Human Detection using HOG Features of Head and Shoulder Based on Depth Map

Qing Tian
College of Information Engineering, North China University of Technology, Beijing, China
tianqingncut@yeah.net

Bo Zhou,Wen-hua Zhao
College of Information Engineering, North China University of Technology, Beijing, China

Yun Wei
Intelligent Transportation System Research Center, Southeast University, Nanjing, 210096, China
Beijing Urban Engineering Design and Research Institute, Beijing, China

Wei-wei Fei
Systems Engineering Research Institute, CSSC, Beijing, China

***Abstract* — Conventional moving objects detection and tracking using visible light image was often affected by the change of moving objects, change of illumination conditions, interference of complex backgrounds, shaking of camera, shadow of moving objects and moving objects of self-occlusion or mutual-occlusion phenomenon. We propose a human detection method using HOG features of head and shoulder based on depth map and detecting moving objects in particular scene in this paper. In-depth study on Kinect to get depth map with foreground objects. Through the comprehensive analysis based on distance information of the moving objects segmentation extraction removal diagram of background information, by analyzing and comprehensively applying segmentation a method based on distance information to extract pedestrian's Histograms of Oriented Gradients (HOG) features of head and shoulder[1], then make a comparison to the SVM classifier. SVM classifier isolate regions of interest (features of head and shoulder) and judge to achieve real-time detection of objects (pedestrian). The human detection method by using features of head and shoulder based on depth map is a good solution to the problem of low efficiency and identification in traditional human detection system. The detection accuracy of our algorithm is approximate at 97.4% and the average time processing per frame is about 51.76 ms.**

***Index Terms*—Human Detection; Depth Map; Hog Features Of Head And Shoulder; SVM**

I. Introduction

Intelligent video surveillance makes use of computer vision and image processing method for motion detection, object classification, tracking of moving objects to image sequence, as well as making understanding and description to the targets' behaviors in surveillance scenarios. With the function of intelligent monitoring system is becoming more and more powerful in recent years, the demands of all walks of life for intelligent monitoring system is rising, so intelligent monitoring system has been widely applied in various fields of life. Due to head and shoulder are less susceptible to interference of other factors (such as human gait, clothing color, environment, etc.), so we propose a pedestrian detection method by using head and shoulder features based on depth map [4] in this paper, which provides a prerequisite for rapidly and accurately detecting moving objects. Using HOG features of head and shoulder to detect human targets reduce the influence of complex background and imaging conditions and other factors, and increasing detection rate and efficiency.

There has been many researches in the past on human detection, and various different methods have been proposed [3][12][13][14]. Due to detection accuracy of human detection based on Integral Histograms of Oriented Gradients need to improve, the problem of vector dimension, QU Yong-yu, [8] presented human detection method based on integral histograms of oriented gradients (HOG) feature, and proposed to use statistical features of gradient histogram, color frequency and skin color to describe human; to select classification ability strong block as final features and use linear SVM classification. The experiment proved that this method can effectively improve detection accuracy. PAN Feng [9] aimed at the shortcomings of the traditional visual surveillance system. This paper proposed a series of methods to solve the difficulty in detecting human under complex background experimental results show that this method is strongly robust and highly accurate.

The technology of depth image pattern recognition is an emerging technology in recent years. Especially Israel PrimeSense company launched Kinect (an external

Manuscript received December 13, 2012; revised February 12, 2013; accepted April 2, 2013.

doi:10.4304/jsw.8.9.2223-2230

device of Xbox 360) for Microsoft in 2010 April. Compared with depth cameras using the technologies TOF technology [18], structured light, three-dimensional laser ranger scanning and so on, the advantages of Kinect depth camera are high resolution rate and low cost. Therefore, more and more scholars pay attention to the new field. Joshua Fabian [20] Integrating the Microsoft Kinect with Simulink: Real-Time Object Tracking Example: this article is to help fully realize the Kinect's potential. Junping Zhang's [21] Predicting Pedestrian Counts in Crowded Scenes with Rich and High-Dimensional Features: In this paper, they describe a system for predicting pedestrian counts that significantly extends the utility of those ideas. Their approach incorporates a richer set of features for statistical modeling. While these features give rise to regression problems in a high-dimensional space, they leverage learning techniques to reduced dimensionality while still attaining high accuracy for predicting the number of pedestrians. Tsinghua University has carried out their research combining Kinect depth map with fast video matting algorithm [5], etc. The method of G.Fanelli [7] also used synthetic depth map and variants of random forest classifier training, what it solved is detection of human head position and orientation, and used head and no-head as positive and negative samples for training.

Many domestic and foreign research institutes and universities made a lot of achievements in human detection. Common detection methods can be grouped into three categories: detection method based on modal [15], detection method based on statistical and machine learning [3][16] and detection method based on image segmentation [17].

This paper makes in-depth research on human objects detection and tracking in the case of static camera, these methods of moving objects detection are easily affected by illumination, pose, occlusion and so on. The method of human detection using head and shoulder features based on depth map is a detection method based on modal, the process can be broadly divided into three parts: acquisition of depth map, extraction of moving objects and detection of head and shoulder features of moving objects. Owing to the pixel value of depth map is only related to the distance from the field of view window plane to object surface, first of all the depth map is independent of color-space. Combining gray value with horizontal and vertical coordinates of depth map[2] can be used to represent coordinates of object in 3D space [23] in a certain space range, so it can be equivalent to carry on pattern recognition in 3D space. Using depth information of the image for moving objects detection, it is equivalent to carry on pattern detection and recognition in monocular 3D space, which can overcome problems of occlusion and overlapping. Kinect camera can not only obtain ordinary two-dimensional image information, but also obtain synchronous depth information (depth map), this greatly improves detecting efficiency and accuracy of the intelligent monitoring system, and it is also a good supplement for ordinary video monitoring system. This paper utilizes Kinect to obtain depth map, because of depth map without shadow and no interference from light illumination factors; we adopt a segmentation method based on distance information to extract moving objects in this paper. In human detection, we use HOG features of head and shoulder features as the detection features which do not easily change and be blocked, and can accurately represent human features. The pedestrian detection method using head and shoulder features based on depth map can not only effectively solve interference of light variation and occlusion problem, but also get a good detection result of high-density [21] passenger flow in complex environment.

II. Acquisition of Depth Map By Kinect

Kinect uses light coding technology in capturing video information. Light coding uses light source to code for space needed to be measured, in fact, actually it is a structured light technology, which is different from the traditional method of structured light, and the light source is the three dimensional deep 'body coding' rather than a pair of periodic variation of the two-dimensional image coding, this light source called laser speckle which is the random diffraction spots formed by the laser irradiate to the rough objects or penetrate the frosted glass the random diffraction spots penetrating. And these speckles have a high level of randomness, which means arbitrary two speckle patterns are different in the space. As soon as playing out structured light, the whole space is marked, putting an object into this space, just look at speckle pattern above the object, we can know position of the object.

The Infrared emitter of depth video capture system of Kinect launch continuous infrared, the laser array which is launched by infrared emitter cover the entire visual range of Kinect. In fact the recognition image of infrared camera is a "Depth Field", where each color of the pixel (pixel value) represents the actual vector distance of from this point to the camera. We can set the effective distance D of the camera through the OpenNI interface program, while the maximum acquisition distance $L_{\max}$ of the camera is a known parameter. The gray image only perform 256 colors (or brightness), that is the pixel information of the gray image can be divided into 0 to 255 grayscale, so we divide the effective distance D (which has been set) into 256 equal parts in this paper. The distance information of the camera corresponds to each pixel of the grayscale in this way, and the gray information of the image can reflect the actual depth information in depth map. Setting the relevant parameters of the image acquisition system, determine the coding method of depth map through the OpenNI interface program, which includes image format, frame rate, save format and so on, and these parameters should be consistent with color image. The depth dates which obtained by Kinect are recorded in pixels, the default is 640*480 pixels. Moving objects detection requires distance determination according to the actual distance, so it is necessary to convert the dates of pixels information to the actual height information. “Fig 1” shows the depth map from Kinect.

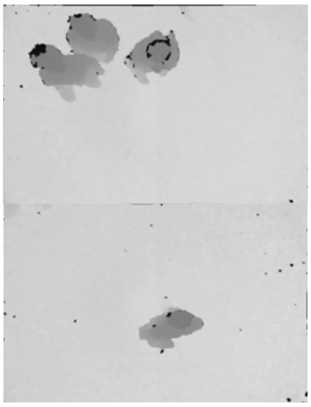


Figure 1. The depth map from Kinect

III. Human Detection based on Depth Map by Using HOG Features of Head and Shoulder

Our algorithm not only utilizes depth information, but also combines with traditional gradient approaches to give faster and more accurate detection. The detection algorithm can also serve as an initial step of the research on pose estimation, tracking, trajectory analysis, calculation passenger flow density and speed or activity recognition by depth information.

Our paper is organized as follows. Section 3 gives an overview of our method and the description of some basic principles which are related to algorithm, and the algorithm flow chart is shown in “Fig 2”. Section 4 discusses the experimental results. Section 5 concludes the paper and gives possible directions for future research. “Fig 2” shows the algorithm flow of HOG Features of head and shoulders.

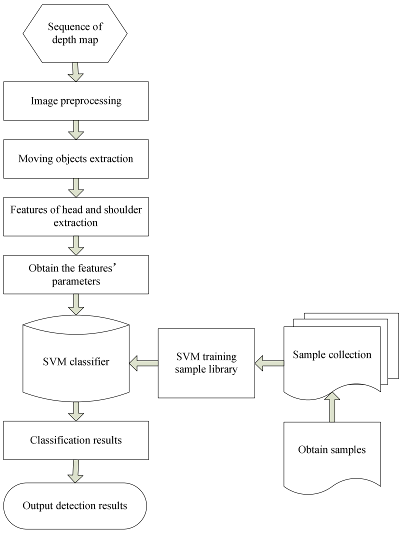
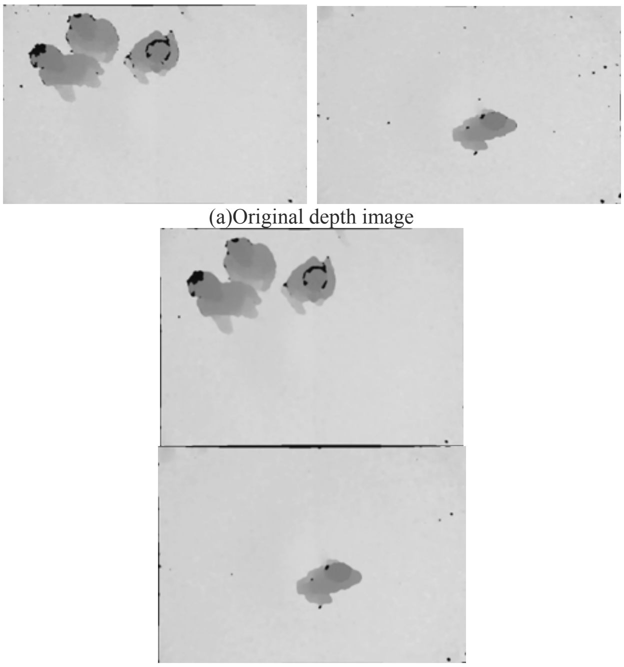


Figure 2. Algorithm flow of HOG features of head and shoulder

A. Image Preprocessing

Image pretreatment process is actually getting rid of useless information, improving the efficiency and speed of the algorithm. Image preprocessing is mainly to eliminate image noise, and smooth the input image to improve image quality, make the image become more clearly, facilitate the subsequent processing and analysis. This paper adopts the median filtering to noise removal, it can filter the depth image noise well which obtained by Kinect, and at the same time can protect edge information well. Median filter is selected for preprocessing and noise filtering. “Figure 3” show original image and the results of median filter. (a) is the original depth image, (b) is the depth map after median filter. It is apparent from the “Fig 3”, the optional median filter achieve a good filtering effect.



(a)Original depth image

(b) The depth map after median filter

Figure 3. The results of median filter

B. Moving Objects Extraction Based on Distance Information Segmentation

How to extract moving objects detection[19] accurately is the first step of pedestrian detection, head and shoulder features information instead of moving objects is to ensure whether any given an image has head and shoulder information. If there are head and shoulder information, considered to be pedestrian target, then continue subsequent processing. Head and shoulder detection positioning is the core step of the pedestrian detection algorithm, and is the key to human detection. We exploited a method based on the distance information segmentation to extract moving objects in this paper. “Fig 4” shows the results of based on the distance information segmentation to extraction moving object. In order to facilitate the data processing, we convert the actual distance data of depth map into the actual pixel values for processing. Through the OpenNI interface program to set the camera effective distance D, and the maximum effective acquisition distance $L_{\max}$ is the known parameters of the camera, depth map pixel values from 0 to 255, we divide effective distance D into 256 equal parts, so each part of the effective distance D corresponding to each pixel level of depth map. Thus we can use the gray information of image to show the actual depth information. We usually get through the practical application scene to determine the camera construction height, and then estimate the distance between head and shoulder and the camera, the next step is to set gray threshold value [24]through conversion relationship of the actual distance and the gray value of depth map. If src(x,y) (Original pixels)> threshold; dst(x,y) (Target pixels) = src(x,y), if src(x,y) (Original pixels)<threshold, dst(x,y) (Target pixels)= src(x,y). Thus separate the pedestrian from the background. On the assumption that gray value of a height for S:

$$S = H / L_{\max} * 255 \tag{1}$$

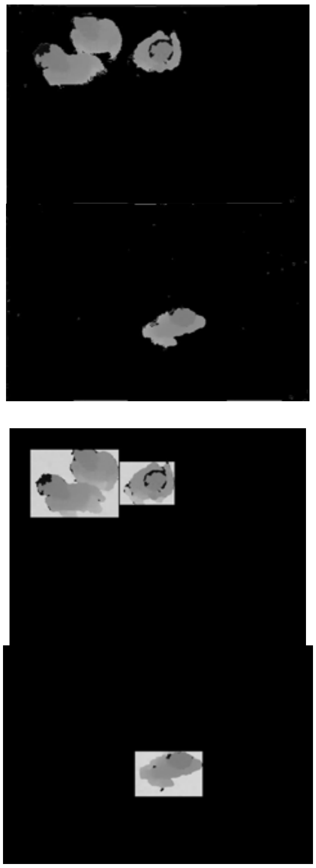


Figure 4. The results based on the distance information segmentation to extraction moving object

The H is a point on the height value of pedestrian, $L_{\max}$ is the maximum effective distance of camera, suppose filtering gray information under 1000 mm (high threshold), we can convert high threshold to gray threshold through formula (2), and then obtain target image.

C. Human Feature Extraction Based on HOG

Histogram of Oriented Gradient (HOG), a method of intensive descriptors that is used for local overlapped images, constitutes features by calculating the local direction of gradient. At present, the approach combining HOG with Support Vector Machine (SVM) has been widely applied to image recognition and achieved a great success especially in human detection. The researcher of French Dalal [6] proposed a classic algorithm using HOG and SVM for human detection at CVPR in 2005. While proposing a lot of human detection algorithms today, they are almost based on the idea of HOG and SVM.HOG features are local descriptors, and human features are constituted by computing local direction of the gradient. In [6], the proposed descriptors could describe well the edge information of human; also the method is robust to illumination variations and small offset. The gradient of the pixel of (x, y) in an image can be denoted as:

$$G_x(x, y) = H(x+1, y) - H(x-1, y) \tag{2}$$

$$G_y(x, y) = H(x, y+1) - H(x, y-1) \tag{3}$$

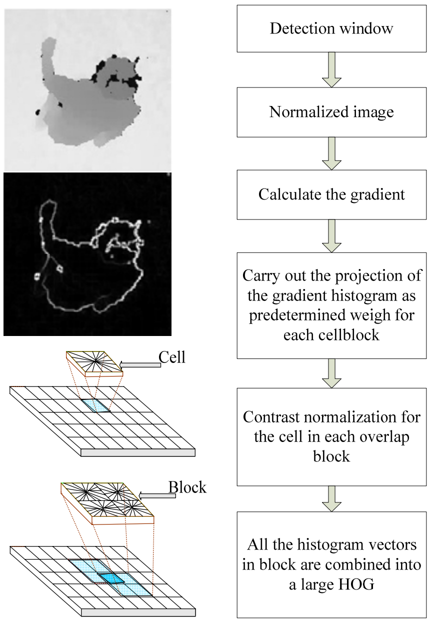

Detection window
Normalized image
Calculate the gradient
Carry out the projection of the gradient histogram as predetermined weigh for each cellblock
Contrast normalization for the cell in each overlap block
All the histogram vectors in block are combined into a large HOG
Cell
Block

Figure 5. The procedure of extracting depth image’s HOG features

Where $G_x(x, y)$ denotes the horizontal direction gradient of input image pixel, $G_y(x, y)$ denotes the vertical direction gradient and $H(x, y)$ denotes the pixel values. Then the gradient magnitude and direction of (x, y) can be represented as:

$$G(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)} \tag{4}$$

$$\alpha(x, y) = \tan^{-1}\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \tag{5}$$

Before classification training of the pedestrian targets, we need describe features of detecting targets, features are critical problems for determining the similarity and classification. A human detection algorithm based on features must be able to separate well. Also the features can be used significantly to discriminate human and non-human even in complex backgrounds and different lighting conditions. While the advantage of HOG feature is that it is based on histogram of oriented gradient. It can not only describe the feature of human contours, but also be not sensitive to light and small offset. We obtain human features by combining the features of all blocks in line. Take the image of 64*64 as an example. “Fig 5” shows the procedure of extracting depth image’s HOG features, we calculate the HOG feature [22] as follows:

1) Input an image which is one of the samples library or the ROLS (the region of interests) in the previous detection phase.
2) Gradient calculation: use the [-1, 0, 1] and [-1, 0, 1] median filter to perform filtering, calculate the vertical gradient and horizontal gradient of the image, and then calculate the gradient direction and gradient magnitude of each pixel.
3) Divide the inputting image into average small cells （(including 16*16 pixels)） and combine four cells into a small block.，one block is constituted by a cell of 2*2.
4) The selection of direction channel: divide $0^o - 180^o$ or $0^o - 360^o$ into n channels averagely. In this paper, we divide $0^o - 180^o$ into nine equal parts, that is, nine channels in total. So there are 4*9=36 features in each block. Then there would be 7 scanning windows in horizontal direction and 7 scanning windows in vertical direction if we scan with 8 pixels. That is, there are a total of 36*7*7=1764 features in the image of 64*64.
5) The acquisition of the histogram: get the statistics of each pixel in each cell of their histogram of orientated gradient. The abscissa of the histogram represents the nine direction channels selected in step 3, and the ordinate represents the summation of the gradient, belonging to a certain direction channel. Thus, we get a set of vectors.
6) The process of normalization: normalize the vectors in blocks in which pixels correspond with the vectors. Currently, the common methods of normalizing include:

$$v^* = \sqrt{\frac{v}{\|v\|_1 + \varepsilon}}$$

7) Form HOG features: combine all the vectors processed above and then form a set of vectors, which are the HOG features.

D. Human Recognition Based on Llinear SVM

This paper selects SVM [6] as classification algorithm. SVM solves two kinds of classification problems that are mainly based on the structural risk minimization principle, finding an optimal separating hyper plane to separate two kinds of data with the largest interval. Suppose linear separable sample sets as:

$$Sample = \{(s_i, f_i) \mid i = 1,2,\cdots,n\} \quad (6)$$

The $s_i \in R^d$, f_i ={+1,-1} as s_i corresponding to the corresponding class label. The general form of the linear discriminant function of D dimensional space g(x) = w·x+k, the corresponding hyper plane equation w·x+k=0. Normalizing the discriminant function g(x) makes the two kinds of samples satisfied with |g(x)|≥1, thus classification interval is 2/ ‖ w ‖ , so to make classification interval surface maximum is equivalent to make ‖ w ‖ minimum. However making sure the hypeplane for all samples should be classified correctly, it must meet with the following formula:

$$f_i = [(w \cdot x) + k] - 1 \ge 0, (i = 1,2,......,n) \quad (7)$$

Hyperplane which satisfies the above two conditions is the optimal hyper plane. Training samples are that of closest points from hyperplane in two groups’ samples, and parallel the optimal hype r plane H_1 ’s and H_2 ’s, that makes equality holds, established in Inequality (7), called support vector. The optimal hyperplane problem can be represented as Inequality (7) to get objective function.

$$\psi(w) = \frac{1}{2} \| w \|^2 = \frac{1}{2} w \times w \quad (8)$$

Minimum value under constraint of Inequality (7). For the linear inseparable samples, introducing slack variable τ_i and penalty factor T, the objective function can be rewritten as follow:

$$\psi(w, \tau_i) = \frac{1}{2} w \times w + T \sum_{i=1}^{N} \tau_i \quad (9)$$

Introduce Lagrange multiplication factor ($\alpha_1, \alpha_2, \cdots, \alpha_n$), convert it to the extreme value problem of constrained quadratic function to solve the optimal hyperplane, the corresponding solution is：

$$w = \sum_{i=1}^{N} \alpha_i s_i f_i \quad (10)$$

Among them, the α_i is nonzero only to s_i , the optimal classification function can be rewritten as follows:

$$m(x) = \sin(w \times x + k) = \sin[\sum_{i=1}^{N} \alpha_i f_i (s_i \times x) + k] \quad (11)$$

We take the method based on the linear SVM [22] combining with HOG features to realize classification and recognition, get optimal hyperplane and obtain SVM decision function through training positive and negative samples. When SVM is offline training, what we should pay attention to is that s_i the input vector of classifier is feature vector acquired in the analysis of HOG features in formula 7, namely, each sample uses a 3780 - dimensional vector to represent. We need to normalize input vector before training, mainly in order to avoid some features of large data range as dominant feature, which affects the small scale features. Also it helps to reduce the complexity of calculation. After training we get a model, which storages SVM classifier learning results. When SVM is on-line identification, firstly do normalization processing for the input target ROLS, then the ROLS window size uniform to the training sample window size, we select the training sample window size to be 64 x 64 in this paper. Then loading linear SVM classifier, judging the ROLS according to the training model obtained from previous classifier, the output of the classifier f_i = 1 is expressed as the pedestrian image, the output f_i= 1 is expressed as delinquent people image. f_i>Threshold of SVM. And save the sample information judged as pedestrian. We use rectangle function of OpenCV to mark the pedestrian who has been detected with rectangular box. Repeat the above operation until all moving targets in the single image are found, thus pedestrian detection of the single image is finished.

IV. Experimental Results and Analaysis

We analyze the quantitative experimental results of our detection method, describe the feasibility of algorithm and compare our approach with other human detection algorithm in this section.

A. Dataset

We collect real-time video by Kinect camera (about containing 1000 moving targets), and process real-time video according to the functional block diagram, synchronously display test results. In order to maintain the real-time requirement of the video image acquisition transmission frame rate in the 20 - 30 frame, and using MPEG4 as image coding and decoding standard. This experiment performed on PC with Intel(R) Core(TM) i7-2600 CPU 3.4GHz , 8G memory and windows7 pro-OS.

B. Experimental Result

In order to prove that the detection method is effective and high efficiency, we use LIBSVM as training tool, we select 3000 positive samples and 12000 negative samples. Figure 6 shows our positive samples; “Fig 7” shows our negative samples. Get SVM decision function after training, and then use it on reality situation.

This system sets up above the top of the stairs in a library hall， there are many pedestrians in this actual applications occasion which has high demand real-time accuracy for the detection system. The experiment results are shown as follows: the total number of pedestrians appearing in the stairs is 235, and the system actual detected the number of pedestrians is 229. “Fig 8” shows the detection results of our algorithm.

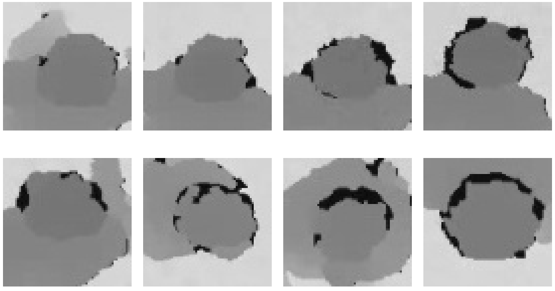
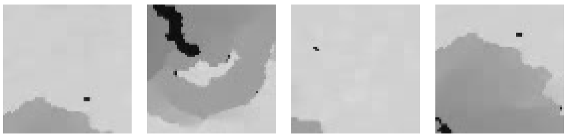


Figure 6. Positive samples



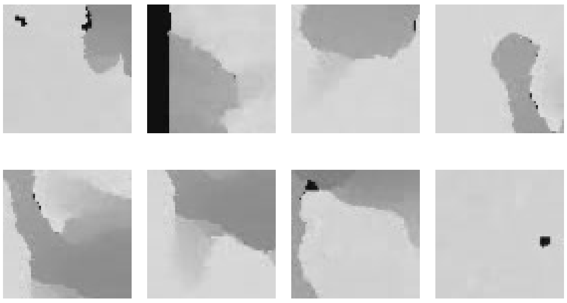


Figure 7. Negative samples

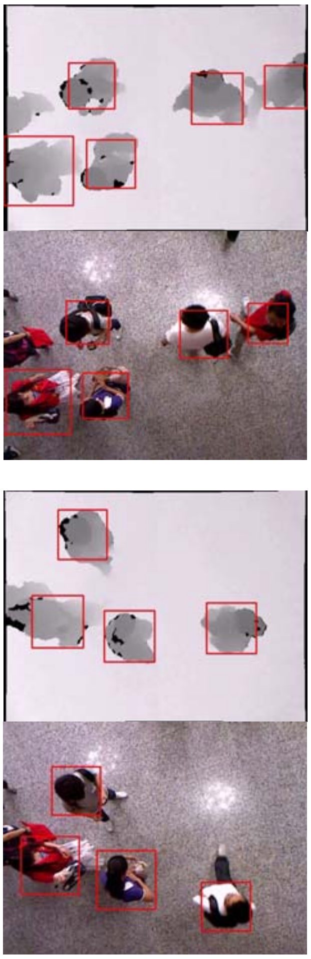


Figure 8. The detection results of our algorithm

We could calculate the detection rate, recall and accuracy through the following formulas:

$$Detection\ \ rate = \frac{TruePos}{TruePos + FalsePos} \quad (13)$$

$$\mathrm{Re}call = \frac{TruePos}{TruePos + FalseNeg} \quad (14)$$

$$Accuracy = \frac{TruePos + TrueNeg}{TruePos + TrueNeg + FalseNeg + FalsePos} \quad (15)$$

The TruePos substitute for the correct detection of the pedestrian image, the FalsePos substitutes for the error detection of the pedestrian image, the TrueNeg substitutes for the correct detection of the wrong Pedestrian image, the FalseNeg substitutes for the error detection of wrong Pedestrian image. Accuracy is the final result of our algorithm. "Table 1" shows the Accuracy of our algorithm, "Table 2" shows the comparison of performances of other detection algorithm.

TABLE 1
THE ACCURACY OF OUR ALGORITHM

FalsePos	FalseNeg	TruePos	TrueNeg
0	9	189	145
Recall		Precision	
95.5%		100%	
Accuracy			
97.4%			

TABLE 2
THE COMPARISON OF PERFORMANCE OF OTHER DETECTION METHOD

Method	Recall	Detection rate	Accuracy
Sho Ikemura	32.9%	90.0%	85.5%
Traditional single HOG	70.6%	96%	90.4%
Our algorithm	95.5%	100%	97.4%

We use the method which combines HOG with depth information to overcome the problem of the low detection rate by traditional single HOG method, we also can reduce false detection rate through the method of fusion detection window. Compared with Xia Lu's experiment environment [10], our algorithm experiment environment has more detection targets, which requires the detection algorithm has better real-time accuracy and stronger robustness. Compared with Sho Ikemura's detecting algorithm [11], our algorithm mainly uses head and shoulder features model as inspection standard, because head and shoulder is not easy to be occluded and change, so head and shoulder information can accurately represent pedestrian characteristics, Support Vector Machine (SVM) classifier overcame the disadvantages of over fitting in traditional methods, experimental results show that this method has strongly robust and highly accurate.

C. Real-time Human Detection

The performance of the computer can guarantee the real-time of our algorithm well. We use the method of moving objects extraction based on height information segmentation while extracting detection target, which can filter interference information rapidly and accurately, reduce calculation when calculate gradient value of HOG features, also improve the detection efficiency and detection accuracy. "Table 3" shows the time efficiency of traditional Hog calculation and our method (HOG combine with depth information).

TABLE 3
THE TIME EFFICIENCY OF TADITIONAL HOG CACULATION AND OUR METHOD(MS)

Method	Total processing time for 500 frames			Average total time for 500 frames	Average time per frame
Traditional Hog	83808	79937	79937	81227.3	162.455
Our algorithm	26715	25424	25512	25883.7	51.7673

V. CONCLUSION

In this paper, we present based on the Kinect depth image head and shoulder features human detection methods for human detection, using depth image to move objects extraction solving the problems of illumination conditions change, complicated background interference, camera dithering, the shadow of the moving objects well. Using moving objects extraction based on distance information segmentation. Through the SVM classifier classify head and shoulder features vector. Judging each slide detection window whether has detection targets through the optimal differential plane. If head and shoulder is occluded, it probably will not be detected. We can handle by other human characteristics or using the aid of the detector for solving this problem. Due to pixel gray values in depth image are only concerned with the distance between viewing window plane and object surface, therefore depth image has the character of color space independence, gray value of depth image combine with the horizontal, vertical coordinate of image, which can indicate objects' coordinates in 3D space in a certain space range. Using image depth information to detect moving objects is equivalent to make detection and recognition in monocular 3D space, so it can overcome occlusion or overlap problems. In this paper our technology innovation place lies in depth image can be reflected the height information of detected target in a certain range, and can remove interference content through setting height threshold. This system also can apply to complex scene high-density passenger flow detection and has received a good detection effect, which also has characteristics of suiting to different scene.

ACKNOWLEDGEMENT

This work is sponsored by Project of Beijing Municipal Commission of Education (No.KM201210009008) and the NSFC (No. 61103113).

REFERENCES

[1] Wang Fei jie, "Real-time Detection and Tracking of Human Based on Head and Shoulder Feature", *Master thesis*, The software department of Jilin university, 2010.

[2] Zhang, Qian , “Reconstruction of intermediate view based on depth map enhancement”, *Journal of Multimedia*, v 7, n 6, pp 415-419, 2012.

[3] Dalal N., Triggs B., “Histograms of Oriented Gradients for Human Detection”, *Computer Vision and Pattern Recognition*, Volume 1, Pages(s): 886-893, 2005.

[4] Qiu Yu, Huang Shan, Wei Yu, “Pedestrian detection and counting in video surveillance”, *Microcomputer information (measurement and control automation)*, Volume 10, Pages(s): 187-188， 2010.

[5] He Bei, Wang Guijin, Lin Xinggang, “Quick matting for vedio based on depth image of the kinect”, *Journal of Tsinhua University (Science and Technology)*, Volume 52, Pages(s): 29-32, 2012.

[6] Ang Yi-bo, Gao Hui, Zhang Mao-jun, “Cascade features based method for pedestrian detection in street scene”, *Joumal of Computer Applications*, Volume 3l, Pages(s): 129-132, 2011.

[7] Fanelli G, Weisi T, Gall J, “Real time head pose estimation from consumer depth cameras”, *Lecture Notes in Computer Science* , Volume 6835, Pages(s): 101-110, 2011.

[8] Qu Yong-yu, Liu Qing, Guo Jian-ming, Zhou Sheng-hui, “HOG and Color Based Pedestrian Detection”, *Journal of Wuhan University of Technology*, Volume 33, Pages(s): 134-138, 2011.

[9] Pan Feng, Wang Xuan-yin, Wang Quan-qiang, “Human detection based on head and shoulder feature intelligent surveillance system”, *Journal of Zhejiang University (Engineering Science)*, Volume 38, Pages(s): 10-14, 2004,

[10] Lu Xia, Chia-Chih Chen and J.K.Aggarwal, “Human Detection Using Depth Information by Kinect”, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2011.

[11] Ikemura, Sho, Fujiyoshi, Hironobu, “Real-Time Human Detection using Relational Depth Similarity Features”, *Lecture Notes in Computer Science*, Volume 6495, Pages(s):25-28, 2011.

[12] Zhu Xudong , Liu Zhijing, Zhang Juehui “ Human activity clustering for online anomaly detection”，*Journal of Computers*, v 6, n 6, pp 1071-1079, June 2011

[13] Dalal Navneet, Triggs Bill, Schmid Cordelia, “Human detection using oriented histograms of flow and appearance”, *Lecture Notes in Computer Science*, vol.3952, pp.428-441, 2006.

[14] Bo Wu, Nevatia R., “Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors”, *Computer Vision*, vol.1, pp.90-97, 2005.

[15] Philomin V., Duraiswami R., Davis L., “Pedestrian Tracking from a Moving Vehicle”, *Proceedings of IEEE International Conference on Computer Vision*, vol.2, pp.350-355, 2000.

[16] Viola P., Jones M.J., Snow D., “Detecting Pedestrians Using Pattern of Motion and Appearance”, *Proceedings of IEEE International Conference on Computer Vision*, vol.2, pp.734-741, 2003,.

[17] Bertozzi M., Broggi A., Chapuis R., Chausse F., Fascioli A., Tibaldi A., “Shape-based Pedestrian Detection and Localization”, *Proceedings of IEEE Intelligent Vehicles Symposium*, vol.1, pp.410-415, 2003.

[18] Van den Bergh M., Van Gool L., “Combining RGB and ToF cameras for real-time 3D hand gesture interaction”, *IEEE Workshop on Applications of Computer Vision*, pp.66 –72, 2011.

[19] Fu, Weina1; Xu, Zhiwen; Liu, Shuai; Wang, Xin,; Ke, Hongchang ,” The capture of moving object in video image”, *Journal of Multimedia*, v 6, n 6, pp 518-525, 2011.

[20] Fabian J., Young T., Jones, J. C. P., Clayton G. M., “Integrating the Microsoft Kinect With Simulink: Real-Time Object Tracking Example”, *IEEE/ASME Transactions on Mechatronics*, vol.99, pp.1-12, 2012.

[21] Junping Zhang, Ben Tan, Fei Sha, Li He, “Predicting Pedestrian Counts in Crowded Scenes With Rich and High-Dimensional Features”, *Intelligent Transportation Systems* , vol.12, pp.1037 – 1046, 2011.

[22] CHENG Guang-tao, CHEN Xue, GUO Zhao-zhuang, “Pedestrian detection method of vision based on HOG features”, *Transducer and Microsystem Technologies*, vol.30, pp.68 – 74, 2011.

[23] Peng Lin, “Body Recognition Based on Depth Images by Learning”, *Master thesis*, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 2012.

[24] Ye, Zhiwei; Hu, Zhengbing; Lai, Xudong; Chen, Hongwei,” Image segmentation using thresholding and swarm intelligence”, *Journal of Software*, v 7, n5, pp1074-1082, 2012.

Qing Tian received his PhD degree in electronic science and technology from institute of electronics, Chinese academy of science, Beijing, China, in 2010. He is a distinguished lecturer at North China University of Technology, Beijing, China. His research interests include electrical engineering, intelligent transportation systems, pattern recognition and computer vision.

Bo Zhou is currently working toward the master’s degree in electrical circuit and system at the College of Information Engineering, North China University of Technology, Beijing, China.

Yun Wei received his master’s degree in intelligent transportation systems from Southeast University, Nanjing, China. He is currently working toward the PhD degree in intelligent transportation systems, at Intelligent Transportation System Research Center, Southeast University. He is also a researcher at Beijing Urban Engineering Design and Research Institute, Beijing. His research interests include intelligent transportation systems and computer vision.

Wen-hua Zhao is currently working toward the master’s degree in information engineering, at the College of Information Engineering, North China University of Technology, Beijing, China.

Wei-wei Fei received his PhD degree in electronic science and technology from institute of electronics, Chinese academy of science, Beijing, China, in 2010. He is a sensor researcher at Systems Engineering Research Institute, CSSC, Beijing, China. His research interests include electrical engineering, communication systems and artificial intelligence.