RESEARCH ARTICLE

Human-computer Interaction System Based on Visual Gesture Recognition

Wencheng Dong
Hebei Xianchao Technology Co., Ltd.

Abstract: Human-computer interaction system is the medium for communicating and transmitting information between people and computers. With the rapid development of computer technology, traditional human computer interaction technologies such as mouse and keyboard, have not met the needs of the development of the times. Instead, people need another human-computer interaction technology which is faster, more natural and comfortable. Gesture-based human-computer interaction is one of the most important technologies in human-computer interaction system. There are problems remained in traditional gesture recognition methods, such as low recognition accuracy and complicated recognition process. In view of the defects above, this paper proposes a gesture recognition algorithm based on deep learning. The algorithm detects joint features of the gesture quickly through gesture estimation and classifies joint feature maps by using convolution neural network, which overcomes the difficulties of segmenting gesture images in complex background and improves the accuracy of recognition results. The experimental results indicate that the method has high recognition accuracy for various gestures at different scales, which reaches 98%. Finally, a human-computer interaction system is designed based on the algorithm, and the application of gesture recognition in the human-computer interaction system is demonstrated.

Keywords: Gesture analysis; Deep learning; Gesture prediction; Robots; Gesture recognition

1. Introduction

With the rapid development of computer and robot technologies, human-computer interaction technology has become an extremely important component in the field of computers and robots. Human-computer interaction is the mutual transmission and exchange of information between people and machines in a certain way, thus completing certain specific tasks cooperatively. After years of development, remarkable progress has been made in human-computer interaction technology, making it an indispensable part in the field of computer science.

Traditional human-computer interaction is generally based on the interaction of the display, mouse and keyboard. People send instruction information to the machine through the mouse, keyboard or other input devices. Then the machine acts strictly according to the instruction and transmits feedback information to people through the display. However, with the development of robot technology and the emergence of intelligent robots, such traditional human-computer interaction mode based on keyboard and mouse is no longer suitable for modern human-computer interaction systems. Considering that users may not be familiar with the operation of computers or are not convenient to give instructions to robots through mouse or keyboard, a simpler and more convenient interaction mode, similar to the natural interaction mode between people, is needed. In human-to-human communication, language is recognized as the most natural and effective way of communication, therefore speech-based human-computer interaction has received extensive attention. In many occasions where touch-control operation is not applicable, speech-based human-computer interaction system plays an irreplaceable role. For example, when drivers use GPS navigation system, speech

doi: 10.18063/phci.v2i1.1112

interaction is the most convenient way. For ordinary users, speech interaction also brings good user experience. The application of speech-based human-computer interaction system has become very common. For example, Siri of Apple, Cortana of Microsoft and Uooglc Assistant of Google have all performed well in natural language processing by means of deep learning. At present, they have been widely used and their performance in the market is very mature. However, human communication needs not only sound, but also more extensive communication methods like body languages and expressions. Gesture-based human-computer interaction has become an important part of human-computer interaction.

Besides language, gesture communication is an important way of interaction between people as well. Gesture communication is a body language that can express human's emotions and intentions naturally and intuitively. Different from languages, gestures are universal. When a person comes to a strange language environment, language communication becomes invalid, but he can communicate with others through gestures. In addition, gesture communication has strong anti-interference ability. For example, in noisy situations, while sound transmission is affected, gesture-based communication can still proceed smoothly. Therefore, gestures are an important way for communication between people apart from languages, and also the main way for human-computer natural interaction.

Gesture recognition is the core of gesture-based human-computer interaction technology. The current gesture recognition technology generally preprocesses RGB images of gestures and then uses model matching to recognize gestures. There are many problems in this kind of gesture recognition method. For instance, it is difficult to segment gesture images in complex backgrounds in the image preprocessing stage; due to different angles of gestures, gesture recognition based on model matching may cause false recognition with low recognition rate; the recognition process is carried out in stages with low efficiency.

This paper summarizes the existing gesture recognition technologies and proposes a novel gesture recognition algorithm based on gesture estimation. The algorithm classifies the images through detecting joint features of the gesture quickly and convolution neural network by combining gesture estimation, which overcomes the deficiency of the gesture recognition technology mentioned above, eliminates the need for image preprocessing, and avoids the difficulties of segmenting gesture images in complex background. End-to-end learning avoids the problem of false recognition in model matching and greatly improves the recognition accuracy and efficiency. This algorithm is applied to design a gesture-based human-computer interaction system—rock paper scissors. The system employs a humanoid robot as an actuator to play the game rock paper scissors with people and judge the interaction results. In addition, LSTM network is used to predict people's behavior intention, which increases the enthusiasm of the machine in human-computer interaction and the game interest.

2. Relevant Work

At present, gesture recognition can be divided into two modes, the mode based on data glove[4] and the mode based on vision[5]. The mode based on data glove needs a special glove, which transmits the hand movement information to the computer for analysis through electrical signal, through which gestures can be identified. This method has the advantages of accurate identification, high precision and good stability, and is applied to related fields of real-time systems and virtual reality. Data gloves are relatively expensive and hinder the natural movement of the hand. Vision-based gesture recognition preprocesses a series of images by catching gesture images, separates hand images, and then recognizes gestures by model matching.Without additional equipment, this mode makes gesture expression more natural, thus becoming the main mode in gesture-based human-computer interaction technology.

The traditional vision-based gesture recognition algorithm is mainly divided into three steps: 1) gesture segmentation, which separates gestures from complex backgrounds; 2) gesture analysis, which analyzes the segmented gesture images to obtain parameters; 3) gesture recognition, which recognizes gestures according to parameters obtained from gesture analysis. In these three steps, gesture segmentation becomes the key of gesture recognition algorithm. Only accurate segmentation can provide the basis for subsequent analysis and recognition. Gesture segmentation should consider the complex background of gesture images, differences in skin colors and light.

Tseng *et al.* separated the gesture images from the background by analyzing the special difference between human skin color and background, then binarized the gesture images, and then used mathematical morphology to carry out eroding and dilating on the binary gesture images to eliminate noise. Before gesture recognition, the previous binary gesture image is eroded again, and finally the finger part image are separated from the palm image to form a separated finger image. Gesture recognition recognizes gestures by counting the number of these separated images.

Lin *et al.* separated the hand image through skin color filter, then converted it into black-and-white image, and finally used fingertip angle algorithm[9] to analyze gestures.

Rui Chen converted RGB images into HSV images to make the acquired hand images more accurate, located the gesture by Camshift-based algorithm, then detected the continuous pixels on the edge of skin color to confirm the hand contour, and finally recognized gesture by feature matching.

All the methods mentioned above need to segment gesture images. Under complex background, the image segmentation effect is poor, which seriously affects the recognition accuracy. In the process of gesture recognition, there are strict requirements on the angle of the gesture as different angles of the same gesture will result in different recognition rates and higher false recognition rates. The gesture recognition algorithm based on deep learning proposed in this paper has greatly decreased the disadvantages above.

3. System Design and Implementation

In gesture interaction, the game rock paper scissors is the most direct and simple interaction mode, in which the interactive parties express their intentions through gestures and recognize each other's behavior. Based on the humanoid robot NAO, this paper designs and implements a human-computer interaction system for the game rock paper scissors to interact with people. The system includes two modules: gesture recognition module and prediction module. The gesture recognition module is the core of the system, which is used to recognize human's gestures. The prediction module uses an LSTM network to predict the intention according to the history data of finger gestures. Because NAO's computing power is weak, gesture recognition and prediction algorithms are completed on the workstation's background processing platform, where gesture recognition and result judgment are made. Then the prediction module is called to predict the next finger gesture of the person, and sends an instruction to the NAO robot to complete the next finger gesture.

3.1 Gesture recognition module

On the basis of convolutional pose machines(CPM), this paper proposes EnhancedCPM(ECPM) algorithm for gesture recognition. ECPM consists of CPM sub-network and recognition sub-network. CPM sub-network quickly detects key points of the gesture to framework figure of gesture feature, and then inputs the feature map into recognition network to accurately classify it. The network structure is shown in the figure. ECPM uses end-to-end training and does not need image preprocessing processes such as gesture image segmentation and skin color detection in traditional gesture recognition methods.

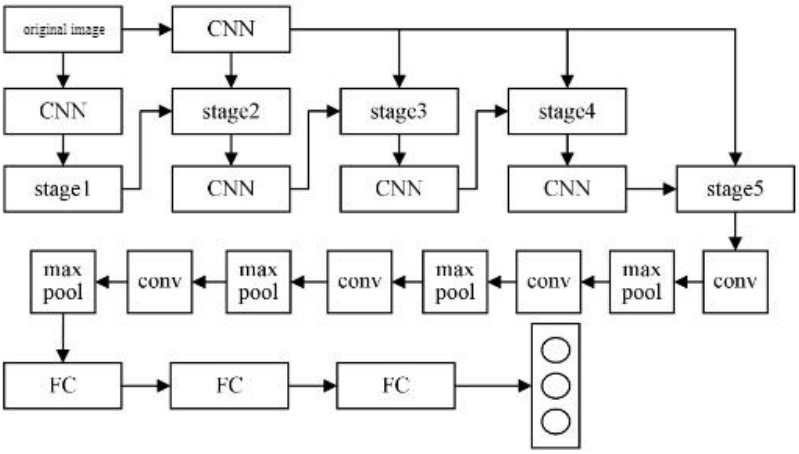

Figure 1. ECPM network structure

CPM network has several stages. The first stage consists of five 9*9 convolution layers and two 1*1 convolution layers. The positions of all joints are directly predicted from the original color gesture images. The obtained results are then passed through a full connection layer to obtain a confidence map of the layer P+1, which is used to predict the output of all joints. Each layer represents the output of one joint, plus a background output.

The structure of the second stage to the fifth stage is the same. The original image passes through a network composed of three 9*9 convolution layers and one 5*5 convolution layer. The obtained feature map is connected with the confidence map output in the previous stage; the result passes through a network composed of three 11*11 convolution layers and two 1*1 convolution layers. Finally, the confidence map of the layer P+1 is output to predict the output of each joint. Each stage of CPM outputs heat maps of predicted positions of all joints, the latter stage refines the positions, and the fifth stage obtains the final heat maps of joint feature , as shown in **Figure 2**.

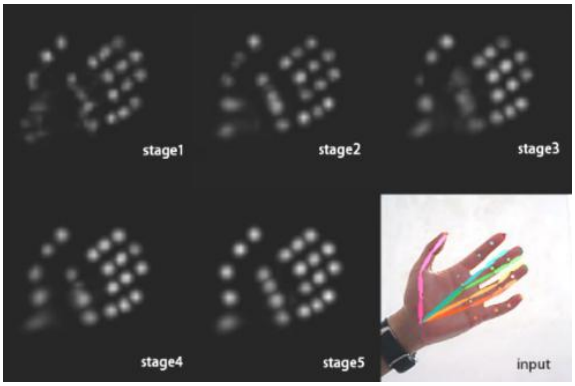

Figure 2. Heat maps of joint feature at different stages

3.2 Prediction module

A person's single finger gesture is completely random and almost impossible to predict accurately, but in a series of gestures, the gesture sequence has a certain tendency that forms a regular sequence.

Long-short term memory network (LSTM) is an improved recurrent neural network (RNN), which adds an input gate, an output gate and a forget gate on the basis of RNN. The added input and output gates, called as cell state, which determines the retained and forgotten information, thus solves the long-term dependence of RNN. LSTM has achieved the best results in a variety of sequence processing tasks, including speech recognition and handwritten numeral recognition. LSTM has developed rapidly in recent years and has been widely applied in processing related tasks of time series .

3.3 System implementation

The process of the human-computer interaction system is shown in **Figure 3**. After the NAO robot is turned on, the system is initialized, the upper computer is connected, and the vision module is started. When ready, human players are invited to play the game rock paper scissors by voice, and the countdown starts. In this process, the pre-trained human behavior prediction model LSTM is called to predict the finger gesture intention of human players. After the countdown, the robot throws the gesture according to the prediction result, and the human player has also completed throwing the action. NAO grabs the gesture image of the human and calls ECPM algorithm for gesture recognition. The recognition result and the robot's own gesture are used together to judge the game result by the decision tree. Each gesture of a human player is recorded in historical information as a data source for predicting subsequent actions.

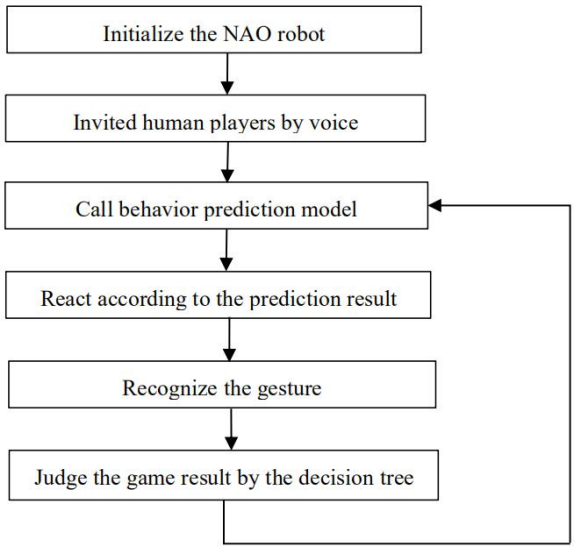

Figure 3. The process of the system

4. Conclusion

This paper proposes a gesture recognition algorithm based on deep learning, which improves the rate of progress of the traditional recognition algorithm based on image segmentation—model matching. The algorithm combines the advantages of gesture joint feature detection and convolution neural network image classification, greatly improving the accuracy of gesture recognition, and the accuracy of the final recognition result reaches 98%. In this paper, a human-computer interaction system based on this algorithm is designed, completing the human-computer interaction process based on visual gesture recognition by means of the game rock paper scissors.

References

1. Agah A. Human interactions with intelligent systems: Research taxonomy [J]. Computers and Electrical Engineering 2000; 27(1): 71-107.
2. Tadeusiewicz R. Speech in human system interaction [C]. Human System Interactions. IEEE 2010: 2-13.
3. Qi J, Xu K, Ding XL. Vision-based on hand gesture recognition for human-robot interaction: A review [J]. Robot 2017; 39(4): 565-584.
4. Panwar M. Hand gesture recognition based on shape parameters [C]. International Conference on Computing 2012: 317-319.