Human-computer interaction system for hand recognition based on depth information from Kinect sensor

Yangtai Shen[1, a] ,Qing Ye[2,b*]

[1]School of Electronic and Information Engineering, North China University of Technology, China

[a]youta_syt@yahoo.co.jp, [b]yeqing@ncut.edu.cn

Keywords: Human-computer interaction; Depth information; Skin color detection; Kinect sensor

Abstract: Over the past few years, people demands human-computer interaction system with more intelligent function, such as gesture recognition and motion recognition. This paper acquires color and depth information from Kinect sensor in order to capture hand motion and interact with the computer game. By using skin color detection, binarization and frame difference on color information, hand 2D information can be extracted and non-moving part will be filtered; Through mapping the hand 2D information onto the depth information, we can extract the depth information of hand; By means of them, 3D coordinates of hand can be calculated; According to the change of coordinates of several continued frames, we can extract the motion characteristic of hand and determine whether or not to meet game's requirement which can achieve human-computer interaction. Experiment results shows that this system can recognize hand motion and interact with game effectively. This system has wide application prospect which can not only utilize in entertainment, but also has wide application value in security and medical treatment.

Introduction

With the progress of science, human-computer interaction has gotten great development on both software and hardware. More intelligent human-computer interaction system tends to an important research field. Previous human-computer interaction has to convert human behavior to machine instruction through keyboard input or so on. Nowadays, people wants to skip this step, which means systems can identify human behavior directly through the capture of human action [1-2], voice [3-4] and expression [5-6] information to achieve human purpose. Smart human-computer interaction systems are able to eliminate physical constraints and provide users more convenient and comfortable experience.

This paper utilizes Kinect sensor in order to achieve somatosensory game. The game content is shooting game which ask player use hand to "hit" targets on the screen. If the change of coordinates meets the range of targets, then system will judge that hit is successful. Kinect sensor attracts domestic and foreign researchers' attention because of less interference under illumination change and complex background situation [7-8]. This paper mainly uses color and depth information to capture hand motion characteristic. Color information are used to locate and trace moving hand; depth information are used to extract motion characteristic on depth. Extracted motion characteristic will be compared with target's judgment range which will make game come to different reaction.

Kinect Sensor

Kinect sensor is developed by Microsoft which starts on sale in October 2010 as Xbox360 host's peripheral equipment and released “Kinect for Windows” in February 2012. In the hardware, Kinect have three camera lens: Maximum support 1280*960 resolution RGB color camera; Infrared emitter; Maximum support 640*480 resolution infrared camera. The infrared emitter and infrared camera constitute depth sensor to capture depth information. Also, Kinect have array microphone which extract voice from four microphones at the same time for speech recognition and sound source localization.

Methodology

This system needs following four modules. Firstly, video image pre-processing is used, which initializes the system and Kinect sensor. Also Gauss filtering and morphological filtering are applied to both RGB and depth image in order to filter noise and useless information. Secondly, we need to recognize hand. To capture moving hand, skin color recognition and binarization have been utilized to capture human body on RGB image and using frame difference to continued frames can extract the moving part of human body which is hand in this paper. Then, hand motion characteristic will be extracted based on depth information through contour recognition and barycenter locating. Last, with both barycenter and depth of hand are extracted, hand 3D coordinate can be calculated. The change of coordinate will be calculated to interact with player by determining whether player hits the target. With all modules above, the realization of somatosensory game can be achieved. System block diagram is shown as Fig.1.

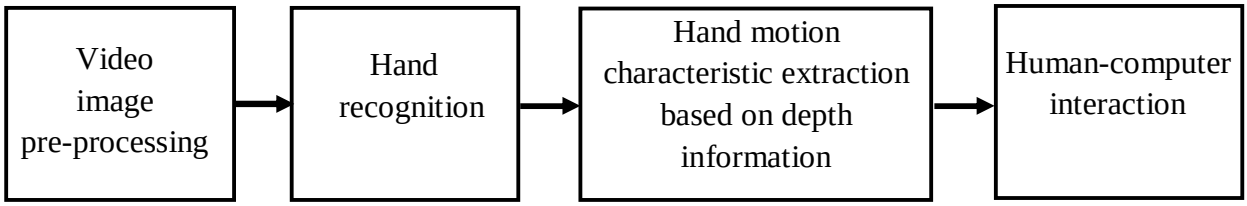

Fig.1 Human-computer interaction system for hand recognition based on depth information from Kinect sensor block diagram

Video image pre-processing

This module includes system and Kinect sensor initialization and denoising which will obtain RGB and depth image. This paper uses Gauss filtering and morphological filtering as main denoising method.

Gauss filter is achieved by calculating weighted average of each pixel and its adjacent pixels, then replace this pixel value with result, which plays the main role of smoothing the overall image. In image processing, Gauss filter has two methods mainly. One of it is convolution in time domain, the other one is go through low-pass filter in frequency domain. Because of the large calculated amount, this paper chooses processing image in frequency domain.

Morphological processing includes erode operation and dilate operation. On the one hand, choosing erode operation first and then dilate the image can remove burr and discrete points which is called open operation; On the other hand, choosing dilate operation first and erode the image will suture the break point effectively which is called close operation. In this system’s scenario, useless noise and burr need to removal, so we choose the open operation.

Hand recognition

In this part, skin color recognition and binarization are utilized to RGB image at first and then using frame difference extract the moving hand.

Skin color recognition and binarization

Skin color recognition finds human skin through transform image from RGB format to Ycbcr format [9]. As Fig.2 shows, there is the skin color characteristic in cbcr domain. If the capture color conform Fig.2, then systems will recognize this pixel as human skin. Transformation of format can bring the advantage of removing brightness which makes the recognition more accurate and reliable. After recognizing the skin color, we turn RGB image into binary image for next processing.

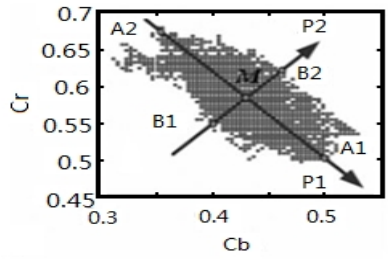

Fig2. Skin color characteristic in cbcr domain [10]

Frame difference

Because of the demand of recognition to moving hand, background and non-moving object will be considered as useless information and need to remove them by utilizing frame difference. Frame difference is making difference with current image and previous image. This principle shows as follows:

If $|I(t)-I(t-1)|<T$, then judge as background (1)

If $|I(t)-I(t-1)|>=T$, then judge as foreground (2)

In this principle $I(t)$ and $I(t-1)$ represent corresponding pixel value at t and $t-1$ moment and T represents threshold value. Due to the short time interval between two adjacent frames, frame difference has better real-time performance which doesn't accumulate background and has advantage of rapid update speed and simple algorithm.

Hand motion characteristic extraction based on depth information

Contour recognition and barycenter locating are main processing in this module. Extraction of depth information is based on the average of depth value in the minimum inscribed rectangle of recognized contour.

Contour recognition

By inputting a binary Image, if a pixel and its adjacent pixels values are *255* then this pixel can be replaced by *0*. When all pixels in the binary image are processed, only the pixels at the edge of contour will not be replaced. As a result, hand contour has been extracted.

Barycenter locating

In the following recognition, system demands hand coordinate in order to recognize hand moving characteristic by calculating the change of coordinate. First of all, 2D coordinate(x, y) need to be extracted. Hand barycenter locating can be extracted according to the information from the contour

recognition. Formula is shown as follow:

$$M_{pq} = \int\int (x^p) * (y^q) f(x, y) dxdy \tag{3}$$

In formula (3), x and y mean pixel coordinate in image, p and q can take as 1, 2, 3 ⊥ ⊥ ∞. If x_c and y_c represent the coordinate of contour's barycenter, then:

$$x_c = M_{10} / M_{00} \tag{4}$$

$$y_c = M_{01} / M_{00} \tag{5}$$

Extraction of depth information

Kinect sensor can achieve the extraction of depth information [11-13], its range of extracting depth information is shown as Fig.3. In Fig.3, default range is suitable for both Kinect for Xbox360 and Kinect for Windows. However, near range is just suitable for Kinect for Windows. This paper selects Kinect for Xbox360, so the range of extraction should be 0.8m to 4m.

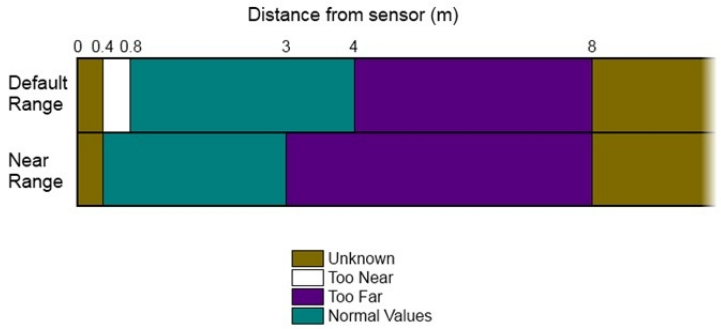

Fig.3 the range of extracting depth information

In order to get depth information, minimum inscribed rectangle of recognized contour should be drawn and mapped to depth image, the average of depth data in the rectangle can be considered as hand depth information which can remove the shadow interference. Since both barycenter and depth have been extracted, we can obtain hand 3D coordinate (*X, Y, Z*).

Human-computer interaction

In this module, system detection is based on the change of 3D coordinate which can extract the characteristic of moving hand. Whether hits the target are determined by the change of the plane coordinate and the depth, only when multi continued frames meet the demand that the system will determine it as hitting the target successfully. As the targets are moving, decision range also need to change follow the real time. This system chooses depth coordinate Z at first to determine. If hand 3D coordinate at this frame is $(X(t), Y(t), Z(t))$ and capture $(X(t+1), Y(t+1), Z(t+1))$ at next frame, after many experiments, we can know that when they meet $200 > Z(t) - Z(t+1) > 50$, we can determine that hand is moving forward to camera; After five continued frames are considered as moving forward to camera, we can determine that this motion is a hit. At the same time, coordinate at this frame will be kept. If the 2D coordinate (x, y) locate in the range of target then system will consider this motion hit the target successfully.

Result and Discussion

This paper investigates hand recognition and track based on depth information and achieves the human-computer interaction system based on depth information under the hardware and software condition of Kinect sensor, VS2010, Opencv and Kinect toolkit functions library. Examples of the actual effect of the system are shown as Fig.4.

A．Example of fail to hit the target 1

B．Example of fail to hit the target 2

C．Example of hit the target

Fig 4. Examples of the actual effect of the system

In Fig.4, the first column are the RGB information extracted from the camera, the second column are the depth information and the minimum inscribed rectangle of recognized contour, the third column are game shown on the screen which is utilized for interaction.

In the Fig.4.A, player hits to the area where targets are not in. Although the depth data meets the demand, system still determines it fail to hit the target because barycenter doesn't locate in the range of target.

In the Fig.4.B, player stops his hand at the area where target is in. Although the barycenter locate in the range of target, system determines it fail to hit the target because hand depth data doesn't meet the demand.

In the Fig.4.C, player hits to the area where one of the target is in. Because both barycenter and depth data meet the demand, system determines player hits this target. At the same time, the extracted data doesn't meet the demand of another target, so system determines player doesn't hit another target.

After a lot of experiments to verify, we can draw the conclusion that this system is able to achieve somatosensory game with higher accuracy and recognize hand motion correctly. This system also has short time for the system to determine and better real- time performance. Whole system utilizes Kinect sensor, combine RGB and depth information and game on the screen. Player can somatosensory game with sense of reality and the sense of substitution.

Conclusion

After testing, this system can use skin color, depth and position and other information to obtain the coordinates of the hand in the three-dimensional. Combining the coordinate of multi continued frames, hand motion characteristic can be extracted, which can be used for determining in order to achieve human-computer interaction. This paper brings out the human-computer interaction system for hand recognition based on depth information from Kinect sensor which has real-time, accuracy and practicability. However, this system still needs to be improved. Because of utilization of skin color recognition, system may make inaccurate determining sometime. In the following improvement, skeleton recognition can replace skin color recognition for more accurate determining. Furthermore, skeleton recognition can determine exact part of human body which can bring player a more intelligent somatosensory system.

Acknowledgment:

This paper is supported by Excellent Talent Training Project of Beijing (2013D005002000002). This paper is the result of 2015 Beijing college students' scientific research and Entrepreneurship Program.

Reference

[1] S.Singh, S.A.Velastin, H.Ragheb, MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods, 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance,,2010,48-55

[2] A.A.Liu, Y.T.Su, P.P.ia, Z.Gao, T.Hao, Z.X.Yang, Multipe/Single-View Human Action Recognition via Part-Induced Multitask Structural Learning, IEEE Transactions on Cybernetics, VOL. 45, NO. 6, JUNE 2015, 1194-1208

[3] V.Mitra, H.Franco, M.Graciarena, D.Vergyri, Medium-duration modulation cepstral feature for robust speech recognition,2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),1749-1753

[4] A.A.M.Abushariah, T.S.Gunawan, O.O.Khalifa, English Digits Speech Recognition System Based on Hidden Markov Models, International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia

[5] B.F.Klare, M.J.Burge, J.C.Klontz, R.W.Vorder Bruegge, A.K.Jain, Face Recognition Performance: Role of Demographic Information, IEEE Transactions on information forensics and security, VOL. 7, NO. 6, DECEMBER 2012, 1789 – 1801

[6] M.A.Lone, S. M. Zakariya and R.Ali, Automatic Face Recognition System by Combining Four Individual Algorithms, 2011 International Conference on Computational Intelligence and Communication Systems, 2011,222 – 226

[7] Y.L.B.Li1, A.S.Mian, W.Q.Liu, A.Krishna1, Using Kinect for Face Recognition under Varying Poses, Expressions, Illumination and Disguise, 2013 Applications of Computer Vision (WACV), 2013, 186-192

[8] G.C.Jaesik Park, Y.W.Tai, Exploiting Shading Cues in Kinect IR Images for Geometry

Refinement, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, 3922-3929

[9] F.Z Chelali, N.Cherabit, A.Djeradi, Face recognition system using skin detection in RGB and YCbCr color space, 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015, 1-7

[10] J.P.Gao, Y.J.Wang, H.Yang, Z.Y.Wu, An Elliptical Model Based on KL Transform for Skin Color Detection, Journal of Electronic & Information Technology, 29(7),2007

[11] M. S.Abd.Manap, R.Sahak, A.Zabidi, I. Yassin and N. Md. Tahir, Object Detection using Depth Information from Kinect Sensor, 2015 IEEE 11th International Colloquium on Signal Processing & its Applications (CSPA2015), 6 -8 Mac. 2015, Kuala Lumpur, Malaysia,160-163

[12] L.Xia, C.C.Chen and J. K. Aggarwal, Human Detection Using Depth Information by Kinect, 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2011,15-22

[13] M.Zeng, Z.Liu, QH.Meng, ZB,Bai, HY.Jia, Motion Capture and Reconstruction Based on Depth Information Using Kinect, 2012 5th International Congress on Image and Signal Processing (CISP 2012), 2012, 1381-1385