

# Image Segmentation Method of RGB Image and Depth Image Based on Kinect

**Xiao Zhiguo** College of Computer Science and Technology, Changchun University, Changchun City, China

**Yang Yongji** College of Computer Science and Technology, Changchun University, Changchun City, China

**Received:** November 14, 2016; **Accepted:** December 5, 2016; **Published:** January 4, 2017

## Keywords

Kinect, Segmentation, Depth Image, RGB Image

The information of existing image segmentation algorithm is too little to achieve the desired results. Kinect depth camera can get the RGB image and depth image of the surrounding scene in real time, which brings a new research method for image segmentation and recognition. This paper proposed an integrating image segmentation method for RGB image and depth image based on Kinect. Integrating color and depth information, dynamic adaptive weighting method simpler and more effective compared to other methods, which provides an accurate criterion and robustness for the subsequent region merging.

## Introduction

The information of existing image segmentation algorithm is too little to achieve the desired results. Kinect depth camera can get the RGB image and depth image of the surrounding scene in real time, which brings a new research method for image segmentation and recognition. This paper proposed an integrating image segmentation method for RGB image and depth image based on Kinect. Integrating color and depth information, dynamic adaptive weighting method simpler and more effective compared to other methods, which provides an accurate criterion and robustness for the subsequent region merging. Image segmentation has always been an extensive subject in computer vision field [1–3], which has been widely used in object recognition, scene understanding, robot navigation, etc.. The image segmentation is bridge and link for low-level and advanced level visual, and the result of image segmentation depends on the pretreatment of low-level visual effect and determines the quality of the final output of various applications in advanced vision. Therefore, image segmentation is the most critical stage of most visual systems [4-6]. Researchers pointed out that the object is defined not only by the consistency of color, texture and other low-level features, but also by the physical connectivity which, in the three-dimensional space, presents the continuity of the depth in the scene [3]. In recent years, with the launched equipment such as Mesa Imaging SR 4K and Microsoft Kinect imaging, depth, as an additional feature, can improve the accuracy of image segmentation, such as when the target and background color are similar, or low contrast edges exists in the image, the depth information can distinguish between various objects and background. Visibly, the integration of depth and color information is an important way to solve the problem of image segmentation. The existing image segmentation algorithms are divided into three categories: one is based on the segmentation of regional color information; the second is based on the segmentation of the edge information; another is based on threshold segmentation. However, due to the less information, the desired results cannot be achieved. Classical regional segmentation algorithms include K mean clustering method [7], region growing method, watershed algorithm, etc. Regional segmentation method can obtain a closed region, which can also achieve very good results without a priori knowledge, but it often is complex and time-consuming. The segmentation of edge information mainly focuses on the improvement of the differential operator. The reference [8] proposed a two order Taylor expansion based on the differential operator to calculate the gradient of the image in a certain direction, which has more accurate edge location and better continuity; the reference [9] proposed an edge detection algorithm based on fuzzy theory which has a strong ability to describe the uncertainty of the image. Even when the edge is not clear, good segmentation results also can be obtained by establishing appropriate membership functions. The edge detection is sensitive to noise, so the edge segmentation algorithm on the one hand is to improve the anti-noise performance of the algorithm, and on the other hand is to form a good edge algorithm because of less continuity in the edge detection algorithm. In the

threshold segmentation, the most important is to determine the threshold, that is, how to quickly and accurately obtain the best threshold. The reference [10] introduces genetic algorithm whose feature can decide quickly to obtain the optimal threshold, and the time is greatly reduced compared with the conventional Otsu method; The reference [11] proposed an improved method based on particle swarm optimization to find the optimal threshold for the largest class of variance, which solves the problem that the multi threshold value is too large; The reference [12] focuses on remote sensing image segmentation threshold selection and fast noise processing algorithm, which carry out anti-noise processing into a neighborhood spatial and gray similarity measure factor, with maximum variance as SFLA fitness function, through the local search and global information exchange SFLA to quickly determine the optimal global threshold image segmentation the. It can be seen that the direction of the research of threshold algorithm is concentrated in the fast searching for the most threshold value.

Kinect depth camera can get the RGB image and depth image of the surrounding scene in real time, which brings a new research method for image segmentation and recognition. This paper proposed integrating image segmentation method for RGB image and depth image based on Kinect, by using segmentation framework of Maximal-Similarity Based Region Merging (MSRM) in reference [13], which makes the results more subjective and improves the accuracy of segmentation, based on the introduction of the depth information. Integrating color and depth information, dynamic adaptive weighting method simpler and more effective compared to other methods, which provides an accurate criterion and robustness for the subsequent region merging. Setting the initial target and the background seed region by using the 3D scene depth information, process of automatically selecting the seed zone replaces the process of artificial interactive marking, which is more robust than the method based on depth discontinuity edge.

## Depth Image and Color Image Alignment

### Calibration of Kinect

#### Internal Parameters of Depth Camera

The transformation relation between the depth camera coordinate and the depth image coordinate is similar to color ones. The distortion model of color camera is the forward model (the world coordinate to the image coordinate), and conveniently, the distortion model of the depth camera is the inverse model (the image coordinates to the world coordinates). The geometric distortion model of the depth camera can be described by the relationship between parallax  $d_k$  and depth  $z_k$ :

$$z_k = \frac{1}{c_1 d_k + c_0} \quad (1)$$

Where,  $c_1$  and  $c_0$  are calibration parameters in the depth camera. If the parallax measurements value  $d$  is calibrated into (1) directly, the depth information produces a fixed deviation in observation process which can be corrected by adding a drift  $z_\delta$ , effectively reducing the re-projection deviation [15], i.e.

$$z_{kk} = z_k + z_\delta(u, v) \quad (2)$$

In this paper, we try to directly modify the original disparity  $d$  to improve the calibration accuracy. A distortion model is constructed, which can offset the deviation caused by the increase of the apparent difference with an attenuation of the spatial offset.

$$d_k = d + D_\delta(u, v) \exp(\alpha_0 - \alpha_1 d) \quad (3)$$

Where,  $d$  is the parallax value obtained from the Kinect without correction,  $D_\delta$ , used to eliminate the influence of distortion, presents the spatial distortion related with each pixel, and  $d_k$  is the disparity value after correction.

Formula (1) (3) can be used to calculate the variation from parallax to depth, and the inverse of these equations can be used to calculate the difference of the re-projection, according to the equation (1):

$$d_k = \frac{1}{c_1 z_k} - \frac{c_0}{c_1} \quad (4)$$

Formula (3) has an exponential relationship, whose inverse is much more complex than that of the formula (1). Therefore, based on the formula of Taylor, formula (3) can be simplified as:

$$d_k = d + D_\delta(u, v) \exp(\alpha_0 - \alpha_1 d) \approx d + D_\delta(u, v) \exp(1 + \alpha_0 - \alpha_1 d) \quad (5)$$

so,

$$d = \frac{d_k - D_\delta - D_\delta \alpha_0}{1 - \alpha_1} \quad (6)$$

Therefore, the depth camera model can be described as  $L_d = \{f_d, p_{od}, k_d, c_0, c_1, D_\delta, \alpha\}$ , in which the first 3 parameters represent the focal length, radial distortion and tangential distortion of the lens, later 4 parameters are used to describe the changing relationship from parallax to depth.

Color camera calibration method has been very mature, which can initialize the camera parameters in Zhang [16]. Simply as follows:

- 1) Extraction of corner points on the color image of the calibration board taken by color camera;
- 2) Calculation of the single stress matrix by using the known corner position in the world coordinate system  $\{W_i\}$ ;
- 3) The camera parameters are constrained by a single stress matrix and are solved by using the linear equation, where the initial value of distortion coefficient is set to 0.

The same method can initialize the distortion parameters of the depth camera. However, due to the depth of the image in the corner of the board is not visible, 4 points are selected manually on the calibration board, and the single stress matrix is calculated between  $\{V_i\}$  and  $\{D\}$  in Zhang [16]. The deviation of these 4 points is very large, which is only suitable for obtaining the initial estimate (including the focal length, the principal point coordinates and the translation vector  ${}^V I_D$ ), and can be used to estimate the depth of each point. Based on the estimated depth values and the measured disparity values, the equation (1) establishes a linear system with an initial value for the depth parameter ( $c_0$  and  $c_1$ ).

## Solving Relative Pose

To depth camera, the coordinate system  $\{W_i\}$  and the coordinate system  $\{V_i\}$  is not aligned. In order to get the camera pose relative to the coordinate system of  $\{W_i\}$ , the coplanar relationship can be used for coordinates  $\{V_i\}$  and  $\{W_i\}$ , extraction of plane equation in each frame ( $n^T x - \delta = 0$  can define the plane, where  $n$  is the unit normal, and  $\delta$  is the distance from the origin) is taken as a constant. Dividing the rotation matrix by column,  $R = [r_1, r_2, r_3]$ , and  $n = [0, 0, 1]^T$  and  $\delta = 0$ , then the plane parameters of the camera coordinate system are

$$n = r_3 \text{ and } \delta = r_3^T t \quad (7)$$

The plane vector of color camera can be connected with the matrix  $M_c = [n_{c1}, n_{c2}, n_{c3}, \dots, n_{cn}]$  and  $b_c = [\delta_{c1}, \delta_{c2}, \dots, \delta_{cn}]$ , the same, the plane vector of depth camera can be connected with  $M_d$  and  $b_d$ . Then relationship from the color camera to the depth ones is

$${}^c P_D' = M_d M_c^T \quad (8)$$

$${}^c t_D = (M_c M_c^T)^{-1} M_c (b_c - b_d)^T \quad (9)$$

Finally, rotation matrix  ${}^c R_D = UV^T$  is the result from the SVD decomposition.

## Nonlinear Optimization

The objective of nonlinear optimization is to minimize the weighted sum of the squared deviation of all the parameters. By calculating the distance from the camera along the ray to the calibration plane, with the estimated original disparity  $\hat{d}$  in the formula (4) (6) and the reciprocal of the variance to be weighted, the final function is:

$$c = \frac{\sum_{\text{Number of pixels}} \left\| \hat{p}_c - p_c \right\|^2}{\delta_c^2} + \frac{\sum_{\text{Number of pixels}} \left\| \hat{d} - d \right\|^2}{\delta_d^2} \quad (10)$$

Where,  $\delta_c^2$  is the variance of the measurement deviation for the color camera and  $\delta_d^2$  is the variance of the measurement deviation for the depth camera. The equation (4) is a nonlinear equation with a very great number of parameters.

Therefore, the formula (10) is modified by equation (5) corrected disparity values  $d_k$  instead of the original measure parallax value  $d$ , by using the equation (4) obtained after correction of disparity estimate instead of the original parallax value, eventually separated parallax distortion parameters and other parameters, i.e.

$$c = \frac{\sum_{\text{Number of pixels}} \left\| \hat{p}_c - p_c \right\|^2}{\delta_c^2} + \frac{\sum_{\text{Number of pixels}} \left\| \hat{d}_k - d_k \right\|^2}{\delta_d^2} \quad (11)$$

## Aligning

P is a point in Kinect original depth image, whose coordinates  $(u_d, v_d)$ , the corresponding depth value is  $z_{raw}$ , the real depth value  $z_d$  related to P can be calculated in the formula (12):

$$Z_d = \frac{1}{c_0 z_{raw} + c_1} \quad (12)$$

$$\begin{cases} x_d = (u_d - cx_d) * z_d / fx_d \\ y_d = (v_d - cy_d) * z_d / fy_d \end{cases} \quad (13)$$

The point P  $(u_d, v_d)$  in the depth image coordinate system is converted to the world coordinate system in formula (13),  $P_d(x_d, y_d, z_d)$ .

Where,  $fx_d$  and  $fy_d$  are the focal length relatively for the x and y in depth camera,  $cx_d, cy_d$  are principal point offset.

Using the formula (14), P  $(x_d, y_d, z_d)$  in the world coordinate system is converted to  $P_c(x_c, y_c, z_c)$  in the color camera coordinate system by using the external parameters, including the rotation matrix R and the T.

$$P_c = RP_d + T \quad (14)$$

Formula (15) converts  $P_c$  in color camera coordinates to Q  $(u_c, v_c)$ , Q in color image is the corresponding point of P  $(u_d, v_d)$  in the depth of the image.

$$\begin{cases} u_c = x_c * fx_c / z_c + cx_c \\ v_c = y_c * fy_c / z_c + cy_c \end{cases} \quad (15)$$

## Region Merging Based on Maximum Similarity Figure 1 Shows This Algorithm

The first, Kinect calibration is aligned RGB image and depth image; secondly, depth image is optimized and de-noised; at last, the initial set on RGB image is obtained by final mean-shift algorithm. Calculating the similarity of color and depth of adjacent areas, total similarity can be a dynamically weighted, according to automatically selecting the seed region of target and background, which, finally, can be segmented by using MSRM [13].

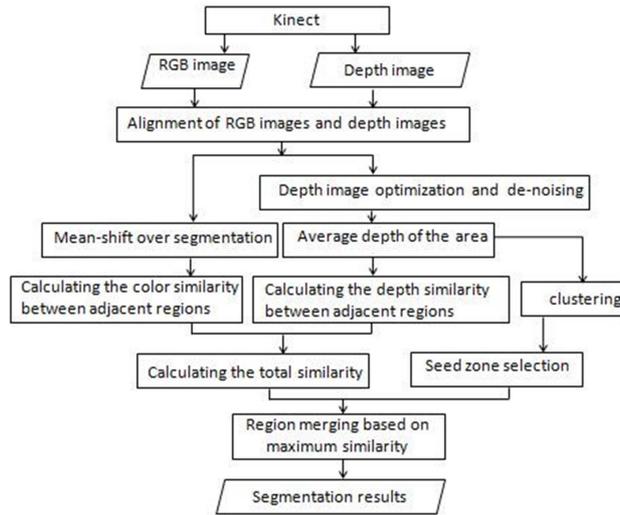


Fig. 1. Algorithm flow chart in this paper.

### Regional Representation and Similarity Measurement

Lab is a color system with device independent and based on physiological characteristics. It is suitable for image segmentation by using a digital method to describe the human visual perception. The L component of Lab is used to represent the pixel brightness range [0, 100], which stands for range from the black to pure white; a [127, -128] stands for range from red to green; b [127, -128] stands for range from yellow to blue. This paper converts the RGB image to Lab space, with the Lab color histogram as descriptors [18-19], and uses Bhattacharyya coefficient [13] to define the color similarity  $S_c$  between two regions Q and R:

$$S_c(R, Q) = \sum_{u=1}^{4096} \sqrt{H_{CR}^u H_{CQ}^u} \tag{16}$$

Where,  $H_{CR}$  and  $H_{CQ}$  are the normalized color histogram of regional R and Q respectively, and the superscript u represents the u element of the histogram.

In this paper, a method for calculating the regional depth similarity is proposed. Define the depth  $D^i$  of a regional i:

$$D^i = \frac{1}{n} (x_1 + x_2 + \dots + x_n) \tag{17}$$

Where, n is the number of pixels in the region i,  $\{x_1, x_2, \dots, x_n\}$  represents the set of the depth value of all pixels in the region.  $S_d$  is defined as the depth similarity of R and Q:

$$S_d(R, Q) = \frac{-|D^R - D^Q|}{\max_{i=1,2,\dots,RN} \{D^i\} - \min_{i=1,2,\dots,RN} \{D^i\}} \tag{18}$$

Where, RN is the number of the initial over segmented regions,  $\max_{i=1,2,\dots,RN} \{D^i\}$  is the maximum depth value of the region, and  $\min_{i=1,2,\dots,RN} \{D^i\}$  is the minimum depth value of the region except for 0. Because of occlusion, shadow and other reasons, part of the pixel depth information is uncertain, which is replaced by value 0 as black. At this time depth information is not reliable, this paper only consider the color similarity. In a region i, if the depth of a part of the pixel is 0, the depth of a part of the pixel value is not 0, then calculate the average depth of the region only consider the non-0 values of those image points. In equation (3),  $S_d \in [-1, 0]$  the bigger R and Q area depth difference is, the smaller their depth of similarity is. total regional similarity S is obtained by dynamic weighting method:

$$S(R, Q) = S_c(R, Q) + W S_d(R, Q) \tag{19}$$

Where, W, a weight factor, reflects contribution to the total similarity from the color similarity  $S_c$  and depth similarity  $S_d$ , when the adjacent color similarity is larger, the more the influence of depth information differences should be considered in

order to distinguish between similar color target and background region, that is, with the increase of the color similarity, the weight of the depth similarity increase too. According to the above principles and the negatives of depth similarity, the dynamic weight value is calculated as follows:

$$W = \frac{A}{0.1 + \exp\left(\frac{-(S_c - B)}{C}\right)} \tag{20}$$

This is a Sigmoid curve, A, B and C respectively, represents its maximum value, offset and steep degree. This paper gives A = 1, B = 0.2 and C = 0.5. According to  $S_c \in [0, 1]$ , figure 2 only draws a portion of the Sigmoid curve.

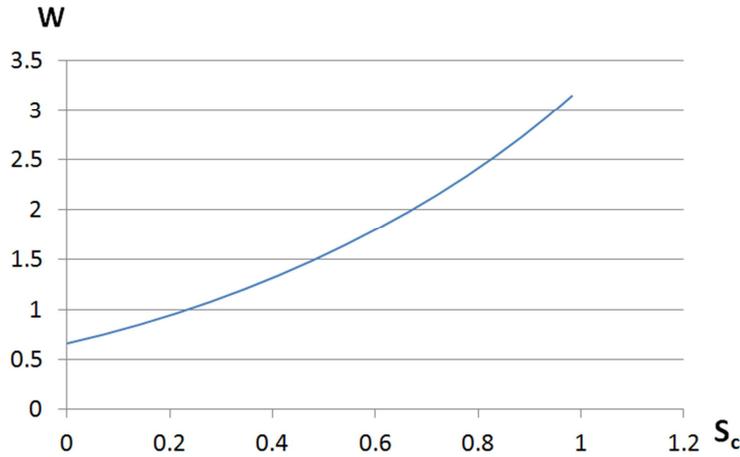


Fig. 2. The curve of the weight factor w with the change of the color similarity  $S_c$ .

### Seed Zone Selection

Before the regional merging, the target and the background area should be selected as the seed point. In many references, the method of manual interaction is directly labeled in the region, and method in the paper can automatically select the seed field according to the depth information of the image. Firstly, the results of over segmentation in RGB images are mapped to the depth image, and then the average depth of each region is calculated by the formula (2), and the  $D = \{D^i\}_{i=1, 2, \dots, R_N}$  is obtained. In general, the target and background in different depth, the over segmentation regions were clustered into two groups by using the K-means algorithm, namely the target class and background class. Then randomly select a part of the region as a seed point. The experimental results show that the more seed points, the better segmentation results, but when the number of the seed region is more than 2.7% of the total area, segmentation accuracy tends to be stable.

In order to ensure the rapidity of recognition, the Euclidean distance is preferred because the geometric meaning is clear and the algorithm is simple.

$$d(A, B) = \sqrt{\sum_{i=1}^n (a[i] - b[i])^2} \tag{21}$$

K-means clustering algorithm uses the deviation sum of squares and criterion function to evaluate the clustering performance.

$$E = \sum_{i=1}^k \sum_{p \in X_i} \|p - m_i\|^2 \tag{22}$$

### Region Merging Image Segmentation Based on Maximum Similarity

MSRM, from reference [13], in essence, is a kind of adaptive region merging algorithm. Iterative algorithm is better in convergence and is not necessary to set the similarity threshold in advance. The key steps in MSRM are similarity measurement, which has a great impact on the accuracy of the final segmentation. This paper accomplishes the image

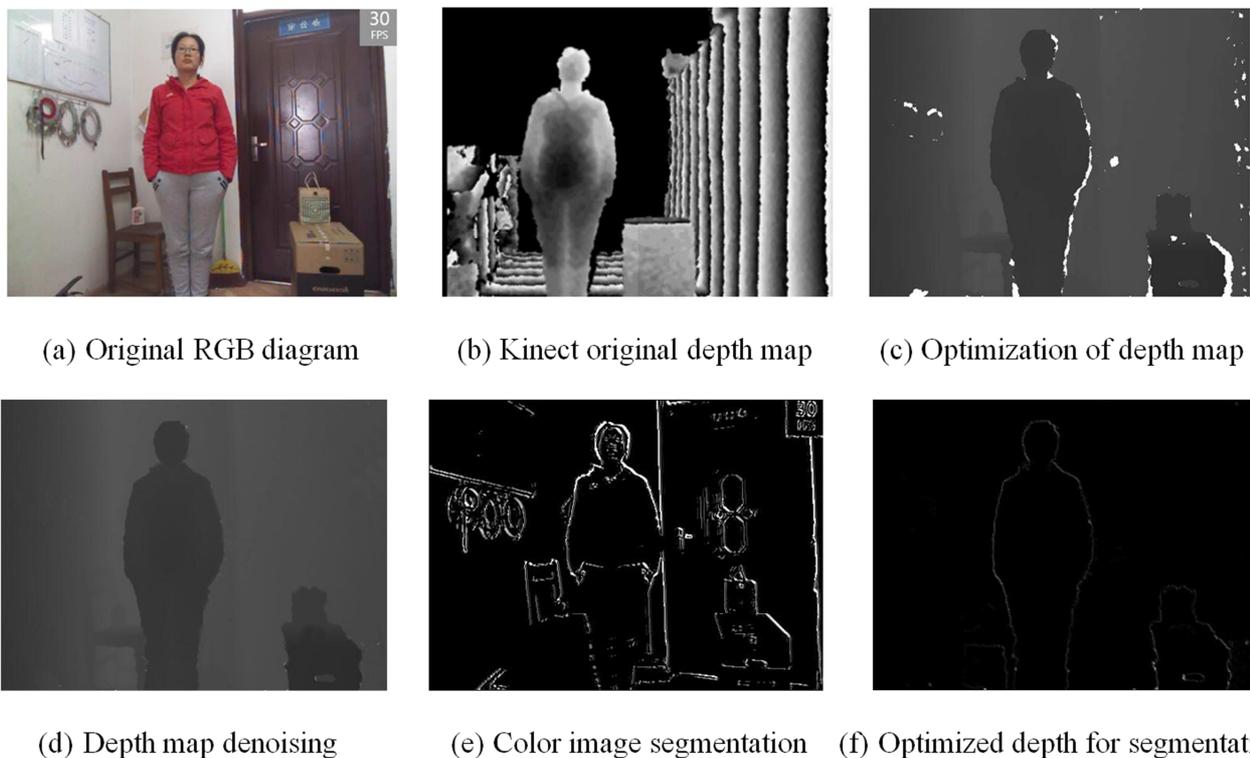
segmentation by using MSRM which has combined with the improved method to get the regional similarity and the seed region.

### Algorithm Complexity Analysis

This algorithm is an unsupervised image segmentation, which does not need offline learning, including over segmentation, the selection of seed region and region merging. Over segmentation uses mean-shift algorithm,  $O(n)$  for the time complexity,  $n$  for the number of pixels in the image. The seed region selection mainly depends on the K-means++ clustering algorithm,  $O(NK_t)$  is the time complexity, where,  $N$  is the number of initial over segmentation region,  $K$  is the number of clusters, and  $t$  is iterations. Generally,  $K \ll N$ ,  $t \ll N$ . Region merging takes MSRM, the time complexity takes  $O(N^2)$ . Based the above factors, the algorithm contains several steps, but the overall algorithm time complexity is not high.

### Experiment and Result Analysis

In order to verify the effectiveness of the proposed algorithm, RGB images and depth images have been shot by Kinect, after optimizing and denoising image, and the final color image and depth image have been segmented by K-means algorithm.



*Fig. 3. Segmentation results.*

According to experiment and analysis, it can be seen that the integrating image segmentation algorithm based on RGB and depth image is very accurate for foreground object segmentation and better than directly to the color image segmentation results. Because Kinect is a depth camera, application can take the foreground image as the integration image to segment and recognise, setting a threshold, the vision image can be segmented and identified using the original classical image segmentation algorithm. ■



### Xiao Zhiguo

He graduated at Changchun University in 2000. Now, he is a teacher of College of computer science and technology. Has been engaged in computer applications on the direction of research. Email: 398837379@qq.com



### Yang Yongji

He graduated at Changchun University in 1997. Now, he is a teacher of College of computer science and technology. Has been engaged in computer applications on the direction of research. Email: 784244306 @qq.com

## References

- [1] ZHOU C B, LIU C C. Semantic Image Segmentation Using Low-Level Features and Contextual Cues. *Computer & Electrical Engineering*, 2014, 40 (3): 844-857.
- [2] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor Segmentation and Support Inference from RGBD Images // *Proc of the 12th European Conference on Computer Vision*. Florence, Italy, 2012, V: 746-760.
- [3] HOIEM D, EFROS A A, HEBERT M. Recovering Occlusion Boundaries from an Image. *International Journal of Computer Vision*, 2011, 91 (3): 328-346.
- [4] R. C. Gonzalez, R. E. Woods. *Digital Image Processing [M]*. 2<sup>nd</sup> Edition. Beijing: Publishing House of Electronics Industry, 2002.
- [5] N. R. Pal, S. K. Pal. A review on image segmentation techniques [j]. *Pattern Recognition*, 1993, vol. 26: 1227-1249.
- [6] R. C. Gonzalez, R. E. Woods. *Digital Image Processing [M]*. 2<sup>nd</sup> Edition. Beijing: Publishing House of Electronics Industry, 2002.
- [7] Yao, Lingyan. *Image classification and segmentation algorithm based on Clustering [D]*. Tianjin: Tianjin University, 2011. 7-8.
- [8] Lv Meng, Su Hongqi, Liu Qichun. A new adaptive edge extraction differential operator [J]. *Data Acquisition And Processing*. 2011, 26 (1): 106-110.
- [9] Fan Guoliang, Su Hongqi, Wang Chunxia. Research of image region segmentation algorithm based on fuzzy enhancement [J]. *Computer Engineering And Design*. 2012, 33 (4): 1463-1466.
- [10] He Chunhua, Hu Yingchun. Automatic threshold image segmentation method based on Improved Genetic Algorithm [J]. *Computer Simulation*. 2011, 28 (2): 312-315.
- [11] Peng Zhengtao, Fang Kangling, Su zhiqi. Fast multi threshold image segmentation based on improved algorithm [J]. *Modern Electronic Technology*. 2011, 34 (6): 10-14.
- [12] Lu Binbin, Jia Zhenhong, He Di. Remote sensing image segmentation method based on improved shuffled frog leaping algorithm [J]. *Computer applications and software*. 2011, 28 (9): 77-79.
- [13] NING J F, ZHANG L, ZHANG D, et al. Interactive Image Segmentation by Maximal Similarity Based Region Merging. *Pattern Recognition*, 2010, 43 (2): 445-456.
- [14] Herrera C, Kannala J. Joint depth and color camera calibration with distortion correction [j]. *pattern Analysis and Machine Intelligence*, 2012, 34 (10): 2058-2064.
- [15] Smisek J, Jancosek M, Pajdla T. *3D With Kinect [M]*. London: Consumer Depth Cameras for Computer Vision, 2013: 3-25.
- [16] Zhang Z. A flexible new technique for camera calibration [J]. *Pattern Analysis and Machine Intelligence*, 2000, 22 (11): 1330-1334.
- [17] Raposo C, Barreto J P, Nunes U. Fast and accurate calibration of a Kinect senso [C] // *Proceedings of International Conference on 3DTV*. Seattle, WA: IEEE, 2013: 342-349.
- [18] Caron G, Eynard D. Multiple camera types simultaneous stereo calibration [C] // *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011: 2933-2938.
- [19] Smisek J, Jancosek M, Pajdla T. *3D with Kinect [M]* // *Consumer Depth Cameras for Computer Vision*. Springer London, 2013: 3-25.