Motion recognition based on Kinect for human-computer intelligent interaction

Xun Pang, Bin Liang*
School of Water Resources and Electric Power, Qinghai University, Xining, China
*Corresponding author e-mail: hdkgigi@163.com

Abstract. At present, human-computer interaction technology has become the focus and hot spot of many scholars. With the deepening of research, facial expression recognition, speech recognition, gesture recognition, face recognition and human motion recognition have become the important content of current human-computer interaction research. In this paper, Microsoft Kinect somatosensory camera is used as the input device of motion, a new method of fast human motion recognition is proposed, and a set of real-time motion recognition and robot control system is designed.

1. Introduction
Generally speaking, the human-robot interaction needs to be mediated by computer, so in a sense, this interaction is actually the human-computer interaction. Man-machine interaction is a technical subject that studies the mutual understanding between human and computer, carries out communication and communication, and completes information management, intelligent service and multi-information processing functions for people to the maximum extent [1][2].

After the man-machine interaction technology experienced the command line interface and the graphical user interface, the natural user interface embodied the user-centered man-machine interaction concept [3]. Traditional way of human-computer interaction in the mouse, keyboard, touch screen is given priority to, although these are often necessary equipment of a computer system, but for contact learned too little children and old people learning ability is not strong, operating the equipment to realize the interactions with the robot becomes a difficult thing, so a natural and intuitive human-computer interaction means to cause the extensive concern of the scientific community [4-6].

2. Acquisition of depth images and bone data

2.1. Introduction of Kinect depth sensor

2.1.1. Kinect hardware structure
Microsoft got a lot of attention when it released a new Kinect depth sensor in 2010. Originally, the Kinect was designed specifically for the body sense game console Xbox360. People could stand in front of Xbox360 and use voice and body movements to control the game, realizing human-computer interaction. There was no need to operate any controller similar to the handle.

Figure 1 shows the appearance of Kinect. It can be seen from the figure that Kinect has three optical elements. In the middle is the RGB camera, which, like a normal camera, captures space scenes to form color images. Depth of the pair is formed on both sides of the image of the main components, including an infrared emitter and a CMOS infrared receiver and the two devices is the most important
components, called the formation of depth image, players contour and skeleton tracking function will depend on the camera, the depth device also includes a set of microphone array, can identify the players voice, a motor, can adjust the pitch Angle.

Fig. 1 External structure of Kinect

2.1.2. introduction of Kinect software features
The Kinect for Windows SDK contains a large number of basic libraries and development tools. Developers can use multiple languages to develop Kinect, including C++, C#, VB and other languages. Through the SDK development kit, researchers can obtain important data such as color data flow, depth data flow, and voice information and three-dimensional coordinates of skeletal points from sensor hardware. NUI (Natural User Interface) is a kind of invisible User Interface. People can interact with the machine in the most Natural way, for example, using language and body movements to realize human-computer interaction. The NUI API is the core of the Kinect for Windows, and through the NUI API users can access the color, depth, skeleton and voice data provided by the Kinect hardware to control the Kinect.

The Kinect NUI API mainly includes the following functions : (1) provide driver program to identify the Kinect device connected to PC; (2) obtain the color data flow and depth data flow from the Kinect camera; (3) Process the depth data and color data and pass them to PC via USB cable; (4) Through the machine learning processing of the depth data, the human body is separated and the bone is tracked to transmit the bone number According to the flow. (5) obtain audio data from the microphone for speech recognition and sound source localization.

The application communicates with Kinect via the NUI API interface. The information transfer diagram is shown in figure 2.

Fig.2 Data stream of Kinect

2.2. Acquisition of depth data
Kinect depth image acquisition is based on a kind of structured Light technology, named as Light Coding by Prime Sense Company. It illuminates the space scene with Light source and encodes the space scene. Is different from common light source device using the light source is called "laser speckle", called light emitter emits the laser emitter surface has a layer of frosted glass, laser penetrates the frosted glass can produce light diffraction, plus an object's surface is rough, the diffraction of random spots formation of the pattern will be different and constantly changing with
distance. As long as the space is hit structure light, the object of different distance will make the structure light produce different patterns, record the speckle pattern, through the light source calibration system, can determine the position information of the space object.

2.3. Bone tracking
Kinect can see the three-dimensional world from the depth image, and on this basis, the most critical step is in the depth map.

The figure is further separated from the human body contours to form a bone chart from the depth chart. Device on the depth of the image of the "brain" "pixel" assessment of the depth of each pixel of the image scanning, through the image processing technology, computer vision technology for each pixel features identification classification, including edge detection, feature extraction and region segmentation to separate the body contour from the environment, the implementation of active tracking or up to 2 players skeleton of passive tracking for seven players. Skeletal tracking functions are divided into three steps:

1. The human body depth map was isolated from the complex environmental background;
2. Divide the human body into 32 parts through machine learning;
3. Considering the overlapping of human body parts, the machine learning method is further adopted, from the front and side

At the same angle, 32 parts are determined as 20 nodes according to the pixel features, and the space of each node is given.

Three-dimensional coordinates, in meters, are shown in figure 3.

![Fig.3 Schematic of human skeleton joints](image)

On the basis of the depth image formed by Kinect, the internal part of the body is segmented by machine learning theory to form bone data, and the three-dimensional coordinates of each node in the Kinect coordinate system are given. Through the study of the feature extraction of specific key points, you can proceed to the description of the action, forming characteristics, and then select the appropriate classification method for action recognition, gesture recognition of this article is based on a device with 20 key points coordinates, as a result, skeletal tracking in action recognition has played a crucial role in [7].

3. Based on the action feature extraction of the node
In the pattern recognition problem, the key step is to extract the features of the identified objects. The appropriate feature selection directly determines the success rate of recognition. In the previous chapter, we introduced the skeletal tracking function of Kinect, which provides the three-dimensional coordinates of 20 human nodes. Kinect's depth image frame rate is 30fps, again, providing a stream of 30 frames per second of bone data. This chapter focuses on the extraction of appropriate motion features from a continuous stream of bone data [8].

In this paper, four key points of the right and left hands and the left and right elbows were extracted as the key points to describe upper limb movements. Four key points of the left and right feet and the right and left knee joints were selected as the key points to describe lower limb movements.
a frame of static bone data, the 3d coordinates of 8 points with a total of 24 dimensions can uniquely represent the current state of human body, and the continuous bone frame sequence can uniquely describe an action. For example, 9 frames of bone data can be extracted from a right leg swing motion, and the human skeleton diagram can be drawn as shown in figure 4

Figure 4 Skeleton schematic of right leg confidence right

Some of the movements are sensitive to changes in the Angle of the nodes, such as one hand waving; Some movements are sensitive to the distance changes of the nodes, such as walking, distance between hands and distance between feet. In this chapter, a distance feature extraction method is proposed for the Angle feature of the node and the distance feature of the node pair. In the recognition stage, the Angle feature and distance feature are integrated to enhance the system robustness and improve the recognition rate of actions [9-10].

Simple actions can be represented by the Angle sequence of several nodes, while complex actions are usually described by the mutual relations between some pairs of nodes, such as applause, and the distance between two pairs of nodes is constantly changing. The walk, the distance between your hands and the distance between your feet are constantly changing. Usually, the distance between two points is measured by the Euclidean distance. Compared with the angular features, such as walking and clapping, the distance features between the nodes are more representative. For example, clapping and jogging, the change of distance between the nodes of the two hands is shown in figure 5. The two curves are the waveforms of two movements of different people. It can be seen from the figure that the features of the same type of movements are very similar and the features of different types of movements are different.

Fig.5 Wave form of hand clap and keeper

4. Quick action recognition
Behavior recognition is regarded as the key research object in the field of pattern recognition and artificial intelligence, and has important application value in the field of security monitoring, target tracking and control of intelligent devices. For example, in public places, the use of cameras to collect crowd behavior video can be used to track the target of abnormal behaviors and actions in crowd by means of pattern recognition image processing, so as to maintain normal order and safety in public places. On the control of intelligent equipment, through motion recognition, you can use the body
movements to control the robot, intelligent air conditioning, physical sense game machine and other intelligent equipment. Instead of the traditional human-computer interaction, action recognition has a broad prospect of application and development, and some achievements have been made. In this paper, the depth sensor Kinect is used as the motion acquisition device, and the bone model of Kinect is used to realize the recognition of 20 movements, and the recognition results are sent to the robot as instructions, so that the robot can complete corresponding functions according to the instructions and realize human-computer interaction.

Similar to the human nervous system, neural network is composed of many neurons. The neural network has strong learning ability and self-organizing ability through weight connection between neurons, which is suitable for processing large-scale parallel data and has strong robustness, fault tolerance and noise resistance. In the field of action recognition, common neural networks include BP neural network and convolution neural network. BP neural network is a feedforward neural network composed of input layer, intermediate hidden layer and output layer. As shown in figure 6, N is the number of neurons in the input layer of the network, L is the number of neurons in the middle layer of the network, and M is the number of neurons in the output layer of the network. First of all, the training sample data input from the input layer, through the weights of connections between layer and layer, the input samples by different weight through the hidden layer neurons, the final output from the output layer classification training as a result, if the results did not achieve the desired effect, then the actual output and the expected value of difference, namely the error information will feedback to continue training network, expectation until you reach the end of the training.

Like BP network, convolutional neural network is a kind of multilayer feedforward neural network, which includes convolution layer, pooling layer and full connection layer. In the process of action recognition, no matter which kind of network is selected, the feature extraction of the action should be carried out first, the input layer data should be obtained, and the output results can be obtained through network training. It is difficult to achieve the effect of fast action recognition with large training samples and multiple intermediate hidden layers.

5. Interactive system construction and experimental analysis
In this chapter, the skeleton model of Kinect is used to realize the fast recognition of 20 kinds of movements, and the robot is quickly controlled to realize human-computer interaction. This paper introduces the construction of the hardware and software environment of the system of action recognition and the construction of the hardware and software environment of the robot. Finally, the recognition rate, recognition time and robustness of action recognition are verified under different feature extraction methods. A simple experimental demonstration is made for the motion control robot, and the results show that the system has achieved good recognition and control effects.

5.1. System robustness verification
The motion control robot system designed in this paper requires that the system can achieve good recognition effect under different circumstances, including the complexity of the environmental background, the intensity of ambient light, the position change of the tester, the size of the tester, and the successful recognition under the interference of multiple people.
Verification experiment of recognition rate

In order to verify the effectiveness of the motion recognition scheme in this paper, the recognition rate verification experiment was carried out. The experimenter is composed of 10 students of different body types in the research room. Standing within the effective field of vision of Kinect, each movement is performed by 10 students, respectively through two feature extraction methods proposed in chapter 3, and the action recognition experiment is conducted by FDTW algorithm. The recognition rate of each movement and the average recognition rate are shown in table 1.

| The Action name               | Vector special extraction recognition rate | Angle distance feature fusion recognition rate |
|------------------------------|--------------------------------------------|-----------------------------------------------|
| Very different raised        | 97%                                        | 97%                                           |
| applaud                      | 96%                                        | 98%                                           |
| The Left legs Left swing     | 95%                                        | 96%                                           |
| Kick in front of the right leg | 96%                                      | 97%                                           |
| Bend over                    | 96%                                        | 98%                                           |
| Squat                        | 97%                                        | 97%                                           |
| 6 kinds of action average recognition rate | 96.21%                                   | 97.53%                                        |

It can be seen from the recognition rate in the table that the action recognition in this paper has a good recognition effect and the average recognition rate is respectively reached 96.21% and 97.53%, basically meeting the requirements of controlling the robot.

Identification time verification experiment

In order to verify the effectiveness of the lower bound function and truncation technique of the proposed FDTW algorithm to improve the speed of action recognition, a recognition time verification experiment was conducted. The general DTW algorithm and the average recognition time of the FDTW algorithm proposed in this paper were respectively calculated under different action types, as shown in figure 7.

![Fig.7 Recognition time verification experiment](image)

It can be seen from the figure that, when the number of action types is no more than 8, the average recognition time of ordinary DTW algorithm is around 0.4s due to the small amount of template data, which can also meet the interaction requirements. However, as the number of action types increases, for example, 20 is designed in this paper, the sequence of the action to be tested needs to match and calculate the similarity one by one with the 20 sequences in the action template. In this case, the calculation amount is very large, and the average recognition time rises to 1.7s, which seriously affects the recognition speed, causing delay when controlling the robot. The tester’s actions are executed, but they are not immediately recognized, which affects the timeliness of the interaction. The first step of the algorithm is
accelerated by the lower bound function, and the average identification time has a significant downward trend. Then, the second step is accelerated by the intermediate truncation technology, and the recognition efficiency of the average identification time drops to about 0.3s, indicating the effectiveness of the FDTW algorithm proposed in this paper.

6. Conclusion
With the introduction of Microsoft Kinect depth sensor, new opportunities have been brought to motion recognition researchers. With its strong bone tracking technology and advantages of being free from interference of natural light, it has made outstanding contributions to the field of motion recognition. In this paper, 20 kinds of custom action recognition were completed based on Kinect bone tracking technology. For the rapid action recognition system in this paper, in addition to the improvement in the matching algorithm, in the aspect of feature extraction, the dimension of feature data is too large, which also causes the problem of slow calculation speed. The principal component analysis (pca) can be used to reduce the dimensionality of feature data, so that the feature data can be simplified without losing the validity of its description, so as to further enrich the action template and meet the needs of society.

References
[1] Yu Tao Kinect application development in the most natural way to talk to the machine [M]. Beijing: Mechanical Industry Press, 2012
[2] Dong Shihai. The progress of human-computer interaction and The challenges it faces [J]. Journal of computer-aided Design & computer Graphics 2004,16(1):1-13
[3] Wang Songlin. Research on gesture recognition and robot control technology based on Kinect [D]. Beijing: Beijing Jiaotong University, 2014
[4] Meng Ming, Yang Fangbo, Yan Qingshan. Human motion detection based on Kinect depth image information [J]. Journal of Instrumentation, 2015, 36(2): 386-393
[5] Liu Yang, Shang Zhaowei. Traffic police gesture recognition based on Kinect skeleton information [J]. Computer Engineering and Applications, 2015, 51(3): 157-161
[6] Wu Zepeng, Guo Lingling, Zhu Mingchao. The Image registration method combining Image information entropy and feature points [J]. Infrared and Laser Engineering, 2013, 42(10) : 2846-2852.
[7] Abdulla W H, Chow D, Sin G. Cross-words reference template for dtw-based speech recognition systems[C]// TENCON 2003. Conference on Convergent Technologies for the asia-pacific Region.
[8] Keogh E, Ratanamahatana C. Exact indexing of dynamic time warping [J]. Knowledge and Information Systems, 2005, 7(3):358-386.
[9] Wei L, Keogh E, Herle h. Atomic wedgie: Efficient query filtering for streaming times series[C]/ IEEE International Conference on Data Mining. IEEE Computer Society, 2005:490-497.