A hand gesture interaction system based on Kinect

Weiqing Li*, Zehui Lu, Shihong Shen

School of Computer Science & Engineering, Nanjing University of Science & Technology, Nanjing, China, 210094

Received 1 October 2014, www.cmnt.lv

Abstract

We introduce a hand gesture interaction system using Kinect, which takes advantage of real-time dynamic motion capture, image recognition and so on, so that people can interact with computer by natural hand gestures. Five kinds of gesture are defined and can be recognized by the system. Kinect-based hand movements and gesture recognition algorithm is studied. A method for hand area image segmentation from the depth map is proposed, using the information of hand joints in skeleton map. We realized a hand gesture recognition algorithm with SVM and tested its stability and robustness. Finally, experimental results verified the feasibility of the algorithm, and a hand gesture interactive demonstration is implemented.

Keywords: hand gesture recognition, Kinect, SVM, interaction

1 Introduction

With the development of computer vision technology, interactive computer-based vision systems have gradually developed. The traditional interaction methods such as mouse and keyboard have been unable to play a facilitating role, in some specific areas. The interaction method based on computer vision will be a very good way. The hand gesture recognition based on computer vision is a hot spot of human-computer interaction research [1]. The hand gesture interaction is rich in expression and contains a lot of information, for example we can get lots of information through different gestures, different locations, and different directions. It is a much natural interaction method.

The Kinect sensor is a horizontal bar connected to a small base with a motorized pivot, and is designed to be positional lengthwise above or below the video display. The device features an RGB camera, depth sensor and multi-array microphone running proprietary software, which provide full-body 3D motion capture, image recognition and voice recognition capabilities [2]. It was developed at Microsoft Research Cambridge in collaboration with Xbox. Kinect gives completely hands-free control of electronic devices possible by using an infrared projector and camera and a special microchip to track the movement of objects and individuals in three dimensions.

In this paper, we propose a hand gesture interaction system based on Kinect. We define some hand gestures that can be used to interact with the system. Then we study the hand movement recognition algorithm, and set up a method of hand area image segmentation from the depth map background by using the information of hand joints in skeleton map. Also, we realize a hand gesture recognition algorithm with SVM. Finally some experimental results verify the validity of this algorithm.

2 Hand gesture definition

The basic idea of Kinect is to segment a single depth image into a dense probabilistic body part, labelling as a per-pixel classification task and then estimate this signal to give high-quality proposals for the 3D locations of body joints. And Kinect's depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions. The sensing range of the depth sensor is adjustable, and the Kinect software is capable of automatically calibrating the sensor based on the player's physical environment, accommodating for the presence of furniture or other obstacles [3]. In order to recognize hand gestures and interact with the system, we define 5 gestures, show in Table 1.

TABLE 1 Hand gesture definition

Gesture	Meaning	Gesture	Meaning
Keeping single finger straight for 1 second	Left click	Spreading out the fingers and keeping moving at the same time	Mouse move
Keeping fist for 1 second	Right click	Spreading out the arms to the two sides	Zoom in
Holding the fingers straight and closing them	Click down	Making hands close to the chest	Zoom out
Spreading out the fingers on one hand for 1 second	Click release	Waving the right hand to right quickly	Forward
Putting hands over the head	Full screen	Waving the right hand to left quickly	Backspace

* ***Corresponding author*** e-mail: li_weiqing@139.com

3 Hand gesture recognition

In this section we will introduce two algorithms: hand movements tracking algorithm and hand gesture recognition algorithm with the depth map, let's discuss them in detail.

The hand movements refer to the gestures of waving our hands to right or left, up and down. We use these gestures to control the system such as zoom in, zoom out, full screen, forward, backspace and so on.

The hand gesture means the finger postures, such as making a fist, spreading out fingers and so on [4,8]. Kinect can't recognize the shape of fingers and palm at present, so we propose a hand gesture recognition method using SVM. The process is shown in Figure 1.

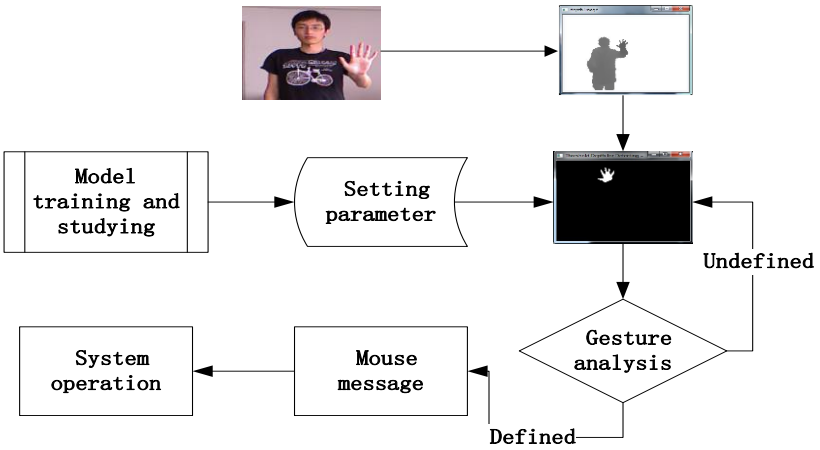


FIGURE 1 Hand gesture recognition process

3.1 HAND MOVEMENTS TRACKING

3.1.1 Joints data pre-processing based on the continuous space-time tracking

Kinect can get three kinds of images, they are color map, depth map and skeleton map. The skeleton map is consisted of twenty joints including our hand joints. The camera's coordinates are (0,0,0), the right direction is the positive direction of the x axis, the upward direction is the positive direction of the y axis and the positive direction of the z axis is facing to the operator. So every frame Kinect captures all the joints' three-dimensional coordinates.

Based on continuous space-time tracking of hand movements' characteristic [5], the movement tracking method can be separated into two parts. The first part is called preprocessing, as Kinect is capable of simultaneously tracking up to six people, but in fact the system can be operated just by one person. So the problem is whose hand gesture should be recognized when there are many people stand before Kinect. We use the bubble sorting algorithm for the head's coordinate on the z axis (if the people is not exist, we will set a maximum value for the coordinate). Kinect will consider the person who has the minimum coordinate as the operator and recognize his gesture.

The second part is called hand movement tracking. We set a sliding window and store 20 frames continuous skeleton image, then the three-dimensional coordinate data extraction and analysis. The sliding window slides backwards and inserts the next frame, the first frame will be deleted. Let's discuss the tracking method in next chapter.

3.1.2 Hand movements' characteristic definition

Hand gesture definition means the gesture's specific function, such as waving right hand to right means going to the next page, waving right hand to left means going back, waving left hand to left and right hand to right at the same time means zoom in, etc. If the left hand thrusts forward and then moving right hand, it means we have clicked the mouse and controlling the movement of the mouse now. These are defined as follows:

1. The definition of click. We get a series of values on the z axis, such as $z_1, z_2, ..z_n$, in a period of time (such as 3 seconds). If $z_1 \succ z_2 \succ ..z_n$ and $\delta_1 \succ z_1 - z_n \succ \delta_2$, then the system will consider the left hand has done the click action. The thresholds δ_1 and δ_2 can be set different values according to different conditions.

2. The definition of forward. We can get a series of right hand's two-dimensional coordinate values such as $x_1, x_2, ..x_n$ and $y_1, y_2 ..y_n$ in a period of time (such as 3 seconds). If $x_1 \prec x_2 \prec x_3 \prec ... \prec x_n$, $x_n - x_1 \succ \delta$ and $\max(y_i) - \min(y_j) < \varepsilon$, then the system will consider the operator has done the forward gesture. The thresholds δ and ε can be set different values according to different conditions. $\max(y_i)$ and $\max(y_j)$ represent the maximum and minimum value on the y axis when we wave our hands.

3. The definition of zoom in. We can get a series of values such as $R(x_1), R(x_2),...,R(x_n)$, $R(y_1), R(y_2),...,R(y_n)$, $L(x_1), L(x_2),..., L(x_n)$ and $L(y_1), L(y_2),..., L(y_n)$. If $L(y_1), L(y_2),...,L(y_n)$, $R(x_1) \prec R(x_2) \prec ... \prec R(x_n)$ and these values satisfy with the formulas as follows:

$$\left|R(x_n) - R(x_1)\right| \succ \delta, \tag{1}$$

$$\left|\max(R(y_i)) - \min(R(y_j))\right| \prec \varepsilon, \tag{2}$$

$$\left|L(x_n) - L(x_1)\right| \succ \delta, \tag{3}$$

$$\left|\max(L(y_i)) - \min(L(y_j))\right| \prec \varepsilon. \tag{4}$$

Then the gesture will be considered as zoom in. $R(x)$ and $R(y)$ are right hand's coordinates (x, y) . $L(x)$ and $L(y)$ are left hand's coordinates (x, y) . We can set different values for and according to different conditions. $\max(R(y_i))$ and $\min(R(y_j))$ represent the maximum and minimum value on the y axis when we wave right hand.

3.2 HAND GESTURE RECOGNITION BASED ON DEPTH MAP

3.2.1 Hand area image segmentation assisted by skeleton map

Except the skeleton map, Kinect can also get colour map and depth map. We can recognize the hand gestures with these two kinds of maps. The colour map has the advantage of clearness, but it is susceptible to interference and only contains the information in two dimensions. On the other hand, the depth map's resolution is lower than colour map, but contains information of 3D. We propose a method of hand seg-mentation from the depth map background using the information of hand joints in skeleton map.

As Kinect can track our hands' coordinates with the help of skeleton map, so we can easily get the precise location of our hands in the skeleton map. The hand's location can be mapped to the depth map with the help of the pixel ratio between the skeleton map and the depth map. As the depth map contains the three-dimensional information, so we can realize the segmentation from the depth map background according to the depth information [6].

For the image segmentation, threshold has a big influence. The definition of depth threshold is:

$$\psi = \min(z) + \delta, \tag{5}$$

$\min(z)$ represents a point's coordinate on the z-axis which is the nearest point to the camera of our body. δ is a experimental value, the range of the value is $\min(z)$ to ψ . This is the most precise value of the palm's location. In our experiment, we set 5cm for δ . In this case, we get the hand segmentation image as Figure 2.

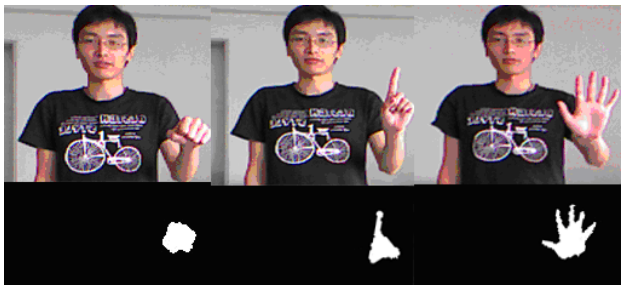


FIGURE 2 Hand area segmentation

3.2.2 Gesture recognition algorithm using SVM

SVM is a classifier, it is firstly proposed by Cortes and Vapnik in 1995. Compared with the traditional method, SVM doesn't depend on the selection of the model. It shows many special advantages in resolving the small sample, nonlinear pattern recognition [7]. The SVM algorithm references kernel function, and then the algorithm converts the practical problem to a high dimensional feature space through the nonlinear transformation. It creates the linear discriminate function in high dimensional space to realize the nonlinear discriminate function in the original space, its particular characteristics can certify the sys-tem has good generalization ability. In addition, SVM solved the dimension problem skillfully. Its algorithm complexity has nothing to do with the dimension of the sample, so the SVM classifier can adapt to the classification problem of different dimensionality. In this section, we use the SVM classifier to recognize the hand gestures according to the limitation of the sample and the accuracy of the recognition.

Firstly, in order to get the contour of the hand, we denoise and smooth the hand image which segmented from the depth map, and remove some isolated points. Then we will calculate some characters of the gesture, such as Circularity:

$$Circularity = \frac{L^2}{4\pi \times AL}. \tag{6}$$

This characteristic describes the closeness between the gesture and circle, the closer the value to one, the more like the circle. L represents the perimeter of the contour of the hand. A represents the area of the contour of the hand. The distribution graph is about the gestures' circularity. We can find that the circularity of the fist most close to one and the circularity of the palm is the biggest. Because the palm's opening in varying degrees so the value of L will be different and then lead to the fluctuant.

Filling Ratio:

$$FillingRatio = \frac{A}{A - R}. \tag{7}$$

This characteristic describes the area ratio of the hand occupies in the rectangle outside. The bigger of the value, the more gathered of the gesture. *A* - *R* represents the area of the smallest rectangle outside. The distribution graph is about the gestures' filling proportion. We can find from the graph that the fist's value is the biggest, and it means the fist has the highest polymerism.

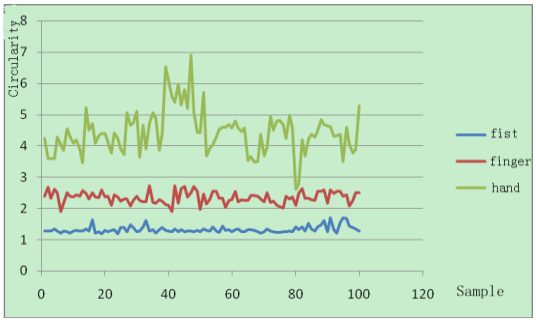


FIGURE 3 Circularity of gestures

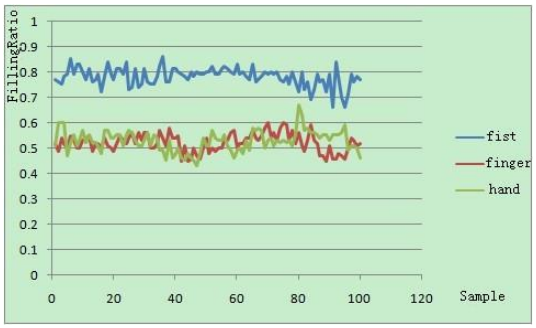


FIGURE 4 Filling ratio of gestures

Perimeter Ratio:

$$PerimeterRatio = \frac{L}{L-C}. \tag{8}$$

This characteristic describes the ratio between the perimeter of the contour of the hand and the mini-mum circumference of the circle outside, the bigger of the value, the more opening of the gesture. *L_C* represents the circumference of the smallest circle outside. The distribution graph shows three gesture's perimeter ratio. We can find that the palm's perimeter ratio is much bigger than the other two gestures.

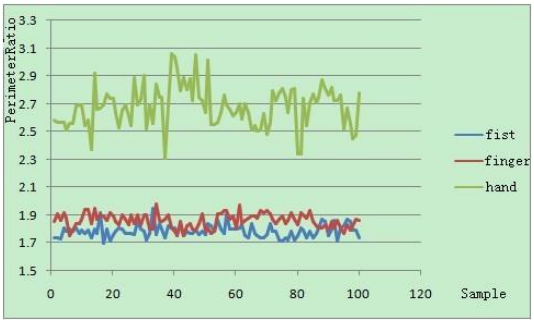


FIGURE 5 Perimeter ratio of gestures

4 Experiment

4.1 HAND MOVEMENTS RECOGNITION EXPERIMENT

Moving the right hand slowly means the movement of mouse, and moving the right hand to right or left quickly means going forward or going back. So it is necessary to set a threshold, if we want to distinguish the forward gesture from the back gesture. Because when we wave our arms, its orbit is an arc, so the co-ordinate on the z axis keeps changing all the time. In order to prevent from confusing with the click behaviour, we should set a threshold on the z axis. The specific value can be changed with different conditions, the method as follows:

As to the click gesture, δ_1 is set to 0.2 and δ_2 is set to 0.1, it means the left hand's range of movement should between 10cm and 20cm in a period of time, then the gesture will be recognized as click.

As to the forward gesture, δ is set to 0.3 and ε is set to 0.1, it means the right hand's range of movement should greater than 30cm and the amplitude should less than 10cm in a period of time, then the gesture will be recognized as going forward.

The gesture of zoom in has the same thresholds with the forward gesture, as the gesture of zoom in includes the forward gesture, so we introduce the priority in order to distinguish these two kinds of gestures. The gesture of zoom in has a higher priority than the forward gesture and the gesture of zoom out has a higher priority than the back gesture.

We perform 100 click operations and 100 forward operations with different thresholds, the recognition rates are showed in Table 2 and Table 3. When we perform the forward gesture, for the ε, a higher value is better, in order to have a better understanding of the gesture, ε should not get a too large value, 0.1 is a right value.

TABLE 2 Threshold of click operation

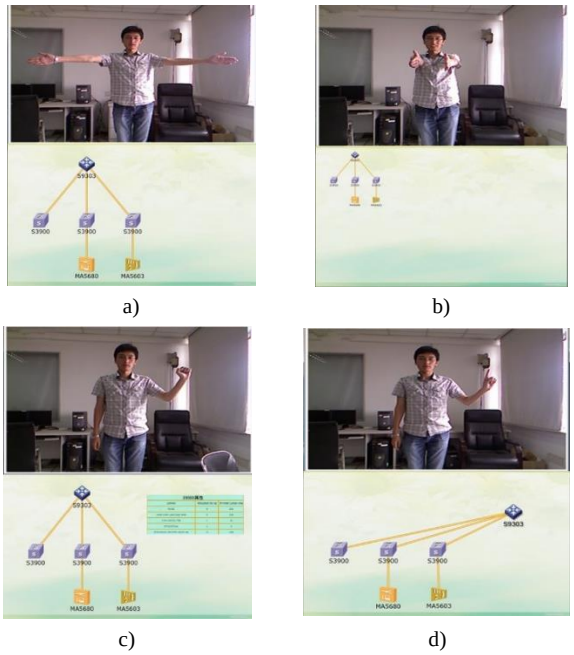
Threshold	$\delta_1 = 0, \delta_2 = 0.1$	$\delta_1 = 0.1, \delta_2 = 0.2$	$\delta_1 = 0.2, \delta_2 = 0.3$	$\delta_1 = 0.3, \delta_2 = 0.4$	$\delta_1 = 0.4, \delta_2 = +\infty$
Recognition rate	0.73	0.91	0.77	0.53	0.12

TABLE 3 Threshold of forward operation

Threshold	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
Recognition rate	0.45	0.68	0.93	0.76

4.2 HUMAN-MACHINE INTERACTION DEMO

Figure 6a shows the gesture before zoom out, Figure 6b shows the gesture after zoom out and Figure 6c shows the gesture of catching and moving. Figure 6d shows the gesture of displaying some parameters.



a) b) c) d)

FIGURE 6 Human-machine interaction demo

5 Conclusions

In this paper, we introduce a hand gesture interaction system based on Kinect. We get the hand gesture image by Kinect, then the system will analyse and recognize both the movement and the gesture. In order to recognize the gesture precisely, we design an algorithm based on SVM. Also, we map gestures to mouse events to interaction with the computer. From the experimental results, we notice that the algorithm has well reliability and strong robustness.

On the other hand, this gesture interaction system has some work to go on, such as we need more gesture definitions and improve the system's sensitivity. These will be considered as future work.

References

[1] Chen F S, Fu C M, Huang C L 2003 Hand gesture recognition using a real-time tracking method and hidden Markov models *Image and Vision Computing* **21**(8) 745-58

[2] Lee S B, Ho Y S 2011 Real-time Stereo View Generation using Kinect Depth Camera *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* 2011 1153-6

[3] Laptev I 2005 On space-time interest points *International Journal of Computer Vision* **64**(2-3) 107-23

[4] Chen Y M, Zhang Y H 2009 Research on human-robot interaction technique based on hand gesture recognition *Robot* **31**(4) 351-6

[5] Lee M, Green R, Billinghurst M 2008 3D Natural Hand Interaction for AR Applications *23rd International Conference Image and Vision Computing* Christchurch New Zealand 6-12

[6] Fujimura K, Liu X 2004 Hand Gesture Recognition using Depth Data *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition* 529-34

[7] Agarwal A, Triggs B 2006 Recovering 3D human pose from monocular images *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(1) 44-58

[8] Zhu Y X, Ren H B, Xu G Y, Lin X Y 2000 Toward real-time human-computer interaction with continuous dynamic hand gestures *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition* 544-51

Authors

Weiqing Li, born in December, 1974, Luoyang, Henan, China

Current position, grades: associate professor at the School of Computer Science and Engineering, Nanjing University of Science and Technologies.
University studies: doctor's degree in Computer Application at Nanjing University of Science and Technologies in 2007.
Scientific interests: virtual reality, human machine interface, computer graphics.
Publications: 20 papers.

Zehui Lu, born in November, 1991, Xuzhou, Jiangsu, China

Current position, grades: academic master candidate in pattern recognition and intelligent system at Nanjing University of Science and Technologies.
University studies: bachelor's degree in Software Engineering at Changchun University of Science and Technologies in 2013.
Scientific interests: virtual reality and simulation systems.

Shihong Shen, born in June, 1987, Cixi, Zhejiang, China

Current position, grades: academic master candidate in computer simulations at Nanjing University of Science and Technologies.
University studies: Master's degree in Computer Science in Nanjing University of Science and Technologies in 2013.
Scientific interests: virtual reality and simulation system.