an important part of human-computer interaction. With the gradual maturity of human-computer interaction systems, the role of gestures in human-computer interaction is becoming more and more important. Due to the variety of gestures in practical use, coupled with factors such as environment and light changes, it makes it very difficult for computers to recognize the correct gestures from image information. Deep learning has achieved certain research results through several years of development, which provides new research ideas for gesture recognition. This paper firstly introduces the common methods and processes of gesture recognition, and then proposes a new convolutional neural network structure by studying convolutional neural networks and adding Inception structure to the network to improve the performance of the network. Finally, experimental validation is conducted, and the experimental results show that the new convolutional neural network achieves an average recognition rate of 97.8% on the test set, which is a good result.

1. Introduction
Human-computer interaction activities have become an essential part of most people's daily life. Especially in the past few years, with the rapid development of machine vision technology, the research on more human interaction technology has become very active and has also made great progress [4]. According to the current development trend of human-computer interaction technology, human-centered human-computer interaction technology will certainly replace computer-centered human-computer interaction technology [5-6]. Vision-based gesture recognition research is exactly in line with this historical trend. The current human-computer gesture interaction is still at a relatively low level, and society urgently needs new technologies to improve the existing human-computer interaction mode and make it more humanized and intelligent. More and more scholars and experts are devoted to this field to study new human-computer interaction modes. Gesture recognition and gesture-based HCI research are becoming the current trend. Major research institutions, software and hardware manufacturers are also devoting themselves to researching this new continent.

2. Common methods and processes of gesture recognition

2.1. Commonly used methods for gesture recognition
At present, a lot of research results have been achieved in gesture recognition based on traditional methods, which can be classified into three categories: methods based on template matching, methods based on artificial neural networks and methods based on probabilistic statistical models [7].
Stencil-based matching methods are based on the selection of matching features, which can be regional features or edge features of the gesture. The feature vectors are extracted from the gesture image to be recognized and compared with the corresponding feature vectors in the saved template, and the distance between the gesture image and the template is calculated to determine the class of the gesture image to be recognized by the minimum distance method [8]. This method is the most basic and primitive matching method in gesture recognition. However, the limitation of this method is that it is more sensitive to information such as the size and rotation of the input image. The recognition rate of this method decreases gradually with the increase of gesture types.

Artificial neural network-based methods were a more interesting method for researchers before the emergence of deep neural networks. Artificial neural networks (ANNs) operate in a distributed manner and have learning classification properties. Theoretically, artificial neural networks can approximate arbitrarily complex nonlinear mappings. However, it is easy to fall into overfitting with this approach. The traditional artificial neural network structure is limited to a shallow neural network, which generally contains only a simple structure of input layer, hidden layer and output layer. Therefore, the ability to learn features is more limited, which also makes the final recognition effect average.

Methods based on probabilistic statistical models generally acquire some features of a gesture and categorize the current gesture by the features in a probabilistic way. A typical representative of this approach is the Hidden Markov Model (HMM). This method includes a dual stochastic process of state transfer and observation output, which requires more parameters and leads to a slower recognition speed. And the intermediate transfer state rate is not known, so only fuzzy calculation can be performed, which also affects the final classification recognition effect.

2.2. Basic process of gesture recognition

First, the image information stream is captured through the camera, followed by pre-processing, such as filtering noisy data (gesture retrieval and segmentation); then it is input to the recognition system, and feature extraction is performed using deep learning for gesture features; finally, the classification algorithm is completed through training and the artificial grammar defined by the system for recognition output [9].

The feature extraction and analysis process is different for static and dynamic gestures; static gestures only need to be judged by various methods on images, while dynamic gestures are a change process, a collection of actions over a period of time, so the system needs to define the start and end timestamps of dynamic gestures. Dynamic gestures are able to become static gestures by some means.

Visual gesture recognition focuses on image processing, and the feature extraction of the gesture directly affects the final recognition result (the gesture recognition mentioned in this paper is the visual gesture recognition) [10]. Traditional gesture recognition for vision contains various algorithms, and how to improve the consistency and robustness of gesture recognition has been the focus of research.

The basic process of gesture recognition is shown in Figure 1.

3. Improved convolutional neural network for gesture recognition algorithm

In this paper, an eight-layer convolutional neural network is designed for the recognition of six hand gestures. Since there are many factors that affect the performance of convolutional neural networks, this section will investigate several of the relevant factors.

In this paper, an eight-layer convolutional neural network is designed, as shown in Fig. 2. Excluding the input, the convolutional neural network has eight layers, including three convolutional layers, three pooling layers, a fully connected layer and a Softmax regression layer. The input of the network is a 224×224 gesture pixel matrix, and the convolutional kernel size of convolutional layer 1 is 5×5, and the number of convolutional kernels is 32, resulting in 32 feature maps. The sampling size of pooling layer 1 is 2 × 2 without overlapping sampling, which corresponds to convolution layer 1, resulting in 32 feature maps. Similarly, the convolutional kernel size of convolutional layer 2 and convolutional layer 3 is 5×5, and the number of convolutional kernels becomes 64, corresponding to pooling layer 2 and pooling layer 3. The number of neurons in the fully-connected layer is set to 500, which is fully
connected to the pooling layer 3. The fully connected layer is followed by a Softmax regression layer containing 6 neurons, which classifies the output features of the fully connected layer and results in six gestures from 0 to 5.

![Image of gesture recognition process](image1.png)

**Figure 1. Basic flow of gesture recognition**

![Image of convolutional neural network structure](image2.png)

**Figure 2. Eight-layer convolutional neural network structure**

In general, the most direct way to improve network performance is to increase the depth and width of the network. However, a large number of parameters can also lead to overfitting and excessive computation problems. The GoogLeNet network proposed by Google team solves this problem by using the Inception structure, which introduces multi-layer perceptrons instead of the generalized linear structure in traditional convolutional neural networks, maintaining the sparsity of the network structure and improving the performance of the network through the dense matrix. The structure of Inception is shown in Figure 3.
In addition to the conventional convolutional layers and pooling layers of convolutional neural networks, a three-level Inception structure is introduced. A fully connected layer and a Softmax layer. Where Stride is the convolutional kernel step. The overall network structure is schematically shown in Figure 4.

4. Analysis of experimental results

The experiments were conducted in a Linux Ubuntu 14.04 environment with an Intel i5 processor, 8 G of RAM, and a NVIDIA GTX950 GPU for CUDA acceleration. Neural networks are built and trained in theano framework, a class library for defining, optimizing, and evaluating mathematical expressions containing multidimensional vectors, developed by Lisa's group at the University of Montreal, Canada. It has the advantage of being computationally efficient and easy to use, and is often used in the construction of deep learning networks.

The experimental dataset consists of two parts, one is the 10 gestures captured by webcam according to the standard American Sign Language 0-9, which is referred to as the W gesture set, as shown in Figure 4-2(a). Each gesture contains 2,200 samples, totaling 22,000 samples.

The training process of the network uses Adam's approach for weight updating, and uses the discard method for normalization to enhance the robustness of the network. The error rate of the network for different N conditions is obtained after 200 complete iterations of training with different number of
nodes, as shown in Figure 5. From the training results, it can be seen that N is more stable and achieves a higher correct rate when 512 is taken as the convergence result compared to other selection terms. Therefore, 512 is taken as the value of the number of hidden nodes N for the gesture classification network.

![Figure 5. Comparison of network error rates for different number of hidden layer nodes N](image)

Neural networks for gesture recognition algorithms are compared. A gesture recognition method combining HOG features and SVM is proposed in the literature [1]. The HOG features of the gesture image are extracted and recognized by SVM classification. In the literature [2], a 32×32 binarized gesture sample is selected and a 7-layer convolutional neural network structure is designed to train the network model for gesture recognition. In the literature [3], a simple 5-layer neural network with 32×32

![Figure 6. Comparison of gesture recognition rates of different algorithms](image)
grayscale gesture samples was chosen for model training. Table 4-2 shows the test performance of different algorithms. Table 4-2 shows that the recognition rate of the method in [1] is relatively low because the HOG features are affected by the gesture rotation and other factors, resulting in poor recognition. The literature [2] and [3] both used the convolutional divine meridian method and achieved good recognition results, but because the designed convolutional neural network is relatively simple, the recognition rate of the trained model test is not as good as the method proposed in this chapter.

5. Conclusion
The development of computer technology has provided more options for human-computer interaction. As a kind of human-computer interaction mode, gesture has good research value and application prospect. There are already some mature applications of gesture recognition technology based on wearable devices. Progress in gesture recognition research is crucial to the development of interaction speed, and improved gesture recognition technology provides a new contactless interaction mode for human-computer interaction. Deep learning, as a branch of machine learning, has become a new research hotspot in recent years. At present, gesture recognition research is still focused on the technical level of algorithms pursuing high accuracy, but in the future, research results in psychology and ergonomics should be integrated to improve algorithm performance.

References
[1] Ren Y., Gu C. C. (2011) Hand gesture recognition based on HOG features and SVM. Science and Technology Bulletin, 27:211-214.
[2] Cai J., Cai J.Y, Liao X. D, et al. (2015) A preliminary investigation of hand gesture recognition based on convolutional neural network. Computer System Applications, 24:113-117.
[3] Wang L., Liu F., Wang B, et al. (2017) A gesture recognition method combining skin color model and convolutional neural network. Computer Engineering and Applications,53:209-214.
[4] Yang L. M., Li Z. H. (2020) Design of human-computer interaction-oriented gesture recognition system. Industrial Control Computer, 33:18-20+22.
[5] Zhang W. W., Gao X., Wang B. Z., et al. (2020) A deep learning network-based gesture recognition system for Raspberry Pi. Information and Computer (Theory Edition), 32:63-66.
[6] Zhang S. J., Zhang Q., Li H. (2020) A review of deep learning-based sign language recognition. Journal of Electronics and Information, 42:1021-1032.
[7] Ku L. Y., Liu L. W., Ye S. X., et al. (2020) Research on access control based on hybrid face and gesture recognition. Science and Technology Innovation Herald, 17:118-121+123.
[8] Liu X. H., Deng B. S., Pei Y., et al. (2020) Research on wearable gesture interaction system and recognition algorithm. Small Microcomputer Systems, 41:2241-2248.
[9] Duan C. X., Bai Y. (2020) Design of robot demonstration and teaching system combined with deep learning[J]. Computer measurement and control, 28(11):164-169.
[10] Cheng R., Shi J. F. (2021) Research on hand gesture recognition algorithm based on convolutional neural network. Electronic design engineering, 29:179-184.