Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6400658, 10 pages
https://doi.org/10.1155/2021/6400658

Research Article

Human Motion Tracking Algorithm Based on Image Segmentation Algorithm and Kinect Depth Information

Zuo Wu

Department of Public Physical, Changchun Humanities and Sciences College, Changchun, 130117 Jilin, China

Correspondence should be addressed to Zuo Wu; wuzuo@ccrw.edu.cn

Received 12 October 2021; Revised 3 November 2021; Accepted 10 November 2021; Published 23 November 2021

Academic Editor: Sang-Bing Tsai

Human motion recognition has an important application value in scenarios such as intelligent monitoring and advanced human-computer interaction, and it is an important research direction in the field of computer vision. Traditional human motion recognition algorithms based on two-dimensional cameras are susceptible to changes in light intensity and texture. The advent of depth sensors, especially the Kinect series with good performance and low price released by Microsoft, enables extensive research based on depth information. However, to a large extent, the depth information has not overcome these problems based on two-dimensional images. This article introduces the research background and significance of human motion recognition technology based on depth information, introduces in detail the research methods of human motion recognition algorithms based on depth information at home and abroad, and analyzes their advantages and disadvantages. The public dataset is introduced. Then, based on the depth information, a method of human motion recognition is proposed and optimized. A moving human body image segmentation method based on an improved two-dimensional Otsu method is proposed to solve the problem of inaccurate and slow segmentation of moving human body images using the two-dimensional Otsu method. In the process of constructing the threshold recognition function, this algorithm not only uses the cohesion of the pixels within the class but also considers the maximum variance between the target class and the background class. Then, the quantum particle swarm algorithm is used to find the optimal threshold solution of the threshold recognition function. Finally, the optimal solution is used to achieve accurate and fast image segmentation, which increases the accuracy of human body motion tracking by more than 30%.

1. Introduction

In the field of computer vision, depth information provides more possibilities for various computer vision applications such as human-computer interaction, three-dimensional scene reconstruction, and 3D printing. In recent years, optical sensor technology has continued to develop, and it has become a reality to use sensors to obtain depth information of a three-dimensional scene. Depth images are similar to grayscale images, and each pixel value indicates the distance between the surface of the object in the scene and the sensor. Traditional in-depth information acquisition devices are usually bulky, complicated in operation, and costly [1, 2]. However, the advent of Microsoft Kinect has changed this situation. Because of its low price, small size, and easy operation, the application of Kinect has rapidly spread to many fields, including medical treatment, games, robotics, and video conferencing.

Although Kinect has a wide range of applications and its advantages have also brought a lot of help to the related scientific research work, it still cannot get rid of some of its own shortcomings, mainly because the depth data collected by it will be missing, resulting in the depth image to be distorted. This problem mainly comes from the limitation of the working principle of Kinect itself. High-quality images are the foundation of excellent computer vision applications. Similarly, high-quality depth images are essential in the related application fields of depth images. The acquisition of

high-quality depth images mainly depends on the performance of hardware devices and effective image enhancement algorithms.

To obtain the image trajectory, Zeng J has extensively expanded the field of image sequence analysis. Generally, the designed system should match the application itself. However, in reality, there is very little information that can be extracted based on the constraints obtained from the application, which makes the algorithm more or less complicated. Ultimately, the expansion of the image depends on the study of time events and events. However, he did not propose a final solution, and the research has not been completed yet [3]. To solve the shortcomings of the instability of the original depth data received by the Kinect device, especially the problems of noise and large holes, Wang Fuwei proposed a depth recovery algorithm that combines the color map to control the filling and filtering of the holes. For the noise at the edge of the image, this is an improved average filtering algorithm to filter the noise. The gray image is used to separate the cavity area and mark it, analyze and determine the position of the cavity point, and fill the cavity with a color card. Compared with the current main depth filter repair methods, this algorithm makes full use of the depth data and color images in the processing process, and it uses a large number of basic image processing algorithms. Experimental results show that the algorithm solves the problems of holes and noise and has a good ability to repair deep data. However, the algorithm is more complicated, the speed of obtaining the results is slow, and some improvements are needed [4]. Li Y believes that target tracking under microscopic spectroscopy is a hot topic in computer vision. Because of its motion characteristics, small area jitter, and difficulty in segmentation, he proposed an improved tracking algorithm for multiple abnormal targets in microscopic images. Based on obtaining the gradient vector flow in the jitter region of the image, the original images during the tracking process of multiple abnormal targets are reconstructed. Through the derivation of the motion process, a more accurate curve of the two-dimensional image composed of the feature points of the image segmentation boundary is obtained. In the process of tracking the contour of the moving target under the shaking state, combined with the statistical theory, the parameter model of the Gaussian distribution of the moving image background is established. Experimental results show that the improved tracking algorithm can effectively improve the robustness of the tracking algorithm [5]. However, the accuracy of the algorithm is not high, and it cannot effectively capture the target [6, 7].

This article outlines the research difficulties and common research methods of human motion recognition, introduces the research background and significance of human motion recognition, and explains the reasons for choosing to use depth information for human motion recognition [8]. Then, it summarizes the principles, advantages, and disadvantages of the three mainstream methods for obtaining depth and outlines the classification basis and main categories of the human motion recognition methods based on depth information, and then, deep research and discussion were carried out from two aspects of human motion recognition based on depth map and skeleton joint points. two aspects of human motion recognition of data comprehensively expound the human motion recognition algorithm based on depth information. Firstly, this method extracts the outline of the human body in the foreground from the depth image, and then, it extracts the body skeleton model according to whether the body has occluded by the limbs. Secondly, it locates the position of each skeleton joint point on the extracted skeleton model. Finally, the method is also used in the MSR. Experiments were conducted on the Action 3D dataset and UTD-MHAD dataset and compared with the previous results.

2. Kinect Data Collection Method

2.1. Introduction to Kinect. The Kinect device was launched by Microsoft in November 2010. Microsoft initially designed and implemented this device to meet people's demand for somatosensory games [9, 10]. Kinect is only used as a control system for somatosensory games. After that, because of its advantages in price and performance, it has attracted people's attention and has been widely used in many fields [11, 12]. At present, the Kinect device has been developed to the second generation, which contains a color camera and a depth sensor, which can capture color images of the current measurement scene. The depth sensor consists of an infrared transmitter and an infrared receiver. The depth, color, and infrared image of the scene can be accessed through Kinect, as shown in Figure 1.

Compared with the first-generation Kinect, the second-generation Kinect has a color camera with better performance parameters. The resolution of the camera is 1920×1080, while the resolution of the first generation is 1280×960. The second-generation Kinect has a larger viewing angle, with a horizontal viewing angle of 70 and a vertical viewing angle of 60; the second-generation Kinect has a higher data transmission speed and supports maximum depth data acquisition at 60 frames per second [13, 14]. The second-generation Kinect depth data also has the advantages of small motion blur and high dynamic range. The fidelity of the second-generation Kinect is three times that of the first-generation Kinect, and hence, it is easier to capture more details of the objects in the scene. However, the depth resolution of the second-generation Kinect is 512×424, which is smaller than the first-generation 640×480, and the acquired depth images inevitably contain a lot of noise as shown in Table 1.

2.2. Human Motion Recognition Method Based on Depth Map. With the application of depth sensors, the ability of the computer systems to obtain a scene's three-dimensional information and low-level visual information has been greatly improved. It is fundamentally different from RGB data. The pixel values in depth maps encode the distance information of the scene rather than the light intensity of a certain color light [15, 16]. The texture information and color information contained in the depth map are lesser than

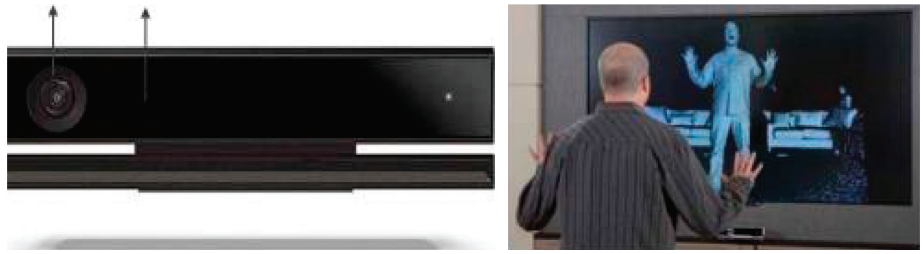


Figure 1: The appearance and structure of the second-generation Kinect and an example of obtaining depth images (the picture comes from https://image.baidu.com/).

Table 1: The comparison of parameters between first-generation Kinect and second-generation Kinect.

Kinect features	Generation Kinect	Kinect II
Perspective	Horizontal 57.5° Vertical 43.5°	Horizontal 70° Vertical 60°
Depth effective value	0.5–5.0 m	0.5–5.0 m
Color camera resolution	1280 × 960	1920 × 1080
Depth camera resolution	640 × 480	512 × 424
Infrared camera resolution	1280 × 960	512 × 424

those contained in the RGB data. An RGB image of a certain frame in an action image sequence and its corresponding depth image [17, 18]. The depth map provides advantages that many RGB images do not have: it can work in a low-light and dim environment and will not be affected by changes in color information and texture information. In a depth map, the depth silhouettes of an observation object can be obtained more accurately and conveniently. However, similar to other data forms, the three-dimensional contour obtained by depth information is also affected by factors such as flicker noise and object occlusion, which are likely to have a negative impact on the effect of motion recognition. Therefore, to solve these problems, many robust feature extraction methods based on depth maps have been proposed [19, 20].

2.3. Human Motion Recognition Method Based on Skeleton Joint Point Data. In addition to depth maps, there is another data source in mainstream modeling methods based on depth information-skeleton joints. As early as 1975, Johansson's research work showed that most of the human motion can be represented by joint positions alone. After that, the human motion recognition method based on the skeleton model began to be widely used. The skeleton joint points can encode the three-dimensional position information of the human body joint points in real time in each frame. Compared with modeling the skeleton structure using RGB data, modeling the skeleton using depth data is more efficient and stable. The basic ideas of these methods are similar. The human body is divided into multiple parts with depth data and similar degree label [21, 22]. The segmentation process of human body parts uses the contribution value of each pixel in the space model to calculate the position information of the three-dimensional connection point, which can be regarded as the classification of each pixel in the depth data. In terms of providing data points of human skeleton joint points, the Kinect camera launched by Microsoft has provided great help to researchers. The first-generation Kinect can obtain the data of 20 skeleton joint points per frame. In the second generation, the function of Kinect is more powerful as it can get the data of 25 skeleton joint points every frame. However, the effect of Kinect is not impeccable. When encountering occlusion or when the human body is facing the camera from the side, the data obtained is often accompanied by noise or even errors [23, 24]. The direct use of skeleton joint point data does not bring good results, and hence, it is necessary to develop a method based on the skeleton joint point data that is robust to occlusion and noise.

3. Kinect Depth Information Tracking Algorithm Correlation Experiment

3.1. Algorithm Evaluation Criteria. Normally, after denoising the image, to verify the effectiveness of the denoising algorithm, it is necessary to evaluate the denoising effect. The evaluation methods of the image denoising effect are generally divided into two types: subjective evaluation methods and objective evaluation methods. The subjective evaluation is mainly based on people's subjective feelings to evaluate the quality of an image, whereas the objective evaluation uses related mathematical models and mathematical formulas to obtain some quantitative indicators using calculations to evaluate the quality of the image. The mathematical model for an objective evaluation can be obtained by calculating the similarity between the normal high-quality image and the denoising processed image. In short, the greater the similarity, the smaller the difference between the two. The pixel value after denoising is closer to the normal value, indicating that the denoising algorithm is more effective. The traditional methods of measuring image similarity include average absolute error, mean square error, signal-to-noise ratio, and so on.

3.1.1. Mean Absolute Error. The calculation method is as follows: firstly, obtain the sum of the absolute value of the pixel difference between the original image and the denoising processed image. Then, divide it by the number of all pixels in the image. It can be seen from the formula that when the value of the average absolute error is smaller, the

performance of the algorithm is better. The formula is shown as follows:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} I(i,j) - G(i,j)}{M * N}. \tag{1}$$

MAE represents the calculated average absolute error. M and N, respectively, represent the length and width of the image. $I(i, j)$ and $G(i, j)$ are the original image and denoised image at point (i, j), respectively [25, 26].

3.1.2. Mean Square Error. The mean square error is a commonly used evaluation index for image denoising algorithms. The calculation method is shown as follows:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} I(i,j) + f(x,y) - b}{M * N}. \tag{2}$$

In the formula, MSE is the calculated mean square error [27, 28]. The signal-to-noise ratio is also a commonly used discrimination method. The difference from the above two methods is that the larger the value, the more effective the denoising algorithm. The calculation method is shown as follows:

$$\mathrm{PSNR} = 10\log_{10}\left[\frac{\sum_{i=1}^{M}\sum_{j=1}^{N} I^2(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[I(i,j) - I\prime(i,j)\right]^2}\right]. \tag{3}$$

The most common noise interferences in an image are the multiplicative noise, Gaussian noise, Poisson noise, and salt and pepper noise. The purpose of our research on image filtering algorithms is to reduce the impact of noise on images, try to preserve the details of the image, and obtain better visual effects. Generally, image filtering can be expressed by the following formula:

$$I\prime(x,y) = \frac{1}{w_p}\sum_{i,j\in\Omega} w(i,j) * I(i,j). \tag{4}$$

In the formula, I is the image after denoising processing, Ω is the neighborhood of the pixel (x, y), usually a rectangular area with (x, y) as the center point, $w(i, j)$ is the weight function of the filter at the point (i, j), w_p is the normalization parameter, and I is the noise image, where

$$W_P = \sum_{i,j\in\Omega} w(i,j). \tag{5}$$

After denoising processing, the pixel value of each pixel can be obtained by equation (5):

$$GI_P = \frac{\sum_{q\in S\cap D_q\neq 0} G_\sigma(\|p-q\|)I}{\sum_{q\in S\cap D_q\neq 0} p-q},$$
$$G_\sigma(\|p-q\|) = \exp\left(-\frac{(p_x-q_x)^2 + (p_y-q_y)^2}{2\sigma^2}\right). \tag{6}$$

3.2. Bilateral Filtering. The bilateral filter proposed by Tomaasi and Mandeuchi et al. is a nonlinear filter that can save the edge information in the image while denoising [29, 30]. The bilateral filter can not only smoothen the image but also preserve the edge information fully, including texture details, on this basis.

$$BI_p = \frac{1}{W_p}\sum_{q\in S} G_\sigma(\|Q-P\|)G_\sigma\left(\left\|I_p - I_q\right\|\right),$$
$$W_p = \sum_{q\in S\cap D_q} G_\sigma(\|p-q\|)\frac{I-b}{p+q}. \tag{7}$$

It determines the degree of influence of pixels in the neighborhood with pixel values different from p on the result.

$$G_\sigma(\|p-q\|) = \exp\left(-\frac{\|p-q\|^2}{2\sigma_x}\right),$$
$$I_p - I_q = \exp\left(-\frac{p-q}{2\sigma^2}\right). \tag{8}$$

Among them, p–q represents the Euclidean distance between the pixel point p and the pixel point q, and their size determines the actual application effect of the bilateral filter. From the aforementioned formulas of Gaussian filter and bilateral filter, it can be seen that the weight coefficient of the Gaussian filter depends on the spatial distance, which is a fixed value in the filter window [31, 32], whereas the weight coefficient of the bilateral filter depends on the space difference and pixel difference. The final filtering effect is determined by both, and hence, the weight coefficient in the filtering window is not a fixed value. Figure 2 shows the design process of the threshold recognition function of this method.

When the neighborhood is small enough, the following formula can be used to estimate the depth of the hole point. The depth value estimation formula of the pixel point p is as follows:

$$D_p = \frac{\sum_{q\in B} w(p,q)\lfloor D_q + \Delta(p,q)\rfloor}{w(p,q)}. \tag{9}$$

It is used to measure the similarity between the point p and the pixels in the neighborhood. This design makes it possible to spread the detailed structure information in the neighborhood $B(p)$ while filling the pixel value of the p point [33].

4. Human Motion Tracking Algorithm

4.1. Image Segmentation Algorithm Based on Graph Theory. Moorer et al. proposed an unsupervised image segmentation algorithm for a superpixel lattice. This method describes an algorithm (greedy algorithm) that can maintain the topological structure of the image and adds constraint conditions (topological information of the image). The definition of a superpixel crystal array is as follows: firstly, divide the image horizontally and vertically. Each path divides the image into two pixels, gradually increasing to 4 superpixels, as shown in Figure 3.

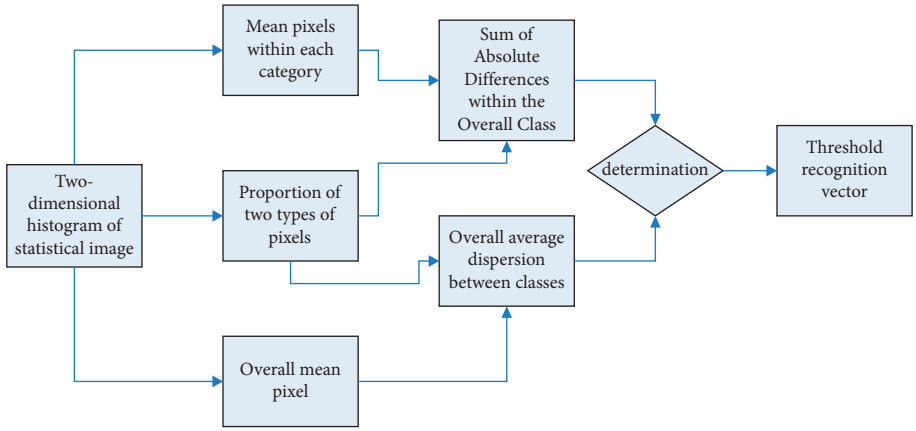
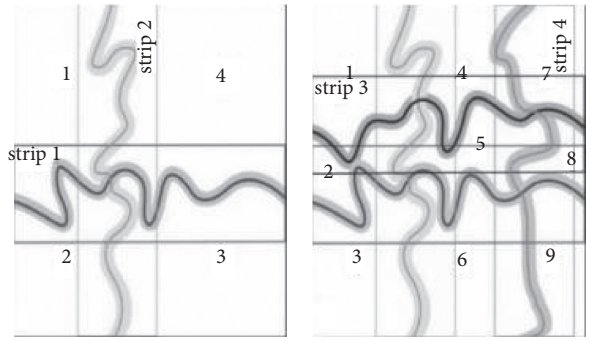


Figure 2: The design process of the threshold recognition function in this paper.



(a) (b)

Figure 3: (a) Divide the image into 4 areas (up, down, left, and right). (b) Add horizontal and vertical paths to divide the image into 9 areas (the picture comes from https://image.baidu.com/).

In the experiment, we implemented the segmentation algorithm and compared it with other algorithms. From the experimental results, we can see that in test 1 and test 2, the algorithm proposed in this article has a higher accuracy rate than other algorithms. In test 3, the accuracy of the algorithm proposed in this paper is slightly lower than that of other algorithms in the literature. The reason for this situation is that in the cross-subject test, the training data and the test data are completed by different testers, although all personnel are advised to express in a similar way as much as possible when the dataset is collected. The action could be the same, but different people's expressions of the same action will still be different. For example, when doing a wave motion, different people may swing at different speeds. When the training data is used as the training data, there is a big difference in the swing speed of the people used as the test data. At this time, the accuracy of the algorithm will be greatly affected.

As shown in Table 2, in terms of operating efficiency, the algorithm proposed in this paper is significantly faster than its speed. However, in terms of accuracy, there is almost no difference between the accuracy of our algorithm and that of other algorithms. In the abovementioned experimental environment, because of the proposed algorithm in the other algorithms, the feature descriptor has a high dimensionality. Hence, the linear discriminant analysis (LDA) is needed for dimensionality reduction. This process takes a lot of time, and some data with important information may be removed as noise. The feature dimensionality proposed in this paper is relatively low, no dimensionality reduction algorithm is needed, and problems such as information loss are avoided at the same time. Therefore, the algorithm proposed in this paper is much better than other algorithms in terms of running time efficiency.

To further verify the algorithm of this paper, we also use MATLAB R2015b to run the experiment on the same PC. For this dataset, we use Leave-one-out Cross Validation (LOOCV), which means that every time all the sequences in the data set run an experiment, one sequence is reserved as the test sequence and all other sequences are used as the training data of the model. This testing method takes a lot of time, but the results obtained are the most reliable.

The experimental results are shown in Figure 4. When the algorithm proposed in this paper recognizes the body movements, such as jogging and squatting, the effect is not ideal. This is because our features do not include these parts at the end of the human body into the extracted information. It is caused by the scope of the algorithm, however, in general, this does not affect the overall effect of the algorithm.

Table 2: The experimental results and comparison of MSR-Action3D dataset.

	Test 1		Test 2		Test 3	
	Other algorithms (%)	This article (%)	Other algorithms (%)	This article (%)	Other algorithms (%)	This article (%)
AS1	86.88	85.31	84.28	84.34	70.32	69.83
AS2	84.48	83.44	85.67	84.56	79.39	78.61
AS3	85.32	87.27	89.97	90.03	61.81	60.93
Overview	85.56	85.34	86.64	86.31	70.52	69.79

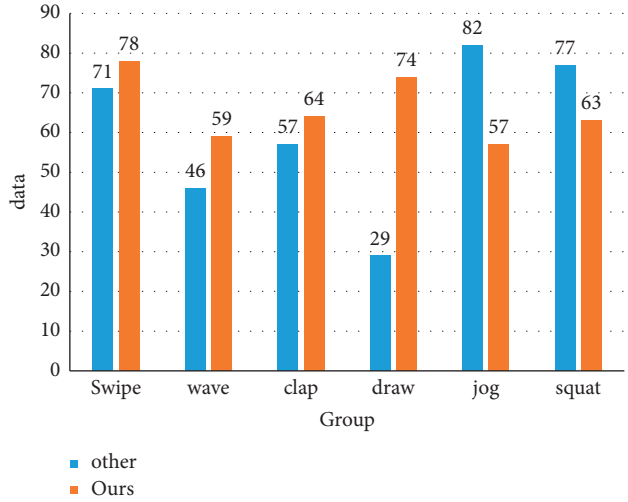


Figure 4: The body motion recognition effect of each algorithm.

The focus of this article is on the detection and recognition of the human body. The acquired depth image does not directly contain the human body information. This requires us to extract the contour of the human body to reduce the interference of the environment and reduce the algorithm at the same time.

Table 3 shows the execution time of each task on different processors at the maximum frequency for the tasks in the application in Figure 5. For example, the values 14, 16, and 19 in the first column indicate that the task n_1 is in the processor at the maximum frequency, with execution time on u_1, u_2, and u_3.

Figure 5 shows the priority values of all tasks in the DAG. Before task scheduling, we need to determine the task scheduling order. HEFT defines the level as the priority of task n and arranges all tasks in a descending order according to the priority.

4.2. Modular 2DUDP Method of Moving Human Body Recognition. The research on the image segmentation method of moving the human body in a complex environment is mainly used to detect the traffic on the bus site. It is mainly used for the recognition of the head in the passenger video-monitoring system to realize the effective management of the vehicle operation status, especially the toll. Since the main target image captured in the video is the human head when the passengers get on and off the vehicle, the segmentation algorithm mainly separates the human head (mainly the top of the human head) from the background for later recognition and tracking. In the in-vehicle video surveillance system, the passengers often bring some personal belongings into the car, and certain items are similar in shape or color to human heads, such as basketballs and pets. At the same time, in this type of monitoring system, the real time monitoring requirements are relatively high, and hence, the use of accurate and fast segmentation methods is the key to realize the segmentation of moving human images. It is also the key to know whether the subsequent target recognition and tracking are effective.

The purpose of moving the human body image segmentation is to segment the moving foreground area. Since the passenger transportation video needs to track the human body movement, it is necessary to identify the human body in the foreground. The moving human body recognition mainly refers to the use of certain methods to classify the detected or segmented human body and nonhuman body.

According to Table 4, it can be seen that for the second image under the traffic video, the MQPSO algorithm proposed in this paper is obtained by the traditional CQPSO algorithm segmentation, and its effect is similar to that obtained by the one-dimensional Otsu algorithm and the two-dimensional Otsu algorithm. In the case of graph effects, the speed of segmentation is further improved. It can be seen that the segmentation image obtained by the MQPSO algorithm proposed in this article is more accurate than the image obtained by the one-dimensional Otsu algorithm and the traditional CQPSO algorithm. For human head segmentation in traffic video, the outline of the top edge of the head is more obvious, and the hair details are

Table 3: The processor's parameters for each subtask in the application.

Task	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
U1	14	13	11	13	12	13	7	5	18	21
U2	16	19	13	8	13	16	15	11	12	7
U3	9	18	19	17	10	9	11	14	20	16

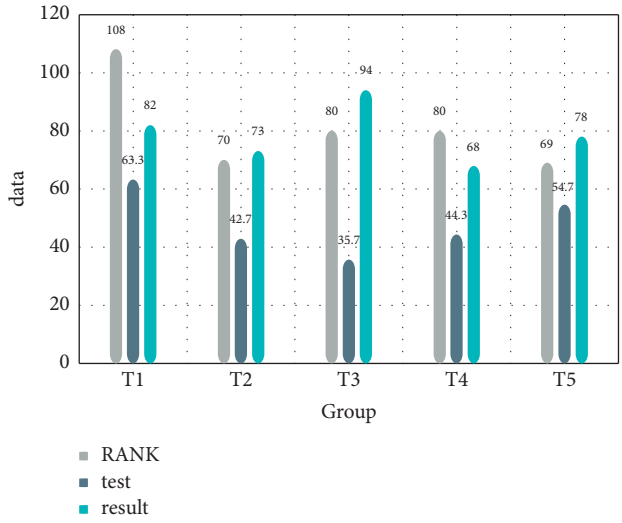


Figure 5: The priority of each task in the algorithm.

Table 4: Performance comparison of the three algorithms.

The second image under the traffic video	Threshold pair	Calculation time (s)
Classical two-dimensional Otsu method	(126, 126)	0.14299
CQPSO	(126, 126)	0.13802
MQPSO	(73, 100)	0.12901

clearer, which is more conducive to target recognition and tracking in the later stage. In the Lena image, the facial features are more obvious, the lower lip of the character is seen, the outline of the hat brim is clearer, and the details of the hat ornaments are also clearer.

As can be seen from Figure 6, the recognition rates of Modular 2DUDP $(3 * 1)$ method, Modular 2DUDP $(3 * 2)$ method, and Modular 2DUDP $(3 * 4)$ method are all above 85%, among them, the recognition rate of Modular 2DUDP $(3 * 2)$ method and Modular 2DUDP $(3 * 4)$ method is relatively stable, tending to over 90%. It can be seen from the experimental results that Modular 2DUDP has a higher recognition rate and better stability. In order to verify the effectiveness of the segmentation algorithm proposed in this paper based on the combination of the improved two-dimensional Otsu method and the quantum particle swarm algorithm, the module 2DUDP method proposed in this paper is effective in the combination of segmentation algorithm and feature extraction algorithm for head recognition. This article uses 50 overhead images extracted when getting in the car for comparison experiments. Firstly, the images of 50 people getting on and off the car are segmented into the human head image according to the above segmentation algorithm. Then, the feature extraction method of this article is used to extract the features of the image. Finally, the Euclidean distance classifier is used to classify and calculate the moving human body. The result is shown in Figure 7.

Here, the disparity data in the Middlebury stereo matching dataset is still used as the actual value of the depth image. To simulate the depth loss and noise of the depth image in the real depth acquisition device, this article first adds noise to the depth image and sets the depth value of some areas to 0. The three representative scenes in the Middlebury dataset, i.e., Art, Laundry, and Moebius are also selected as experimental subjects. We compare the processed image with the previously executed FMM algorithm and depth image restoration algorithm (with the proposed C-means clustering segmentation as a guide) and calculate the mean square error between the restored image and the real depth image. Perform quantitative comparative analysis.

From the mean square error comparison in Table 5, it can be seen that the mean square error between the depth map obtained by the FCM clustering segmentation-guided

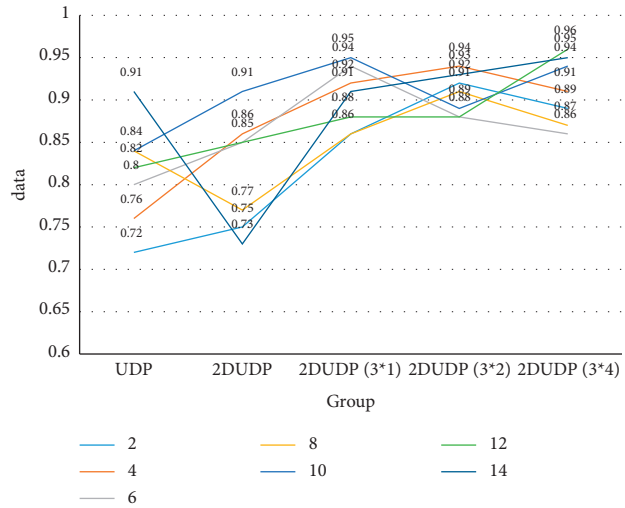


Figure 6: Experimental results on Yale face database.

Figure 7: FMM repair results and details (the picture comes from https://image.baidu.com/).

Table 5: MSE evaluation of Art, Laundry, and Moebius scenes.

Human body	FMM	C mean-FMM	FCM-FMM
Art	4.0673	2.4297	0.9048
Laundry	0.3945	0.3009	0.1512
Moebius	0.7807	0.1949	0.2271

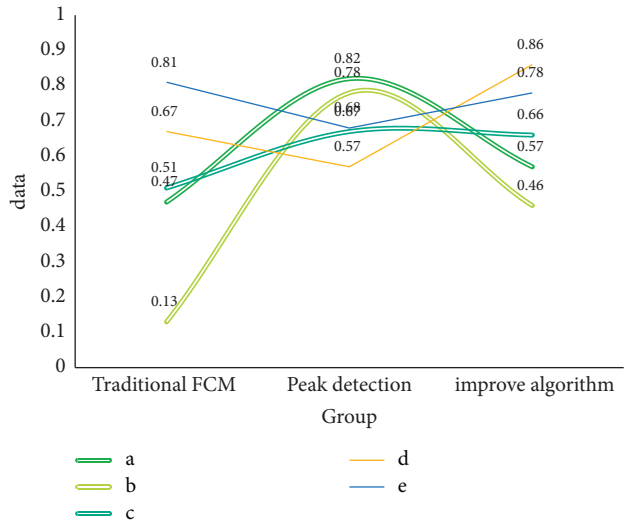


Figure 8: The comparison of segmentation quality.

repair algorithm and the real depth map is smaller, which shows that the algorithm has more advantages. In the Art scene, the focus to be repaired is the edge of the object that contains the hollow area. Although the FMM algorithm repairs the edge of the object, it appears blurry at the edge. The repair algorithm guided by C-means clustering segmentation repairs the edges of the image. There are still some differences in the edge position of FCM. In comparison, the repair algorithm guided by the FCM clustering segmentation has a clearer edge. It can not only recover structural information but also separate the target object from the background clearly at the boundary between the foreground and the background. The transition is obvious. It is observed that upon considering the ambiguity among the pixels in the color image using FCM clustering and segmentation, the introduction of the structure information of the object can help us better restore the missing depth information.

As shown in Figure 8, this paper uses the established graph model to presegment the original image. Using the hierarchical clustering on the obtained similarity matrix W, the initial clustering center of the FCM algorithm is obtained. In the process of finding the initial cluster center, the influence of the peak value of the histogram, the distance between the peaks, and the associated pixels on the peak is comprehensively considered, which effectively avoids the problem of local optimal solution. It is observed from the two quality indicators of interregion gray contrast GC and Bezdek partition coefficient VPC involved in the figure that the improved algorithm in this paper is better than the traditional FCM algorithm and the peak detection FCM algorithm.

5. Conclusions

The emergence of depth images provides new directions for many computer vision applications. Depth images play a very important role in object recognition, three-dimensional reconstruction, scene understanding, and other applications. Kinect has gained the attention of more and more researchers, and its application fields have become more and more extensive. This paper aims at the problem of the lack of depth information in the depth image obtained by the Kinect device (i.e., the void area). Based on the current research trends, new ideas are proposed to solve this problem so that the depth image can be applied to computer vision systems. Aiming at the problem of the hollow area in Kinect depth images, a depth image restoration algorithm based on color image clustering and segmentation is proposed. The algorithm uses the same type of pixels near the hole pixels to fill the waiting area according to the repair sequence of the fast-moving algorithm. C-means clustering is used to segment the color images obtained by Kinect synchronization to highlight the structure information of the objects in the target scene. Repair the pixels. By continuously filling the edge hole points, the effective pixel value is continuously diffused into the hole to complete the repair of the entire hole area. Although the algorithm in this paper has achieved a good repair and repair effect, the repair time is slightly longer than that of the traditional algorithm as the algorithm in this paper needs to cluster and segment the depth images collected by Kinect. For real time computer vision applications, the processing efficiency is not high, and the GPU parallel computing can be considered later to improve the processing speed and real time performance. In the future, on the basis of the algorithm research in this article, we can further study the related algorithms when multiple moving human bodies appear at the same time and contribute our own strength to human motion tracking and recognition.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] F. Xiao, "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy," *Information Fusion*, vol. 46, pp. 23–32, 2019.

[2] W. Elsayed, M. Elhoseny, S. Sabbeh, and A. Riad, "Self-maintenance model for wireless sensor networks," *Computers & Electrical Engineering*, vol. 70, pp. 799–812, 2018.

[3] J. Zeng, J. Huang, Z. Yu, and Y. Zhang, "Research on human motion detection and tracking algorithm based on adaptive dynamic video image scaling technology," *Revista de la Facultad de Ingenieria*, vol. 32, no. 4, pp. 455–463, 2017.

[4] F. Wang, L. Wang, and Y. Fu, "Research on depth image restoration algorithm based on RGB-D," *Modern Electronic Technology*, vol. 042, no. 002, pp. 143–146, 2019.

[5] Y. Li, G. Li, and R. Li, "Research on the tracking algorithm for multiple abnormal targets of micro spectral image," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 21, no. 3, pp. 755–761, 2018.

[6] G. T. Zhang, Z. Gao, and H. Zhang, "Human action description algorithm based on depth motion trajectory information," *Guangdianzi Jiguang/Journal of Optoelectronics Laser*, vol. 28, no. 1, pp. 100–107, 2017.

[7] X. Hu and D. Li, "Research on a single-tree point cloud segmentation method based on UAV tilt photography and deep learning algorithm," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 99, p. 1, 2020.

[8] H. Song and M. Brandt-Pearce, "A 2-D discrete-time model of physical impairments in wavelength-division multiplexing systems," *Journal of Lightwave Technology*, vol. 30, no. 5, pp. 713–726, 2012.

[9] H. Pang, Q. Xuan, M. Xie, C. Liu, and Z. Li, "Research on target tracking algorithm based on siamese neural network," *Mobile Information Systems*, vol. 2021, no. 4, pp. 1–11, 2021.

[10] S. Y. Cho, S. Y. Lee, J. H. Lim, and S. J. Park, "Simultaneous motion tracking and localisation of a person based on the integration of multiple IMUs and depth camera," *IET Radar, Sonar & Navigation*, vol. 11, no. 11, pp. 1679–1688, 2017.

[11] Y. Zhang, K. Wang, J. Jiang, and T. Qiyun, "Research on intraoperative organ motion tracking method based on fusion

of inertial and electromagnetic navigation," *IEEE Access*, no. 99, p. 1, 2021.

[12] Y. Zhang and X. Chen, "The deformed gesture tracking algorithm based on feature space segmentation modeling," *Jiqiren/Robot*, vol. 40, no. 4, pp. 401–412, 2018.

[13] L. Li, X. Li, B. Ouyang, S. Ding, S. Yang, and Y. Qu, "Autonomous multiple instruments tracking for robot-assisted laparoscopic surgery with visual tracking space vector method," *IEEE*, no. 99, p. 1, 2021.

[14] J. Li, Z. Li, Y. Feng, Y. Liu, and G. Shi, "Development of a human-robot hybrid intelligent system based on brain teleoperation and deep learning SLAM," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1664–1674, 2019.

[15] H. Jiang, M. Wang, D. Liu, and Z. Siwang, "CTrack: acoustic device-free and collaborative hands motion tracking on smartphones," *IEEE Internet of Things Journal*, vol. 8, no. 19, p. 1, 2021.

[16] N. Gu, D. Wang, Z. Peng, T. Li, and S. Tong, "Model-Free containment control of underactuated surface vessels under switching topologies based on guiding vector fields and data-driven neural predictors," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–12, 2021.

[17] A. Gantala, N. Telagam, G. V. Kumar, P. Anjaneyulu, and R. Murali Prasad, "Content-based image retrieval using genetic algorithm retrieval effectiveness in terms of precision and recall," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 9, no. 18, pp. 2020–2028, 2017.

[18] M. S. C. Zamwar, D. S. A. Ladhake, and M. U. S. Ghate, "Human face detection and tracking for age rank, weight and gender estimation based on face images using raspberry pi processor," *International Journal of Engineering Research in Africa*, vol. 07, no. 05, pp. 16–21, 2017.

[19] H. Luo, "Research on moving object detection and tracking algorithm based on optical flow method in complex background," *Revista de la Facultad de Ingenieria*, vol. 32, no. 15, pp. 419–422, 2017.

[20] S. Dian, H. Fang, T. Zhao et al., "Modeling and trajectory tracking control for magnetic wheeled mobile robots based on improved dual-heuristic dynamic programming," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 1470–1482, 2020.

[21] L. Zhang, Z. He, M. Gu, and H. Yu, "Crowd segmentation method based on trajectory tracking and prior knowledge learning," *Arabian Journal for Science and Engineering*, vol. 43, no. 12, pp. 7143–7152, 2018.

[22] W. Sun, Z. Pang, W. Huang, J. Yonggang, and D. Yongshou, "Vessel velocity estimation and tracking from Doppler echoes of T/R-R composite compact HFSWR," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, no. 99, p. 1, 2021.

[23] Y. Shen, J. Zhu, H. Liu, Y. Cui, and B. Zhang, "Rapid target plant image mosaic based on depth and color information from Kinect combining K-means algorithm," *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering*, vol. 34, no. 23, pp. 134–141, 2018.

[24] Y. Du, N. Yang, and M. Dong, "Application of target detection algorithm based on depth learning in remote sensing image classification," *Revista de la Facultad de Ingenieria*, vol. 32, no. 4, pp. 179–187, 2017.

[25] F. Zhen and J. Yang, "Research on information processing mechanism of human visual and auditory information based on selective attention," *Journal of Changchun University of Science and Technology (Natural Science Edition)*, vol. 041, no. 004, pp. 127–131, 2018.

[26] J. Deng and L. Cao, "Research on human motion test based on biomechanical sensors using electromyography and pressure detection systems," *Journal of New Materials for Electrochemical Systems*, vol. 22, no. 2, pp. 98–101, 2020.

[27] M. Zhang, J. Yi, and S. Qian, "Research on human action recognition algorithm based on LBP feature," *Jiangxi Science*, vol. 035, no. 006, pp. 940–946, 2017.

[28] Y. Tian, M. Yao, and M. Pan, "Human's mouth location based on YCbCr complexion detection and AdaBoost cascade connection method," *Computer Application Research*, vol. 034, no. 003, pp. 933–935, 2017.

[29] Q. Sun, F. Hu, and Q. Hao, "Human movement modeling and activity perception based on fiber-optic sensing system," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 6, pp. 743–754, 2017.

[30] A. Booranawong, N. Jindapetch, and H. Saito, "An autonomous RSSI filtering method for dealing with human movement effects in an RSSI-based indoor localization system," *Journal of Electrical Engineering & Technology*, vol. 15, no. 5, pp. 2299–2314, 2020.

[31] A. Halili and M. Fetaji, "A survey and assessment of intelligent control technologies and algorithms for helping human movement at national borders," *IFAC-PapersOnLine*, vol. 52, no. 25, pp. 81–86, 2019.

[32] H. El-Hussieny and J.-H. Ryu, "Inverse discounted-based LQR algorithm for learning human movement behaviors," *Applied Intelligence*, vol. 49, no. 4, pp. 1489–1501, 2019.

[33] S. Chadsuthi, B. M. Althouse, S. Iamsirithaworn, W. Triampo, K. H. Grantz, and D. A. T. Cummings, "Travel distance and human movement predict paths of emergence and spatial spread of chikungunya in Thailand," *Epidemiology and Infection*, vol. 146, no. 13, pp. 1654–1662, 2018.