Research on Cooperative Control of Human-Computer Interaction Tools with High Recognition Rate Based on Neural Network

Shi Heng
School of Astronautics
Beijing University of Aeronautics and Astronautics
Beijing 100191, China
e-mail: sh_shiheng@outlook.com

Dong Yunfeng
School of Astronautics
Beijing University of Aeronautics and Astronautics
Beijing 100191, China
e-mail: sinosat@buaa.edu.cn

Abstract

Human-Computer Interaction (HCI) is the vital technology of Virtual Reality. The control using several different HCI tools together provides a natural and effective method of human-computer interaction. This paper provides a way of cooperative control with multi-recognition tools. The separate tools, including speech recognition tools, hand gesture recognition tools, posture recognition tools, are developed firstly. Then assign the cooperation program receives the recognition result sent from the three independent tool through UDP protocol. A neural network that fits the habits of the operator is trained. Process the three recognition results using the trained BP neural network. Higher recognition rate compared with a single HCI tool is accomplished. Operators can communicate with the machine much more naturally and effectively.

Keywords: Human–Computer Interaction; Virtual Reality; Neural Network; Cooperative Control

1. Introduction

Virtual reality technology is one of the most important scientific and technological achievements in the 1990s. It can stimulate the behavior of the real world, and make real-time response to user's location, attitude, language, etc. With some certain interactive device, participants can communicate with the objects in virtual environment in a way close to natural, so that a real-time interaction can be established between participants and virtual environment, producing an experience similar to the real environment. With the development of virtual reality, human-computer interaction was increasingly applied in all walks of our life. Such as speech recognition, hand gesture recognition, image recognition, eye tracking recognition, skeleton scanning recognition, and so on. They were put into use widely at home and abroad for their different advantages.

Speech recognition technology has been widely used in healthcare, military, telephony and other domains [1], including higher fields such as aerospace. The Mars Polar Lander of NASA employed speech recognition system in its lander [2]. As to data glove, such as CyberGlove, 5DT Data Glove, these new glove products were also applied widely in the fields of pattern recognition, virtual reality, etc [3]. Moreover, the new motion-sensor device called "Kinect" by Microsoft was taken by a lot of developers to develop the functions of body recognition in the field of virtual reality since its SDK was released in 2011. K. K. Biswas and Saurav Kumar Basu developed a posture recognition program based on Kinect in 2011 [4].

In the traditional research of HCI, the single interactive tool is usually employed. Such as the isolated word recognition system by Jianxiong Wu and Chorkin Chan [5] and the gesture recognition program using Kinect by Madabhushi and Aggarwal [6]. However, it is defective to apply a single existed recognition tool into practical use to realize interaction. First of all, the security is insufficient. It is unreliable to be used in the fields like aerospace. For example, an astronaut manipulates the devices in space by a speech recognition tool. Even though the recognition rate of the tool is 99%, once the "1%" touches off an error operation, the consequence would be disastrous. Moreover, the decline of recognition rate is inevitable when the device is interrupted in noisy environments. Secondly, the convenience of use is insufficient too. When we visit a bakery store, we are usually not allowed to take a cake out from the showcase. Instead, we have to ask the store attendant to get the cake for us. In this case, we usually say "this one" by pointing with a finger at the cake we want instead of spelling out the name of the cake. The store attendant will then nod to confirm that he/she understands what we mean. Not only people can express his real opinion in this way, but also the attendant will understand him easily. Under the trend of diversification in human-computer interaction, the cooperation of several interaction patterns is more humane than a single one.

But cooperation is not just simply put several HCI tools into use together. Shunji UCHINO developed a real-time VR interaction system between Avatar with speech and gesture recognition, which accomplished a natural communication interface in virtual reality [7]. But the cooperative control that he used is just a simple combination of two tools. The final program didn't realize higher recognition rate. If a certain algorithm is used to improve

978-1-4799-6854-1/14 $31.00 © 2014 IEEE
DOI 10.1109/ICVRV.2014.6

the cooperation strategy of the interactive tools, the several tools could complement each other's advantages, and a higher efficiency could be accomplished. This article describes a way of combination of the speech recognition tool, hand gesture recognition tool and the posture recognition tool together, which aims to achieve higher reliability and practicality.

2. METHODS: PRIOR WORK

N. Abe applied HCI technology in virtual assembly in engineering [8]. He pre-assembles the fittings with the help of virtual reality technology to verify the correctness of the assembly process, which plays an important role in the development of products. This is very representative in HCI applications. Therefore, the final control objectives of the cooperation program this article discussed are the following controlling and viewing orders: “Left”, “Right”, “Up”, “Down”, “Zoom out”, “Zoom in” and “Break”.

2.1. Speech Recognition Tool

The control instructions specified are simple. They could be defined and distinguished by short vocabularies, instead of complex and lengthy presentation. So what the speech recognition tool uses is limited isolated vocabulary.

In this article, the speech recognition program portion employed Microsoft Speech SDK 5.1 to carry out second development, which achieved the recognition of voice order of limited vocabulary. The principle is as follows. First of all, create the grammar file of the speech order according to the vocabularies needed to recognize. Secondly, after recognized by the recognition engine, send the code to the target program though RTI_UDP tools. Finally, process the information received in the receiving and processing function.

The final program could achieve the goal that a non-specific operator could use the wireless microphone say a limited word of ten to operate the OSG graphical model. The recognition rate can reach 90%.

2.2. Gesture Recognition using Data Glove

Hand gesture recognition portion is based on 5DT data glove. It recognizes the gesture by dynamic gesture recognition [9]. Using the method of finding key frame, it searches the initial and final state of each hand gesture, which involves the information of the position and attitude of the palm and stretching information of fingers, and so on. The basic principle can be summarized as follows: within a certain time interval, once recognize the initial and final state of a certain hand gesture, it is considered to have been made the kind of hand gesture.

In order to identify the one-handed action, the attitude angle of right hand palm is set as state parameter. The program constantly updates the state information of hand. For example, considering the consuming time of the dynamic gesture “Left” is in the range of 0-3 seconds set the time interval of 3 seconds. If the program did not find the initial and final state in the time interval, it would be turned into the next finding stage, and so on. The program flow chart is shown in Fig. 1. Because of the error is existed, as long as it is within the corresponding angle of plus or minus 15 degrees, such condition is identified. The variable “possible” represents the possibility of a certain hand gesture, whose initial value is 0. After obtaining the initial state of hand complied with a certain order, the variable “possible” will be set to “i”. “Left”, “Right”, “Up”, “Down” represents “1”, “2”, “3”, “4”. If there is a corresponding final state within the given time, output the order of the dynamic hand gesture. If there is no corresponding state recognized within the given time, reset the “possible” to 0. The variable “time” is to record the time of system, so that the time interval of the hand gesture could be collected.

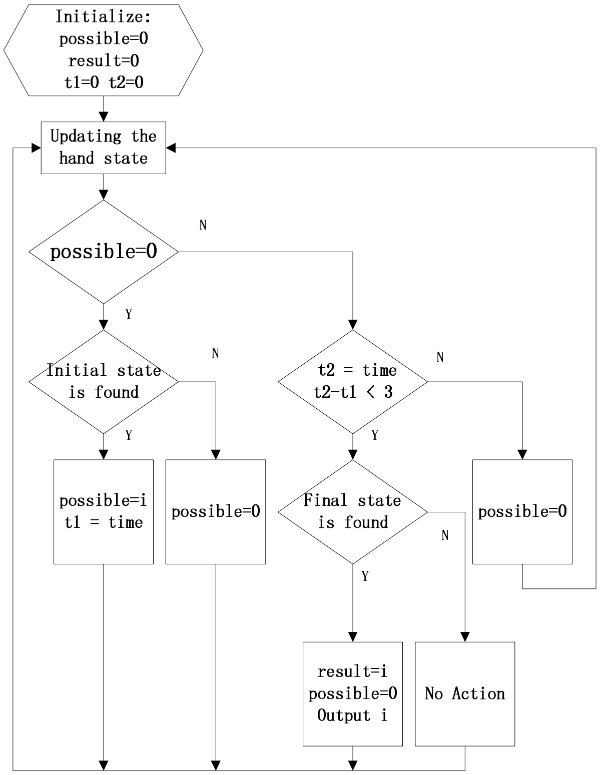


Figure 1. Program flow chart of hand gesture recognition tool

For the sake of recognizing the four kinds of hand gesture, define the state parameter as hand palm attitude angle, including yaw angle, pitch angle and roll angle. Each of them ranges from -180 degree to 180 degree. The four kinds of hand gesture are defined by the initial and final state of the attitude angle, which is shown in Table 1. In addition, define the fist gesture as “Stop”. The recognition rate of this program is more than 95% by extensive testing.

Table 1. The initial and final attitude angle of hand palm(unit:deg)

	Initial Yaw Angle	Initial Pitch Angle	Initial Roll Angle	Final Yaw Angle	Final Pitch Angle	Final Roll Angle
Left	90	0	-90	45	0	-90
Right	90	0	90	135	0	90
Up	90	0	0	90	45	0
Down	90	0	-180	90	-45	180

For example, the gesture “Left” is shown in Fig. 2.

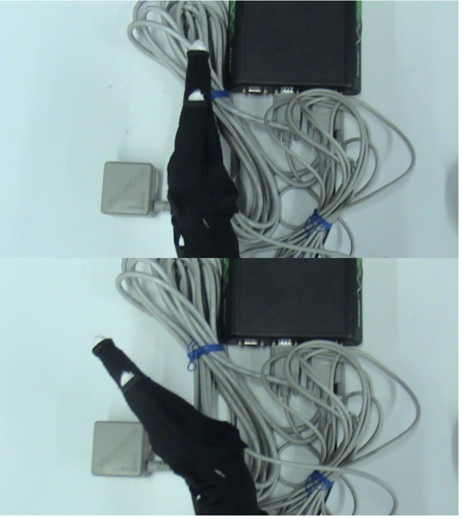


Figure 2. Initial and final hand gesture state of command “Left”(Top view)

2.3. Posture Recognition using Kinect

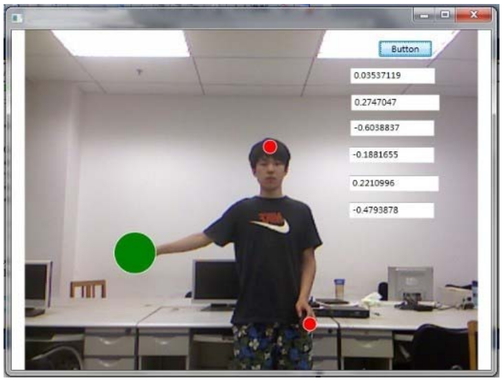
Posture recognition portion uses Kinect to get the position of the operator’s head, left hand and right hand. By judging the relative position of the three points, the static action could be recognized. The definition of each action is shown in Table 2.

Table 2. Definition of actions and the way to judge(unit: m)

Command	Definition	Judging method
Left	Stretch out left arm left	leftHand.X < head.X-0.5
Right	Stretch out right arm right	leftHand.X< head. X-0.5
Up	Raise right arm up	rightHand.Y> head.Y + 0.2
Down	Stretch out right arm down	rightHand. Y< head.Y - 0.7
Zoom out	Cross left and right arm in the chest	rightHand.X - leftHand.X < -0.1
Zoom in	Stretch out both arms	rightHand. X -leftHand. X >0.9
Break	Hands put together	$(\text{rightHand.X -leftHand.X})^2 + (\text{rightHand.Y -leftHand.Y})^2 < 0.02$

The posture “Left” is shown in Fig. 3.

By testing, the recognition rate of the tool could reach 100% while the action is up to standard. But it may causes some undesired operation in some cases.



0.03537119

0.2747047

-0.6038837

-0.1881655

0.2210996

-0.4793878

Figure 3. Interface of the posture recognition tool

2.4. Combination

In this study, User data protocol (UDP) is employed to realize the processing of real-time information of the cooperation program. Compared with TCP protocol, UDP has the following advantages. Firstly, the transmission speed is faster. Secondly, UDP can send data to multiple computers simultaneously.

The operating process of the cooperation program is shown in Fig. 4.

Register a UDP channel for each of the three interactive tools’ program. They can send data to the receiving unit separately. The seven control commands that could be sent are replaced by numbers from 1 to 7. The receiving unit continuously scans the input. When it receives an instruction code, the timer starts to count 2 seconds. Receive the instruction code sent from the other two channels during the timing. If there is no signal input, replace it by 0. Then, send the instruction codes received from the three channels to the processing unit. In this unit, data input is processed by the trained BP neural network. This portion is the key to realize high recognition rate in the cooperation program. Data output is the final recognition command. Send it back to the operating unit, the corresponding action will take place.

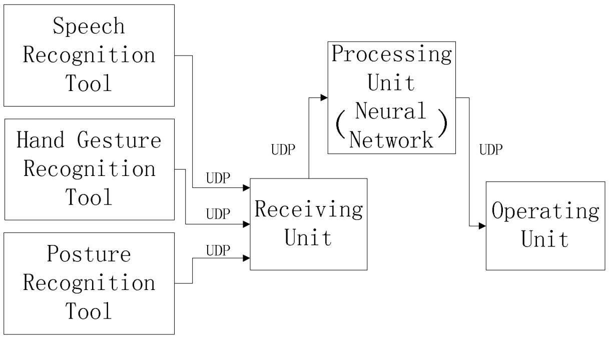


Figure 4. Flow chart of the cooperation program

3. Neural Network Training

3.1. Training Method

Neural network training is the key to achieve the final scheme. Without this portion, many unexpected problems are likely to emerge. For instance, the operator says a word or performs an action that has nothing to do with the program, but it is mistakenly captured by the interactive program and causes the sudden operation. Another case is that the device cannot stop when the operator want it stop owing to the noise. In order to achieve the goal of higher recognition rate, the basic rule of training the network is as follows: when two of the three interactive tools have sent the same instruction code, then perform the operation; when only one of the three interactive tools has sent the order "Break", then perform the emergency stop operation at once. The training of network is done through Matlab neural network toolbox.

The structure of network training portion is same as the first part of the cooperation program. The receiving unit receives recognition data from the three interactive tools during the two seconds after getting the command of start training. Save the data received as a set of training samples in the database. Then the operator defines the commands he issued just now. That is the completion of teaching the machine.

3.2. Achievement in Matlab

3.2.1. The establishment of training samples

Since there are both seven commanding orders of the speech recognition tool and the Kinect posture recognition tool, and five commanding orders of the hand gesture recognition tool based on data glove, the input of the cooperation program has 19 characteristic values altogether. So the input can be replaced by a 19-dimensional vector every time. While there is a certain commanding order appears, assign "1" to the corresponding position; or assign it as "0". The specific defining method is as follows.

```
p=[Speech Left; Speech Right; Speech Up; Speech Down; Speech Zoom Out; Speech Zoom In; Speech Break; Hand Left; Hand Right; Hand Up; Hand Down; Hand Break; Kinect Left; Kinect Right; Kinect Up; Kinect Down; Kinect Zoom Out; Kinect Zoom In; Kinect Break];
```

For example, once during a receiving time interval of two seconds, the operator speaks the command "Up", perform the dynamic hand gesture of "Left", and put two hands together as the posture order "Break". The input sample right now is:

```
P=[0;0;1;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;1];
```

Theoretically speaking, all of the possible training samples can be listed up while inputting, forming a matrix of 19×245.Therefore, the recognition rate of 100% could be reached by such training method. But it is unrealistic in practice. First of all, there is no need to define some uncommonly used combination. What's more, there might be errors in some of the recognition results. For example, the speech recognition program may mistakenly take "Up" as "Left" in the noisy environment. Though the input of the processing unit is still the instruction code of the command "Left", it should be set as the code of "Up" in the output of the neural network. Thus, the trained neural network is suitable for the operator, and higher reliability could be accomplished. In this study, the number of samples is selected as 25, 50 and 100. Finally, test the recognition rate respectively.

3.2.2. Defining the target output matrix

For the input of each sample, define a 7-dimensional vector as the ideal output. Assign "1" to the corresponding position when such command is about to take place; or assign it as "0". The definition form is similar to the one of training samples.

```
t=[Left; Right; Up; Down; Zoom Out; Zoon In; Break];
```

3.2.3. Establishment and training of neural network

BP neural network (Back-Propagation Network) has good characteristics of predicting the complex nonlinear system. It can effectively describe the complex nonlinear characteristics that are uncertain, multi-input and so on. 80 percent to 90 percent of models of the artificial neural network adopt BP network or its variations. It is also the core part of feedforward network and reflects the most essential part of the artificial network. For the need of integrating many kinds of independent recognition means, the neural network is designed as a two-layer BP network. The elements of the output vector are composed of 0 and 1. Since the output of the function "logsig" ranges from 0 to 1, which is just satisfied with the requirements of the bool value, choose S-type logarithmic function "logsig" as the transfer function of the output layer. The increase of the number of hidden layer could improve the ability of nonlinear mapping of the BP neural network. But if the number exceeds a certain value, the performance of network would decrease instead. However, a single hidden layer BP neural network can approximate any continuous nonlinear function. So the network of single hidden layer is adopted. According to the empirical equation, assign the number of the hidden layer neuron of the network as 11. In accordance with the general design principles, choose S-type tangent function "tansig" as the transfer function of the hidden layer. Choose "trainlm"(Levenberg-Marquardt) as the arithmetic function. The target error is taken as 0.01. The specific realizing statement is as follows.

```
net=newff(minmax(p),[11,7],{'tansig','logsig'},'trainlm');
net.trainParam.goal=0.01;
net=train(net,p,t);
```

4. Experiment Result

When the number of samples is selected as 100, the neural network could be trained after 20 times by testing. The training result is shown in Fig. 5.

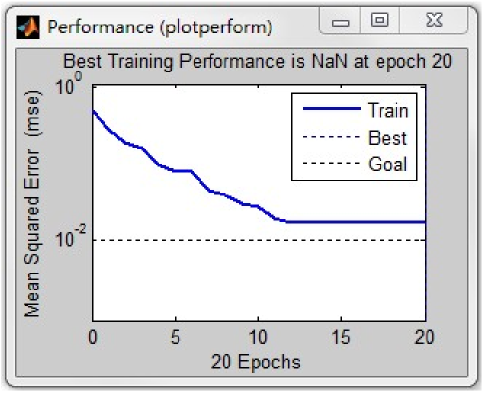


Figure 5. Convergent curve of the neural network training

The test of the training result could be achieved by network simulation. Translate the testing actions into the same matrix form as the training samples, p_test. Employ the statement: hardlim(sim(net,p_test)-0.5), which could output the result of the network simulation.

Test the trained neural networks whose number of samples is 25,50 and 100 respectively. Due to limited space, 10 random simulation results are listed in Table 3.

Table 3. Result of the test (false results have been underlined)

Experimental Action			Output		
Speech	Hand	Kinect	25 Samples	50 Samples	100 Samples
Left	Left	None	Left	Left	Left
Left	None	Break	Break	Break	Break
Up	None	Up	Up	Up	Up
Down	Down	Break	Down	Break	Break
Break	Right	Right	Right	Break	Break
Zoon in	Up	Zoom in	None	None	Zoom in
Left	Right	Up	Up	None	None
Up	Down	Down	Down	Down	Down
Right	Up	Right	Up	Right	Right
Zoom out	Break	Zoom out	Zoom out	Zoom out	break

Based on large numbers of experiments, the conclusion can be drawn that the accuracy of using 25 input samples is less than 50%; the accuracy of using 50 input samples can reach 80%; the accuracy of using 100 samples could be 99%. It can be inferred from the experimental results that the increase of training samples of the neural network is beneficial to improve the recognition rate. Keep increasing the number of the training samples till saturated. The machine will be very familiar with the intentions and habits of pronunciation and actions of the operator. Then the recognition rate of 100% could be accomplished. Even if a certain kind of single recognition tool makes mistakes, the cooperation program will output the true result at last.

5. Conclusion and Future Work

In this article, a research on the cooperation of several human-computer interaction tools is carried out. On the basis of that the speech recognition tool, the hand gesture recognition tool based on data glove, and the posture recognition tool based on Kinect could control the target VR model program, assign the cooperation program receives the recognition result sent from the three independent tools through UDP protocol. Optimize the fusion algorithm by neural network. Higher recognition rate, the effect of convenience and humanization are achieved by training the network. The conclusion can be drawn after the work as follows. Firstly, compared with the interaction of the traditional way that define all the actions into statements, the interaction by training the neural network is more suitable for practical applications. Secondly, the cooperation program of many human-computer interaction tools could achieve higher reliability and convenience. It can be better used in more fields as a supplement of the traditional human-computer interaction. In the future research, the improve of the neural network training will be considered, which could make the system learn the operating habits of different operators better and faster.

References

[1] The Planetary Society. http://www.planetary.org/programs/projects/planetary_microphones/mars_microphone.html, 2012. [Online; accessed 22-May-2012].

[2] K. K. Biswas, Saurav Kumar Basu, “Gesture Recognition using Microsoft Kinect®”, Proceedings of the 5th International Conference on Automation, Robotics and Applications, pp. 100 – 103, 2011.

[3] J.Wu, C.Chan, “Isolated Word Recognition by Neural Network Models with Cross-Correlation Coefficients for Speech Dynamics”, IEEE Transactions on Paitern Analysis and Machine Intelligence, Vol. 15, pp. 1174 - 1185, 1993.

[4] A. Madabhushi, J. K. Aggarwal, “Using head movement to recognize activity”, Proceedings of 15th International Conference on Pattern Recognition, vol. 4, pp. 698 – 701, 2000.

[5] S. Uchino, N. Abe, K.Tanaka, T. Yagi, H.Taki, S.He, “VR Interaction in Real-time between Avatar with Voice and Gesture Recognition System”, 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), 2007.

[6] N. Abe, T. Amano, K. Tanaka, J.Y.Zheng, S. He, H. Taki, “A Training System for Detecting Novice's Erroneous Operation in Repairing Virtual Machines”, ICAT, pp.224 - 229, 1997.

[7] W.D. Deng, Y.F. Dong, “Dynamic Hand Gesture Recognition Based on Data Glove and Application in Human-computer Interaction”, The Sixth China system modeling and Simulation Technology Forum, 2011.

[8] Y. F. Dong, S. M. Chen, “Dynamic Simulation Technology of the Satellite Attitude Controlling”[M], Science Press, 2010.