2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI) | 978-1-6654-3881-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CEI52496.2021.9574581

Design of Human-computer Interaction System Using Gesture Recognition Algorithm from the Perspective of Machine Learning

Jiawei Li
Applied Mathematics
Beijing Normal University, Zhuhai
Zhuhai, China
yy13532285812@163.com

Abstract—**The human-computer interaction (HCI) system is the hub of conversation between man and machine. The previous HCI system is challenging to meet the needs of modern society with the continuous progress of computer technology. A new HCI technology different from the past needs to be put forward to solve the problem that the development of computer technology and outdated HCI technology lag behind the times. Machine learning aims to build a behavior method that can obtain information from data and predict the data. Machine learning features are used and integrated into gesture algorithms through the basic principle of a finger-guessing game. Gesture estimation is adopted to detect joint gesture features efficiently, and a convolutional neural network is employed to plan and process the joint features to solve the bottleneck of poor segmentation of gesture images in different environments. HCI system based on gesture recognition algorithm is designed from the perspective of machine learning. The final experimental results show that this method has good recognition accuracy for the performance of different scales of various gestures, and the accuracy of recognition results can reach 70%. It suggests that the system is theoretically conditional on making the recognition result reach the most authentic level. It is demonstrated that gesture recognition is of excellent use value in this HCI system, and the HCI system established here also has an efficient reference value.**

Keywords- machine learning; gesture recognition algorithm; HCI system; finger-guessing game

I. INTRODUCTION

Computer technology and robotics technology continue to improve with the progress of the times. As a hub connecting the two, human-computer interaction technology has become a key role in computers and robotics. In short, the so-called human-computer interaction technology is that humans and machines communicate in a specific way and then cooperate to complete some specific tasks under the premise of ensuring quality and efficiency [1]. HCI technology has also proliferated with the progress of the times and has its place in computer science. In this regard, the traditional HCI system is reviewed. This exploration is based on building a new HCI system through a gesture recognition algorithm from machine learning.

Present gesture recognition is generally divided into two kinds. One is the recognition method based on data glove; The second is based on vision [2]. Tseng et al. (2017) [3] studied the difference between human skin color and background, segmented the gesture image from the background, and then processed it in the way of binarization. The later binary gesture graphics are processed for corrosion, expansion, and noise reduction through mathematical morphology. Before gesture recognition, the previous gesture binary image is repeatedly corroded. Then, some finger and palm images' segmentation are completed, showing a different finger image state than before. The recognition principle is to analyze the number of these scattered images and then perform gesture recognition. Cheng (2017) et al. [4] segmented the hand image using the skin color filter, transformed the image into black-and-white form, and finally calculated and processed the gesture through the fingertip angle algorithm. All the methods mentioned above need to segment the gesture image. The image segmentation result is unsatisfactory because the background environment is not very suitable, which has a tremendous negative impact on the recognition accuracy; the gesture angle will also have essential requirements during the gesture recognition process. Even if the gestures are the same and the angles are different, it will also negatively impact recognition accuracy, and the probability of false recognition will be very high.

The most crucial part of the HCI technology based on gesture is gesture recognition. Present gesture recognition technology usually processes the RGB (red, green, blue) diagram of gesture in advance and then uses model matching to recognize gesture [5]. Many shortcomings will accompany this kind of gesture recognition method. For example, the difficulty of segmentation of gesture images in complex backgrounds is relatively considerable in image preprocessing: gesture recognition based on model matching will lead to misrecognition of the gesture direction, thereby affecting the recognition rate: the recognition process will be divided into parts, and the recognition speed will not be soon. The proposed gesture recognition algorithm improves the above shortcomings based on deep learning. The relevant technologies currently used for gesture recognition are used for reference, and gesture recognition algorithms based on previous pose estimations are found. The algorithm uses gesture estimation to discover gesture joints' characteristics quickly and uses convolutional neural networks to classify and process images, thereby improving the existing disadvantages of the gesture as mentioned above recognition technology. It does not need to process images and avoids some difficulties when gesture images are difficult to process when the background

978-0-7381-4649-2/21/$31.00 ©2021 IEEE

environment is lacking. At the same time, end-to-end learning can avoid the disadvantages of model matching misidentification and significantly bring the recognition result closer to the origin. In addition, the recognition speed can be increased simultaneously. Finally, the algorithm was adopted to build a gesture-based HCI system-finger guessing game. The executors of the system use humanoid robots to play finger guessing games, communicate information with human players, and judge the results of interaction: LSTM (Long Short Term Memory Network) is used to predict human behavior intentions, improve the machine's initiative in HCI, and increase interaction Interest in making machine learning play a role in the HCI system [6].

II. System Design and Implementation

With the continuous optimization of machine learning technology, data can be "predicted". From a human point of view, the more people have experienced the past, the more likely people are to correctly judge the future. There is a positive correlation between the two. It is often said that people with "experience" have an advantage in the workplace over employees who are "new to the workplace"; experienced people get more rules than others. How to use machine learning features to add new impetus to the design of the new HCI system? We introduce finger guessing games to deepen the design gradually.

Finger-guessing game is considered to be Introduce finger the most direct and simple form of interaction in gesture interaction. In this guessing game, the interactive parties can use gesture recognition to express their intention and recognize each other's words and behaviors.

The most crucial platform of this exploration is NAO (Biped humanoid intelligent robot). An HCI system to simulate the finger-guessing game is built to interact with –the man and the machine in the guessing game [7]. There are two points in this system: gesture recognition and prediction module. The most crucial module of the system is gesture recognition, which is adapted to identify player gestures; the prediction module uses the LSTM network to predict the boxing purpose according to the previous data of players' boxing.

For NAO robots with insufficient computing power, gesture recognition and prediction algorithms must be run on the workstation's back-end processing platform. That is, gesture recognition and win and lose decisions need to be performed in the workstation. Then, the prediction module is projected to guess the fist's next gesture and send the action command to the NAO robot for the following fist action. Then, the whole process was completed. Figure 1 shows the system architecture diagram [8].

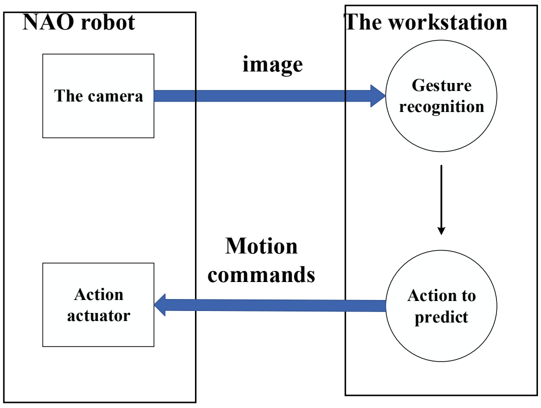

Figure 1. Architecture of robot boxing guessing system.

A. *Gesture Recognition Module*

The enhanced convolutional pose machines (ECPM) algorithm for gesture recognition is proposed based on the convolutional pose machines (CPM) [9]. CPM sub-network and identification sub network jointly construct ECPM. The process is that the CPM subnet quickly detects the critical features of the gestures, obtains the gesture feature skeleton block diagram, and then transmits the feature map to the recognition network, thereby classifying the detected feature skeleton diagram more accurately. ECPM adopts an end-to-end model and does not require a series of image preprocessing processes such as gesture image segmentation and skin color detection in traditional gesture recognition methods. Figure 2 shows its network structure.

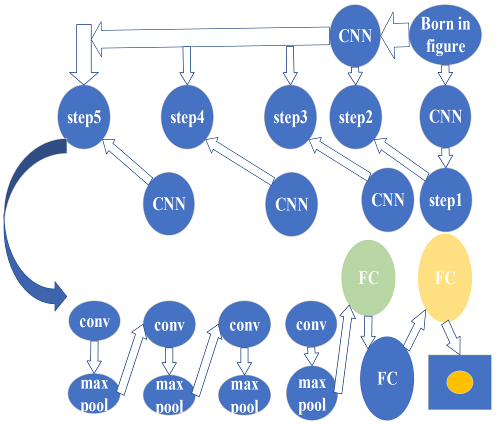

Figure 2. ECPM grid structure diagram.

The CPM network has many stages. Its first stage consists of five 9*9 convolutional layers and two 1*1 convolutional layers. In the initial color gesture image, the relevant specific directions of the joints are preliminarily detected. After the results are obtained, the confidence map of the P+1 layer is obtained through the fully connected layer to predict the output of other joints. A layer represents only the output of a joint and then attaches a background output.

Phases 2 to 5 are composed similarly. The original image passes through a network composed of three 9 * 9 convolutional layers and one 5 * 5 convolutional layer. The confidence map output in the previous stage is connected after the characteristic map is obtained; then, the interconnected network is composed of three 11 * 11 convolutional layers and two 1*1 convolutional layers. Thus, the confidence map of the P+1 layer is output to infer the output of other joints. In all stages of CPM, the heat map of the theoretical orientation of all joints is output, and the position is accurate later. The final joint characteristic heat map is obtained at stage 5 [10].

In order to avoid the disappearance of gradient, the output in each stage in Fig. 2 will use a loss function to minimize the error of predicted joint position and ideal joint position. The best position confidence diagram of each joint position is recorded as:

$$b^{p}*(Yp)=Z \tag{1}$$

$$f_t = \sum_{P=1}^{P+1} \sum z \in Z \left\| b_t^p(z) - b_*^p(z) \right\|_2^2 \tag{2}$$

p goes through the joints at each position, and *z* is the position of the corresponding joint. Then, the sum of the loss function at each stage is:

$$F = \sum_{t=1}^{T} f_t \tag{3}$$

All the network parameters in each stage use the standard random gradient descent method. In the second and subsequent stages, the network weight values corresponding to the convolutional layer are shared to obtain the shared image feature map results.

The primary convolutional neural network is the end of ECPM. It consists of four convolutional layers and the four largest collection layers. Then, the joint function image output characteristics in the fifth stage are recorded and transmitted through three fully connected layers. Finally, if the output dimension of the fully connected layer is 3, it corresponds to three different gestures, and each gesture image is mapped to a related type.

B. Prediction Module

Statistically, there is no rule for human players to punch at random, and there is no way to budget accurately. If the data sample size of boxing is large enough, the data are analyzed and processed. The boxing sequence of individual players will have some sequence characteristics. The law of this group of sequences may become the theoretical and statistical basis for predicting their boxing.

The recurrent neural network (RNN) is improved to obtain LSTM, which has one more input, one more output and forget gate under the premise of RNN [11]. The added input and output are cell states, which often affect some information retained and forgotten to the greatest extent so that the disadvantages of long-term dependence of RNN no longer exist. LSTM shows the best state in various sequence data tasks, including speech recognition and handwritten numeral recognition. The development speed of LSTM is considerable in recent years. It will be used more when analyzing relevant data of time series [12]. Fig. 3 presents the LSTM network structure and the internal structure of a single cell.

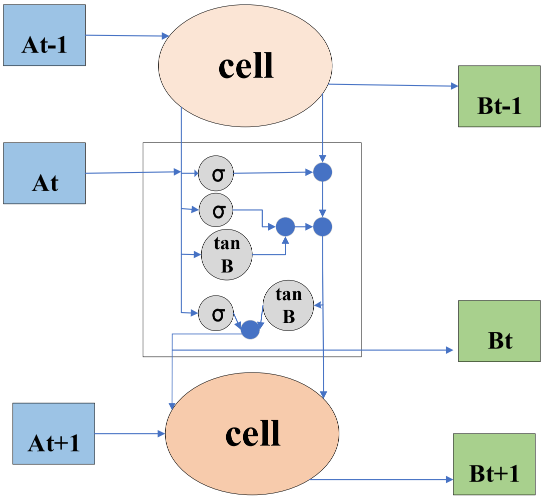

Figure 3. LSTM network structure and cell internal structure.

The sequence of human player actions is recorded as a time series a A{A$_1$, A$_2$,…,A$_t$,…}. In the data, the *t*-th action in the action execution is represented by A$_t$. Stone, scissors, and paper are represented by 0, 1, and 2. An LSTM network is built, which covers four hidden layers and one output layer. 1 is the value set as the timestep. The following output prediction value 0, 1 or 2 is the category predicted in the next punch. The tanh is adopted as the activation function, the action sequence is taken as the dataset, and the loss function uses Mean-Square Error (MSE). The expression of mean square error is:

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{4}$$

In (1), $\hat{\theta}$ is the predicted value, and θ is the true value [13].

C. Result Judgment Design of Finger-guessing Game

In the finger-guessing game, there will be three different gestures: scissors, paper, and stone. There are four corresponding results: paper wins stone, stone wins scissors, scissors win paper and a draw. There are nine groups of all gesture result states of finger-guessing robot and player, and the interaction result is relatively single. With the robot as the decision node, the boxing action is the state node, and the final result is the result node. Tab. 1 displays the winning and losing combination.

TABLE I. WINNING AND LOSING COMBINATION OF FINGER-GUESSING GAME

Robot People	Scissors	Stone	Cloth
Scissors	A draw	Robot wins	People win
Stone	People win	A draw	Robot wins
Cloth	Robot wins	People win	A draw

D. Result Judgment Design of Finger-guessing Game

The designed man-machine game system will take the NAO humanoid robot produced by France Aldebaran Robotics as the platform and transmit the collected image data to the workstation. After the data are processed, the workstation sends the control command to the NAO humanoid robot to correct the response.

NAO humanoid robot is 57cm high, with 25 degrees of freedom on the body, and can efficiently complete many kinds of actions. It is also equipped with two high-resolution cameras with a maximum resolution of 1280*960, with functions such as wireless network image transmission, voice output, multi-language, different platforms, and multi-language programming. The workstation is equipped with E5-2620 v4 Dual CPU, 64 G memory, 500 G SSD, Quadro P5000 display adapter and Ubuntu 18.04 operating system [14].

After the NAO robot starts up, it will "reset" the system, connect the upper computer, and open the vision module. After completion, other players will be invited by voice to start guessing, and then it starts to calculate the time. During this period, it will run the trained LSTM human behavior to predict the boxing trend of the model to human players, to make a series of inferences. After the set time expires, the robot will punch with the estimated results. When the NAO robot obtains the human gesture image, it will call the ECPM algorithm to recognize the gesture. The detection results will be determined according to the instructions of the robot and table 1 above. All fist gestures of players will be stored in "memory" and become the source of prediction data. Fig. 4 presents the flow of the HCI system.

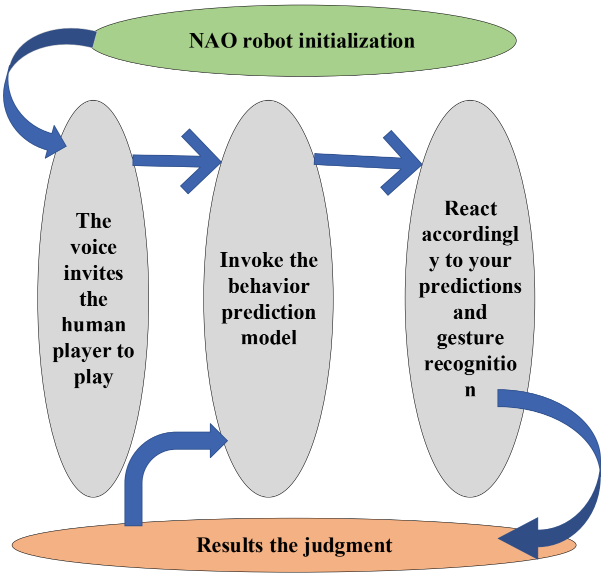
NAO robot initialization
The voice invites the human player to play
Invoke the behavior prediction model
React accordingly to your predictions and gesture recognition
Results the judgment

Figure 4. System flow chart.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In reality, two players play a finger-guessing game. The boxing sequence of one of them is recorded, totaling 150 groups. Stone, scissors, and paper are represented by 0, 1, and 2, respectively, forming a time series {0,1,2...}. The first 100 sequences are taken as the train set, and the last 50 sequences are taken as the test set. The epoch of training is adjusted to 100. The results of the prediction are plotted as shown in Fig. 5.

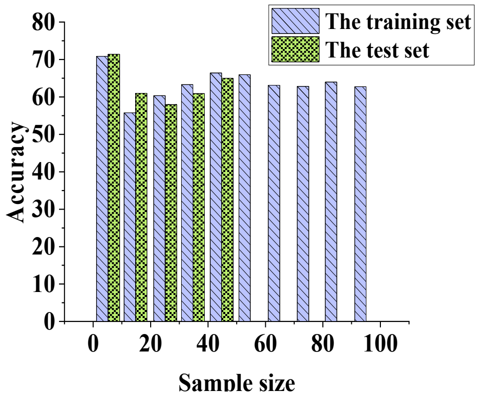

Figure 5. Comparison histogram of prediction accuracy of gesture sequence.

Figure 5 displays that the prediction accuracy can be as high as 70% when data is less than 10. However, the accuracy will decline slightly with the continuous improvement of the data; it gradually increases and tends to be stable after a minor fluctuation. The accuracy is stable and greater than 60% when the later data is greater than 50. It can be concluded that the prediction accuracy on the train set can reach 62.37%, and the prediction accuracy on the test set can reach 64.44%. The results prove that it has good stability [15].

IV. Conclusion

Driven by the trend of the times, computer technology and robot technology continue to improve. As the hub connecting the two, HCI technology has become a crucial role in computers and robots. Hence, it is essential to make the information exchange and transmission between humans and machines faster and more accurate to cooperate to complete specific tasks. Based on the traditional HCI system, from the perspective of machine learning, the prediction module is established through the gesture recognition algorithm using the humanoid robot NAO platform and the LSTM network. After the completion of the system process, a new HCI system is constructed. The gesture recognition algorithm under the premise of deep learning explored optimizes the process of the previous recognition algorithm based on image segmentation - pattern matching. This algorithm extracts the two significant advantages of gesture standard features of gesture estimation, fast detection and image classification of the convolutional neural network and improves gesture recognition accuracy. The accuracy of the final recognition result can be as high as 70% or more. The human-computer interaction system based on visual gesture recognition is completed by guessing the finger game. Improving the human-computer interaction system from the perspective of machine learning has a good reference significance. However, the small sample size in the experimental prediction module may cause the problem of small samples. Increasing the amount of sample data in data prediction will make the experiment of the prediction module more convincing. In principle, the experiment relies on the basic principle of the finger-guessing game, which is relatively single from the perspective of machine learning. It can be enriched in the subsequent development process to improve the learning speed and quality of the machine.

References

[1] X. Wang, Y. Wang and Y. Qi, "Data Exchange of Relay Transmission and Information Scheduling in Shared Networks," IEEE Access, 2019, 7, pp.1-1.

[2] X. Wang and K. Yan, "Immersive human-computer interactive virtual environment using large-scale display system," Future Generation Computer Systems, 2017, 96(JUL.), pp.649-659.

[3] K. T. Tseng, S.F. Lo, D.S.Chan and C. B. Yang, "Efficient merged longest common subsequence algorithms for similar sequences," Theoretical Computer ence, 2017, 708,pp. 75-90. *(references)*

[4] M. Cheng, E. Klopfer, L. Rosenheck and C. Lin, "Analyzing gameplay data to inform feedback loops in The Radix Endeavor," Computers & education, 2017, 111(AUG.),pp.60-73. *(references)*

[5] J. Wen, J. Zhang and F. Wang, "Dynamic Hand Gesture Recognition Based on Short-Term Sampling Neural Networks," IEEE/CAA Journal of Automatica Sinica, 2021, v.8(01), pp.114-124.

[6] Z. Huang, R. Li, K. Driggs-Campbell, A. Hasan and K. Shin, "Long-Term Pedestrian Trajectory Prediction Using Mutable Intention Filter and Warp LSTM," IEEE Robotics and Automation Letters, 2020,99, pp.1-1.

[7] S. Benferhat, K. Tabia and M. Ali. [Lecture Notes in Computer Science] Advances in Artificial Intelligence: From Theory to Practice Volume 10350 || NAO Robot, Transmitter of Social Cues: What Impacts? 2017, 10.1007/978-3-319-60042-0(Chapter 62), pp.559-568.

[8] L. Braubach, M. Lama, L. Burgueño, N. Moha, M. Oriol, J. M. Murillo and N. Kaviani. [Lecture Notes in Computer Science] Service-Oriented Computing – ICSOC 2017 Workshops Volume 10797 || Power-Based Device Recognition for Occupancy Detection. 2018, 10.1007/978-3-319-91764-1(Chapter 14):174-187. Y Liu, X Yuan, X Gong. Conditional Convolution Neural Network Enhanced Random Forest for Facial Expression Recognition. Pattern Recognition, 2018, 84, pp.251-261.

[9] M. H. Calp and M. A. Akcayol, "Optimization of Project Scheduling Activities in Dynamic CPM and PERT Networks Using Genetic Algorithms," Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 2018, 22(2), pp.10.

[10] B. D. Unluturk and I. F. Akyildiz, "An End-to-end Model of Plant Pheromone Channel for Long Range Molecular Communications," IEEE Trans Nanobioscience, 2017, 16(1),pp.11-20..

[11] S. Chandraprabha, G. Saravanan,C. Dinesh and A. M. Ibrahim, "LSTM Model Based Wind Speed Forecasting," International Journal of Future Generation Communication and Networking, 2020, 13(4), pp.1726-1731.

[12] Q. Miao, M. Gong, J. Liu, J. Song,C. Ying and X. Ge, "RBoost: Label Noise-Robust Boosting Algorithm Based on a Nonconvex Loss Function and the Numerically Stable Base Learners," IEEE Transactions on Neural Networks & Learning Systems, 2017, 27(11), pp.2216-2228.

[13] R.Tcp/Ip, M. Technical, C. Systems, M. T. Publishing, C.N.T. Tin and Q.T. Mạng, "CCIE Professional Development: Routing TCP/IP," Tailieu Vn, 2005, 21(3),pp.3-392.

[14] C. M. Own, F. Sha and W. Tao, "Triplet Decoders Neural Network Ensemble System and T-Conversion for Traffic Speed Sequence Prediction," IEEE Access, 2019, 99, pp.1-1.

[15] A. Norman, P. Telfer, H. Kuchel, J. Taylor and E. Tanaka, "Increased genomic prediction accuracy in wheat breeding using a large Australian panel," Theoretical & Applied Genetics, 2017, 130(2), pp.1-13.