

Robust hand gesture detection by fusion of depth and colour information using kinect

Shuai Yang^{*}, Prashan Premaratne, Peter Vial, Qasim Alshebani

School of Electrical Computer and Telecommunications Engineering, University of Wollongong, North Wollongong, NSW, Australia

Received 1 October 2014, www.cmnt.lv

Abstract

Microsoft Kinect camera has drastically changed the world of human computer interaction based computer vision, due to its low cost and high quality of depth information for visual images. This has made the depth data to become common place at a very low cost allowing myriad of computer vision related application including hand gesture recognition. Hand gesture recognition research suffered severely from the clutter and skin tone regions in any background. With the availability of depth information, background clutter and skin toneregions which are not part of the hand gesture can be removed improving the performance of any classification strategy. This article discusses a novel hand detection strategy based on Kinect camera by combining depth and colour image information. In the detection procedure, the Kalman filter is applied to the study to achieve a good detection result. The experiment results show this detection method is reliable and stable in the clutter background, and works well in various light conditions.

*Keywords:*Kinect, human computer interaction, depth information, Kalman filter, fusion

1Introduction

As a major method in human communication such as in sign language [1], hand gesture recognition has been applied to the Human- Computer Interaction (HCI) area for a long time [2, 3]. Compared to the traditional inputs such as keyboard and mice, hand gestures are more natural and flexible. They have a great potential in the area of digital control, real-time input and communication among disabled [4, 5]. In recent years research has focused on vision- based hand gesture recognition and control as the contact-type devices strongly diminish the flexibility of hand movements. The tremendous development of computational ability of many hardware devices, such as smart phones, has allowed computer vision the opportunity to play an important role in human computer interaction [6, 7]. Although, the research on the vision-based hand gesture analyses have made rapid progress, the reliability and practicality of the hand gesture recognition systems are still problematic. The biggest challenge in advancing the field lies on reliably tracking hand gestures in order to recognise dynamic gestures under complex lighting conditions and background clutter [8, 9].

The recognition systems under background clutter usually require the background to be free of skin tones. This usually require the people to wear colour markers or use fixed colour background, then machines analyse and segment hand gestures by detecting the marked colour [10]. These contact type devices such as colour gloves or markers worn on hand limits the flexibility and reaction time of the system. However, these methods have ~~advantages in terms of accuracy~~. The skin colour

detection is based on the characteristics of the spatial distribution of the skin colour in colour space to convert the image to the corresponding colour space to do the threshold segmentation [11]. The skin colour detection can directly isolate the skin colour area from the image, but there are some disadvantageous in current technology, such as the gesture and skin colour area cannot overlap as the segmentation will be affected by background clutter [12, 13].

Nowadays, with the development of cameras, there are two kinds of the most adopted devices for hand gesture detection and recognition, one is Time of Flight (TOF) cameras [14], and another one is Microsoft Kinect. Both devices can produce the depth image which is also known as the range image. The information of this kind of image records the distance of each point of the space between the object surface and the camera. The grey scale of each pixel on depth image is only related to the distance of each point between the object and camera, so depth data have the 3D characterises of an object in the space, which the grey scale image and colour image don't have. It can be used to accurately extract the foreground from background in computer vision. The TOF cameras record the depth data by measuring the flying time of the light between the object (the hand) and the sensor. In fact, it calculates the time elapsed between the sent pulse and the reflection of it off the object when received by the receiving sensor. Compared to the traditional 2D camera, the TOF cameras can easily extract foreground from background, so it has advantages of object tracking and analysis. However the disadvantage is that the TOF cameras are expensive and lower resolution. In 2010, Microsoft launched a 3D camera peripheral

^{*} *Corresponding author's* e-mail: yangshuaiwoll@163.com

somatosensory for the Xbox360 [15]. The Kinect uses structured light coding techniques to obtain depth information of captured images. The Kinect camera includes an RGB camera, an infrared camera and an infrared emitter. The Infrared emitters can emit near-infrared laser, when the laser irradiate rough object, it will produce a high degree of randomness diffraction spots, called laser speckle. Laser speckle will vary patterns according to the distance of objects. When the laser speckle irradiate to the entire space, it means that the space has been marked. Infrared camera is used to receive space markers and pass the markers to the core chip of the Kinect. The processor produces the depth image by analysing the laser speckle pattern. Compared to the TOF cameras, Kinect is much cheaper with higher resolution, besides, the Kinect has an additional graphic processor, so there is no extra computation for the computer. It achieves the real-time gesture analysis under a comparative low configuration [16].

These features make Kinect become a popular tool in the domain of movement recognition. The hand gesture recognitions based on it usually use the depth information, which was produced by Kinect. There are two common ways for using the depth information. First is that using depth information instead of colour information, which means transferring the depth information of the hand area to 2D image, and then applies the traditional recognition methods to the 2D image, this kind of methods actually take advantage of the depth information to get a relative robust recognition, however, it wastes of depth data while converting 3D depth information to 2D information, and it will easily effected by the finger occlusion problems [17]. The second way is totally using depth information, which means transferring the depth information to the 3D pixel cloud [7]. Then simulate the gesture motion in virtual space, and calculate the 3D information of each point. This kind of method is more accurate than the former method, but the drawbacks are obvious, the computation is too large for a normal computer, if count the hand gesture recognition system model, it cannot be a real-time method under the usual current hardware [18]. In the recent years, there is one recognition model which is a fusion by depth image and colour image. The common way is that apply the depth information to the colour image to use the depth information to isolate the hand gesture from colour image. Then put the processed colour image to the traditional recognition model. This kind of methods has more advantages than the former methods, it uses the accurate position information to get the gesture, and then apply the traditional system to the colour image to save the resources. It uses the depth information one time during gesture detection phase, although, it save the computation time, because the detection phase only use the depth data, the object, which has the same distance away from the camera with your hand, will highly effect the recognition results [19,20].

Hand detection is one of the most important phases of hand gesture recognition. It highly affected the recognition accuracy [21]. So in this paper, a novel hand

gesture detection method is proposed, which is a combination of depth data and colour information, which uses both colour information and depth information during hand gesture detection. The key idea of this method is the multiple threshold settings to isolate the useful information from the depth image or the colour information alternately to achieve a better hand gesture.

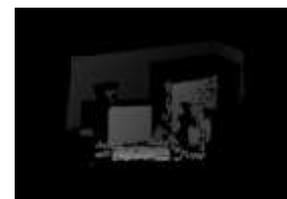
This paper is organised as follows, section 2 discusses the basic characteristics of the camera and the calibration of the Kinect. Section 3 presents the method to set the threshold and do the first phase segmentation. Section 4 describes the detailed procedure to further remove some useless part. Section 5 illustrates the method to apply the threshold to hand detection to achieve a further clear hand. Section 6 analyses the methods and meaning to do the region growing and corrosion, it's not the necessary step for general hand detection, but for detailed high quality hand detection, it's quite suitable. Section 7 is the last step of the system, the Kalman filter is chosen due to its high toleration for the sudden noise, and high performance for a continuous recognition. Section 8 is the conclusion of the entire paper to talk about the achievement of this system.

2 Characteristics of Kinect camera and calibration

The Kinect sensor can achieve depth data and RGB colour image at the same time. It can also track object movement. The left lens is an infrared emitter with a common RGB colour camera in the middle and a 3D depth sensor is on the right. Kinect has focus tracking function with the base motor can rotate Kinect by around 270 degree. It also has an array of microphones. This allows Kinect to capture a colour image, 3D depth image and audio as shown in Figure 1a and 1b. Compared to an ordinary camera, the Kinect has a CMOS infrared sensor, which is used to estimate the environment by using black and white spectrum. The pure black is on behalf of infinity faraway, pure white means infinity close, the grey area between black and white is corresponding to the distance between the point and camera. It collects every point in the space to form a comprehensive depth image of the surrounding environment. The sensor generates a depth image at 30 frames per second to rebuild the surrounding environment [6].



a) the colour image produced by Kinect



b) the depth image produced by Kinect

FIGURE 1 Images produced by Kinect

Compared with traditional cameras, the Kinect has many advantages, it work in real-time, and the depth data,

which is sent to the next step process without additional computation. Besides, the depth data from Kinect will not be affected by the light condition and clutter in the background. The depth camera generates depth data even at low lighting conditions. Compared to the traditional hand detection methods, Kinect doesn't require the colour markers or fixed colour background. Even with overlap of two skin colour areas, it will not affect the detection result [14].

To use the Kinect camera to produce depth data, there are two usual methods; first is using the Microsoft SDK to achieve the data and another is to use Microsoft virtual studio to input the libraries of OpenCV and OpenNi, and then achieve the depth data from Kinect. Before producing the depth image and colour image at the same time, the camera should be calibrated. Because there is some distance between the depth camera and colour camera. The depth-Generator Get-Alternative View-Point-Cap sentence can use to adjust the view of two cameras to achieve the same image in virtual studio as shown in Figure 1a and 1b.

3 First time thresholding

Threshold of hand detection phase is very important to divide the points into different groups by their different characteristics. Because of the drawbacks of depth image and colour image in the hand detection phase, the threshold only applied to colour image or depth image may lead to different kinds of inaccuracies. In this paper, thresholding will apply to the colour image and depth image for multiple times in order to achieve a clear hand gesture.

The first step is to apply the threshold to the depth data as the grey scale of each pixel on depth image is only related to the distance, so that the point which is closer to the camera is much brighter than a distant point. This step is used to exclude the obvious background in depth data as visible in Figure 2. Any skintone object in the background will not affect the hand gesture which is in the foreground when using this depth thresholding simplifying age-old background separation problem in computer vision.



FIGURE 2 a) The original depth image, b) The image after first time thresholding, which had been removed the background

A proper threshold can lead to a good segmentation result. In this research, the wide spread Ostu method was applied to look for a proper threshold by using grey level histograms [23]. The main idea is to select a threshold

from the histogram which was derived from discriminant analysis point of view. The optimum threshold is determined by the discriminant criterion which maximizes the discriminant measurement of separability of the resultant. A threshold, T , is set, all the points, which their grey scale values are lower than T , will be dropped. In this phase, it will reduce the majority noisy signal in the image to achieve a relative clear and smaller area.

4 Overlapping depth image on colour image to remove the background

After segmentation using depth information, the foreground is identified in the depth information. This information can be utilized to threshold the colour image to remove the background. It is a process very similar to Logic AND operation where the foreground image which contain the hand region will preserve the area in the colour image. Everything else will be discarded in the colour image thereby obtaining the hand gesture in full colour. In this phase, the image should be converted to HSV format, which is more convenient for image analysis [24]. The above process can be mathematically described as follows: Assuming the total number of pixels in this phase is N , and the H , S , and V represents hue, saturation and brightness respectively. The system should require three constrains to achieve the colour thresholding. First, by choosing the skin colour area and second requiring the saturation to be not white, and thirdly, the skin colour area should be bright in case of choosing the other object which has a similar colour with skin. So the constrains can be summarized as follows:

$$y = \begin{cases} -10 < Hy < 10 \\ Sy > ths \\ Vy > thv \end{cases} \quad y \in N . \quad (1)$$

And then, the threshold should be set again to isolate the skin colour area. After this process, only the skin colour area is keep in the image.

5 Total thresholding

After the above steps, the elements in the image are clear, but there are still some unnecessary artefacts in the image, such as another arm. Final segmentation can remove this as shown in Figure 3b. These constraints can be mathematically expressed using the d_{min} referring to the shortest distance, d_{max} referring to the longest distance, y is a point in the image, G_y referring to the grey scale value, D_y is the distance between y and camera.

$$y = \begin{cases} d \min < Dy < d \max \\ Gy > \lambda \\ y \in N \end{cases} \quad (2)$$



FIGURE 3 a) After applying the threshold to the input image, which is fusion of the depth image and colour image, only the skin colour area is kept. b) After the final phase processing, the detected hand looks like this

6Morphological filtering for smooth edges

In order to remove the jagged area that is the result of the imperfections of the above thresholding requires certain morphological filtering to achieve a smooth edges for effective hand gesture recognition. A process known as region growing as shown in Figure 4a is very effective at producing a smooth edge. Before the process, a seed pixel of the image must be settled as a start point, and then the seed pixel will absorb a set of pixels [25], which have the similar characteristics with the seed pixel, in the neighbouring regions. The pixel of the image, which has no similar characteristics with any other pixels, will be settled as the new seed pixel. This process repeats until there is no pixel to absorb, it means the system finish the region growing phase.

In mathematics, image expansion is that doing the convolution between image (called A) and core (B). The core can be any shape or size, it has an additional reference point. Generally, a core is a solid square or a disc with a reference point. Expansion is a method, which is used to get the local maximum value, and then assign the value to the reference point specified pixel to make the highlight area of the image gradually increase [26].

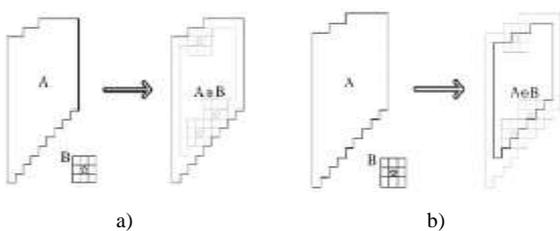


FIGURE 4 a) image expansion process from A to B b) image erosion process from A to B

Erosion phase of an image is on the contrary of expansion phase as shown in Figure4b. It is used to record the minimum value of the pixel in the core region. The system will calculate the minimum value of a pixel in the area covered by B while the core B is doing the

convolution with the image, and then place the value on the reference point [27].

For this image expansion procedure, the set of seeds should be founded, assuming that the Gx is the grey scale value of the point x , the new threshold is represented by λ_{new} , the set of seed is S , the point in this area, which will be absorbed, could be summarized as follows:

$$x = \begin{cases} Gx > \lambda_{new} \\ |x - s| > 1 \end{cases} \quad x \in s \quad (3)$$

For this hand detection system, all these procedure are used to achieve a smooth contour, reliable hand shape to improve the precision of detection rate. The corrosion phase is mainly used for removing some fixed useless point of the image, because the final image is much smaller than before, so the noise filter is very important, and it will be easily affected the detection results.

7Hand detection

In the OpenCV library, there is one filter which is widely used, because it has many advantages than other methods, such as good tracking ability, edge detection and so on. It is the Kalman filter. The main task of it is to track the value of a variable. The tracking is based on the equation of motion of the system to make a prediction. The prediction may have error, so that the Kalman filter uses another measuring instrument to measure the value of the variable, the measurement may also have the error. But these two values have different weight ratio, the Kalman filter is based on these two values to do a series of iterations to track the target [28].

In this paper, this detection and tracking system is attempted by two fundamental formulas. First, we need to introduce a system of discrete control process. The system can be described by a linear stochastic differential Equation:

$$X(k) = AX(k-1) + BU(k) + W(k) \quad (4)$$

The measurement from the system:

$$Z(k) = HX(k-1) + V(k) \quad (5)$$

In the above two Equations, $X(k)$ is the system state at the k time, $U(k)$ is the control amount at the k time. A and B are two system parameters. $Z(k)$ is the measurement value at k time, H is the parameter of the measurement system. $W(k)$ and $V(k)$ are noises [28].

In this system, a fusion of colour and depth information is collected for Kalman filter, which keeps the detection system working well under the different lighting conditions and background clutter as shown in Figure 5. Even there are other people in the background, the system is not affected. With the use of the Kalman filter, the system will keep work accurately when some

short term gestures appear in the scene, these useless gesture will not affect the detection and detection accuracy unless these gestures keep in the scene for a longer time than the main user.

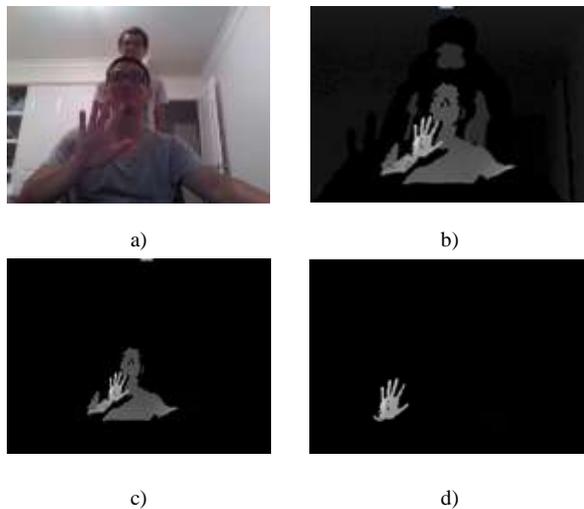


FIGURE 5 These four images show the performance of the system

The specific model chosen for the hand detection depends on the specific application and circumstance, for

References

- [1] Premaratne P, Yang S, Zou Z, Vial P 2013 Australian SignLanguage Recognition Using Moment Invariants *Lecture Notes on Artificial Intelligence* 7996:509-14
- [2] Premaratne P, Ajaz S, Premaratne M 2011 Hand Gesture Tracking and Recognition System for Control of Consumer Electronics *Springer Lecture Notes in Artificial Intelligence (LNAI)* 6839 588-93
- [3] Premaratne P, Nguyen Q, Premaratne M 2010 Human computer interaction using hand gestures in *Advanced Intelligent Computing-Theories and Applications - Communications in Computer Information Science* 93 381-6
- [4] Premaratne P, Safaei F, Nguyen Q 2006 Moment Invariant Based Control System Using Hand Gestures: Book *Intelligent Computing in Signal Processing and Pattern recognition Springer Berlin / Heidelberg* 345 322-33
- [5] Khan R, Ibraheem N 2012 Hand Gesture Recognition: A Literature Review *International Journal of Artificial Intelligence & Applications* 3(4) 161-74
- [6] Premaratne P, Premaratne M 2014 Image Matching using Moment Invariants *Neurocomputing* 058
- [7] Premaratne P, Ajaz S, Premaratne M 2013 Hand Gesture Tracking and Recognition System Using Lucas-Kanade Algorithm for Control of Consumer Electronics *Neurocomputing Journal* 116(20) 242-9
- [8] Doliotis P, Stefan A, McMurrough C, Eckhard D, Athitsos V 2011 Comparing Gesture Recognition Accuracy Using Color and Depth Information *Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments* 20
- [9] Molina J, Escudero-Viñolo M, Signoriello A, Pardàs M, Ferrán C, Bescós J, Marqués F, Martí J M 2011 Real-time user independent hand gesture recognition *Machine Vision and Applications* (2013) 24:187-204 (in Spanish)
- [10] Wysocki S G, Lamar M V, Kuroyanagi S, Iwata A 2012 A Rotation Invariant Approach On Static- Gesture Recognition Using Boundary Histograms And Neural Network *International Journal of Artificial Intelligence & Applications (IJAI)* 3(4)
- [11] Ibraheem N, Hasan M, Khan R, Mishra P 2012 Comparative study of skin color based segmentation techniques *Aligarh Muslim University AMU, Aligarh India*
- [12] Oprisescu S, Rasche C, Su B 2012 Automatic static hand gesture recognition using ToF cameras *Signal Processing Conference (EUSIPCO) Proceedings of the 20th European* 2748:2751
- [13] Premaratne P 2014 Human Computer Interaction using Hand Gestures *Springer Publication ISBN-10 9814585688*
- [14] Lai K, Konrad J, Ishwar P 2012 A gesture-driven computer interface using Kinect *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on* 185-8
- [15] Hasan M M, Misra P K 2011 Brightness Factor Matching for Gesture Recognition System Using Scaled Normalization *International Journal of Computer Science & Information Technology* 3(2) 35-46
- [16] Kolb A, Barth E, Koch R, Larsen R 2010 Time of Flight Cameras in computer Graphics *Computer Graphics Forum* 29(1) 141-59
- [17] Kuno Y, Sakamoto M, Sakata K, Shirai Y 1994 Vision-based human computer interface with user centred frame in *Proc IEEE/RSJ/Int. Conf. IROS* 3 2023-9
- [18] Premaratne P, Nguyen Q 2007 Consumer electronics control system based on hand gesture moment invariants *IET Computer Vision* 1-1 35-41
- [19] Yang S, Premaratne P, Vial P 2013 Hand Gesture Recognition: An Overview *5th IEEE International Conference on Broadband Network and Multimedia Technology*
- [20] Zou Z, Premaratne P, Premaratne M, Monaragala R, Bandara N 2010 Dynamic hand gesture recognition system using moment invariants *ICIA/S 2010 5th International Conference on Information and Automation for Sustainability IEEE Computational Intelligence Society Colombo Sri Lanka* 108-13
- [21] Premaratne P, Safaei F 2008 Feature based Stereo Correspondence using Moment Invariant *Proceedings of the IEEE International Conference on Information and Automation for Sustainability* 104-8

- [22] HerathDC, KroosC, Stevens CJ, CavedonL, PremaratneP 2010 Thinking head: towards human centred robotics *11th International Conference on ControlAutomation, Robotics and Vision (ICARCV)Singapore* 2042-7
- [23] Otsu N 1979 *IEEE Trans Systems Man and Cybernetics*9(1) 62-6
- [24] PremaratneP, Nguyen Q.PremaratneM 2010 Human computer interaction using hand gestures in *6th International Conference on Intelligent Computing, ICIC* 381-6
- [25] Chang Y L,Li X 1994 Adaptive image region-growing *IEEE Trans Image Processing* 3(6) 868-72
- [26] IkonomakisK N, ZervakisP M, VenetsanopoulosA N 1997 Region growing and region merging image segmentation in *Proc 13th IntConf Digital Signal Processing*1299-302
- [27] Tabb M..AhujiaN 1997 Multiscale image segmentation by integrated edge and region detection *IEEE Trans Image Processing* 6(5) 642-55
- [28] Chen R,Liu J 2000 Mixture Kalman filters *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*493-508

Authors	
	<p>Shuai Yang, 1992, China.</p> <p>Current position, grades: PhD student in the School of Electrical Computer and Telecommunications Engineering, university of Wollongong.</p> <p>Scientific interest: Human computer interaction using hand gestures.</p> <p>Publications: 3 papers.</p>
	<p>Prashan Premaratne, 1972,Australia.</p> <p>Current position, grades: PhD supervisor in the School of Electrical Computer and Telecommunications Engineering, university of Wollongong, a Senior Member of IEEE, the author of the book titled 'Human Computer Interaction Using Hand Gestures by Springer Publishers.</p> <p>Scientific interest: Image Processing Computer Vision and Radar Signal Processing.</p> <p>Publications: over 60.</p>
	<p>Peter Vial, 1962, Australia.</p> <p>Current position, grades: a PhD supervisor in the School of Electrical Computer and Telecommunications Engineering, university of Wollongong, a Senior Member of IEEE, member of APESMA, member of the Emerging Networks and Applications Laboratory within ICTR and the UOW ACORN representative 2008-2011.</p> <p>Scientific interest: wireless communication systems and network management.</p>
	<p>Qasim Alshebani, 1972, Australia.</p> <p>Current position, grades: PhD student in the School of Electrical Computer and Telecommunications Engineering, university of Wollongong.</p> <p>Scientific interest: face recognition.</p> <p>Publications: 3 papers.</p>