Depth Image Acquisition Method in Virtual Interaction of VR Yacht Simulator

Qin Zhang(✉) and Yong Yin

Dalian Maritime University, Dalian, China
Zhangqin_dmu@163.com

Abstract. The paper investigates depth image acquisition to realize virtual natural interaction of the VR yacht simulator. Because the existing image de-noising algorithms have the disadvantages of poor robustness and unstable edge of image, an improved bilateral filtering algorithm is proposed. Firstly, the algorithm obtains the depth information in the scene to generate depth image by using Kinect sensor. Secondly, the background removal is applied to keep the hand data from the depth image. Finally, median filtering and bilateral filtering are used to smooth the depth image. Experimental results show that the proposed method can obtain better depth images after removing background, thus the algorithm has good robustness.

Keywords: Kinect · Depth image · Background removal · Median filter
Bilateral filter

1 Introduction

At present, yacht industry is developing rapidly in China. The ability training and valuation of yacht operators are a key part of yacht industry economy chain. If part of the yacht's actual operation training is completed in the full task yacht simulator, it will significantly improve training efficiency, save training costs and avoid the risk of training. With the development of virtual reality technology, it has the advantages of much cheaper, easier network and smaller occupied space as a complement to full mission simulator. So it will gradually occupy a certain share of market.

Virtual natural interaction technology is the key technology of yacht simulator based on VR. The key point of realizing virtual natural interaction is to acquire and process 3D scene information. In computer vision system, 3D scene information is possible for computer vision applications such as image segmentation, object detection and object tracking. However, the restoration of scene depth information has the disadvantage of poor robustness. In this paper, an improved depth image de-noising method is proposed to improve the robustness of depth image de-noising algorithm.

Y. Zhao et al. (Eds.): ICIG 2017, Part III, LNCS 10668, pp. 82–92, 2017.
https://doi.org/10.1007/978-3-319-71598-8_8

2 Related Work

With the increase of demand, the 2D images with similar color of foreground and background cannot solve the problem in object tracking and pattern recognition. Therefore, the importance of depth images is represented. In 2010, Microsoft's Kinect sensor, due to its good depth image features and appropriate price, attracted the attention of intelligent surveillance, medical diagnosis, human-computer interaction and other fields [1].

The research which uses the depth image to detect human body and removes the background is usually based on the method of human body detection commonly used in ordinary image, mainly based on the two methods that human body feature matching and background subtracting.

This method based on human feature matching is mainly to extract features that describe the profile information of human body to detect human body in the image. Methods Spinello [2] described that based on statistical learning methods, the color image gradient direction histogram features human detection algorithm is proposed. And it has obtained good results in the measurement range of the whole sensor. Lu, Xia et al. [3] established the 3D head surface and 2D contour model, and then matched the depth image with the model to detect the human body. The disadvantages of this algorithm are that the feature classification result is greatly affected by the training samples, and the computational complexity of the feature generation and matching process is relatively high.

The core of the background subtraction method is to set up the background model, and then classify the pixel whether it is the background one or the foreground one. The common algorithms include average background model [4], hybrid Gauss model (GMM) [5], Code-Book [6] and ViBe algorithm [7].

ViBe algorithm is a fast motion detection algorithm proposed by Olivier [7]. The background model is set up with the surrounding pixels for background extraction and object detection. Ye [8] tested the Vibe algorithm with 3 sets of public data sets, and proved that the algorithm has the characteristics of small computation, high processing efficiency and good detection result. In this paper, Vibe algorithm is used to separate the human body from depth image.

Because the infrared measurement of scene illumination changes to a certain extent will lead to instability in the depth data and infrared reflectance characteristics of surface material and other reasons, on the edge of the object and the occluded areas are often prone to cavities. It caused a lot of noise. A variety of algorithms have been proposed for smooth image de-noising in depth image. Vijayanagar [9] used Gauss filter to estimate the hole depth value, which is a filtering method based on time-domain, this method only uses the hole depth information of surrounding pixels while ignoring the gray information. So the de-noising result is not great. Yang et al. [10] estimate the depth of the center through the distribution of the effective depth values. The method is efficient for de-noising and smoothing, but the algorithm has high complexity. In this paper, bilateral filtering and median filtering are combined to smooth noise reduction, and it improves the robustness of the algorithm.

3 Depth Image Acquisition

In order to enhance realism of VR yacht simulator interaction, 3D data acquisition and modeling of hand is necessary in real time. The main research work of this paper is to acquire 3D scene information to generate depth image. Then extract human body information from the scene and separate the hand and body information. Finally, process the depth image de-noising and smoothing.

3.1 Depth Image

The method of acquiring depth image can be divided into two classes: passive ranging and active ranging method.

(1) Passive ranging method
Passive ranging method is the most commonly used binocular stereo vision, two images of the same scene obtained by two distant cameras. The stereo matching algorithm finds the two images corresponding to the pixel. And then according to the principle of triangulation calculates the parallax information, and parallax information can be converted to display the depth information of the object in the scene. Based on the stereo matching algorithm, the depth image of the scene can be obtained by shooting a set of images from different angles in the same scene. In addition, the depth information of the scene can also be estimated indirectly through the analysis of the image's photometric characteristics, shading features and so on.

Although the disparity map obtained by stereo matching algorithm can obtain roughly 3D information of the scene, there is a large error in the parallax of some pixels. The method of obtaining parallax images in binocular stereo vision is limited by the length of the baseline and the matching accuracy of the pixels between the left and right images. The range and accuracy of the parallax images are limited.

(2) Active ranging method
The active ranging sensor device needs to emit energy to complete the collection of depth information. This ensures that the acquisition of depth images is independent of the color image acquisition. The methods of active ranging include TOF (Time of Flight), structured light and laser scanning. Kinect V2 uses TOF active ranging method to obtain depth data.

3.2 Obtain Depth Image with Kinect Sensor

The Kinect sensor is a RGB-D sensor which can acquire color data(RGB) and depth of value (depth) at the same time. The depth data obtained by Kinect is valid in the range of 0.5 m–4.5 m, with a resolution of 512 × 424, and each pixel is 16-bit (see Fig. 1). This data represents the distance from the depth (IR) camera to the object, expressed in millimeters (mm).

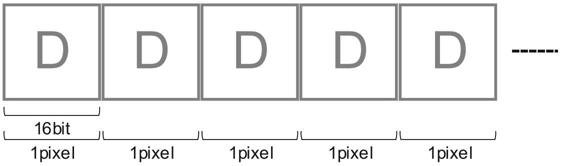


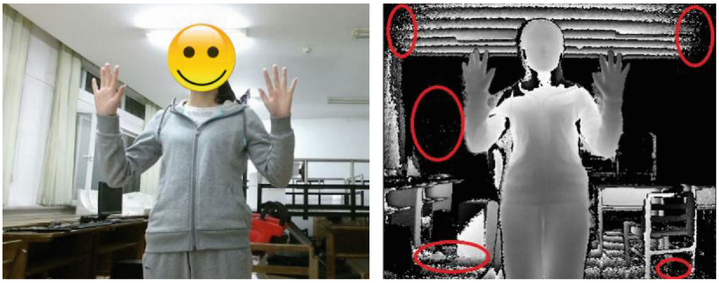
Fig. 1. Data format of depth image

Since OpenCV can only display 8-bit image data, it is necessary to convert the depth data of 16bit into image data within the 8bit (0–255) range. In depth image, the distance from the depth (IR) camera to the object is converted to the corresponding gray value. The relation between the pixel value and the true distance in a depth image (see Eq. (1)):

$$d = K \times \tan(\frac{d_{gray}}{2842.5} + 1.1863) - \theta \tag{1}$$

Where d is as actual distance, d_{gray} is grayscale of image, $K = 0.1236$ m, $\theta = 0.037$ m.

As shown in red circles of Fig. 2(b), there are two main problems due to the way to get the depth of the distance: one is that the depth of the image pixel values corresponding to a stationary object which will have some changes or even disappear in the sequence of adjacent frame images; the other one is that edges of object in the depth image which are not stable and prone to shake violently. In order to obtain a better quality depth image, both issues need to be further addressed.



(a) RGB image

(b) Depth image

Fig. 2. Sampling data from Kinect. (Color figure online)

3.3 Fast Background Removal Algorithm

Kinect provides a simple segmentation method between human body and background. The human detection algorithm is used to add ID to the detected pixels with the BodyIndexFrame data. This method supports the identification of 6 at most. Due to the BodyIndexFrame data and the depth data have the same spatial coordinates, we can use Eq. (2) to quickly get the depth data after the removal of the background.

$$d(x,y) = \begin{cases} 0, & \textit{others} \\ d(x,y), & g(x,y) \in \{0,1,2,3,4,5\} \end{cases} \tag{2}$$

Where $d(x,y)$ is depth data, $g(x,y)$ is BodyIndexFrame data.

However, the method has low detection accuracy, the body data maybe appear to be empty and missing, and human contact or adjacent objects will be false positive as part of the human body (see red circles as in Fig. 3).

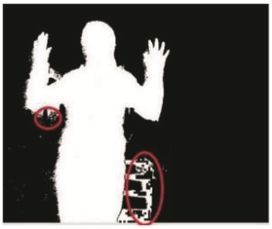


Fig. 3. Removal result of background image using Kinect SDK (Color figure online)

In order to solve the above problems, this paper uses ViBe algorithm to separate the background. ViBe algorithm is a kind of sample random clustering method. The algorithm idea is specific for each pixel storing a sample dataset, and sample value is the pixel's past value and its neighborhood values. And then each new pixel value will be compared with every pixel in sample dataset to determine whether the point belongs to the background. The core of the algorithm is the background model initialization, pixel classification and model updating.

(1) Initialization of background model

The initialization of a generic detection algorithm requires a certain length of video sequence, which usually takes a few seconds and greatly affects the real-time performance of the detection. The initialization of ViBe model is accomplished only by one frame of images. ViBe model initialization is the process of filling the sample dataset of pixels. But because the spatial and temporal distribution information of pixels cannot be contained in an image, the near pixels have similar temporal and spatial distribution characteristics. The value of the pixel $v(x)$ is at the location x of the image in the given color space, $M(x)$ is the background sample dataset at the location x. $M(x)$ contains n pixel values selected from the neighborhood of the pixel $x, i = 1, 2, \ldots \ldots, n$, named the background sample value:

$$M(x) = \left\{v_1, v_2, \ldots \ldots, v_n\right\} \tag{3}$$

(2) Pixel classification

The pixel x in the current image is classified by comparing the pixel value $v(x)$ with the corresponding model $M(x)$ in the background model. The distance between the new pixel value and the value of the sample dataset is calculated. If the distance is less than the threshold R, the approximate number of samples is increased. If the approximate sample number is greater than the threshold N, the new pixel is considered as the background, as shown in Eq. (4).

$$\left\{S_R(v(x)) \cap \{v_1, v_2, \ldots \ldots, v_n\}\right\} \geq \min \tag{4}$$

Where $S_R(v(x))$ indicate that the pixel x is the center, R is the area radius, $N = 20$, $\min = 2$.

(3) Model updating

Model updating allows the background model to adapt to ever-changing in the background, such as changes in illumination, background changes and so on. The updating method adopted in ViBe algorithm is foreground spot counting method and random sub sampling. Foreground spot counting: the pixels are counted, and if a pixel is continuously detected as foreground times, it is updated to the background point. Random sub sampling: for each new image frame, it is not necessary to update the sample value of each pixel. when a pixel is classified as background, it has $1/\varphi$ probability to update the background model.

$$P(t, t + dt) = e^{-In\left(\frac{N}{N-1}\right)dt} \tag{5}$$

When a sample value to be replaced is selected, a sample value is updated randomly so that the life cycle of the sample value in the background model is exponentially monotonically decreasing. Since it is a random update, the probability $(N-1)/N$ that such a sample value will not be updated at t time. If the time is continuous, then the probability that the sample value remains after the dt time passes is as Eq. (5).

The Eq. (5) shows that whether a sample value is replaced in the model is unconcerned with time, and the random strategy is appropriate.

This is the result of fast removing background by using the ViBe algorithm (see Fig. 4). It can be seen that in different positions, the algorithm can effectively remove the noise of the background, access to the depth of the human body data.

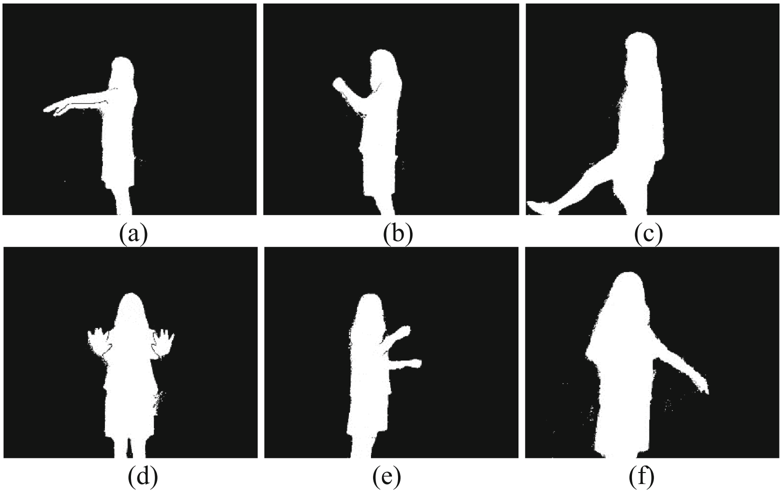


Fig. 4. Background removal result using ViBe algorithm

In the virtual interaction of a VR yacht simulator, the hand is always in front of the body. Thus, the depth value of the hand is always smaller than the depth value of the body in the depth image. According to this characteristic, the position of the shot can be calculated from the depth image and displayed selectively:

$$dst(x,y) = \begin{cases} src(x,y), \; if \; src(x,y) > threshold \\ \quad 0, \quad other \end{cases} \tag{6}$$

Using this method to get hand depth data in Fig. 5 shown here:

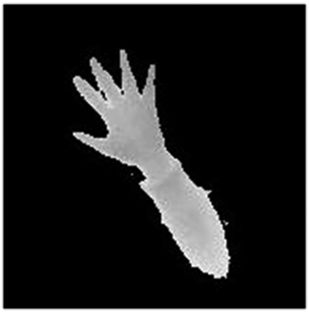


Fig. 5. Hand depth image using ViBe algorithm and threshold method

3.4 Improved De-noising Algorithm of Depth Image

(1) Bilateral filtering

When measuring the scene illumination changes to a certain extent, it will lead to unstable the depth data. And because of the surface material of the infrared reflection characteristics and other reasons, on the edge of the object and the occluded areas are prone to cavities in the depth image which obtained by Kinect sensor.

Because the depth values can be considered as continuous in a relatively small field, holes can be filled with valid depth values around the holes. Bilateral filtering is a nonlinear filtering method, which is a compromise between image spatial proximity and pixel similarity. This method takes into account both spatial information and gray similarity, thus keeps the edge information of the image and achieve the effect of noise reduction. It has the characteristics of simplicity, non-iteration, locality and so on.

There is the Bilateral filtering Eq. (7):

$$W_{ij} = \frac{1}{K_i} e^{-\frac{(x_j - x_i)^2}{\sigma_s^2}} \bullet e^{-\frac{(I_j - I_i)^2}{\sigma_r^2}} \tag{7}$$

Where W is the weight, i and j are pixel index, K is the normalization constant, σ_s is spatial standard deviation of Gaussian Function, σ_r is range for the standard deviation of the Gaussian Function, I is the pixel gray value.

According to the characteristics of exponential function, e should be a $f(x)$ monotonic decreasing function, so the weight will decrease and the filtering effect will be smaller when the gray range is large (such as edge). In general, in the region where the pixel gray level is too mild, bilateral filtering has the effect similar to the Gauss filter. While in the larger gradient of the image edge, the bilateral filter has the effect of retention.

For digital image, the filtering process is eventually represented as a form of convolution using a template, which in fact indicates that the output pixel values depend on the weighted combination of the values of the domain pixels (see Eq. (8)).

$$h(i,j) = \frac{\sum_{k,l} f(k,l)\omega(i,j,k,l)}{\sum_{k,l} \omega(i,j,k,l)} \tag{8}$$

The weight coefficient $\omega(i,j,k,l)$ (Eq. (11)) depends on the product of the domain kernel (Eq. (9)) and the range kernel (Eq. (10)).

$$d(i,j,k,l) = e^{-\frac{(i-k)^2 + (j-l)^2}{\sigma_s^2}} \tag{9}$$

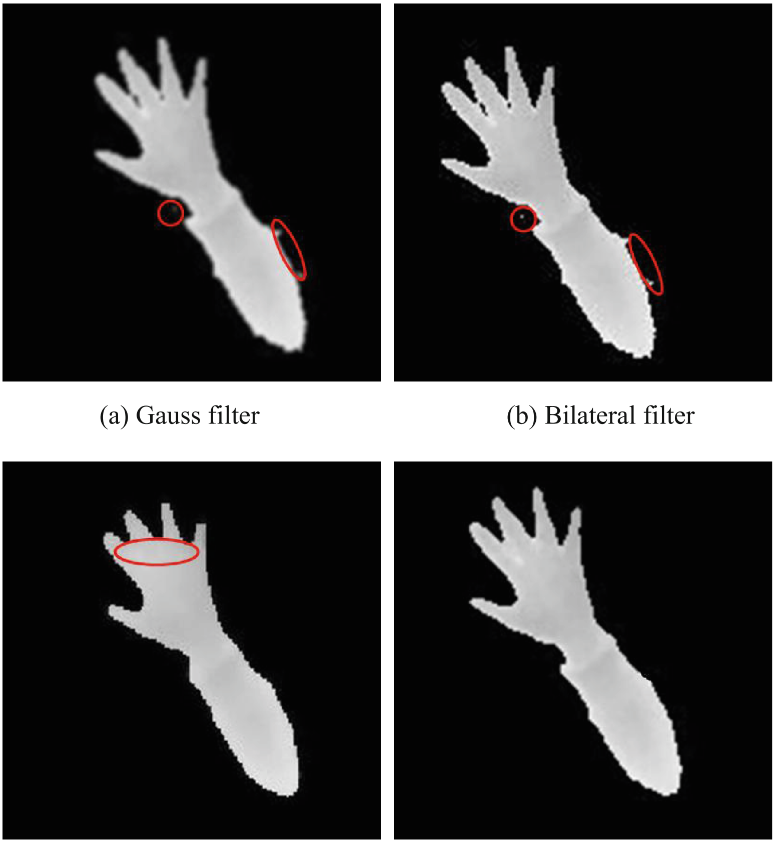
$$r(i,j,k,l) = e^{-\frac{\|f(i,j) - f(k,l)\|^2}{\sigma_r^2}} \tag{10}$$

$$\omega(i,j,k,l) = e^{-\frac{(i-k)^2+(j-l)^2}{\sigma_s^2}-\frac{\|f(i,j)-f(k,l)\|^2}{\sigma_r^2}} \tag{11}$$

(2) Improved de-noising algorithm of depth image
Following a hole repaired in using the bilateral filter, there are still grayscale isolated points which cannot be repaired and cannot be effectively used by the right pixel around the gray value to fill. Therefore, the median filter is used to smooth the depth image. The median filter replaces the value of each pixel with the median pixel value in the neighborhood. For larger isolated points, median values can be chosen to avoid the effects of these points.

There is the Bilateral filtering Eq. (12).

$$D(i,j) = \underset{\wedge_{\kappa}}{Med}(f(i,j)) \tag{12}$$



(a) Gauss filter (b) Bilateral filter

(c) Median filter (d) Bilateral filter+ Median filter

Fig. 6. Image smoothing result (Color figure online)

Where $\wedge_K$ is 3 × 3 filter window, $f(i,j)$ is the center pixel, $D(i,j)$ is the intermediate pixel values for the neighborhood.

Through median filtering, gray value deviation and the surrounding pixels will be larger median replacement, and eliminate the noise to achieve the purpose of de-noising the image.

Literature [9] Gauss filter de-noising effect is shown in Fig. 6(a), and it demonstrates that the part of the hand in the region is not correct in depth value, the repair effect is not very good as red circle shown in Fig. 6(a). Bilateral filtering results as shown in Fig. 6(b) is that the depth value is not homogeneous, and there are some isolated noises. The smoothing effect of the median filter is shown in Fig. 6(c), the gap between some fingers are disappear. After comparison, in this paper the proposed method of bilateral filtering and median filtering smoothing is better, as shown in Fig. 6(d).

4 Result

The process of the improved depth image smoothing algorithm in this paper is shown in Fig. 7.

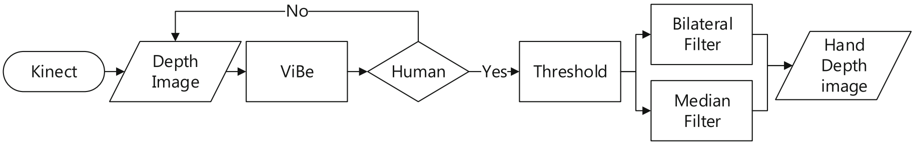


Fig. 7. Working process

The development environment of this paper is VC++ 2012 and Kinect for Windows SDK V2.0. The PC configurations used are Inter(R) Core(TM) i5-4590 CPU @ 3.30 GHz quad core processor, 8 GB memory, NVIDIA GeForce GTX 900 graphics card, Xbox ONE Kinect 2 sensor device. The experimental results are shown in Fig. 6(d).

In this paper, depth data is obtained by Kinect sensor. Hand depth data is acquired through ViBe algorithm and threshold method. And using bilateral filtering and median filtering apply to de-noise the depth image. Experimental results show that this method can effectively acquire high quality depth images in real-time.

5 Conclusion

In this paper, an improved method of depth image acquisition and processing is proposed in virtual interaction of yacht simulator. A single channel depth image obtained by a Kinect sensor's infrared camera has a one-to-one correspondence between the gray value and the distance from the camera to the object. Based on this characteristic, we use ViBe algorithm and threshold method to segment the background. After de-noising and smoothing by using bilateral filtering and median filtering, we obtain a clear and seamless high-quality depth image, which lays a solid foundation for the virtual interactive operation of the later VR yacht simulator.

In the future, the depth data obtained from this paper can be converted into 3D point cloud data, and a 3D hand model can be created in combination with Kinect RGB images to provide real-time 3D data for the virtual natural interaction of the yacht simulator.

Acknowledgment. The authors would like to acknowledge the support from the Fundamental Research Funds for the Central Universities [No. 3132016310] and Traffic Youth Science and Technology Talent Project [No. 36260401].

References

1. Han, J., Shao, L., Xu, D.: Enhanced computer vision with Microsoft Kinect sensor: a review. IEEE Trans. Cinematics **43**(5), 1318–1334 (2013)
2. Spinello, L., Arras, K.O.: People detection in RGB-D data. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3838–3843 (2011)
3. Xia L., Chen, C.-C, Aggarwal, J.K.: Human detection using depth information by 3 Kinect. In: 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 15–22 (2011)
4. Zhu, T., Lu, L.B., Jin, G.D.: Obstruction on-line detection algorithm based on Kinect in-depth technical. Electron. Des. Eng. **22**(12), 176–179 (2014)
5. Amara, A., Moats, T., Aloof, N.: GPU based GMM segmentation of Kinect data. In: 56th International Symposium ELMAR, pp. 1–4 (2014)
6. Murgia, J., Meurie, C., Ruichek, Y.: An improved colorimetric invariants and RGB-depth-based codebook model for background subtraction using Kinect. In: Mexican International Conference on Artificial Intelligence, pp. 380–392 (2014)
7. Barnich, O., Van Droogenbroeck, M.: ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans. Image Process. **20**(6), 1709–1724 (2011)
8. Yu, Y., Cao, M.W., Yue, F.E.: ViBe: an improved ViBe moving target detection algorithm. Chin. J. Sci. Instrum. **35**(34), 924–931 (2014)
9. Vijayanagar, K.R., Loghman, M., Kim, J.: Refinement of depth maps generated by low-cost depth sensors. In: Soc Design Conference, pp. 355–358. IEEE (2013)
10. Yang, N.E., Kim, Y.G., Park, R.H.: Depth hole filling using the depth distribution of neighboring regions of depth holes in the Kinect sensor. In: IEEE International Conference on Signal Processing, Communication and Computing, pp. 658–661. IEEE (2012)