ICIC Express Letters
Part B: Applications
Volume **7**, Number **7**, July **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **1567–1574**

HUMAN DETECTION USING COLOR AND DEPTH INFORMATION BY KINECT BASED ON THE FUSION METHOD OF DECISION TEMPLATE

Yang Shi, Xiucheng Dong, Decai Shi and Qiuyan Yang

School of Electrical Engineering and Electronic Information
Xihua University
No. 999, Jinzhou Road, Jinniu District, Chengdu 610039, P. R. China
Shi_yabc@tom.com

Received December 2015; accepted March 2016

Abstract. *As it is hard to fulfill the people detection tasks in complex environment using a common camera, this paper presents a Multi-Features human detection method based on color and depth image provided by Kinect. Firstly, we calculate the Histogram of Oriented Gradient (HOG) and Histogram of Oriented Depths (HOD) features of color and depth data. Then, the Support Vector Machine classifiers were trained based on the two features. Finally, Decision Template (DT) is calculated to match the decision profile of new incoming objects by the Squared Euclidean distance similarity measure. The method mentioned above is implemented in the weak light and complex environment. Experimental results show that the method has better robustness in these environments with the help of the depth images than only using color images. The detection effect has improved and the false detection and the miss rate were reduced.*
Keywords: Kinect, Depth image, Human detection, Classifier fusion, Decision template

1. **Introduction.** Human detection is a significant research problem that has wide application in security system, people behavior study, etc., and there has been much research in the past few years in human detection [1,2]. However, detecting human is still a challenging problem due to variations in clothing, posture, illumination and complexity of the background.

It is generally known depth map represents the 3-D information which is an important feature for researchers to recognize objects. However, most of the existing depth map sensors, such as TOF camera, are expensive and lack friendly application interface, so human detection on depth map is rarely applied in practice. Now, the Kinect launched by Microsoft is cheap and very easy to use so it can be used in human environment. In [3], the authors proposed a new method of human detection using Support Vector Machine (SVM) algorithm via extracting the new designed Local Ternary Direction Pattern (LTDP) feature descriptor only based on depth image collected by Kinect sensor. Xia et al. [4] proposed a method of human detection approach based on depth image by using a two-stage model containing a 2-D head contour model and a 3-D head surface mode.

Unfortunately, since the Kinect mainly depends on speckle method, the depth map captured by Kinect often contains much noise. So the detector trained from the depth image is not likely to be ideal due to its unstable quality. Some researchers have done lots of works to cover this shortage. In [5], the authors presented a depth feature called HOD, and introduced Combo-HOD that detect people in RGB-D data. And in [6], it has presented a people detection system for mobile robots using an RGB-D and thermal sensor fusion. So Multi-Features can be an effective pathway to improve this recognition. For the higher efficiency and flexibility, we select parallel multiple classifier system and make use of the fusion rule of decision templates. The DT method proposed by L. I.

Kuncheva [7] could apply to Multi-Features classifier fusion, and it was widely adopted in this field [8]. DT is insensitive to poorly trained individual classifiers and can achieve good and stable performance without strict probabilistic conditions.

In this paper, we present a human detector that combines color and depth detection results. The Kinect includes depth sensor, RGB camera, four microphones and power. It captures both color image and depth data, and thus we can get the HOG and the HOD features from Kinect. Then, we train the SVM detectors with these features. From these detectors and the training data, the DT matrix can be calculated. Finally, we compute the Squared Euclidean distance between the DT matrix and the decision profile matrix to evaluate the similarity and obtain the Classify Label. And we experiment with our methods in a complex environment and a weak light environment to verify these performances. Figure 1 shows the system overview.

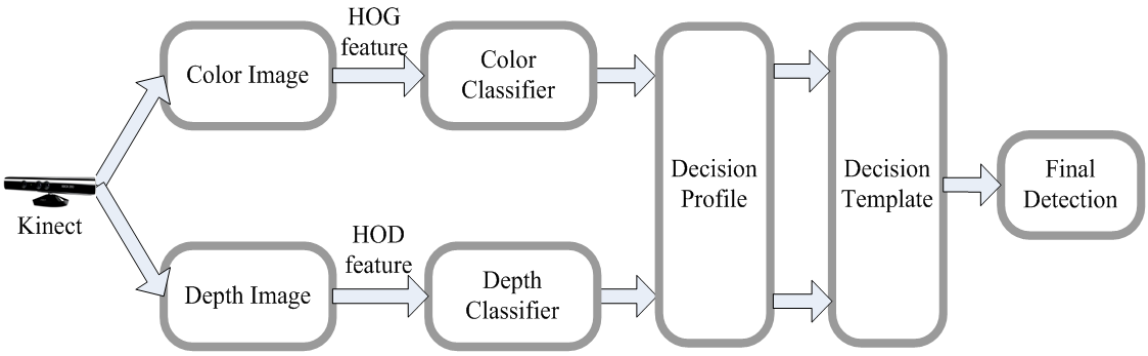


FIGURE 1. System overview

The remaining parts of the paper are organized as follows. Section 2 presents the algorithm of the feature vector extraction and the methods of detection combination and similarity evaluation. Main results of HOG detection, HOD detection and detection combination are given in Section 3. Finally, conclusions are stated in Section 4.

2. **Methods of Feature Extraction and Detection Combination.**

2.1. **Feature vector extraction.** HOG introduced by Dalal and Trigs [1] is currently one of the best performance and widely used methods for visual people detection. This algorithm extracted the features by calculating the histogram of image's gradient direction. The algorithm could keep better invariance in the geometric and optical deformation. Subtle movements also can be ignored and will not affect the detection. Based on the idea of HOG, Spinello and Arras introduced HOD as a novel person detector for dense depth data. HOD follows the same procedure with HOG for the depth image. These algorithms consider a subdivision of a fixed window into cells, compute descriptors for each cell, and collect the oriented depth gradients into 1D histograms. Each cell has 9 orientation bins in 0°-180°. Four 8×8 pixel cells also form a block to collect and normalize the histograms to L2-Hys (Lowe-style clipped L2 norm) unit length and to achieve a high level of robustness with respect to depth noise. We can compute the gradients by $[-1, 0, 1]$ gradient filter with no smoothing and gradient directions in each bin as Equation (1). In the equation, $G_x(x, y)$, $G_y(x, y)$ and $H(x, y)$ stand for the horizontal gradient direction value, vertical gradient direction value and gray or depth value of pixel(x, y). The function value $G_y(x, y)$ and $\alpha(x, y)$ stand for the gradient direction value and the gradient direction. The resulting HOG and HOD features are used for training some soft

linear SVM.

$$\begin{aligned} G_x(x,y) &= H(x+1,y) - H(x-1,y) \\ G_y(x,y) &= H(x,y+1) - H(x,y-1) \\ G(x,y) &= \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \\ \alpha(x,y) &= \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \end{aligned} \tag{1}$$

2.2. **The combination of color and depth detection.** When the HOG and HOD SVM descriptors are classified, the information is ready to be fused by DT method. Combining classifiers is an approach to improve the performance in classification particularly for complex problems. Suppose D is a single classifier. Let $x \in R^n$ be a feature vector and $\{1, 2, \ldots, c\}$ be the label set of c classes. We can assume that all c degrees are in the interval $[0, 1]$. The output of D is signified by $\mu_D(x) = [\mu_D^1, \mu_D^2, \ldots, \mu_D^c]^T$. The decision of D that is assigned to the input x is typically made by the maximum membership rule:

$$D(x) = K \Leftrightarrow \mu_D^k(x) = \max\{\mu_D^i(x)\} \quad i = 1, 2, \ldots, c \tag{2}$$

Now let $\{D_1, \ldots, D_L\}$ be a set of classifiers and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be the set of class labels. We denote the output of the ith classifier as $D_i(x) = [d_{i,1}(x), \ldots, d_{i,c}(x)]^T$ where $d_{i,j}(x)$ indicated the support that classifier D_i gives to the supposition that x comes from class ω_j. At the measurement level, $d_{i,j}$ is the discriminant value (likeness or distance) or probability-like confidence of ω_j. We construct $\widehat{D}$, the combination output of the L classifiers as:

$$\widehat{D}(x) = F(D_1(x), \ldots, D_L(x)) = \left[\mu_D^1, \ldots, \mu_D^c\right]^T \tag{3}$$

where F is called aggregation rule. The L classifier outputs for an input pattern x can be arranged in a decision profile matrix ($DP(x)$) as shown in Equation (4):

$$DP(x) = \begin{bmatrix} d_{11}(x) & \cdots & d_{1j}(x) & \cdots & d_{1c}(x) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1}(x) & \cdots & d_{ij}(x) & \cdots & d_{ic}(x) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{L1(}x) & \cdots & d_{Lj}(x) & \cdots & d_{Lc}(x) \end{bmatrix} \tag{4}$$

DT_i for class i is the average of the decision profiles of the elements of the training set X, labeled in class i. Thus, $DT_i(X)$ of class i is the $L \times c$ matrix $DT_i(X) = [dt_i(k,s)(X)]$ whose (k,s)th element is computed by:

$$dt_i(k,s)(X) = \frac{\sum_{j=1}^N Ind(x_j, i) d_{k,s}(x_j)}{\sum_{j=1}^N Ind(x_j, i)} \quad k = 1, \ldots, L, \quad s = 1, 2, \ldots, c \tag{5}$$

where $Ind(x_j, i)$ is an indicator function with value 1 if pattern x_j belongs to class ω_i, and 0, otherwise. To simplify the notation $DT_i(X)$ will be denoted by DT_i. After constructing DT matrices, in testing phase, when $x \in R^n$ is submitted for classification, the DT scheme matches $DP(x)$ to DT_i, $i = 1, 2, \ldots, c$, and produces the soft class labels:

$$\widehat{\mu}_i(x) = S(DT_i, DP(x)) \quad i = 1, \ldots, c \tag{6}$$

where S is interpreted as a similarity measure. The idea of the DT combiner is to remember the most typical decision profile for each class ω_j, called the decision template, DT_j, and then compares it with the current decision profile $DP(x)$ using some similarity measure S. The closest match will be labeled x. The higher the similarity between the decision profile of the current $x(DP(x))$ and the decision template for class i (DT_i), the higher the support for that class ($\widehat{\mu}_i(x)$). The decision profile matrix for each particular x_i

is an $L \times c$ matrix. The measure of similarity is based on the Squared Euclidean distance $(DT(E))$:

$$S(DT_i, DP(x)) = \frac{1}{L \times c} \sum_{i=1}^{L} \sum_{j=1}^{c} [DT_j(i,k) - d_{i,k}(x)]^2 \tag{7}$$

where $DT_j(i,k)$ is the (i,k)th entry in decision template DT_j.

Then, we can get the predicted class label ω_j of x by the rule:

$$\widehat{i} = \arg\max_i (1 - S(DT_i, DP(x))) \quad i = 1, \ldots, c, \tag{8}$$

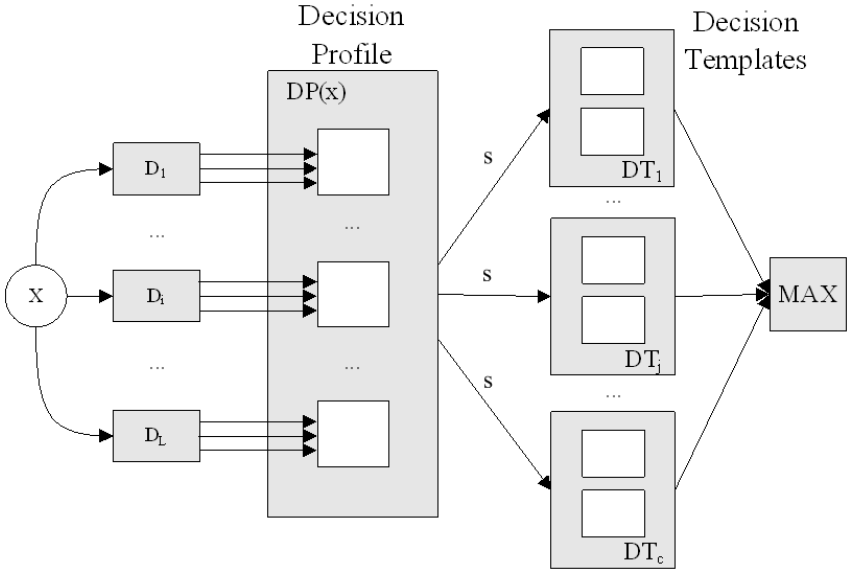
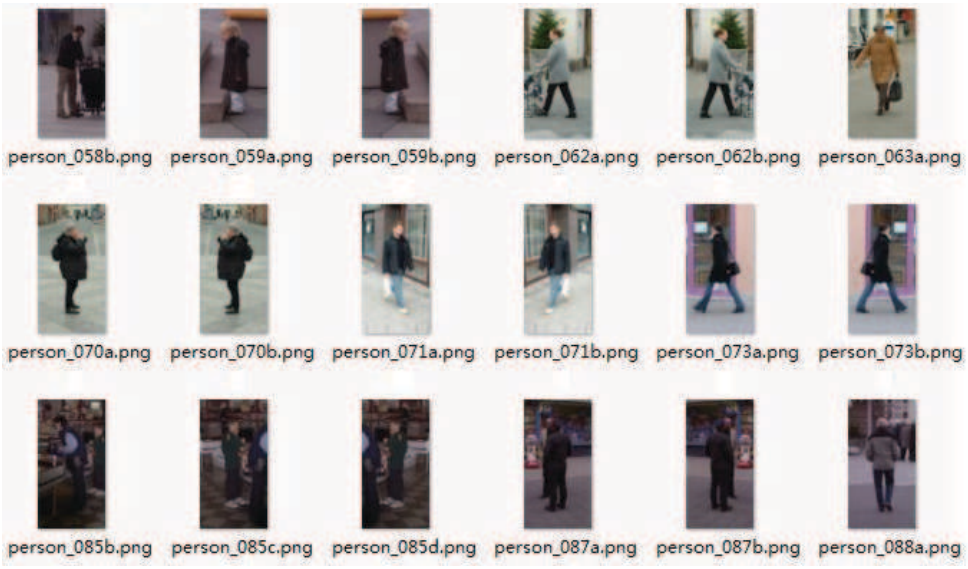


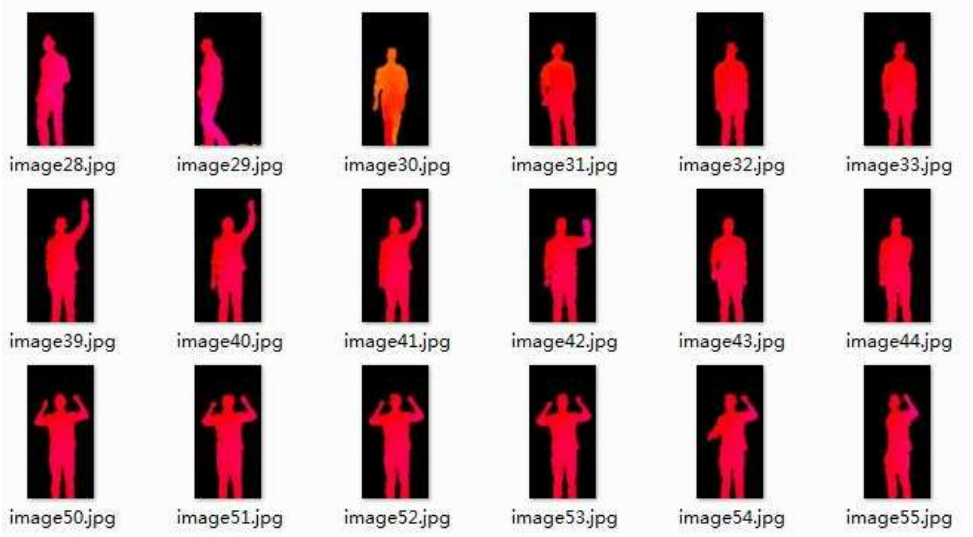
FIGURE 2. Architecture of the decision templates classifier fusion scheme

3. **Experiment Results.** In order to verify the effectiveness of the combination of HOG and HOD in this paper, the experiment is investigated under the environment of C++. We present the experimental evaluation of our approach carried out using data collected with Kinect in two scenarios: a corridor being used as the weak light environment and an office being used as the complex environment. To obtain positive examples in both scenarios, the Kinect was operated in the two environments (the corridor and the office). The color and depth images collected by Kinect were hand-labeled as positive examples if people were visually detected in the image and as negative examples, otherwise. The color dataset is composed of 4762 positive examples and 28632 negative examples. The depth dataset is composed of 3589 positive examples and 18936 negative examples. The set of positive examples contains people at different positions and dressed with different clothes in a typical manufacturing environment. The set of negative examples is composed of image with no human presence and containing other objects, such as machines, tables, chairs, and walls. Figure 3 shows some database samples.

In Figure 4, the first column is the detection results of color image. The second column is the detection results of depth image. The third column is the overlapping of the two results. The last column is fusion results of the two detections based on DT method. The first and second rows are detected in the office as a complex environment. The other two rows are detected in the corridor as a weak light environment. The first row shows that the false positive can be effectively filtered out. The second row shows that the missed person in color image can be detected and the false positive in color can be filtered out. The third row shows the mussy fault detections can be filtered out and the person that is too far to recognize in the depth image also can be detected in the result. The last row



(a)



(b)

FIGURE 3. The database samples, (a) color positive samples, (b) depth positive samples

shows that not only the false positive can be filtered out but also the person missing in the depth image can be detected in the result.

In order to reduce the influence of the numbers of sample dataset, in the experiment, we collocate the color and depth sample dataset into three training datasets. The quantity of the first training dataset is a third of the sample dataset. The quantity of the second training dataset is two-thirds of the sample dataset. The third training dataset is the same with the sample database. The parameter setting of the SVM detectors is based on the Equal Error Rate (EER) point of the Receiver Operating Characteristic (ROC) curve.

We collected 200 samples for each environment, in order to evaluate the fusion method. The result is shown in Table 1 with the aforementioned three kinds of datasets. The test samples were detected respectively by HOG, HOD and the DT fusion method.

It can be obviously seen in Figure 5, either the false positive or the miss rate can be effectively reduced. The method mentioned in this paper can reduce 56.7% false positive and 46.5% miss rate in average compared with the method using HOG or HOD alone.

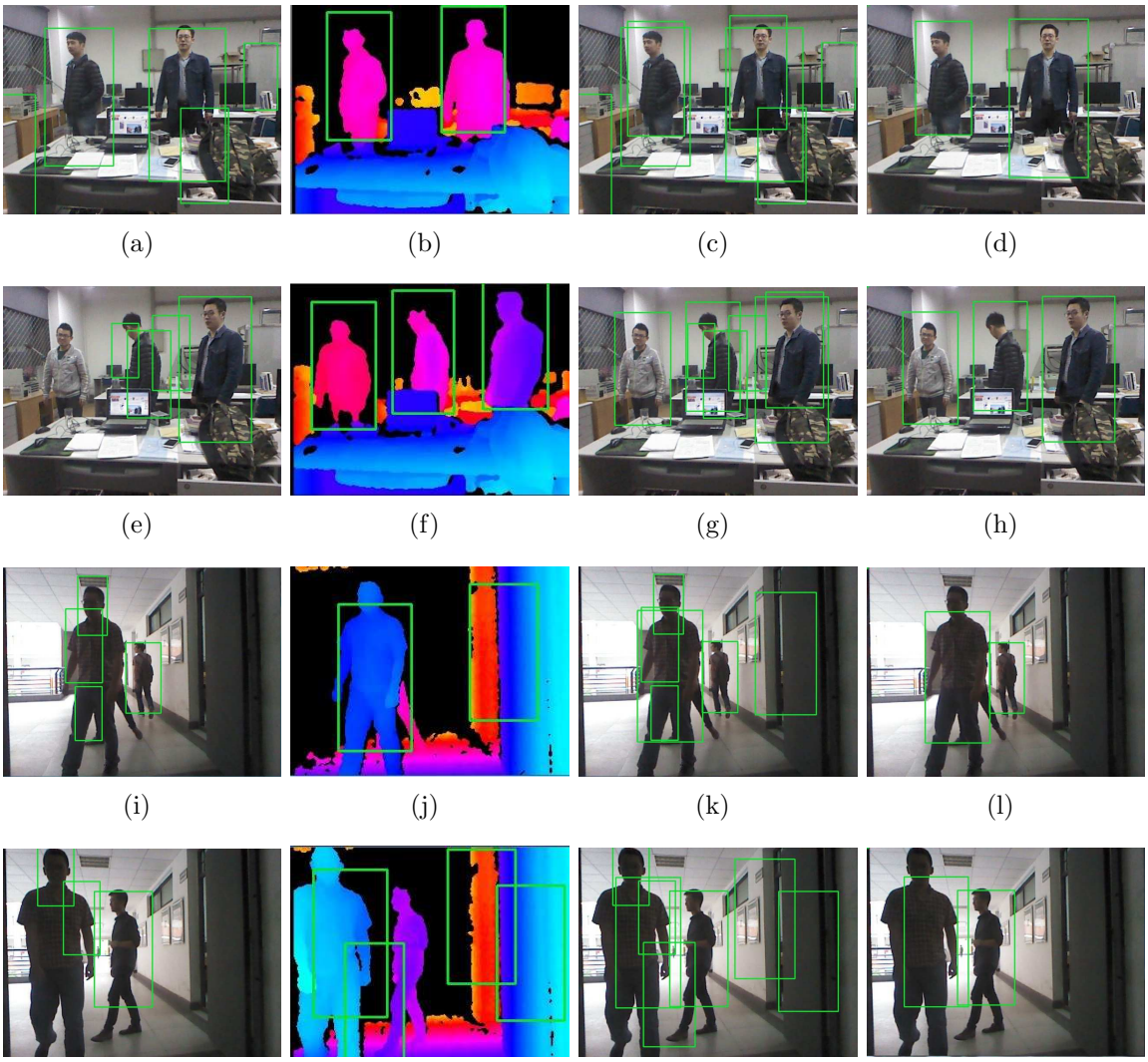
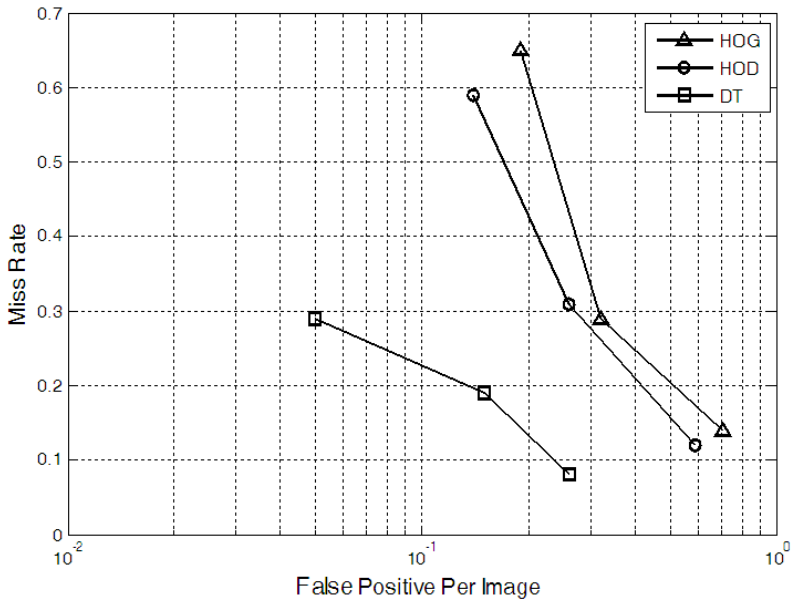


FIGURE 4. The results of the combination of color and depth detection

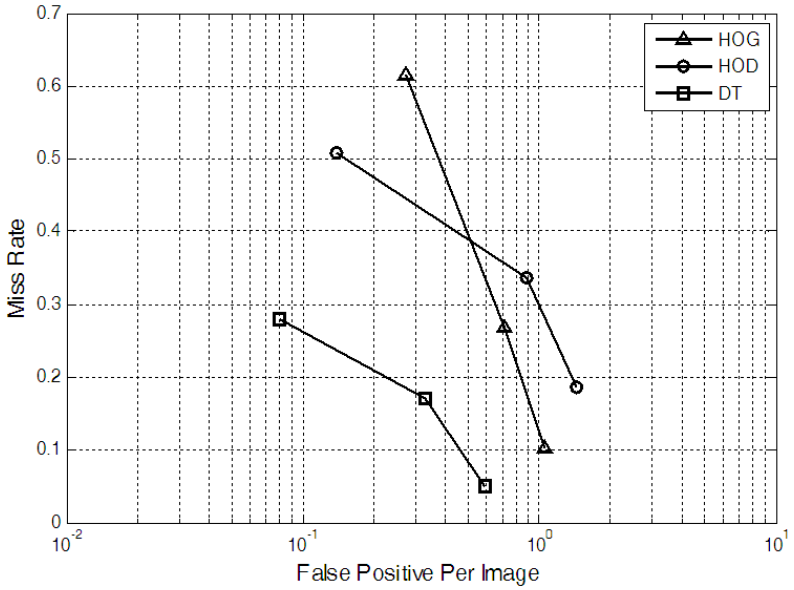
TABLE 1. Miss rate and false positive per image in the detections

Results	Algorithms	Complex environment			Weak light environment		
		Dataset1	Dataset2	Dataset3	Dataset1	Dataset2	Dataset3
False Positive Per Image	HOG	0.19	0.32	0.71	0.27	0.72	1.05
	HOD	0.14	0.26	0.59	0.14	0.88	1.44
	DT	0.05	0.15	0.26	0.08	0.33	0.59
Miss Rate	HOG	0.65	0.29	0.14	0.62	0.27	0.10
	HOD	0.59	0.31	0.12	0.51	0.34	0.19
	DT	0.29	0.19	0.08	0.28	0.17	0.05

4. **Conclusions.** In this paper, we proposed a new method of human detection using depth images to enhance the detection effect by color images via the Kinect sensor. With the help of the HOG and HOD features we trained color and depth image human detectors, and adopted DT method to combine these detectors, and used the Squared Euclidean distance to evaluate similarity between DP and DT. Experiments results on the collected dataset showed that, the proposed method can effectively detect people both in complex or weak light environment. It also can be seen in the experiment, that the false positive and the miss rate are effectively reduced compared with the HOG or HOD. Study in the



(a)



(b)

FIGURE 5. False positive per image and miss rate curves, (a) the experimental results in a complex environment, (b) the experimental results in a weak light environment

multi-classification based on DT in order to identify different postures of human is the next step.

Acknowledgments. This work is partially supported by Fund of the major cultivation project of Education Bureau of Sichuan Province, China (No: 13ZC0003) and partially supported by Fund of the Chunhui plan project of Chinese Ministry of Education (No: Z2012028). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.886-893, 2005.

[2] F. Xu and F. Xu, Pedestrian detection based on motion compensation and HOG/SVM classifier, *Proc. of IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol.2, pp.334-337, 2013.
[3] Y. Shen, Z. Hao, P. Wang, S. Ma and W. Liu, A novel human detection approach based on depth map via Kinect, *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.535-541, 2013.
[4] L. Xia, C. Chen and J. K. Aggarwal, Human detection using depth information by Kinect, *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.15-22, 2013.
[5] L. Spinello and K. O. Arras, People detection in RGB-D data, *Proc. of IEEE International Conference on Intelligent Robots and Systems*, pp.3838-3843, 2011.
[6] L. Susperregi, B. Sierra, M. Castrillón, J. Lorenzo, J. M. Martínez-Otzeta and E. Lazkano, On the use of a low-cost thermal sensor to improve Kinect people detection in a mobile robot, *Sensors (Switzerland)*, vol.13, no.11, pp.14687-14713, 2013.
[7] L. I. Kuncheva, J. C. Bezdek and R. P. W. Duin, Decision templates for multiple classifier fusion: An experimental comparison, *Pattern Recognition*, vol.34, pp.299-314, 2001.
[8] A. Sajedin, R. Ebrahimpour and T. Y. Garousi, Electrodiogram beat classification using classifier fusion based on decision templates, *Proc. of the 10th IEEE International Conference on Cybernetic Intelligent Systems*, pp.7-12, 2011.
[9] A. P. Gritti, O. Tarabini, J. Guzzi, G. A. Di Caro, V. Caglioti, L. M. Gambardella and A. Giusti, Kinect-based people detection and tracking from small-footprint ground robots, *Proc. of IEEE International Conference on Intelligent Robots and Systems*, pp.4096-4103, 2014.
[10] J. Liu, Y. Liu, G. Zhang, P. Zhu and Y. Chen, Detecting and tracking people in real time with RGB-D camera, *Pattern Recognition Letters*, vol.53, pp.16-23, 2015.