Complex Background Gesture Recognition Based on Convolutional Pose Machines

Ganzhou Liao[1, a], Xia Zeng[2, b]

[1]Guangzhou University Sontan College, Guangzhou 511370, China;

[2]Guangzhou University Sontan College, Guangzhou 511370, China.

[a]aimxu@qq.com, [b]278354772@qq.com

Abstract. Gesture recognition is a research hotspot in HCI (Human-Computer Interaction). The camera-based gesture recognition method has become the focus of research in this field due to its low cost. Due to the lack of three-dimensional coordinate information and depth information, image-based gesture recognition has been a difficult point. Therefore, this paper applies the Convolutional Pose Machine method which is more mature in human body pose estimate in the field of gesture recognition. In this paper we collect gesture data and create a gesture database for training and testing. From the results of the test, this paper has certain practical significance.

Keywords: CPM, CNN, gesture recognition, Joint.

1. Introduction

Gesture recognition is an important method of HCI and an important research topic in the computer field. Using gesture recognition, we can remove the hardware of HCI, such as mouse and keyboard, and further enhance the ability of HCI. There are many ways to implement gesture recognition, such as data glove collection method [1], gesture recognition method based on Kinect depth camera [2], computer camera captures and processes the image, etc.

In these methods, the computer camera can obtain a higher recognition rate with simple hardware. Therefore, the acquisition and processing of gesture images has become a hot topic in the research of gesture recognition, and various methods have emerged in an endless stream. Such as skin color segmentation and feature extraction methods [3], such as Sift based static gesture recognition [4], etc.

However, there is a serious problem with these methods, that is, the robustness to illumination is poor, and the recognition rate is seriously affected in complex backgrounds.Although some algorithms can be used to improve these problems, but lack of versatility, this paper proposes the use of CPM (Convolutional Pose Machines) [5] method, gesture training, and gesture estimation to obtain better results.

2. Algorithm Interpretation

Pose Machines is a serialized prediction framework. It is first used for human body attitude prediction. The general input is a human body posture map, and the output is N heat maps, representing the response of N joints. The Convolutional Pose Machine, based on the Pose Machine, uses convolutional neural networks to learn image features and image-depenent spatial models to estimate human pose. The human posture here mainly refers to detecting the joints of the human body and the relationship between the joints, and finally obtaining the posture of the human body.

Convolutional Pose Machine is mainly used in the estimation of human body posture. This method is applied in this paper and applied to static gesture recognition. The human body posture and gesture posture are very similar, The most similar point is that both the human body and the hand gesture can be thought of as a connector between joints. Therefore, the hand can detect the joint and furthermore, can detect the static gesture. The algorithm is applied to gesture recognition below for detailed description.

3. Gesture Recognition Algorithm

3.1 Palm Joint Freedom Description.

To apply the CPM algorithm to gesture recognition, the description of the palm joints is a prerequisite. According to the anatomical structure, the palm has a total of 15 slender metacarpal and phalanx, and 8 carpal bones. It can be divided into 16 joints with a total of 26 DOF (Degrees of Freedom), of which 15 metacarpals and phalanges correspond to 15 joints. Total 20 DOF, 8 wrist bones form the wrist joint, the wrist joint has 6 DOF, a total of 26 DOF, as shown in Fig. 1.

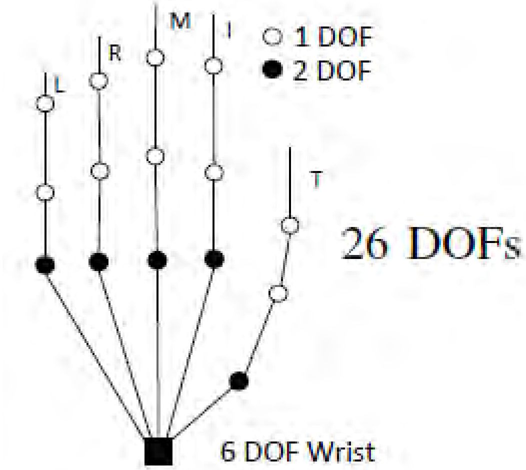

Fig. 1 Hand joints and degrees of freedom

3.2 Algorithm Description

CPM takes advantage of the deep convolutional neural network and the spatial modeling of the Pose Machine framework. CPM generally uses a multi-step approach. The process is as follows: First. Calculate the heat map of each joint point of the network prediction according to the local image evidence; Second. Accumulate the heat maps of all scales corresponding to each joint point in turn; Third. According to the accumulated heat maps, if the maximum value is greater than specific threshold, the location of the maximum is the predicted joint point location.

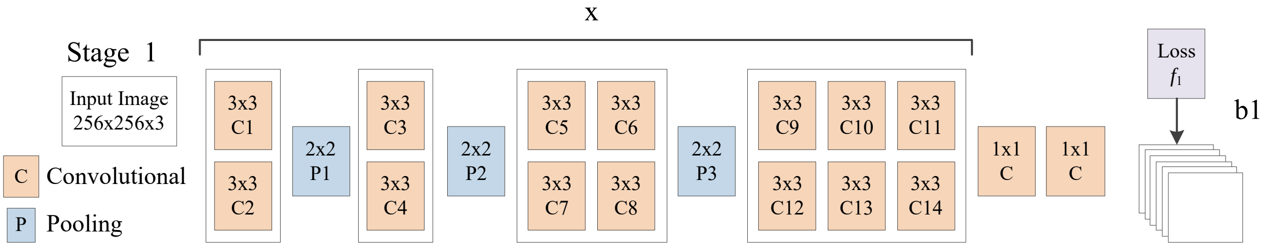

Fig. 2 Stage 1 Network Diagram

The most critical stage is stage 1, which is use to predict the joint points, is shown in Fig. 2. The input image is processed, and multiple heat maps are output. The convolutional neural network is used to implement stage 1. The input original image is x through a CNN model. After testing, the convolution calculation is carried out using the structural models consisted of 14 convolutional layers and 3 Pooling layers. Finally, 2 1x1 convolutional layers are used to output one belief map. If the palm has *p* joint points, then the belief map has *p* layers, each layer represents a heat map of a palm joint point.

The belief map and the training tag calculate the loss function of this stage and store it, adding the loss of each layer at the end of the network as total loss for reverse transmission.

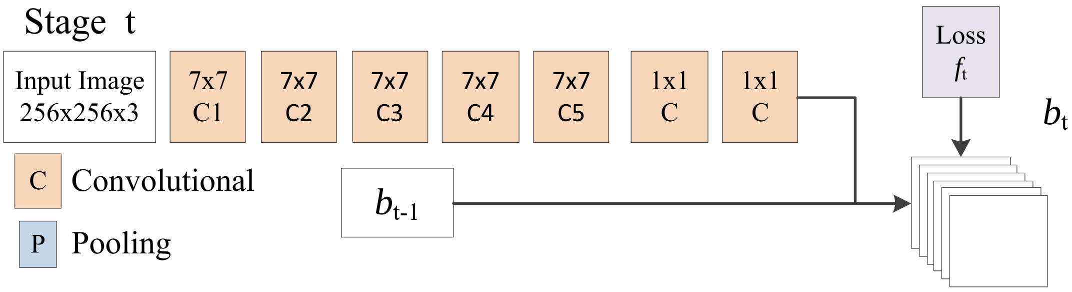

Fig. 3 Stage t Network Diagram

Stage 2 and the following stages have the similar structure, so there are generally called stage t, including stage 2.

In the case of stage 2, the input of the network generally includes two contents: 1. The belief map of the previous stage output; 2. The processing result of the original image, the processing operation here is based on stage 1, which is the x part of Fig. 2. The complexity of the network model will be different according to the number of stages. Generally, the larger the number of stages, the complexity of network model is lower. This paper uses a 4-layer stages, so the complexity of each stage is lower than stage 1. When the 1 and 2 parts are added together, the belief map of stage 2 is obtained. The belief map still has to calculate the loss of the stage with the training tag as stage 1.

It can be inferred that the method of stage t including stage 2 are similar. The reason for using multi-stage is that intermediate supervision can be realized to avoid the situation where the Gradient Disappears.

After 4 stages, we can get the output picture shown in Fig. 4. As we get from the figure, one finger has 4 points, representing 4 degrees of freedom, and 1 point at the wrist represents 6 degrees of freedom of the wrist. Through this picture, you can get the hand image without background, solving the most difficult background interference information in gesture recognition, which lays a solid foundation for the next stage of gesture recognition.

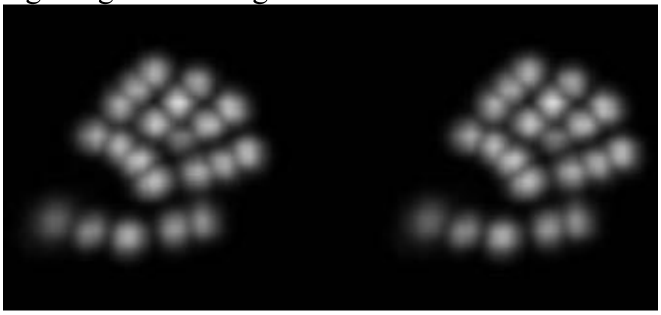

Fig. 4 Output gesture images

4. Gesture Recognition

4.1 Gesture Data Collection.

In order to test the recognition effect, using the computer camera captures gesture pictures of hand, a total of 6 gestures are entered, which are 0, 1, 2, 3, 4, 5 of the gesture, 600 pictures for each gesture, a total of 3000 pictures, as shown in Figure 5 below.

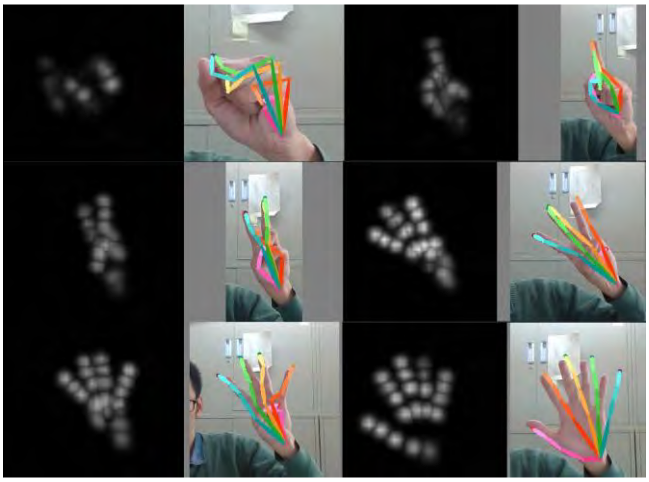

Fig. 5 CMP series picture

Table 1 Label format definition

Image data	gesture	label
00000.jpg~00499.jpg	fist	0
00500.jpg~00999.jpg	one	1
01000.jpg~01499.jpg	two	2
01500.jpg~01999.jpg	three	3
02000.jpg~02499.jpg	four	4
02500.jpg~02999.jpg	five	5

In this paper, the serialization module ‘pickle’ is used to make the dataset, and the label is defined. The label format is shown in Table 1.

By pickleing the image data and the label, a persistent sequence file is generated, which contains the image data and the label information in one file, and the pickle tool can be used to extract the image data and the label for training.

4.2 Training Model

This model uses a 6-layer structure. The first layer of the convolutional layer uses 32 5×5 convolution kernels, using a 2×2 maximum pooling layer, and the activation function uses the relu function; the second layer is the same as first layer; layer 3 convolution layer using 128 3 × 3 convolution kernels, using 2 × 2 maximum pool The activation layer uses the relu function; and the forth layer is the same as third layer. Because the parameter quantity is relatively large, the double Dense layers are used, the first level Dense layer output is 512 types of gesture, and the second level Dense layer outputs the final 6 types of gesture as we mention above. As shown in Fig. 6.

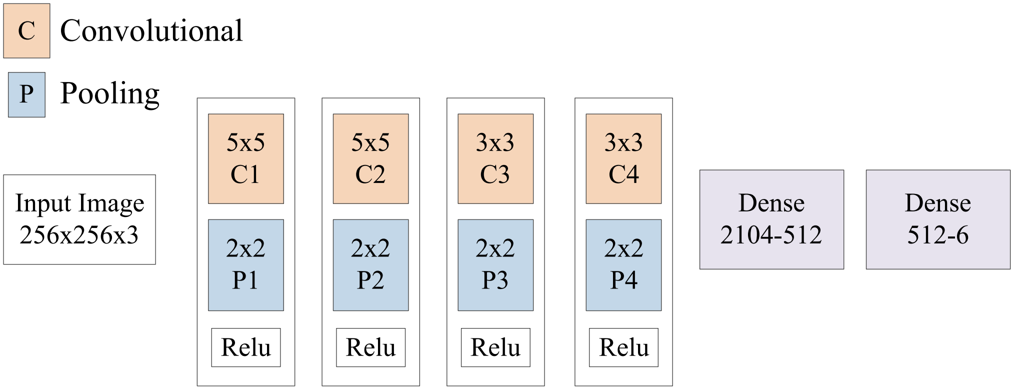

Fig. 6 Gesture recognition network structure

4.3 Test Results

All 3000 data is selected 80% of image data and label as training image data randomly, and the remaining 20% are used as test image data and test label. All training data, labels and test data, label are substituted into the training model set which is described above for training and testing. The results are shown in Fig. 7.

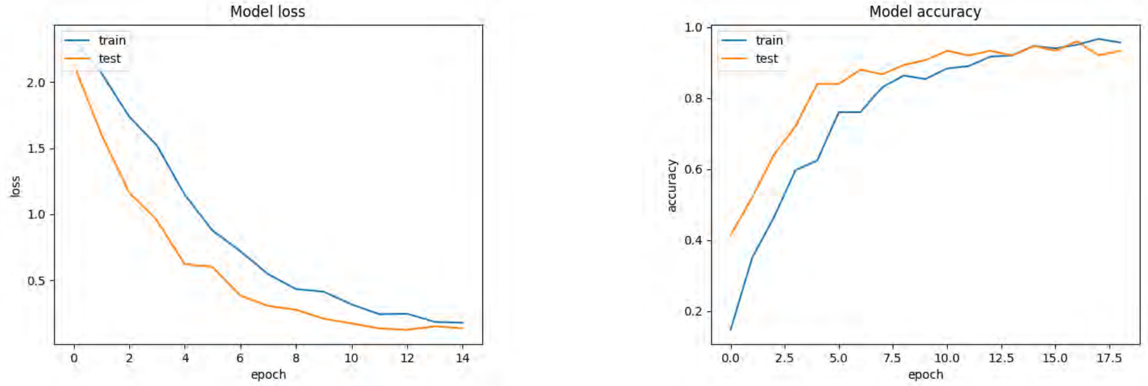

Fig. 7 Result of training and testing

It can be seen from the figure that as the number of epoch increases, the loss of training data gradually decreases, and the accuracy gradually increases; the loss of test data gradually decreases, the accuracy gradually increases, and the final recognition rate can reach about 95%.

5. Summary

We use the method of Convolutional Pose Machine to obtain images of gesture joints in complicated backgrounds in this paper. Using this gesture image, we can make data sets and obtain better recognition results in the case of medium data volume of 3000 images. However, there are still deficiencies, such as the acquisition of the image of the gesture joint, in a particularly complicated background, it will lead to recognition errors, thus obtaining the wrong gesture joint image; the use of the center map method to force the acquisition of the gesture image, resulting in the original In the image, the human hand must be at the center of the image to obtain a better recognition effect; the recognition efficiency of the system needs to be improved. Despite the shortcomings, the ideas presented in this paper still provide new ideas for gesture recognition after all. Being solve the above problems, the system is bound to have better application.

Acknowledgments

Fund project:Provincial key platforms and major scientific research projects in Guangdong universities,project number:2016KQNCX218

Bibliography:Ganzhou Liao(1983-),:MA.Eng, Lecturer, Research direction:Machine Learning, Computer Vision, Pattern Recognition

Xia Zeng(1984-),:MA.Eng, Lecturer, Research direction:Pattern Recognition, Signal processing and recognition

References

[1]. Sturman D J, Zeltzer D. A survey of glove-based input[J]. Computer Graphics & Applications IEEE, 1994, 14(1):30-39.

[2]. Arici T. Introduction to programming with Kinect: Understanding hand / arm / head motion and spoken commands[C]// Signal Processing & Communications Applications Conference. 2012.

[3]. Wang R Y. Real-time hand-tracking with a color glove[J]. Acm Transactions on Graphics, 2009, 28(3):1-8.

[4]. Wang C C, Wang K C. Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction[M]// Recent Progress in Robotics: Viable Robotic Service to Human. 2008.

[5]. Wei S E, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[J]. 2016.