Gesture Recognition Method Based on Attention Mechanism for Complex Background

GENG Jianchun¹*, Geng Jiannuan² and Wen Lili³

¹Department of Information Engineering, Wei Fang Engineering Vocational College, Weifang 262500, China
²Library, Hebei University of Science and Technology, Shi Jiazhuang 050018, China
³College of Teacher Education, Wei Fang Engineering Vocational College, Weifang 262500, China

*Corresponding author’s e-mail: geng.jianchun@163.com

Abstract: Human-computer interaction technology is still an unaccomplished goal due to complex background and changeable lighting conditions. A two-stage gesture recognition method is proposed to improve the accuracy of gesture recognition. The first stage combines attention mechanism and atrous spatial pyramid pooling (ASPP) which aims to segment gesture from complex background. The second stage uses a two stream convolutional neural network, combines features from the original picture and the output of the first stage, aims to realize the classification of gestures. It can recognize gestures more accurately compared with other gesture recognition methods.

1. Introduction

As part and parcel of communication, gesture can be used not only for people to people communication, but also for human-computer interaction. Compared with traditional human-computer interaction methods, such as mouth or keyboard, gesture is a more nature and intuitive communication way. In addition, hand gesture can also be used for military actions,diving or other circumstances when talk is not applicable. Gesture recognition technology is the basis of human-computer interaction based on gesture and can be divided into wearable device based gesture recognition technology and machine vision based gesture recognition technology. Machine vision based gesture recognition technology is of lower cost and has more extensive application prospects compared with the former one.

The goal of hand gesture recognition is semantic interpretation of hand gestures. In general, Gesture recognition technology based on machine vision consists of two stages, which are hand gesture segmentation and classification. It is the extraction of hand image from the overall image that hand gesture segmentation should finish. It is a binary classification, which aims to divide all pixels in the picture into two categories, hand section and non-hand section. The tasks of classification is to classify hand sections according the position, gesture of the hand, and match each hand gesture with its semantics. With the development of deep learning, convolutional neural network(CNN) shows good results in image processing[1-3]. Deep learning models are used for image segmentation, object detection, image classification etc[4-7]. Many researchers begin to study its usage on hand gesture recognition recent years. Kang used CNN to extract features from the fully connected layers, and recognize sign language based on deep image. This model was tested on the classification of 31 alphabets and numbers using a subset of collected depth data from multiple subjects and achieved 99.99% accuracy for observed signers.
and 83.58% to 85.49% accuracy for new signers[8]. Oyedotun and Khashman proposed a stacked denoising autoencoder that are capable of learning the complex hand gesture classification task with lower error rates. The model was tested on the whole 24 hand gestures obtained from the Thomas Moeslund’s gesture recognition database[9]. Chengqi chose FCN (Fully Convolutional Networks) as basic model and proposed a static gesture recognition method[10]. The method combines the known hand shapes of gestures in the classification process, achieved high recognition rate and strong robustness of static gesture recognition in complex background. All the above methods achieved good results on hand gesture recognition, but there are still lots of problems to be solved. Among all problems, the central one is that recognition effect is poor for hand gestures with complex background and environments.

In this paper a new recognition based on attention mechanism is proposed, aims to reduce the interference of background and illumination conditions on gesture recognitions results and make gesture recognition technology more practical.

2. model structure

A two stage convolutional neural network model is proposed in this paper, which contains three convolutional neural networks. The first stage is a deep convolution neural network, which is used to segment hand region from images and the output of this network is a grayscale image that has the same size with input image. The second stage is a two stream convolutional neural network, uses both the original color image and the output of the first stage as input. The two stream convolutional neural network extracts feature maps from input images respectively and the feature maps are fused followed by a softmax function, which aims to classify the hand gestures. In the second stage, features of the color image and grayscale image are both used and make the classification more accuracy. Overall structure of the model is shown in Fig 1.

![Figure.1 overall structure of proposed model](image)

2.1 segment hand gesture

Most semantic segmentation methods are based on FCN (Full Convolutional Network). The main feature of FCN is to remove the full connection layer and adopt an efficient end-to-end training mode. This paper uses this idea for reference, and the proposed network model mainly includes three parts, which are FCN, Attention module and ASPP (Atrous Spatial Pyramid Pooling). FCN uses the idea of parameter sharing, which reduces the amount of parameters of the network and makes the network have deeper level. Attention module makes the network pay more attention to the part including gesture information, and reduces the influence of environmental factors on segmentation results. ASPP extracts gesture information features on different scales, so that the network can adapt to different sizes of gesture area.

The first stage is based on VGG-16 network, the last full connection layer is removed, and attention module and ASPP are added to increase the understanding and expression ability of the model. Structure of the first stage is shown in Fig4. CBAM[11] was taken as the attention module in this model and was
added to VGG-16 network at layer 2, 3-4, 5-7 and 8-13. Since the weight of the attention module is between [0,1], the output response of the feature map will be weakened by directly multiplying with the main branch, and the value of each point of the final output feature map will become very small if there is multi-layer superposition structure, and the segmentation effect will be worse. In order to avoid the weak response of feature map, this model uses the mechanism of RestNet for reference, transforms the weight of attention module into [1, 2] interval, and then multiplies it with the main branch, so that multiple attention modules can be superimposed in the network. The function of attention module can be expressed by formula 1.

$$H_{ic} = (1 + M_{ic}) \ast F_{ic}$$  \hspace{1cm} (1)

Where \(M_{ic}\) denotes outputs of attention module, \(F_{ic}\) denotes feature map get from main branch.

In order to extract the features of gesture region on different scales, the feature maps of VGG-16 network output was processed by ASPP module, as shown in Fig 2. The image is segmented robustly on multiple scales by ASPP, and the input feature maps are sampled in parallel at different sampling rates. In this paper, the sampling rates take rate={3,6,9,12}. Feature extraction is carried out on four scales, and the feature map is restored to the same size as the input feature map by bilinear interpolation. Finally a 1*1 convolution layer is used to get the output with the same size as the input image.

![ASPP Diagram](image_url)

**Figure 2 ASPP**

### 2.2 Hand gesture classification

The second stage of the model consists of two convolutional neural networks with the same structure to extract gesture features based on shape and appearance respectively. Structure of the second stage is shown in Fig2-1. Each convolutional neural network consists of multiple convolution layers, pooling layer and two fully connected layers. Their inputs are the original gesture image and the grayscale image get from the first stage respectively. The output results are fused by adding, and the classification results are obtained by softmax activation function.

### 3. training

In this paper, OUTHANDS[12] and HGR1 datasets are used to train and test the model. OUTHANDS is a static public dataset specially used for training and evaluating gesture classification and detection. It is one of the few datasets containing gesture segmentation information. This dataset contains 10 categories of static gesture images, a total of 3150 gesture images taken under different backgrounds, the image resolution is 640*480, of which 2150 are used as training set and 1000 as test set. HGR1 contains 899 gestures of 25 categories. Since only a small number of images is contained in HGR1, its training and evaluation of the model is limited. This dataset is mainly used for comparative analysis with other models.
The training and testing of the model are implemented on the Ubuntu 16.04 system and tensorflow-2.0.1 deep learning network framework. The main hardware device is an NVIDIA gpu-rtx2070 graphics card with 8GB video memory.

The model is trained in two stages. In the first stage, the input is the original image, and the label is the segmented gray image. Epochs = 50. Due to the limitation of the device, the image size in the dataset is first scaled to 256 * 256 pixels, and the size of the training batch is batch = 4. At the same time, data enhancement measures such as random clipping, rotation, movement and random sample scrambling are adopted to avoid over fitting. The data enhancement parameters are shown in Table 1.

The classification part of the second stage training model is trained on the basis of the first stage model. The training times were epochs = 50 and the training batch size was batch = 4.

| Rotation angle | Clipping range | Moving range |
|---------------|----------------|--------------|
| ±40           | 30%            | 15%          |

### 4. Evaluation

We evaluate the model on Outlands and HGR1 datasets. In order to compare the recognition effect with other similar models, we introduce F-score as the evaluation index of the model effect. F-score is calculated according to formula 2.

$$ F = \left(1 + \beta^2\right) \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} $$  (2)

Where Precision denotes accuracy rate, recall denotes recall rate, $\beta$ is a factor that balances accuracy rate and recall rate, it takes 1 in this paper.

The segmentation effect of this model is compared with that of FCN-8s, PSPNet and deeplabv3, and the comparison results are shown in Table 2. The segmentation accuracy of this model is close to that of deeplabv3 and PSPNet, but the network structure is simpler and the segmentation time is shorter.

Figure 3 shows the segmentation results of the model for gesture images. It can be seen from the figure that for gesture images with complex background, the model retains the gesture information in the original image after segmentation, and eliminates the interference of background and environmental factors.

| Model            | F-score | Time (ms) | size (MB) |
|------------------|---------|-----------|-----------|
| FCN-8s           | 0.9559  | 63        | 537       |
| PSPNet           | 0.9702  | 50        | 318.      |
| DeepLabv3        | 0.9735  | 43        | 302       |
| This model(stage 1)| 0.9668  | 30        | 173       |
Figure 3 segmentation results of proposed model (A. input images, B. ground truth, C. segment results of proposed model)

The gesture classification effect of this model is compared with restnet-50 and densenet-121. The comparison results are shown in Table 3. It can be seen from data in the table that F-score of this model is significantly higher than that of other models.

| Model          | F-score  | Input size (pix) |
|---------------|----------|-----------------|
| This model    | 0.8621   | 256×256          |
| RestNet-50    | 0.8138   | 224×224          |
| DenseNet-121  | 0.8281   | 224×224          |

5. Conclusion

In this paper, a two-stage convolutional neural network model based on attention mechanism is proposed to solve the problem of static gesture recognition under complex background and illumination conditions. The experimental results show that the model achieves good results in static gesture segmentation and recognition.

References

[1] Hassan B, Ahmed R, Hassan T, Werghi N. SIP-SegNet: A Deep Convolutional Encoder-Decoder Network for Joint Semantic Segmentation and Extraction of Sclera, Iris and Pupil based on Periocular Region Suppression[J]. 2020..
[2] Zhang Di. Research on Hand Gesture Recognition Technology Based on Machine Learning[D]. Nanjing University of Posts and Telecommunications, 2019.
[3] Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.
[4] Ori M, Egvi S. Efficient semantic segmentation with pyramidal fusion[J]. Pattern Recognition, 2020:107611.
[5] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
[6] Li G, Zhang C, Lei R, Li X. Hyperspectral remote sensing image classification using three-dimensional-squeeze-and-excitation-DenseNet (3D-SE-DenseNet)[J]. Remote Sensing Letters, 2020, 11(2):195-203. 
[7] Dong, G., Yan, Y., Shen, C., & Wang, H. (0). Real-time high-performance semantic image segmentation of urban street scenes. IEEE Transactions on Intelligent Transportation Systems, PP(99), 1-17.

[8] Kang B, Tripathi S, Nguyen T Q. Real-time Sign Language Fingerspelling Recognition using Convolutional Neural Networks from Depth map[C]. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015.

[9] Oyedotun O K, Khashman A. Deep learning in vision-based static hand gesture recognition[J]. Neural Computing and Applications, 2017, 28(12):3941-3951.

[10] Cheng Qi. Static Gesture Recognition in Complex Background Based on Convolutional Neural Network[D]. Shandong University. 2019.

[11] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. CBAM: Convolutional Block Attention Module[C]. IEEE Conference on European Conference on Computer Vision, 2018:3-19.

[12] Matilainen M, Sangi P, Holappa J, Silven O. OUHANDS database for hand detection and pose recognition[C]. International Conference on Image Processing Theory Tools & Applications. IEEE, 2017.