2018 International Conference on Computer, Communication and Network Technology (CCNT 2018)
ISBN: 978-1-60595-561-2

A Gesture Recognition Method Integrating RGB and Depth Image Features

Kang WANG[1,2], Zhi-quan FENG[1,2,*], Chang-sheng AI[3], Ying-jun LI[3] and Rui HAN[2]

[1]School of information Science and Engineering, University of Jinan, Jinan 250022, China

[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, China

[3]School of Mechanical Engineering, University of Jinan, Jinan 250022, China

[*]Corresponding author

Keywords: Static gesture recognition, Dual-channel convolutional neural network, Characteristics of the fusion.

Abstract. The key to the success of gesture recognition often depends on the extraction and representation of the gesture features. Compared with the traditional gesture recognition methods, the dual-channel convolutional neural network that we constructed in this paper can extract the features of RGB and depth images of the same gesture, making the gesture feature mining and extraction more accurate, and gesture recognition more robust. Firstly, the features are learned from the two channels respectively by convolutional neural network. Then, those features are mapped to the fusion layer, which the fusion features are used for classifier training. Finally, the model recognition rate can reach 98.11%, through the collected gesture database verification.

Introduction

With the development of artificial intelligence technology, more and more human-computer interaction technology is researched and applied, such as, eye movement, language, expression, gesture and limb [1]. As an intuitive, natural and concise feature, gestures have attracted a great deal of attention.

The extraction of gesture features is one of the key factors in the success of gesture recognition. Deep learning is good at extracting more and more abstract feature forms which have better generalizations from the original images. Moreover, it has been applied to machine vision, natural language processing, speech recognition, semantic analysis and other fields, and has achieved outstanding performance. Deep learning is a multilevel nonlinear hierarchical machine learning method and its main form is deep neural network at present. Convolutional neural network has a typical and extensive structure. The convolutional neural network which is composed of multiple convolutional layers, pooled layers, and full-connected layers. It adopts strategies such as local links and weight sharing to reduce the number of learning parameters in the network and reduce the complexity of the model. Moreover, the model has certain degree of invariability to translation, distortion, and scaling. In the early 1990s, Prof. LeCun and others proposed a multi-layer neural network and applied it to hand-written digital recognition. The proposed LeNet has reached the commercial level [2][3]. In 2012, Alex proposed the AlexNet network structure model and obtained the champion of ILSVRC 2012. Then the network models such as VGG, GoogLeNet, etc. are applied to multiple research fields.

On the basis of these basic knowledge network models, ChaoRen Yi et al. proposed a multi-channel convolutional neural network based on image gradient information [4]. JiaWen Feng et al. proposed using different size convolutions and constructing a dual-channel convolutional neural network to obtain multi-scale information of the image in order to fully extract image features. However, for some deep static gesture images, similar gestures are easily generated due to the occlusion problem, and inaccurate recognition may occur for RGB images with complicated

background and blurred appearance. Therefore, in the case of a small number of samples, we can't get well recognition effect by still use this characteristic extraction method. So this paper presents the fusion of RGB and depth image characteristics of gesture recognition method, and dual channel convolution neural network model based on CaffeNet network structure is constructed.

Convolutional Neural Network

Convolutional neural network (CNN) is a neural network that is specially designed to process data which is similar to a grid structure. Including the input layer, hidden layer, and output layer [5]. The hidden layer consists of convolutional layers and pooled layers (down sampled layers). Generally, several convolutional layers and pooling layers are alternately set, that is, one convolution layer links one pooling layer, which is alternately performed. Among them, the convolution layer is used for feature extraction. Each neuron is linked with the local receptive field of the upper neuron, and the weighted sum is obtained by using the corresponding connection weight and local input information, and then the bias value is used to obtain the input value of neurons. The pooling layer can be used as a secondary extraction of the features of the convolution layer, and when the convolution output is pooled, the feature is invariant to a small amount of translation [6]. The entire hidden layer is a learning process from local vision to global vision.

Architecture of Proposed Method

In the process of processing gesture images using a convolutional neural network, it is often performed in a single channel. For a single channel network, the input image often has grayscale, color, and depth maps. In depth image, its internal texture is not clear, but it is easy to shield the influence of the surrounding environment. For color image, the internal texture is clear, but it is easily affected by the complex background. For dual-channel convolutional neural networks, the article [7] uses two different convolution kernels to build two independent channels, and fully extract the features of the input image. And the article [8] proposes a feature learning method based on spatial self-coding and principal component analysis, and applies the established dual channel neural network to gesture recognition of color information. The double channel convolutional neural network constructed in this paper can extract and fuse the features of RGB and depth images of the same gesture. We select color images and depth images with same content but different types as data input for different channels, each channel is processed according to a single channel method, and a fusion layer is established at the fc7 layer, so as to form a dual channel network, as shown in Figure 1.

In the double channel convolutional neural network, each channel refers to the Caffe-Net model, including seven layers. The first five layers are convolutional layer, and the last two layers are full-connection layer. Among them, the core of the convolution layer is connected to all the kernel maps in the former convolution layer, and the neurons in the full connection layer are connected to all the neurons in the previous layer. The response normalized layer follows the first and second convolutional layer. The largest Pooling layer follows the normalized response layer and the fifth convolutional layer. The output of each convolution layer and the full connection layer will be calculated by ReLU operation.

Taking channel-one as an example, the first convolutional layer uses 96 convolution kernels of size 11×11×3 and a step size of 4 pixels to convolve a depth image of size 70×70×3. The second convolution layer needs to normalize the response of the first convolution layer and the result after pooling as its input and convolve it with 256 convolution kernels of size 5x5x48. There is no pooling layer and normalized layer between the third, fourth, and fifth convolutional layer, which are connected by convolution kernels. Among them, the output of the second convolution will be convoluted with 384 convolution kernels of size 3×3×256 in the third convolution layer and convoluted in the third and fifth convolution layers with 384 convolution kernels of size 3x3x192 and 256 convolution kernels of sizes 3x3x192 respectively. The output size of the fifth convolutional

layer is 1×1×256, and the size of the full-connection layers (fc6, fc7) is 1×1×2048, which means there are 2048 neurons linked to the previous layer. For the second channel, since the background of the RGB image is more complex and contains more information than the depth image, so the full connected layers fc6, fc7 select 4096 neurons for connection.

The feature-mapping layer is used to replace the last full-connection layer (fc8) of the original network. And the output of each channel's fc7 layer is mapped to this layer, forming a feature vector consisting of 6,144 neurons. Then a label distribution that covers 6 categories will be generated when the feature vector was send to the soft-max layer. In the entire model training method, a stochastic gradient descent (SGD) method was adopted to accelerate the training speed and reduce over-fitting.

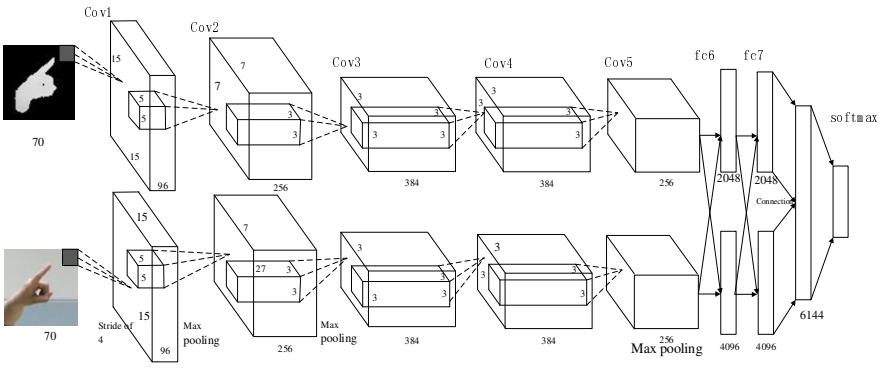

Figure 1. Figure headings.

Experiment and Analysis

The model proposed in this paper is trained on the Intel(R) Xeon(R) CPU E5-2620 v3 @2.40GHz 2.40GHz 2-way processor, Win10 64-bit operating system, and NVIDIA Tesla K40m graphics card. The operator used Kinect camera to capture 6 static gestures of 120 experimenters. Each gesture contained 3600 deep and 3600 RGB images, which totaled 43200. Among them, 36000 images are used for model training, and the rest images are for testing.

Section Headings

The original gesture image in the database are color images with 310*310 pixel and depth images with 90*90 pixel. In depth images, not only hand information but also most body information. And for color images, the proportion of hand information is small and contains a lot of background information. Due to the small amount of data, the performance of classification with the convolutional neural networks not satisfied. Therefore, the experimental data needs to be preprocessed first as follow steps:

Step1:The image is cropped with a certain size block diagram, so that the sample image contains as much complete hand information as possible, while reducing background information.

Step2:Taking binary process for the depth images with a settled threshold to filter out or reduce the interference of depth information of other parts except the hand. Gesture Picture Preprocessing

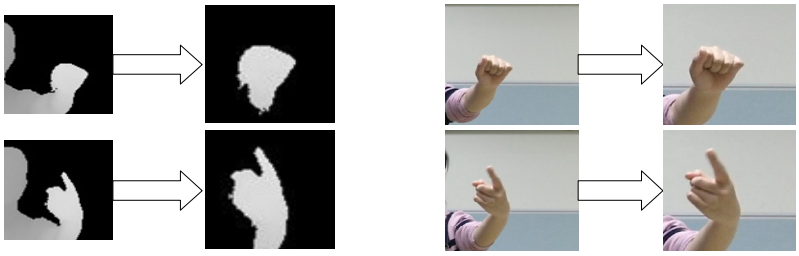

Figure 2. Figure headings.

In the process of processing depth images, the threshold need to be measured repeatedly until get an appropriate value, so we can make sure the image contains as much hand information as possible, and the interference of the limb information can be well filtered out too. The size of all the processed images is 70*70 pixels, and the information of hand will account for 40%~70% of the entire image information. Six gestures such as Figure 3.

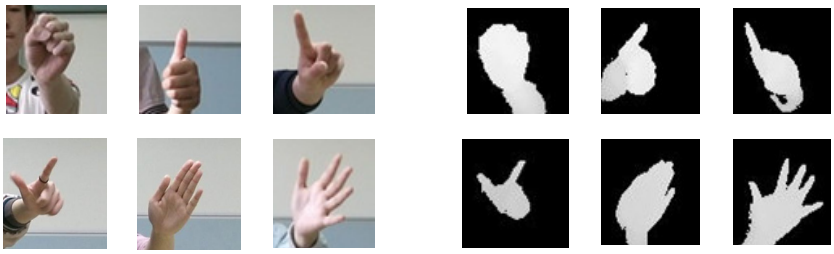

Figure 3. 6 different gestures: (a) color images, (b) depth images.

Double Channel Convolution Neural Network Gesture Recognition

This paper gives a comparison of gesture recognition performance between single and dual channel neural networks. At the same time, compared with other neural networks for gesture recognition.

Comparison of Gesture Recognition between Single and Double Channels. According to the principle of double channel convolutional neural network model, the experiment can get the gesture recognition effect of each channel separately. Design channel 1 to enter a 70×70 pixel depth image. Channel 2 enters a 70*70 pixel color image of the same action. Modify the double channel configuration file and network structure, select the optimal basic learning rate: 0.001, and the maximum number of iterations is 20000. At the same time, the network structure model of Channel 1 is modified to include only 2048 neurons in the fc7 layer without affecting the accuracy of the model, which not only reduces the parameters of the model training, but also reduces the scale of the mixed features. Channel 2 still contains 4096 neurons in the fc7 layer. The two channels are independently trained and do not affect each other. Feature maps containing depth information and color information are obtained in the feature mapping layer. Each feature map is a 1*6144-dimensional vector. All the feature maps of the training set are combined in rows to form a feature matrix, then the softmax classifier is trained. The feature matrix obtained from the test set is used to detect the softmax classification effect.

Table 1. Single and dual channel gesture recognition effect.

Model structural	Train loss value	Accuracy value(%)
Channel 1	0.0029329	96.55
Channel 2	0.0031122	96.61
Dual channel	0.0012635	98.11

During the experiment, the difference between the value of the function loss and the accuracy of the single and dual channel convolutional neural network is compared. For the dual channel neural network, the L2 regularization is added to the loss function of the softmax classifier. The term is used to punish excessively large parameter values and solve the numerical problems caused by parameter redundancy in softmax regression. Selecting the optimal weight attenuation factor λ=1.00e-07 makes the double channel neural network achieve higher accuracy: 98.11%. Moreover, the value of the cost function during training was 56.9% lower than that of channel 1 and 59.4% lower than channel 2, which was significantly better than that of single-channel train. The experimental results show that using double channels for feature extraction and fusion of different types of pictures of the same gesture will obtain more abundant feature information than the single channel network model. Classifying these fusion features can achieve a higher static gesture recognition accuracy.

Comparison with other Models. In order to better reflect the effectiveness of the method, several popular network structures and traditional gesture recognition models are selected, and experiments are carried out on the data set provided in this paper. The experimental results are shown in the following table:

Table 2. Comparison with other models.

Model structural	Test loss value	Accuracy value(%)
AlexNet (coler picture) [9]	0.10495	97.89
AlexNet (depth picture)	0.13694	97.44
CNN-SVM classifer [10]	0.14280	95.81
Double channel CNN of deferent kernal [7]	0.15277	94.86
Our	0.10095	98.11

As shown in Table 2. Adjust AlexNet's network parameters (base learning rate is 0.01, batch_size is 60), after 30000 iterations times to achieve stability, the accuracy of 97.89% and 97.44% is obtained respectively on the RGB and the depth image data sets. The CNN-SVM classifier [10] uses a combination of LeNet-5 convolutional neural network feature extraction and SVM classification. The accuracy of the model on the depth image dataset can reach 95.81%. Based on the LeNet-5 network model, JiaWen Feng uses different convolution kernels to construct a dual-channel neural network for static gesture recognition. The classification accuracy on the depth image data set is 94.81%. The double channel convolutional neural network model proposed in this paper based on RGB and depth images achieves a classification accuracy of 98.11% on the provided data set. Based on the above experimental results, the analysis leads to the following conclusions:

(1) Double channel convolutional neural networks constructed with RGB and depth images of the same gesture can obtain more abundant hand features. These features, to some extent, make up for the deficiency of single channel convolution on the same type of image extraction. Thus a higher accuracy of gesture recognition is achieved.

(2) The convolutional neural network model proposed in this paper is an extension of the traditional network models. The single channel is used for feature extraction independently. Channel 1 adopts an appropriate dimension reduction strategy. Finally, the fusion feature is classified. Use supervised learning to train. This is a more effective static gesture recognition method.

Conclusion

In this paper, a dual-channel convolutional neural network model is proposed. In this model, each channel include 7-layer network structure and the RGB image and depth image will be trained and extracted feature separately. For the depth image with more redundant information, by reducing the number of neurons in the full link layer, we can reduce redundant information and reduce the dimension of fusion features. The experimental results show that the dual-channel convolutional neural network based on RGB and depth images can classify six gestures well. The method can improve the classification accuracy when the depth image texture is not clear or the RGB image background is complicated. At the same time, the network model proposed in this paper still need to improve. For example, the model still employs a supervised learning style, requiring a large number of labeled images during training. In the future, network training can be conducted in a semi-supervised manner while reducing the workload. In addition, the dual-channel model is may be used in the field of dynamic gesture recognition.

Acknowledgement

This paper is supported by the National Key R&D Program of China (No. 2016YFB1001403) and the National Natural Science Foundation of China (No. 61472163).

References

[1] FengJun Zhang, GuoZhong Dai, et al, A Survey of Human-Computer Interaction in Virtual Reality, Chinese Science: Information Science 46 (2016) 1711–1736.

[2] Y. LeCun, et al, Comparison of learning algorithms for handwritten digit recognition. Proceedings of the International Conference on Artificial Neural Networks, Paris, 1995, pp. 53-60.

[3] Y. LeCun, et al, Learning algorithms for classification: a comparison on handwritten digit recognition. Proceedings of the Neural Networks: the Statistical Mechanics Perspective, Pohang, Korea, 1995, pp: 261-276.

[4] ChaoRen Yi, et al, Multi-channel convolutional neural network image recognition method, Journal of Henan University of Science and Technology (2017).

[5] Y. LeCun, et al, Deep learning. Nature, 521(7553), 625-660. (2015)

[6] Lan Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning. Posts and Telecommunications Press. Beijing, Chain (2016).

[7] JiaWen Feng, et al, Application of Dual Channel Convolution Neural Network in Static Gesture Recognition. Computer Engineering and Applications (2017).

[8] ShaoZi Li., Bin Yu, et al, Feature learning based on SAE–PCA network for human gesture recognition in RGBD images. Neurocomputing , 151, 565-573. (2015).

[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. In NIPS, pp. 1106–1114. (2012).

[10] X.X. Niu, et al., A novel hybrid CNN-SVM classifier for recognizing handwritten digits, J. Pattern Recognition 45, 1318-1325. (2012).