Research on Static Gesture Recognition Based on Deep Learning

Min Zhang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1565364293@qq.com

Pingping Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1341369601@qq.com

***Abstract*—With the continuous development and progress of the times, the ways of human-computer interaction have become more and more diverse. In order to reduce the spread of the new crown virus, gesture recognition has become a hot topic in the field of human-computer interaction in recent years. Traditional gesture recognition is affected by the environment and database, etc., with poor robustness and low accuracy. In order to improve the recognition rate of static gestures, this paper proposes to establish a deep learning model using CNN convolutional neural network, and a static gesture recognition method based on template matching. By establishing a palm template diagram, the gesture image to be recognized is matched with the template diagram based on the feature point, and the image is rotated after matching, and the template based on the grayscale value is matched again, so as to extract the gesture part. Through experimental proof, the algorithm can effectively improve the gesture recognition rate, the recognition accuracy rate reached 93.17%, and the recognition speed is faster.**

Keywords-Static Gesture Recognition; Template Matching; Deep Learning; Convolutional Neural Networks

I. INTRODUCTION

Gesture recognition, as the name suggests, refers to the recognition of the palm and arm parts of the human body and the reading of the meaning expressed by the gesture. Gestures are generally divided into two forms: dynamic and static [1]. Dynamic gesture refers to the trajectory of palm movement, this one trajectory change represents a combination of gestures, for the recognition of combination gestures, that is, dynamic gesture recognition, common dynamic gesture recognition acts on: control the synchronous movement of the robot arm and the human gesture, control the page jump through the gesture movement, etc. Static gesture refers to a single gesture, the recognition of a single gesture can read continuous information, common static gesture recognition acts to assist deaf people and ordinary people to communicate, through gesture reading for simple verification. Static gestures are an extremely convenient human feature with the advantage of being simple and easy to read.

At present, the commonly used gesture acquisition methods are: reading gestures through data gloves [2], but the data gloves collect data, and the final recognition rate is easily affected by the device. Reading gestures by optical marking, which transmits the position of the human hand

and the dynamic changes of the fingers to the system screen through infrared, also achieves good gesture recognition, but still requires complex devices to support the intervention of external devices. The Bayesian attention model [3] is used for gesture detection and then the support vector machine (SVM) is used to complete gesture recognition, which is too complex and less practical.

Based on the gesture recognition method of the hidden Markov model [4], this method has achieved good results, but the real-time and accuracy rate need to be improved. Deep learning [5][6] has made great achievements in the field of computing. In neural networks, large network frameworks such as R-CNN [7], Faster R-CNN[8], YOLO[9], and SSD[10] can be well used to recognize gestures, but the amount of computation in large networks is too large to use gesture recognition. Therefore, gesture recognition based on computer vision has become the mainstream research direction. This experiment will capture the gesture image through the camera, realize the processing of the gesture image in the computer, segment the gesture profile, extract the gesture part, and then take the method of deep learning, establish a deep learning model and identify it, and realize the effect that the computer can "recognize" the gesture.

II. ESTURE RECOGNITION ALGORITHM

A. Gesture image acquisition

The gesture image is captured through the camera, and the gesture image is collected into a computer, that is, the gesture image is converted into a matrix of pixels stored in the computer, and pixels are a matrix composed of numbers. In a computer, the operation of an image is, in essence, an operation of a matrix of pixels.

B. Overall system framework

As shown in Fig. 1, the overall framework of the system is divided into four parts: image preprocessing, gesture segmentation, gesture feature extraction and hand recognition. The gesture image is captured through the camera, and the color image is passed into the computer, which requires certain processing of the image to make the overall image easier to extract the gesture. The main role of gesture segmentation is to separate the hand area from the background, and a large part of the performance of this part affects the main performance of the system; Gesture feature extraction is to reduce the depth of the recognition network of hand extraction contour features, reducing the training time; The hand recognition network identifies what kind of gesture is from manually extracted features.

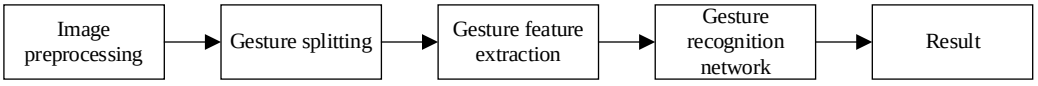

Figure 1. System framework

C. Gesture image preprocessing

The gesture image is captured through the camera, and the color image is transmitted to the computer, and the image needs to be processed to make the overall image easier to extract the gesture, and then extract the gesture part of the image through a certain algorithm, and then judge the extracted gesture to complete the recognition of the gesture.

1) Gesture image grayscale

Color images in the computer, generally stored in the form of RGB three-color channels, grayscale of the image, refers to the image changed to a single channel form stored in the

computer, can effectively reduce the amount of computing during image processing, improve processing efficiency [11].

Grayscale images have equal values in all three color channels of RGB. The grayscale method of the image used in this experiment is weighted averaging. The weighted averaging principle is as follows:

$$f(R,G,B) = Wr * R + Wg * G + Wb * B \quad (1)$$

$f(R,G,B)$ represents the grayscale, R, G, B, represent the value of the point on the three-color channel, Wr, Wg, Wb,respectively represent the weights of R, G, B.

2) Image enhancement processing of gesture images

Image enhancement refers to the specific processing of the image to make the image suitable for a specific occasion. In this experiment, the static gesture recognition needs to make the features of the gestures in the image more prominent, so as to improve the recognition efficiency and effect of the gestures.

In the gesture image captured by the camera, there will be many noise points that interfere with the recognition of the image, and the image smoothing processing is required, which can effectively reduce the interference of the noise points in the image and make the recognition of the image more accurate. The image smoothing method used in this experiment is: median filtering. Set the convolutional kernel with a size of 3×3, sort the 3×3 area centered on the pixel, and assign the median value to the pixel. It can be very effective in smoothing impulse noise while protecting the sharp corners of the image without affecting profile detection.

D. The gesture area splits the network

The hand area processing part is to manually extract features by dividing the hand area into hand areas. Because convolutional neural networks can avoid complex pre-processing of images, directly input the original image, and can identify changing patterns, tolerate image distortion, and have strong robustness [12]. The role of the convolutional layer of the gesture recognition part of the convolutional neural network is to extract the characteristics of the gesture in the picture data, if the picture is processed before the picture is sent to the neural network for discrimination, so that the feature extraction of the neural network requires less convolutional kernels, so as to improve the recognition rate while optimizing the network structure.

The flow of hand area feature extraction is shown in Fig. 2.

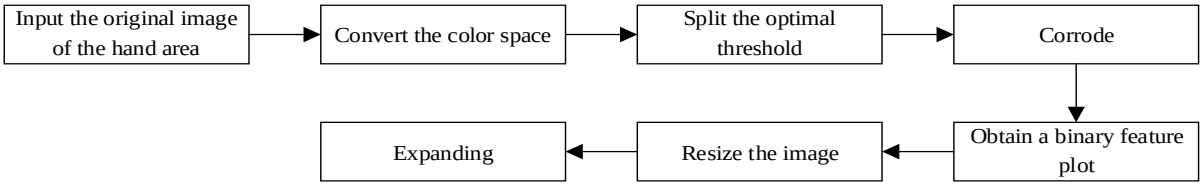

Figure 2. Feature extraction flowchart

The skin color pixels of the human body are in the Ycrcb space, the skin pixels basically meet the elliptical distribution, each pixel is analyzed, whether it is within the ellipse where the skin color pixels are located, if in the ellipse, it means that the pixel is the skin pixel and extracted. Where Y represents the brightness, Cr and Cb can be represented as the difference between the

blue component, the red component and the brightness, respectively, representing the chroma.

In this experiment, RGB pattern images are often converted to Ycrcb color space by the following formula and then processed:

$$Y = 0.299R + 0.587G + 0.114B \quad (2)$$

$$Cr = -0.147R - 0.289G + 0.463B \quad (3)$$

$$Cb = 0.615R - 0.515G - 0.100B \quad (4)$$

Then by adjusting the thresholds of Y, Cr, Cb, you can segment the gesture from the image.

E. Gesture feature extraction

The static gestures studied in this article have one thing in common – the palm of the hand. Only the palm part of the template matching, get the palm part in the image and the angle of deflection relative to the horizontal direction, distinguish the gesture area, only identify the gesture part, and then eliminate the impact of the non-gesture part (such as the face, etc.), the red line contains the area of the palm template and the gesture area respectively. For example: Fig. 3, Fig. 4.

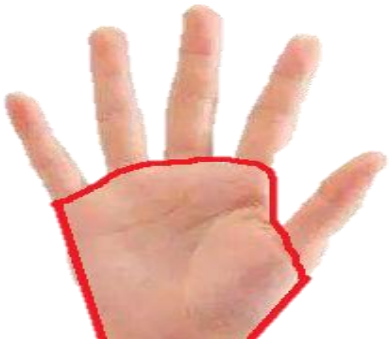

Figure 3. Palm area

Figure 4. Gesture area

This paper can match the palm area in the gesture image through SIFT features, calculate the palm area through the feature point coordinates, enlarge the palm area to a certain extent, increase a certain pixel in the horizontal and vertical directions, that is, include the entire gesture area, identify the gesture area, and effectively avoid the interference caused by the non-gesture part.

Skin tone areas in an image are extracted through skin color recognition to remove the influence of non-skin color parts. The sift algorithm then matches the template of the palm image and the input image, and the angle of deflection of the two images is calculated by the sift algorithm for the direction given by the feature point. Using the principle of image rotation, the pixel matrix is deflected to overcome the defects of the template matching algorithm based on grayscale values. Then the palm is matched with the image by the square difference matching method, the coordinates of the palm area are obtained, and the gesture area can be obtained by stretching the area. The flowchart of the template matching module is shown in Fig. 5.

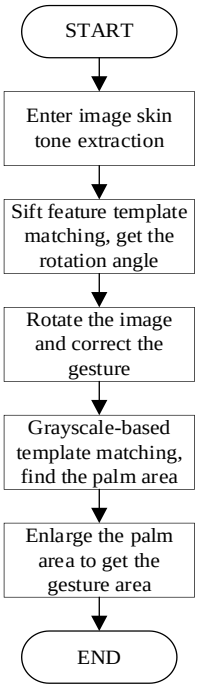

Figure 5. Template matching flowchart

The image rotation algorithm used in this experiment is to rotate the gesture image at a specified angle in the center. Gesture images are stored in a computer in the form of a matrix of pixels, so a rotated image is a matrix rotation algorithm. Therefore, if the width of the original figure is w, the height is h, and the rotation angle is θ, the following formula can be obtained:

$$x_1 = \frac{x_0 - w}{2}; y_1 = -y_0 + \frac{h}{2} \quad (5)$$

F. Gesture recognition network

Static gesture recognition neural network diagram as shown in Figure 6, through the input layer input gesture image, after each convolutional layer of the input gesture image convolution processing, connect a pooling layer, the input gesture image for local feature extraction work, after the processing of two layers of convolutional layer and two layers of pooling layer, through the processing of two layers of fully connected layers, the image features are classified. Through local connection and weight sharing, the construction of CNN convolutional neural network is realized.

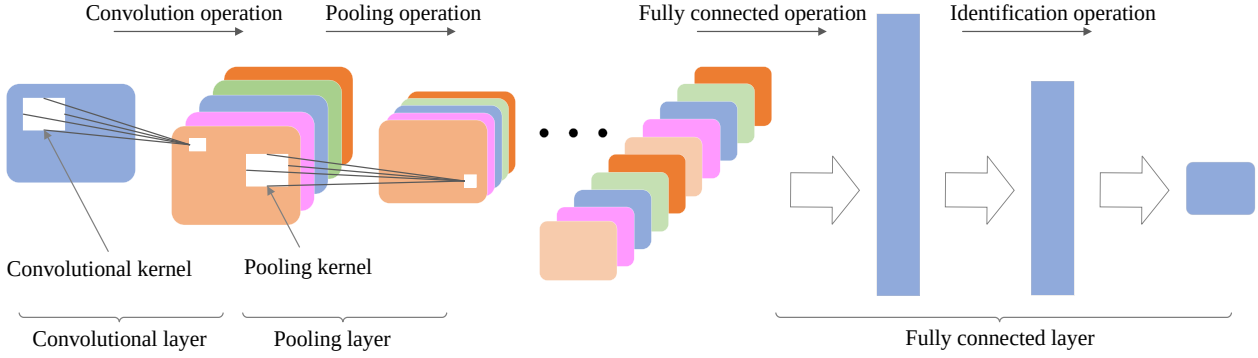

Figure 6. Static gesture recognition neural network framework

The construction of the convolutional layer is the core of the entire CNN construction [13], in the image recognition, set the size of the convolutional core, convolution operation on the input image, that is, the convolution core slides on all pixels on the image, and the pixel at that location is convoluted, and different features can be extracted for the image through the setting of the convolutional core. In the convolutional layer setting for static gesture recognition, the convolutional kernel is set to 2×2 size, the step

size is set to 1, and the mathematical principle of convolution is as follows:

$$h(x) = f(x) * g(x) = \int f(t) * g(x-t) d_t \quad (6)$$

Suppose a gesture image with 64 × 64 pixels, after convolutional and pooling, use the above equation to calculate, the output size is 32×32.

After the convolutional layer, the static gesture recognition model construction needs to add a pooling layer, the main role of pooling is to reduce the parameters and calculation times in the network, which can effectively improve the training efficiency of the static gesture training set and prevent over fitting during the training process.

In a static gesture recognition neural network, the relu function is used to make network training more efficient and increase the nonlinearity of the network compared to the sigmoid function.

The fully connected layer mainly implements the function of classification in the static gesture recognition neural network. The essence of full connection is to linearly transform the feature space acquired by the convolutional layer into another feature space, and connect the two layers of the full connection layer after the static gesture recognizes the convolutional layer in the neural network, and the weighting operation of the feature can effectively reduce the influence of feature location on classification.

III. THE TRAINING OF THE NEURAL NETWORK AND THE FINAL RESULT

A. Static gesture recognition preprocesses module test results

In order to verify the effectiveness of the recognition method proposed in this paper, this paper collects six kinds of gestures in life scenarios, namely digital gesture 1, digital gesture 2, digital gesture 3, digital gesture 4, digital gesture 5 and gesture good. After acquiring a certain number of gesture images through the camera, each picture is preprocessed to make the image easy to identify the gesture, and the size meets the input size of 89×89 pixels required by the deep learning model. After the image is processed, the effect is shown in Fig. 7.

Figure 7. Image preprocessing effect

Through skin color recognition, the skin tone of the human body is extracted from a complex background. Through the method of template matching, the coordinate position of the gesture in the image is found, and the non-gesture part is shielded, and then the extraction and pre-processing function of the gesture under the complex background is completed.

B. Static gestures identify deep learning model test results

In the training of the deep learning model of the static gesture recognition convolutional neural network, the recognition accuracy calculated by the model recognition verification set continues to increase with the increase of the number of trainings, and the final recognition accuracy is stable between 96.9% and 100%. The effect during training is shown in Fig. 8, Fig. 9 and Fig. 10.

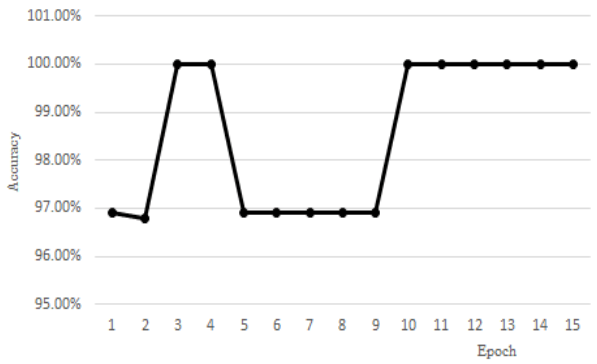

Figure 8. Curve of the accuracy change of the training set

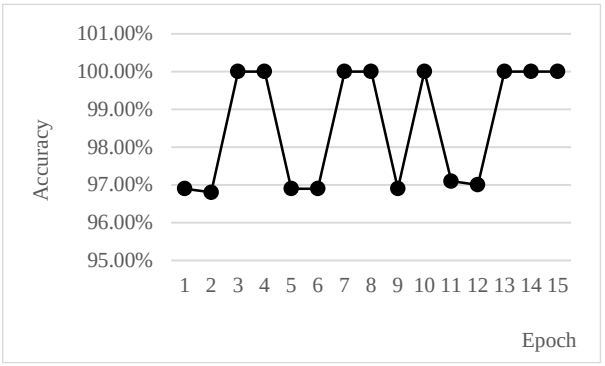

Figure 9. Curve of the change in the accuracy of the test set

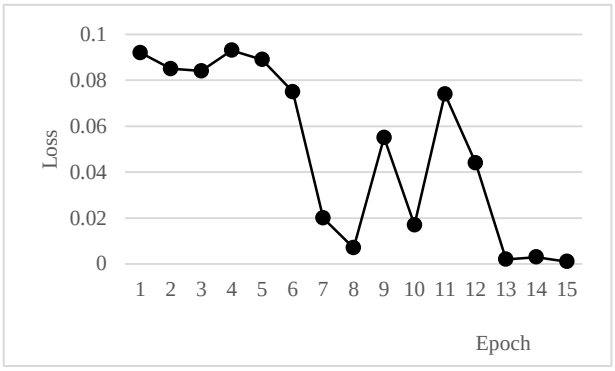

Figure 10. Loss change curve

This training will take the average accuracy of 10 independent experiments as the final experimental results, and the following table lists the recognition rates of six gestures.

TABLE I. GESTURE RECOGNITION RATE TRAINING EFFECT

Gesture category	Recognition rate
Digital gesture 1	95%
Digital gesture 2	99%
Digital gesture 3	96%
Digital gesture 4	83%
Digital gesture 5	86%
Digital gesture good	100%

From Table 2.1, it can be seen that the average recognition performance of the gesture recognition neural network model in this experiment reached 93.17%, which effectively recognized static gestures.

IV. CONCLUSION

Aiming at the problems of large number of parameters of gesture recognition model based on deep learning, slow training speed and high equipment requirements, which increases the cost, this paper proposes a gesture recognition system based on convolutional neural network. Using the template matching algorithm based on feature points, the gesture is finally effectively recognized by image acquisition, image preprocessing, gesture segmentation, etc. Experimental results show that the method can complete the recognition of static gestures, the recognition accuracy rate has been improved, and the recognition speed has been significantly improved, which has high real-time performance compared with other algorithms and has good generalization.

REFERENCES

[1] WENG H L, ZHAN Y W. Vision-based hand gesture recogni-tion with multiple cues [J]. Computer engineering & science, 2012, 34(2):123-127.

[2] L ü N, Yang Y J, Xu T. Sparse decomposition for data glove gesture recognition [C]. Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Piscataway, NJ: IEEE, 2017: 1-5.

[3] Pisharady P K,Vadakkepat P, Loh A P. Attention based detection and recognition of hand postures against complex backgrounds [J]. International Journal of Computer Vision, 2013, 101(3): 403-419

[4] Dai Y K, Zhou Z H, Chen X, et al. A novel method for simultaneous gesture segmentation an recognition based on HMM [C]. Proceedings of the 2017 International Symposium on Intelligent Signal Processing and Communication Systems. Piscataway, NJ: IEEE,November 6-9,2017: 684-688.

[5] SERMANET P, KAVUKCUOGLU K, CHINTALA S,et al. Pe-destrian detection with unsupervised multi-stage feature lear -ning [C]// 2013 IEEE Conference on Computer Vision and Pat-tern Recognition. Portland:IEEE, 2013:3626-3633.

[6] ZHANG C, ZHANG Z. Improving multiview face detection with multi-task deep convolutional neural networks [C]// 2014 IEEE Winter Conference on Application of Computer Vision. Steamboat:IEEE, 2014:1036-1041.

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[8] Ren S,He K,Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]. International Conference on Neural Information Processing Systems. [S.l.]: MIT Press, 2015: 91-99.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: unified,real-time object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[10] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector [C]. European Conference on Computer Vision. Springer International Publishing, 2016: 21-37.

[11] WU Yaoling. YCrCb color space face detection algorithm based on the design and implementation [D]. Chengdu:University of Electronic Science and Technology of China,2013.

[12] PENG Yaqin, CHENG Xiaogang. An optimized deep learning algorithm of convolutional neural networks [J]. Modern electronics technique, 2016, 39(23):179-181.

[13] SAXE A M, PANG W, KOH Z, et al. On random weights and unsupervised feature learning [C]// Proceeding of 2011 In-ternational Conference on Machine Learning. Bellevue:ACM, 2011:1089-1096.