Depth and Skeleton Information Model for Kinect Based Hand Segmentation

Zhenhao Huang, Zeke Xu, Zhiyuan Li, Zhuoxiong Zhao, Dapeng Tao

School of Electronic and Information Engineering,South China University of Technology

Abstract

With the rapid development of the low-cost Microsoft Kinect, hand segmentation has been a resurgence of broad interest. This is because the depth and skeleton information provided by the Kinect opens an innovative way for hand segmentation. In this paper, we propose a new scheme for hand detection and segmentation based on depth and skeleton information. We conduct experiments on our new collect RGB and depth image pairs. The results demonstrate the robustness and effectiveness of our proposed model.

Keywords: Human-Computer interaction, hand segmentation, depth information, skeleton information, Kinect

1. Introduction

With the continuous development of computer technology, Human-Computer interaction has been gradually making its entry into our lives. The traditional ways of interaction, like mouse keyboard, remote control interaction is gradually replaced because of its lack of intelligence and convenience. It cannot satisfy the general user. Nowadays the natural Human-Computer interaction receives significant attention, such as facial recognition, speech recognition, and hand-gesture recognition. The traditional human-computer interaction is based on instruction. Users send a request and instruction to the computer. As an executive terminal, the computer outputs the discrete-time flow. User experience is restricted by the interaction device. The gesture-based interaction passes the information in a natural way, meeting people’s needs. Hence, gesture interaction is an important interactive way of Human-Computer interaction and one of the hotspots in the present research. In order to interact with hand gesture, the hand in the image must be segmented.

At present, several ways of hand segmentation are as follows:

The color-based hand detection has the advantages of simplicity and quick operation. It also cannot be affected by the deformation. In the study of skin detection, many color space has been used, such as RGB, YCbCr, and HSV and so on. RGB is the basic and the most commonly used color space. But it is closely related to illuminate, which causes great errors when it is not in an ideal surrounding of illuminate. The experimental result is better in YCbCr color space, because the light intensity variable Y is calculated separately. L. Zhiguo *et al.* [1] proposed hierarchical optimization method into the particle-filter-based tracking frames to improve the efficiency of particles sampling from the hidden state space. But the feature of the image is extracted based on skin color model. It didn’t work well under complicated illumination conditions. F. Baoling *et al.* [2] effectively segmented the hand in bad illumination conditions. But it could not distinguish hand from the face, because the color of hand is the same as that of face.

Another way of hand detecting and tracking is based on depth information. Kinect is a 3D camera, from which many data can be obtained. These data includes:

1）Color information: The data captured by the RGB camera, whose sampling rate can be set manually.

2）Depth information: The data captured by an infrared camera, the sampling rate can be set manually as well.

3）ID of players: Kinect can capture the data of six persons. The data can be classified according to player's ID.

4）Skeleton information: The data of joints of skeleton.

Therefore, depth information in the scene can be recovered. The object and background can be separated. This method overcomes the disadvantages of skin detection. It will not be affected by illumination and improves the accuracy of detection. [3] segmented the hand with depth information and the experimental result is satisfactory, but his algorithm has a relative high complexity. G. Chuang *et al.* [4] detected the hand position with skeleton information got from Kinect, and recognized the hand gesture with the changing depth value. Because he got the position of the hand instead of segmenting the hand, the device could only recognize the simple gesture like moving the hand forward or backward.

The skin color information can be combined with the depth information. Y. Wen *et al.* [5] extracted the skin color information from the RGB image and found out the point which had the minimum depth value to determine the position of the hand. And then distinguished different hands via K-Means-clustering algorithm, calculated the minimum convex hull point set to divide the hand contour via Graham's Scan algorithm. This method works well, but it has some limitation: The hands must be in front of the body and there is nothing having the similar color with skin in the space which has smaller depth value than the hand. D. Haijin [6] segmented the hand with depth value, but there is also the limitation of hand's position.

In this paper, a new method is proposed. The skeleton information can be obtained from Kinect SDK. In this way, the approximate position of the hand can be extracted from the joint's information. With the position information, we can use dynamic-threshold to segment the hand from the image. In a word, this method not only has the advantage of depth information, being unaffected by illuminate and surroundings, but also overcomes the disadvantage. The hand's position is freedom. It is no longer having the limitation that the hand must be at the forefront. What's more, the limitation of hand gesture decreases greatly after using this method.

2. Hand segmentation based on skin color and depth information

2.1. Hand detection and segmentation based on skin color

The selection of the color space is very important for skin detection. The skin detection technique is relatively mature and simple. The common way is to build a skin-color-clustering, a threshold is set to distinguish whether the pixel belongs to skin-color pixel and filter those non-skin-color pixels. This method has some obvious disadvantages:

1）*Some errors have occurred:* Those pixels whose color is similar to skin cannot be filtered effectively.

2）*Affected greatly by illumination:* One method to reduce the illumination is to build skin-color-clustering in YCbCr color space, because the light variable, Y, can be considered separately. His detection result is better than the ordinary method but it does not satisfy our need. “Fig. 1” is the experimental result of skin-color-clustering in YCbCr color space, as in (1):

$$\begin{cases} Y > 80 \\ 85 < Cb < 135 \\ 123 < Cr < 138 \end{cases} \qquad (1)$$

Fig. 1 shows the skin detection result. It is not surprising that confusions occur between the skin and the furniture. In addition, the left part of the face is missing for the strong light intensity.

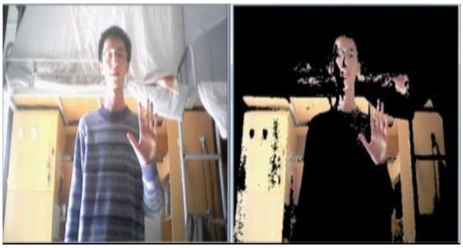

Fig. 1: Skin detection in YCbCr color space.

3） *The hand cannot be distinguished from other skin tissue, as in “Fig.1”:* The face detection is a 2-D morphological segmentation method, detecting the face according to the feature of skin color, contour, heuristic characteristics like jaw and hair. Relative to the face, hand has more degrees of freedom, which means the algorithm will be very complicated if the method of 2-D morphological segmentation is adopted.

2.2. Hand detection and segmentation based on depth information

Considering the convenience and practicality in the human-computer interaction, hand is in front of the body in most cases. Hence, after searching the minimum of Kinect’s depth information, the pixel which is closest to the camera can be found out from depth information and the position of the hand is determined. After that, with an appropriate threshold, the whole hand can be segmented. In the ideal case, hand detecting and tracking work well. The experimental result is in “Fig. 2” The advantage of this algorithm is simple and fast. But there are some obvious disadvantages:

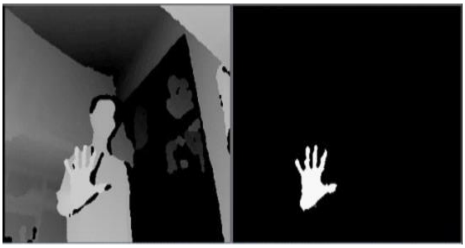

Fig. 2: Hand at the forefront.

1）*The position of the hand is limited:* On the one hand, the hand must be in the forefront and the body should not be too close to the hand. Otherwise, there will be some errors, affecting the hand detection. On the other hand, the distance between hand and Kinect should be in the range of 80cm to 4m because of device limitations.

2）*The experimental result will be affected by the noise and background:* Despite that hand is put at the forefront, the object which has the same depth value will be recognized as the hand, as in “Fig. 3”.

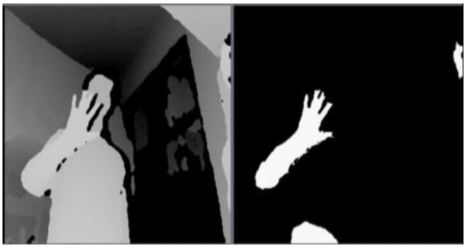

Fig. 3: Hand segmentation affected by noises.

3）*The hand gesture is limited:* The hand cannot be detected when having too much deformation.

4） *Some noise should be filtered:* These noises are created by the device itself. According to the image-forming principle of Kinect, Kincet receives the reflected infrared ray make the stereotaxic of the room. The camera can detect the movement of the body with the help of infrared ray. Because of the Kinect's limitation, the infrared ray reflected from those objects in the distance can't be obtained. So Kinect set the depth value of this object to be 0, which means that gray value is 255. It will be recognized as the object which is close to the Kinect. In order to filter these noises, the gray value must be set to be 0, as in "Fig. 4".

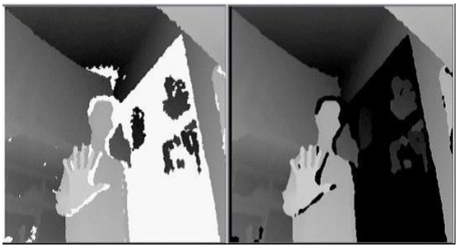

Fig. 4: The image on the left is before filtering the noise. The right one is after filtering the noise.

The method, proposed as follow, overcomes these disadvantages effectively.

3. Hand segmentation based on depth and skeleton information

Being able to extract the skeleton information is an important characteristic of Kinect. With the help of Prime Sense, Kinect capture player's hand gesture and compare the captured image with the mannequin-model that saved in Kinect. The object matching the mannequin-model will be created as skeleton model. Then the system transforms the model into a virtual character. By detecting the key parts, the joints, of the character, the movement of the character is triggered. With this virtual skeleton model, the system can recognize 25 key parts of the human body. Some simple gesture like standing and sitting can be recognized as well. Hence, an improved algorithm is proposed in this paper based on depth and skeleton information.

The algorithm proposed is as follows:

1) Extract the depth data stream.
2) Extract the skeleton data stream, locate palm and wrist of the model, the 2-D coordinate of hand can be obtained.
3) Extract the depth data of palm and wrist, Dep_1 and Dep_2 .
4) Calculate the geometric distance L_0 between palm and wrist. Via debugging, an adjusting coefficient Δd can be got.
5) Figure out the planar threshold D_c , as in (2).

$$D_c = L_0 + \Delta d \qquad (2)$$

6) Calculate the depth distance between palm and wrist D_0 , as in (3), and calculate the depth threshold D_d , as in (4).

$$D_0 = Dep_1 - Dep_2 \qquad (3)$$

$$D_d = |D_0| \qquad (4)$$

7) Set wrist to be the center point, $2D_c$ as the length of a square, D_d as the depth threshold, the square as the planar threshold and segment the hand gesture from the image with the help of the dynamic threshold.

4. Experimental Result

Because Kinect SDK has provided a skeleton tracking interface, the position of palm and wrist can be obtained from the skeleton data stream. It is robust. The hand can be detected even it is occluded. The skeleton point is captured and the corresponding depth information can be obtained afterwards. With the planar and depth thresholds, the hand is segmented, as in "Fig.5" and "Fig.6".

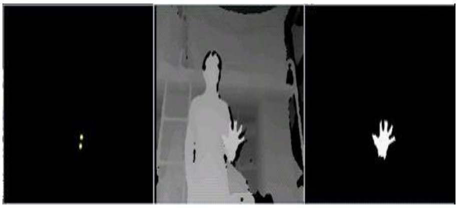

Fig. 5: Hand is at the forefront. The image on the left is the skeleton point, the middle one is depth image, the right one is hand that is segmented.

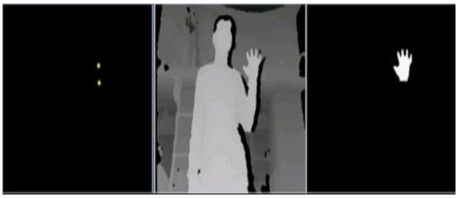

Fig. 6: Body is at the forefront. The image on the left is the skeleton point, the middle one is depth image, the right one is hand that is segmented.

The proposed algorithm effectively makes up for the limitation of pure depth data. The advantages are as follow. First, the position of the hand is relative freedom. Owing to skeleton information, the hand is no longer limited to be at the forefront. Second, because the position of the skeleton has nothing to do with illuminate, anti-noise property significantly enhances. Third, depth threshold is adaptively chosen according to the depth data of joint. The threshold changes adaptively when the shape of a hand and the angle of rotation change. Hence, anti-deformation property significantly enhances.

Besides, this algorithm has the classical advantage of depth information. It will not be affected by illumination. Those objects whose color are similar to skin will not be mistaken for the skin color objects like the hand and the face.

To sum up, the algorithm, hand segmentation based on depth and skeleton information, can detect the hand in complex background. What's more, it overcomes the disadvantage of conventional method of depth information. There is no limitation of position of the hand, which is a breakthrough.

5. Conclusion

In this paper, the common algorithm of hand detecting is discussed and a new algorithm based on depth and skeleton information is proposed. From the comparison between different algorithms of hand segmentation, the algorithm proposed in this paper has a great advantage and robustness. The experimental result is satisfying.

6. References

[1] L. Zhiguo, L. Yan, X. Xin, "Research on Fast 3D Hand Motion Tracking System," *Journal of Computer Research and Development*, Vol.49 No.7, pp. 1398-1407, 2012.

[2] F. Baoling, W. Ming, D. Yingdi, "Hand Gesture Segmentation Based on Skin Color Detection Technology," *Computer Technology and Development*, pp. 105-108, 2008.

[3] Chen. Zihao, "Hand Detection and Tracking Based on Depth Maps," *MS thesis. South China University of Technology*, 2012.

[4] G. Chuang, "Dynamic Gesture Recognition Based on 3D Kinect," *Electro-optic Technology Application*, Vol.27 No.4, pp. 55-58, 63, August 2012.

[5] Y. Wen, C. Hu, G. Yu, C. Wang, "A robust method of detecting hand gestures using depth sensors," *Haptic Audio Visual Environments and Games (HAVE), 2012 IEEE International Workshop on. IEEE,* pp.72-77, 2012.

[6] D. Haijin, "3D Vision-based Hand Tracking and Its Application in the Human Computer Interaction," *MS thesis. Nanjing University*, 2011.