Advanced Human Detection Using Fused Information of Depth and Intensity Images

Gyu-Hong Lee, Dong-Suk Kim and Chong-Min Kyung

Abstract Human detection systems have been applied to many applications such as intelligent vehicles and surveillance cameras with increasing demands on safety and security. The scope of previous works has been confined usually in color (or intensity) images. In this chapter, we present a complete human detection system using the information on both depth and intensity images. First, we apply a segmentation algorithm to a depth image. Then we merge the segmented regions and generate Region-Of-Interests (ROIs) which may contain a human, considering experimentally determined horizontal overlap and aspect ratio, respectively. Second, we use a newly proposed feature descriptor, Fused Histogram of Oriented Gradients (FHOG), to extract feature vectors from the ROIs applied in both depth and intensity images. Finally, we check the presence of humans in the ROIs with linear SVM. Following the basic principles of Histogram of Oriented Gradients (HOG), we develop this FHOG descriptor to utilize both gradient magnitudes of depth and intensity images. With our datasets obtained from Microsoft Kinect sensor, the FHOG descriptor and overall system achieve a miss rate of 1.44 % at 10^{-4} FPPW and of 10.10 % at 1 FPPI, respectively. The computing time of proposed system is also significantly reduced. Experimental results show our system is able to detect humans accurately and fast.

Keywords Human detection · Pedestrian detection · Segment-based ROI generation · RGB-D data

G.-H. Lee (✉) · D.-S. Kim · C.-M. Kyung
Smart Sensor Architecture Lab, ITC Building (N1) #314, Korea Advanced Institute of Science and Technology, 291 Daehak-Ro, 305-701 Yuseong-Gu, Daejeon, Republic of Korea
e-mail: ggamid79@naver.com

C. Kyung (ed.), *Theory and Applications of Smart Cameras*,
KAIST Research Series, DOI 10.1007/978-94-017-9987-4_12

1 Introduction

Human detection is one of the most interesting research topics in computer vision. It has traditionally been developed for robotics because the human detection algorithm can extend the perception ability of a system. Further, with growing interest on safety and security, human detection has received considerable attention for surveillance cameras and intelligent vehicles. However, human detection in visible-spectrum images has great difficulty in applying it to those real applications because of insufficient detection performance and high computational cost. According to the pedestrian detection benchmark [1], pedestrian detection performance in visible spectrum images marked over 0.15 of miss rate at 1 false positive per image (FPPI) and took more than 6 s to process a single frame.

It is difficult to address the problems using visible-spectrum images alone, because there are a lot of discouraging factors to build a robust human detector in visible-spectrum images such as illumination changes, complex background, and various human clothing. Further, computing time for human detection is not a negligible issue. Running a heavy algorithm in real time requires additional hardware resources such as GPU (Graphic Processing Unit). This leads to the degradation of power efficiency of overall system. So, the final goal of our research is designing human detection system satisfying both accuracy and computing times. Many researchers have been tried to reach the goal and find various methods for detecting humans. Owing to the development of affordable RGB-D cameras such as Microsoft Kinect and Mesa SR4000, depth information has become a new clue for designing advanced human detection system. (Details in Sect. 2)

In this chapter, we propose complete human detection system using both depth and intensity images. Our contributions are as follows:

- We generate segment-based ROIs on depth images. It reduces the number of candidate windows to classify whether a given image contains human or not. As a result, both false positives and computing times can be reduced.
- We develop a new feature called FHOG which is based on Histogram of Oriented Gradients (HOG). By fusing the depth and the intensity information, contours of human are intensified and thus detection rate can be increased.

The rest of this chapter is organized as follows. Section 2 describes the related work. Section 3 presents our ROI generation method and new feature descriptor, FHOG. Section 4 shows experimental results compared with other approaches.

2 Related Work

A typical human detection system starts detecting regions that are highly likely to contain humans. Then, features that describe a human are derived from the windows and the descriptors are classified by a pre-trained classifier. The advent of affordable RGB-D cameras which provide reliable depth information brought advantages to the human detection systems. The advantages are apparent for two modules: feature extraction and ROI generation.

Typically, features have been mainly extracted from intensity (or color) images. They use texture or gradient information in the image. Haar-like feature is a representative feature of texture-based method [2, 3]. It considers the difference of sum of intensity values in pre-determined rectangles. It works well for face detection but does not work satisfactorily for detecting humans. For human detection, HOG is the well-known descriptor [4]. It focuses on the discontinuities in image intensity. Local gradient orientation histograms are computed and normalized to make the feature more robust to the illumination changes in the image. The shape of head, shoulder, and legs is the most fundamental feature of a human in the HOG descriptor.

After depth information is widely used, new features are introduced which utilize depth information. In [5], Spinello et al. proposed Histogram of Oriented Depths (HOD) descriptor which locally encodes the direction of depth changes. They showed that detection accuracy increases because depth data is not affected by illumination changes unlike visible-spectrum images. Fusing features of intensity and depth images were proposed in [6]. They simply concatenate HOG features with HOD features. As a result, detection performance is improved, but it needs more computing time because feature dimensions are increased.

In visible-spectrum camera-based human detection system, fixed-size window is densely scanned through entire intensity (or color) image to extracting ROIs of including human [4]. To cover various heights of humans, images are scaled up or down, so the number of ROIs is significantly increased. Hence, it decreases overall speed of human detection system. To solve this problem, Q. Zhu et al. capture salient features of humans automatically and discriminate the appropriate regions [7]. In [5], they distinguish compatible scales likely to fit a height of human from a predetermined scale map and test the scaled windows.

Unlike intensity information, depth information is advantageous for extracting ROIs. In [8], they utilize graph-based segmentation algorithm on depth image to generate ROIs by merging segments based on their location.

We take a similar approach with B. Choi et al.'s methods. In our approach, the segments are merged perpendicularly as overlap ratio and ROIs are generated based on the aspect ratio of human body. Then the proposed FHOG descriptor is applied to these ROIs for classification.

3 Proposed Method

The proposed system consists of three stages as shown in Fig. 1. In the first stage, image segmentation is applied to depth images and ROIs are generated by combining the segments. Then features are extracted using fused information of intensity and depth images within ROIs, and finally ROIs are classified using a

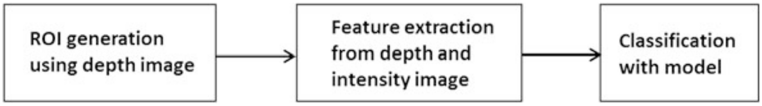

ROI generation using depth image
Feature extraction from depth and intensity image
Classification with model

Fig. 1 Flows of the proposed human detection system

linear Support Vector Machine (SVM). Depth images are obtained from Microsoft Kinect sensor. This sensor provides both color and depth images. To use both images in the same image coordinate, image rectification is performed offline.

3.1 ROI Generation Using Depth Images

In depth image, intensity variation within same object is smaller than that of gray (or color) image because depth is not affected by textures on object. So the same object tends to have similar depth values. The characteristic becomes a motivation to utilize image segmentation to depth images for clustering similar depth regions.

(1) Depth image segmentation

We use the mean shift algorithm to segment depth images. The mean shift algorithm is a mode seeking algorithm that was made popular for image segmentation by Comaniciu et al. [9]. The size and the number of segments are decided according to the parameter set. The spatial, range and minimum size parameters for mean shift segmentation are determined experimentally to separate human body from background in depth images. As a result, labeled segments are acquired. Then we find the left-uppermost coordinate and the right-lowermost coordinate of each segment and calculate the width and height. The information is used to merge segments.

(2) Segment merging

In this step, we merge segments to obtain human candidates. Ideal results of the segmentation are that human is represented in one segment. However, human is usually separated into several segments (see Fig. 2). To elicit an intact human candidate from these segments, the horizontal overlap ratio (r_o) is used to combine pairs of segments. The horizontal overlap ratio is defined as

$$r_o = \frac{\mathrm{hlength}(a \cap b)}{\mathrm{hlength}(a \cup b)} \tag{1}$$

where $\mathrm{hlength}(a \cap b)$ and $\mathrm{hlength}(a \cup b)$ are horizontal intersection and union of pairs of segments, respectively. If horizontal overlap ratio between two segments is greater than a threshold, two segments are merged as depicted in Fig. 3. We set the threshold value as 0.35.

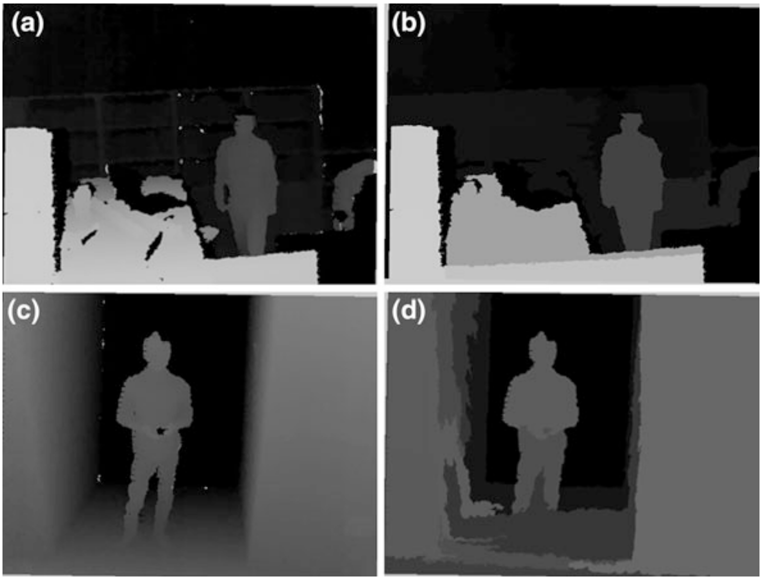


Fig. 2 Image segmentation results. (**a**) and (**c**) are depth images. (**b**) and (**c**) are segmented results of (**a**) and (**c**), respectively. A human in (**b**) is separated as one segment, but (**d**) is not

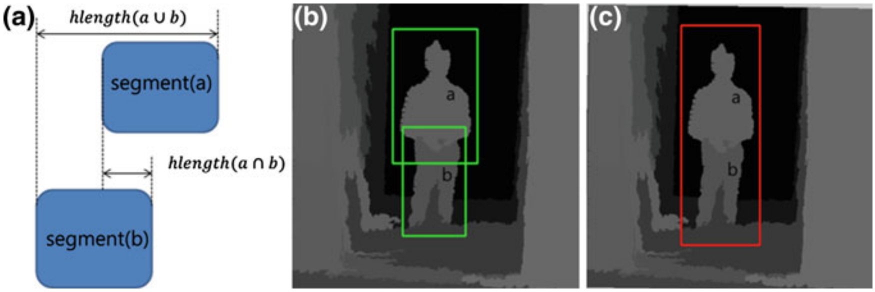


Fig. 3 Concepts for merging segments. **a** Parameters for calculating horizontal overlap ratio, **b** before merging, and **c** after merging

(3) ROI generation

A large set of candidates are generated from previous steps (Fig. 4a). Here, we investigate the aspect ratio (r_a) of candidates to filter out impractical candidates. The aspect ratio of a segment (a) is defined as

$$r_a = \frac{\text{width}(a)}{\text{height}(a)} \tag{2}$$

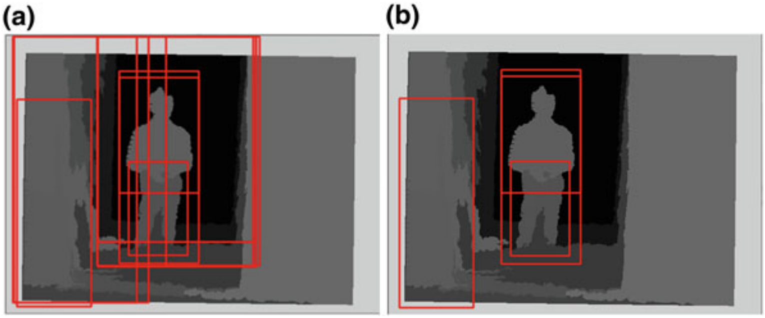


Fig. 4 ROI generation results. **a** Example of candidate generation, **b** final result of ROI generations

where, width (a) and height (a) are the width and height of candidates (a), respectively. All the candidates of satisfying the predetermined aspect ratio are selected as ROIs. The aspect ratio is determined as the ratio of human body (between 0.25 and 1 in our system). The ROIs are confined in a bounding box with their coordinates acquired in step 1. As illustrated in Fig. 4b, ROIs of containing human are successfully generated with a small number of total ROIs.

3.2 *Feature Extraction from Depth and Intensity Image*

We propose a new feature extraction method which is an extended version of HOG for human classification in depth and intensity images. In this section, we introduce our feature called FHOG after a summary of the HOG descriptor.

(1) Histogram of Oriented Gradients (HOG)

HOG is the most popular feature for human detections. HOG feature expresses a sample image on the basis of its local shape and appearance using histograms of gradient orientation. It computes gradient magnitude and orientation in a fixed-size window called detection window. Then it builds histograms with orientation bins for each cell which is densely subdivided regions in the detection window. Votes of the histograms are accumulated into the orientation bins. The histograms are normalized within a group of cells, which is called a block. The normalization process is necessary to make the feature more robust to the effect of illumination changes. Finally, extract the HOG descriptor by a feature vector concatenation of all the normalized histograms.

(2) Fused Histogram of Oriented Gradients (FHOG)

We developed a new feature extraction method for human classification in depth and intensity images based on HOG descriptor. Our descriptor takes advantage of the information from both depth and intensity image. The reasons for using both images are as follows: Generally, intensity image gives detailed information of an object because it has abundant textures. But it is vulnerable to illumination changes, and also when the background is complex, the rich textures can increase the false positives. In contrast, depth image is robust to illumination changes and can alleviate the effect of the complex background. However, it is sensitive to low return signals and may give insufficient data for detecting humans. Thus, the feature of using both complementary images can be very powerful and promising.

In a depth image, the gradient magnitude of a human contour appears relatively stronger because textures of object and background are ignored. So we can obtain amplified gradient magnitudes around human contours by adding the gradient magnitude of depth image to the intensity image (Fig. 5a). The amplified gradient magnitude (M_s) at pixel(x, y) can be defined as:

$$M_s(x, y) = M_i(x, y) + \omega \cdot M_d(x, y) \tag{3}$$

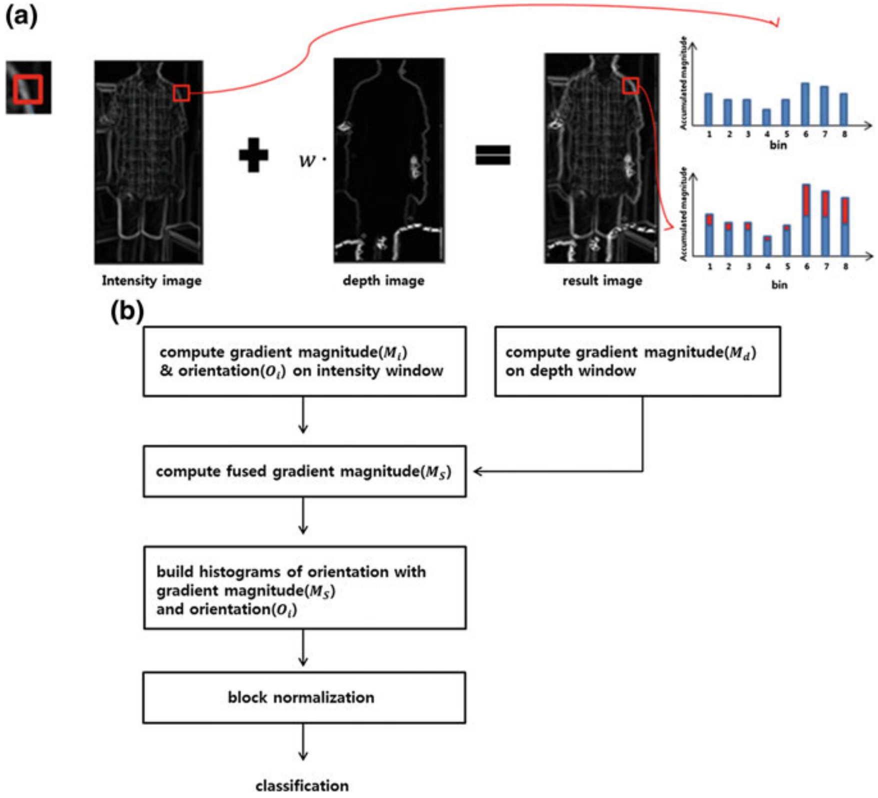

(a)
w ·
Intensity image
depth image
result image
Accumulated magnitude
1 2 3 4 5 6 7 8
bin
Accumulated magnitude
1 2 3 4 5 6 7 8
bin
(b)
compute gradient magnitude(M_i) & orientation(O_i) on intensity window
compute gradient magnitude(M_d) on depth window
compute fused gradient magnitude(M_S)
build histograms of orientation with gradient magnitude(M_S) and orientation(O_i)
block normalization
classification

Fig. 5 **a** Procedure used for extracting fused histogram of oriented gradients, **b** an overview of our feature extraction method

Table 1 FHOG parameters

Cell size	12 × 12 pixels
Block size	2 × 2 cells
Overlap of block	1
The number of bins	9 (unsigned)
Vote method for histogram	Gradient magnitude
Normalization factor	L2-norm

where $M_i(x, y)$ and $M_d(x, y)$ are the gradient magnitude at pixel(x, y) of intensity and depth image, respectively, and ω is a weighting factor to control the amount of effect of depth's gradient magnitude. We use the amplified gradient magnitude (M_s) to build histograms for each cell. The overview of our feature extraction procedures is illustrated in Fig. 5b.

In a similar method, S. Wu et al. extract features by combining depth and intensity images [6]; while their method increases the dimensions of feature vectors, our method maintains the dimensions of feature vectors although the descriptors are more discriminative. As a result, we can save memory space and decrease the execution time. The FHOG parameters which are used in this experiment are described in Table 1.

3.3 Classification with Model

We use linear SVM for classification. The linear SVM is a binary classifier looking for the most suitable hyperplane as decision function defined as

$$h(x) = \sum w_i x_i + b \tag{4}$$

The optimal $h(x)$ is sum of the inner product of the feature vectors x_i and the weight vectors w_i. Here, w_i and b are obtained from supervised learning with a training set. The sign of $h(x)$ decides whether the features are in-class or out-of-class. We use two sets of training examples which are externally and internally obtained from Kinect sensors.

4 Experimental Results

The proposed human detection system was tested on two different datasets that are externally and internally obtained from Kinect sensors. In this section, we introduce these two datasets and show the evaluation results in terms of the performance of ROI generation, feature extraction, and overall system.

4.1 Dataset

(1) Public dataset

The first dataset is RGB-D dataset provided by Spinello et al. [5]. The dataset has been taken in a university hallway using three vertically mounted Kinect sensors. It includes three sequences of videos and a total of 1648 people in 1088 frames are labeled. As shown in Fig. 6, people in this dataset are upright and completely visible or partially occluded. We use this dataset to compare the performance of our ROI generation method with the method proposed by B. Choi et al. [8]. 1000 positive examples and 4500 negative examples are extracted in depth images for training, and 200 depth images are used for testing (Fig. 6).

(2) Our dataset

We have used Kinect sensor mounted at a height of 1.7 to 2 m from the ground to collect our own RGB-D dataset. The dataset was taken in various indoor places (such as university hallway, laboratory, underground parking-lots, and classroom). Our dataset includes people who are upright and fully visible or partially occluded. We use the dataset to evaluate the performance of feature extraction and overall system. To evaluate feature extraction method, 1025 positive samples and 5000 negative samples are used for training and 1325 positive samples and 11774 negative samples are used for testing. To evaluate the overall system performance, 1175 images containing 1316 people are used. Figure 7 shows some examples of our dataset.

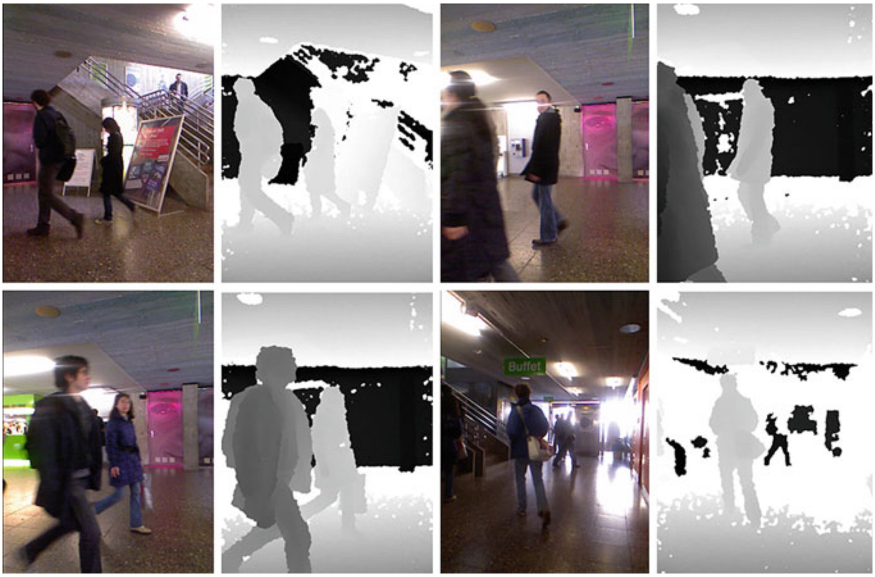


Fig. 6 Examples of public dataset

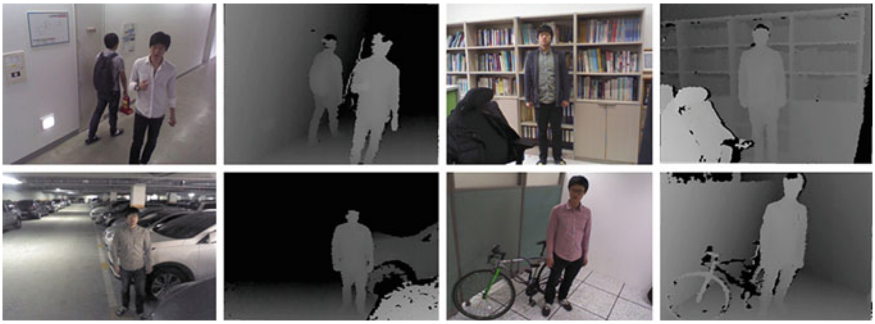


Fig. 7 Examples of our dataset

4.2 Evaluation of ROI Generation Method

In this section, we evaluate our ROI generation performance by comparing to the B. Choi et al.'s method which is based on graph-cut algorithm [8]. They use the HOD descriptor and it is tested on the public dataset which was mentioned in previous section. To compare the performance, we implemented the HOD algorithm and used the same dataset. To quantify performance, we plot Equal Error Rate (EER) curves. The EER is the matching point between recall and precision. In the EER curves, if the matching point is located on the right-uppermost areas, it can be considered that the accuracy of the detection system is relatively high. Our method achieved an EER of 87 %, which performs slightly better than the 84 % of graph-based ROI (Fig. 8).

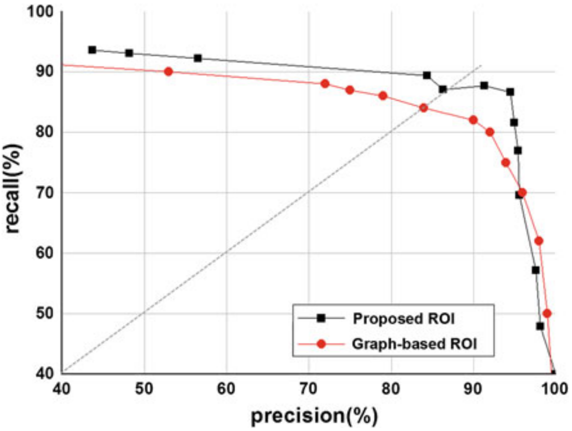


Fig. 8 Equal Error Rate (EER) curves

4.3 Evaluation of Feature Extraction

We compare the performance of our FHOG feature with HOG [4], HOD [5] and HOG-HOD [6]. These features are tested on our dataset and we see the per-window performance. Detection Error Tradeoff (DET) curves on a log–log scale are used to evaluate the performance of features, *i.e.,* miss rate (1-recall) versus false positives per window (FPPW). First, we test our FHOG feature to determine the optimal weighting factor (ω) by varying the value from 1 to 100. As shown in Fig. 9a, when the ω is 10, FHOG achieved the lowest miss rate (1.44 %) at 10^{-4} FPPW. For the other features, HOG, HOD, and HOG-HOD achieved a miss rate of 10.94, 8.31, and 6.72 % at 10^{-4} FPPW, respectively (Fig. 9b). It seems that using depth images for detecting human is helpful to improve the detection rate. Further, FHOG reduces the miss rate by 5.28 % as compared to the HOG-HOD. This means that our fusion method strengthens the features better than simple feature concatenation approach.

Fig. 9 Detection Error Tradeoff (DET) curves. **a** The effect of ω, **b** different descriptors on our dataset

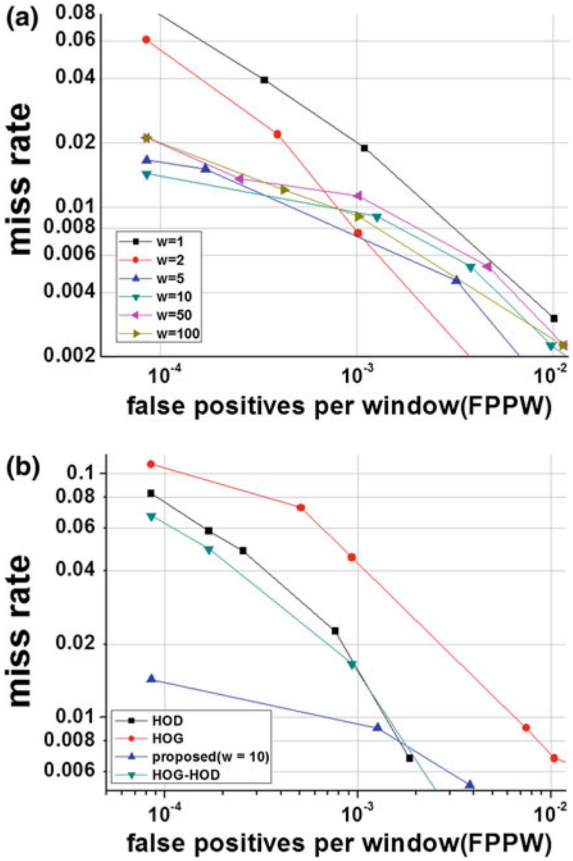
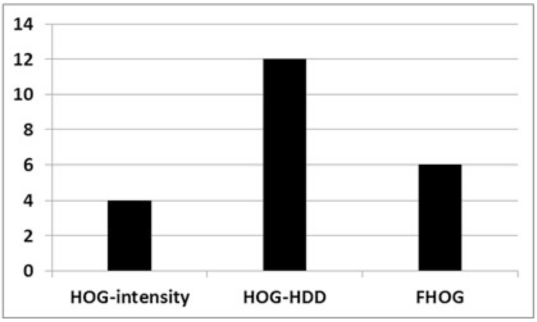


Fig. 10 Computing time of each descriptor



Further, we evaluate the computing time of each features. Figure 10b shows that FHOG is computed faster than HOG-HOD, but slower than HOG (or HOD). Interestingly, our feature can detect better to the partially occluded human (see Fig. 11). This result indicates that fused feature of depth and intensity works robustly on complex background and in the case that the contours of human are lost in an image.

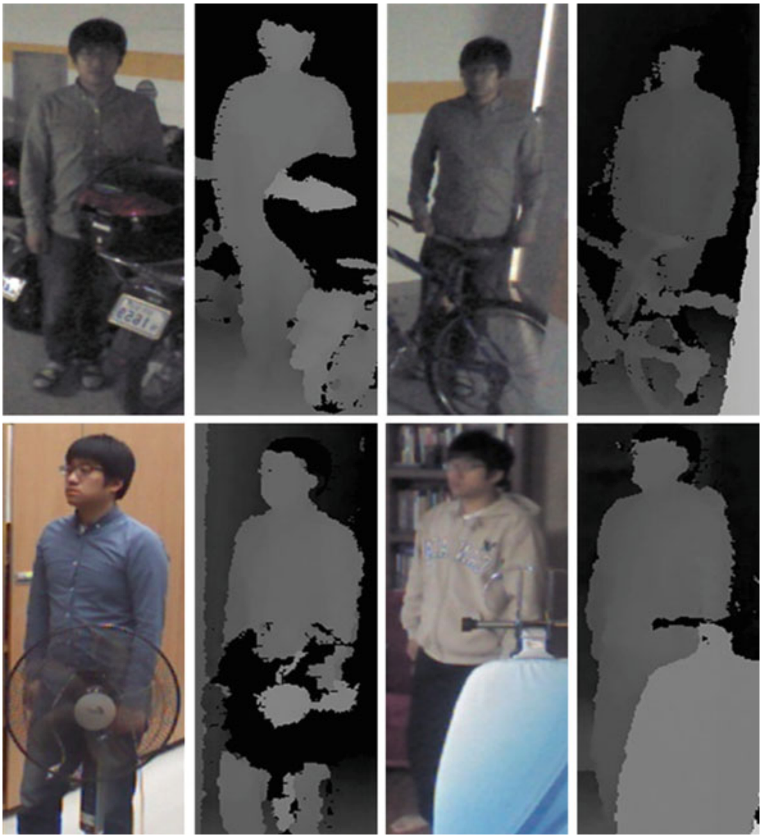


Fig. 11 Examples of detecting a partially occluded human

4.4 Evaluation of Overall System

In this section, our human detection system is compared against a reference system on our data set. Ideally, comparing our system to the system proposed in [8] is more precise since they use a similar strategy (graph-based ROI + HOD) to ours. However, we did not compare our system to [8] since the original implementation of graph-based ROI was not available to us. So we designed a system using FHOG descriptor and sliding window technique for extracting ROIs. To avoid the circumstances that the ROIs do not contain humans, we scanned the image as densely as possible. We plot DET curves by the miss rate versus false positives per image (FPPI) to evaluate per-image performance. As shown in Fig. 12a, our system performs better than the reference system. Further, our system takes 2.65 s to process a frame, while the reference system takes 6.95 s. The experiments were conducted on a computer with an Intel core i5 processor (Fig. 12b). Figure 13 shows the examples of human detection.

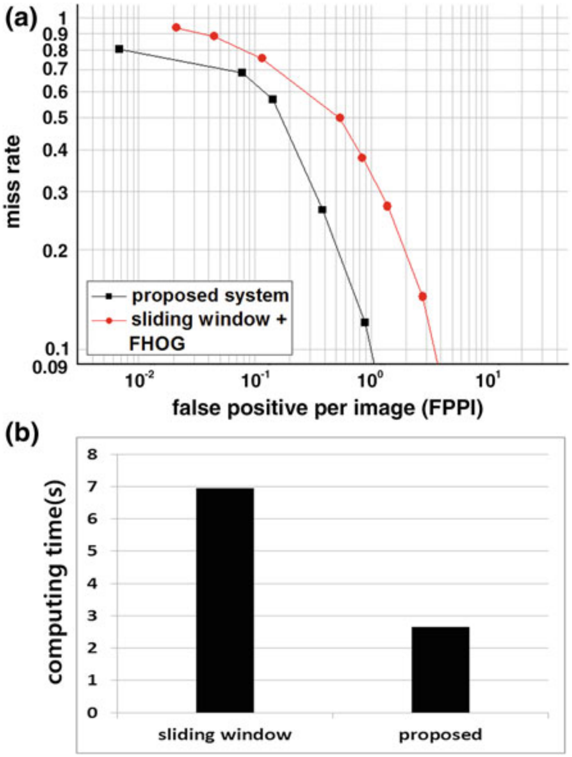


Fig. 12 **a** DET—human detection performance comparison results, **b** computing time of each system

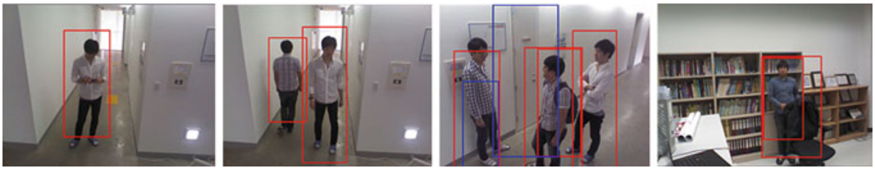


Fig. 13 Examples of human detection. Red boxes—true positives, blue boxes—false positives

5 Conclusions

In this chapter, we introduced an advanced human detection system using depth and intensity images. First, we applied image segmentation to depth images and generated feasible ROIs in consideration of the predetermined aspect ratio. This process significantly reduces the false positives and the computing times. Further, a new descriptor (FHOG) that fusing depth and intensity images is proposed for feature extraction. The FHOG achieved a recall of 98.56 % at 10^{-4} FPPW and it takes about 6 ms for processing a detection window (48 × 96 pixels). Further, the FHOG worked well for detecting partially occluded person. The overall system (mean-shift based ROI + FHOG) achieved a recall of 89.90 % at 1 FPPI and it significantly reduces the computing time by 61.87 % compared to the reference system (Sliding window + FHOG).

Acknowledgments This work is supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as the Global Frontier Project.

References

1. Dollar P, Wojek C, Schiele B, Perona P (2009) Pedestrian detection: a benchmark. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009, pp 304–311. IEEE
2. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vision 57(2):137–154
3. Viola P, Jones MJ, Snow D (2003) Detecting pedestrians using patterns of motion and appearance. In Proceedings ninth IEEE international conference on computer vision, 2003, pp 734–741. IEEE
4. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, vol 1, pp 886–893. IEEE
5. Spinello L, Arras KO (2011) People detection in RGB-d data. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), 2011, pp 3838–3843. IEEE
6. Wu S, Yu S, Chen W (2011) An attempt to pedestrian detection in depth images. In 2011 Third Chinese conference on intelligent visual surveillance (IVS), pp 97–100. IEEE

7. Zhu Q, Yeh M-C, Cheng KT, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 1491–1498. IEEE
8. Choi B, Mericli C, Biswas J, Veloso M (2013) Fast human detection for indoor mobile robots using depth images. In: IEEE international conference on robotics and automation (ICRA), 2013, pp 1108–1113. IEEE
9. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619