A Simple Manner of Dynamic Gesture Recognition Based on Kinect

Fenggang Li[1,2,*], Xiangfei Jiang[1] and Xiaobo Xia[3]
[1]School of management, Hefei University of Technology, Hefei 230009, China
[2]Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009, China
[3]Anhui Antai Technology Co. LTD, Hefei 230088, China
[*] Corresponding author

***Abstract*—The paper develops a simple method of dynamic gesture recognition based on the Kinect which is a new sensor from Microsoft in the environment of VS2010 combining Kinect for Windows SDK v1.8. Kinect sensors can track human bodies within their effective scope in real-time and obtain the depth of the corresponding information and bones at the same time. Firstly, we separated each dynamic hand gesture to be a combination of several micro-gestures which were from predefined eight micro-gestures with eight different directions. Then Dynamic Time Warping algorithm was adopted to find the most similar template gesture, then the system realized corresponding control instruction of the matched template gesture, and consequently, we can achieve somatosensory interaction by recognizing dynamic hand gestures, and the total recognition rate isn't less than 90%.**

Keywords-Kinect; Somatosensory interaction; depth information; gesture recognition

I. INTRODUCTION

HCI (Human-Computer Interaction) refers to the process of the communication between people and Computer through the Computer input devices and output devices. Mouse and keyboard are the initial input devices of HCI, and along with the rapid development of the computer hardware and artificial intelligence technology, the concept of natural interaction arises at the historic moment driven by social needs. The ways of natural interaction mainly include multi-point touch, post recognition, gesture recognition and speech recognition and so on. They are more natural intuitive and in line with the human daily communication habits [1].

Gesture recognition has attracted the attention of scholars from all walks of life and become the research hot spot gradually. Initial gesture recognition was achieved based on some worn equipments. For example, the American scholar Lee and Xu utilize CyberGlove gloves to extract gesture signal data, and developed a sign language recognition system to control the robots[2]. By the mid 1990s, with the improvement of computer computing ability, the progress of computer vision processing technology, gesture recognition based on vision research rises gradually. For example, Zhu who is from Tsinghua University presents dynamic isolated gesture recognition technology based on visual whose recognition rate of 120 gestures from the template is no less than 88%[3]. In 2010, Microsoft's Kinect Sensor was published, and the characteristics of its low cost and high efficiency make it preferred in gesture recognition technology research. For example, Raheja, Chaudhary and Singal use the Kinect sensor to realize the track of the fingers and the palm center; Bhattacharya[4], Bhattacharya and Czejdo utilize support vector machine (SVM) and the decision tree algorithm for finishing gesture recognition in aviation flight[5]; Luo, Xie and Zhang apply Kinect sensor for depth information combined with hidden Markov model (HMM) to identify five dynamic gesture trajectories to control the movement of wheelchair[6]. This paper firstly puts forward the concept of "micro-gestures" to break down of dynamic hand gesture trajectories which is from Kinect for depth information and then combines with dynamic time warping (DTW) algorithm to recognize dynamic gestures. This method can reduce the complexity of the algorithm and improve the efficiency of gesture recognition effectively.

II. OVERVIEWS OF THE KINECT

Kinect was initially a Somatosensory peripheral device which was developed for the home video game console (Xbox 360) by Microsoft in 2010, and then the Kinect for windows version was published. The latter had more visual range, and mainly used for application development. Kinect can exactly identify the user's body, and the users can control people or things of the game follow their inclinations through their movements or postures, and which makes users integrate successfully into the game environment. Then, Kinect for Windows SDK (Kinect SDK) was released and attracted a lot of developers in the development of somatosensory interactive applications, which could easily implement real-time human body tracking, gesture recognition, image recognition, speech recognition and other somatosensory interaction functions. Kinect makes people no longer need wear various types of sensors on the body, and the operating costs fell sharply, so the development of human-computer interaction based on Kinect gradually becomes the research hotspot[7].

In July 2014, Microsoft released the second generation of Kinect sensor (called Kinect V2). Compared with the first generation, its performance and price has certain advantages, but its higher requirement for the hardware configuration is still an obstacle for part of the developers. All research in this article is based on Kinect V1, and its appearance is shown in Figure 1. The middle of the lens is a common RGB camera

which is used to obtain the color image information, and its two sides are the infrared transmitter and infrared CMOS camera respectively, and they form a 3D depth sensor, so someone vividly calls them "three eyes". Left and right sides of Kinect sensor is the microphone array which is used for

FIGURE I. THE APPEARANCE OF KINECT

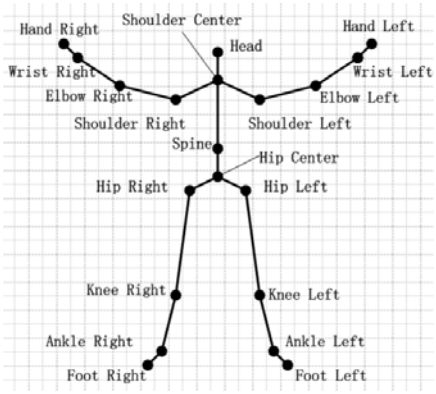

FIGURE II. HUMAN BODY SKELETON MODEL

voice input and recognition. The nether motor can drive it around 27° rotating to obtain the best visual angle of observation.

Kinect's core technology of obtaining the depth image is Light Coding which uses the light source to achieve spatial encoding. The light source refers to the diffraction speckle formed by the laser shines on rough surface or through the frosted glass random, and it will transform random pattern depending on the distance so as to make any the object in the space respectively corresponded to the unique speckle pattern. By this way, we can achieve the effect of the "code" and end up with obtaining the corresponding depth images. Human body recognition is built on the basis of depth image. Kinect assesses depth image in pixel level, and scans depth image pixels point by point. Then it will identify 20 different human body skeleton nodes and build human body skeleton model[8], as shown in Figure 2. Developers can obtain users' real-time location 3D information of any skeleton nodes within the effective scope with Kinect for Windows SDK, then calculate the continuous time motion vectors of each node which is the key feature data of posture recognition and gesture recognition.

III. ACHIEVING DYNAMIC GESTURES RECOGNITION

The main steps of dynamic gesture recognition include hand segmentation, hand tracking, feature extraction and classification. Developers can realize the real-time track of human body each bone node and acquire human body skeleton information which includes 3D coordinate information in the hands at the same time by using Kinect SDK, and by default, users' hands will stay in front of the bodies when they gesticulate. In other words, the hands stay the least distance to Kinect sensors. Therefore, the method of hand segmentation is combined the depth threshold discrimination with human skin detection in this paper. Firstly, the Kinect detects object (hand) nearest to sensor and intercepts the part within the depth threshold, then makes use of human skin detection methods based on the HSV to segment the hands[9]. This paper disassembles the dynamic hand gestures into some predefined micro-gesture sequences, and then encodes each kind of micro-gesture. We can translate dynamic hand gestures into coding sequences by this way, and find the best match for dynamic gesture templates by comparing their coding sequences to achieve the recognition finally.

A. The Confirmation of Start Frame and End Frame

According to three-dimensional coordinate information of human skeleton nodes from Kinect depth image we can calculate the displacement distance (Euclidean distance) of each node between two serial frames. However, human cannot stay static as machines completely, and we hypothesize a stationary state if the distance was less than a certain threshold (generally is 2 cm), and then we think that the coordinate of skeleton node between two serial frames is the same, or else select the current location information directly.

There are some septal areas between the dynamic hand gestures in the human-computer interaction, and it means that there is starting point and end point in each gesture. Liu calculates the Euclidean distance of the hand node between consecutive frames, and if the node is all the time in a stationary state between M (M > = 5) consecutive frames and the M + 1th frame set in motion, then it will set the M + 1th frame as the starting frame; If the node is in a stationary state between 5 serial frames after it has moved N (N > = 10) serial frames, the end frame will be the N + 1th frame[10]. Moreover, we can apply the *GetHandState ()* method provided in Kinect SDK v1.8 to judge whether the operator's hand is in the fisted state. Therefore, in order to reduce restrictions in the recognition process, the paper adds the necessary prerequisite for the determination method of the start frame and end frame on the basis of above method (fisted state), and this is saying that the above method will take effect when the hand is fisted.

B. Micro-gestures

Micro-Gesture is made by the decomposition of each dynamic gesture, and in other words, it is the main components of dynamic gestures. The study defines eight kinds of micro-gestures according to the vector change of hand node between consecutive frames and the improved ideas of 8-direction chain code, and then encodes them with special method as shown in Figure 3.

In Figure 3, we get the improved 8-direction chain code in the (b) from the original in the (a), and the X-Y plane is divided into 8 regions according to the angle (each is 45°) which are one to one correspondence with the direction chain codes in the (a) respectively. The coordinates vector of skeleton hand node between two consecutive recognizing frames —this paper regards every four frames as a recognizing frame, namely starting frame and end frame of each micro-gesture— will fall in a region of the (b) in the process of recognizing a dynamic gesture, and the code of micro-gesture is as same as the corresponding direction chain code of the region. The traditional encoding is defective when calculating the local matching distance. For example, the Euclidean distance of 0 and 1 is different from that of 0 and 7 but they are adjacent in Figure 3, and that is obviously inconsistent with the facts. Therefore, in order to improve the accuracy and the convenience of calculation, the paper takes a new coding of the 2-dimension vector coordinates. As shown in the (c) of Figure 3, in order to avoid negative number in the coding we firstly set the center of the plane to (1, 1) and take the distance of starting frame and end frame of each micro-gesture in all directions to be the unit of length. For example, "0-chain code" indicates a horizontal move to the right in the (a), and it means that it will add 1 to the coordinate of X-direction and that of Y-direction remain the same, so its code is (2, 1). By this way, a complete dynamic gesture will be made up of some micro-gestures whose codes are unique.

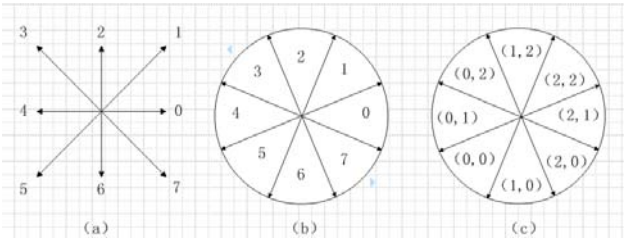

FIGURE III. ENCODING OF MICRO-GESTURES

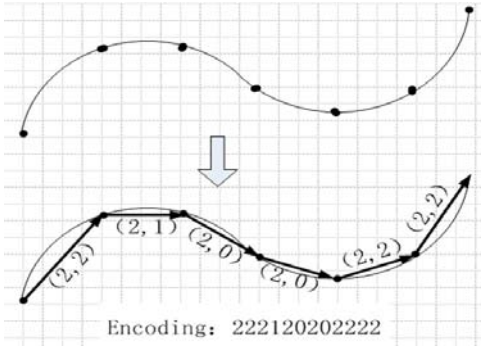

FIGURE IV. ENCODING OF THE DYNAMIC GESTURE

The curve of a gesture trajectory is shown in Figure 4. We set every four frames to be a recognizing frame, a total of seven recognizing frames, namely six micro-gestures, and finally acquire the encoding of the dynamic gesture. Because the total number of frames from the starting to the end frame cannot be always divisible by 4 when we confirm the recognizing frame, we will take the last frame to be the end frame of the last micro-gesture if the remainder is 1, and it means that the last micro-gesture will be made up of 5 frames. Otherwise, the remainder frames will be set as the last micro-gesture. For example, the last micro-gesture shown in Figure 4 is made up of only three frames.

C. Custom Templates of Dynamic Gestures

We define six simple dynamic gestures in the experiment, and select five users to participate in the entry of gestures template, and everyone makes each gesture for 5 times according to the action descriptions, then we get a total of 150 template sequences. There are six trajectories of different gestures as shown in Figure 5, and Table 1 shows the functions and the brief descriptions of the corresponding gestures.

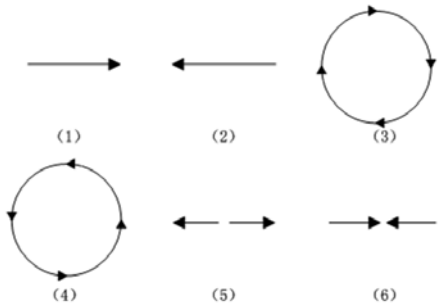

FIGURE V. TRAJECTORIES OF TEMPLATES

TABLE I. CUSTOM TEMPLATES OF DYNAMIC GESTURES

NO	Function	Descriptions
1	Next	Right hand moves to the right horizontally
2	Previous	Right hand moves to the left horizontally
3	Clockwise rotation	Right hand rotates for a circle clockwise
4	Anticlockwise rotation	Right hand rotates for a circle anticlockwise
5	Zoom in	Both hands translate inside-out
6	Zoom out	Both hands translate outside-in

D. Matching Dynamic Gestures

The gesture recognition based on HMM (Hidden Markov Model) and that based on DTW (Dynamic Time Warping) are by far two popular methods of dynamic gesture recognition, and they both have applied in the field of speech recognition widely. HMM is a statistical model in which the system is assumed to be a Markov process with two states which are the hidden state and the observable state. The basic idea is to calculate probability of from different observable states to a certain hidden state by utilizing some complex formulas, and there are many characteristics, such as time scale invariance and automatic segmentation. DTW is based on the idea of dynamic programming (DP), and it can find out an optimal path from the starting frame to the end frame in the gesture sequences, and the lengths of two sequences to be matched don't have to be the same. There is not much difference between HMM and DTW when the trajectories are relatively simple, but the DTW doesn't need us to provide lots of training

samples, and using it can effectively solve the problem of matching dynamic gestures with different length of time. Therefore, this paper adopts DTW to match dynamic gestures and templates.

The basic idea of DTW is to find out a nonlinear time distortion function between two sequences with different length of time, and then to adjust the length of the gesture sequence to be consistent with that of the template sequence by compression or extension as far as possible. We set T = {T (1), T (2), T (3),...,T (M)} as template sequence and R = {R (1), R (2), R (3),..., R (N)} as a gesture sequence to be matched, and M or N is the sequence number of the recognizing frame, and the component of R or T can be a number or a smaller component, but the dimension of each component must be equal so that we can directly calculate the similarity (Euclidean Distance) between them. According to the particularity of encoding of micro-gestures in this experiment, we set the dimension of each component as 2. As shown in figure 5, the encoding of the dynamic gesture sequence to be matched is T = {(2, 2), (2, 1), (2, 0), (2, 0), (2, 2), (2, 2)}.Therefore, if the partial distance between each micro-gesture T(i) and R(j) is set as d (i, j), we can use the two-dimensional Euclidean distance formula to calculate it. Hereinto, x_1 and y_1 is the coordinate of X-direction and Y-direction of the micro-gesture T(i), x_2 and y_2 is that of the micro-gesture R(j):

$$d(i, j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (1)$$

In order to find out the optimal sequence between T and R we need to construct a two-dimensional matrix W of M * N, and matrix element W[i, j] is the DTW distance between the subsequence {T (1), T (2), T (3),...T (i)} of T and the subsequence {R (1), R (2), R (3),..., R (j)} of R, and we set the distance as D(i, j). According to Bellman principle, we needn't exhaust all possible matching paths from left to right and, from top to bottom when calculating the optimal sequence D_{min}. In short, Bellman principle indicates that the optimal path from node A through node B and then to node C is the series connection of the optimal path from node A to node B and that from node B to node C, and it effectively reduces the complexity of calculation. Finally, we calculate D_{min} (M, N) is the DTW distance between R and T as follows a formula. Hereinto, 1<i<=M and 1<j<=N:

$$D_{min}(i, j) = d(i, j) + \min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\}. \quad (2)$$

E. The Process of Dynamic Gesture Recognition

In the whole process of gesture interaction, if the system finds that the user is in a state of fist and has correctly identified the starting frame, it will capture gesture trajectories

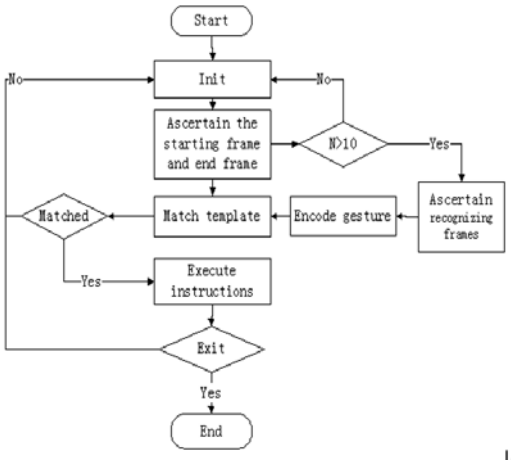

FIGURE VI.THE FLOW DIAGRAM OF GESTURES RECOGNITION

In the depth image until it has identified the end frame. When all recognizing frames are determined we can get the micro-gestures by linking the recognizing frames end to end. After that, we will acquire the gesture sequence after encoding each micro-gesture. Then, calculate the DTW distance to the template sequences, and the template sequence with minimum distance will be matched unless the minimum distance is greater than the preset threshold. Finally, the system achieves the corresponding action of matched template sequence. N is the total number of frames of the sequence in Figure 6.

IV. RESULTS

In order to further understand the Kinect's ability to capture depth information and to detect recognition efficiency of various gestures more accurately, the experimental results of Han show that the recognition scope of Kinect is not influenced by the height of the place (in the vertical range of 0.8 m to 1.2 m), and it can quickly capture the subtle variations of depth data in the horizontal range of 1.0 m to 4.0 m in front of Kinect[7]. Therefore, this experiment selects five experimenters which are never known about Kinect sensor to complete six different gestures according to relevant text descriptions on the basis of the above conclusions, and everyone experiments 10 times for each gesture. Finally, we calculate the recognition rate of each gesture. Some experimental screenshots are shown in Figure 7 and the results are shown in Table 2. The experimental results in Table 2 show that the recognition rate is more than 90%, which can

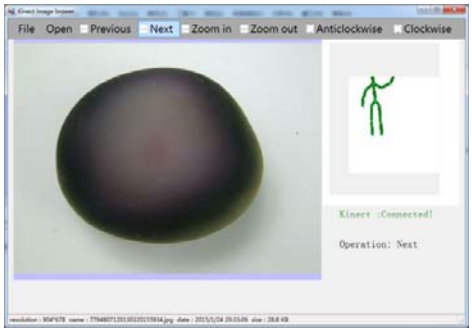

FIGURE VII. THE EXPERIMENTAL SCREENSHOTS

TABLE II. THE RECOGNITION RATE OF DYNAMIC GESTURES

Gestures	Times of test	Times of recognized	Recognition rate
Next	50	45	90%
Previous	50	47	94%
Clockwise rotation	50	48	96%
Anticlockwise rotation	50	47	94%
Zoom in	50	43	86%
Zoom out	50	44	88%
Total	300	274	91.33%

Satisfy the basic requirements. At the same time, because this method doesn't have to operate every frame of the sequence, and it greatly reduces the number of iterations of the DTW algorithm and improves the recognition efficiency.

V. CONCLUSIONS

As the link between the virtual world and the real world, motion-sensing technology connects them together directly, and the emergence of Kinect sensor directly shortens their distance and promotes the development of the somatosensory interactive technology. The paper makes a study of dynamic gesture recognition utilizing Kinect sensor with low cost and high efficiency, and it can easily obtain experimenters' depth image and color image within the scope and calculate the coordinate information of human skeleton nodes which is the input feature value of gesture recognition. This paper achieves segmentation of dynamic gesture by the way of defining the micro-gesture, which not only ensures ideal recognition accuracy, but also reduces the time complexity of DTW algorithm. However, there are also some insufficiencies in the study. For example, a large amount of noise and the slight shaking of the hands or sensor may interfere with the performance of the system in the process of dynamic gesture recognition, and it's a main issue for future work.

ACKNOWLEDGMENT

This work is supported by the merit aid project of the returned overseas talents of Ministry of Human Resources and Social Security, Natural Science Young fund of China (No.71301041), and I would like to thank all the people for helping finish the paper.

REFERENCES

[1] Tao Yu, Actual combat of the application development of Kinect: Dialogue with the machine in the most natural way, China Machine Press, 2013.

[2] C. Lee, Y. Xu, Online, interactive learning of gestures for human/robot interfaces, Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on, vol 4, IEEE, 1996, pp. 2982-2987.

[3] Yuanxin Zhu, Guangtuo Xu and Yu Huang, Appearance-based dynamic hand gesture recognition from image sequences with complex background, Journal of Software 11 (2000) 54-61.

[4] J.L. Raheja, A. Chaudhary and K. Singal, Tracking of fingertips and centers of palm using kinect, Computational intelligence, modelling and simulation (CIMSiM), 2011 third international conference on, IEEE, 2011, pp. 248-252.

[5] S. Bhattacharya, B. Czejdo and N. Perez, Gesture classification with machine learning using kinect sensor data, Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on, IEEE, 2012, pp. 348-351.

[6] Yuan Luo, Yu Xie and Yi Zhang, Design and implementation of a gesture-driven system for intelligent wheelchairs based on the Kinect sensor, Robot 34 (2012), 110-113.

[7] Xu Han, The human behavior recognition research and system design using Kinect, Ph.D. Dissertation, Shandong University, 2013.

[8] Guobin Wu, Bin Li and Jizhou Yan, The development practices of human-computer interaction using Kinect, Posts & Telecom Press, 2013.

[9] Wei Tian and Zhenquan Zhuang, Self-adaptive skin color detection based on HSV color space, Computer Engineering and Applications 40 (2004), 81-85.

[10] Yang Liu, Gesture recognition technology research based on kinect, Chongqing University, 2014.