

# Stereo-based region of interest generation for real-time pedestrian detection

Joohee Kim · Maral Mesmakhosroshahi

Received: 15 July 2013 / Accepted: 2 September 2013 / Published online: 26 September 2013  
© Springer Science+Business Media New York 2013

**Abstract** Pedestrian detection is one of the major goals in advanced driver assistance systems (ADAS) which has become an active research area in recent years. In this paper, we present a stereo based pedestrian detection system by fusing the depth and color data provided by a stereo vision camera on a moving platform. The proposed method uses an adaptive window for region of interest (ROI) generation using dense depth map. The extracted candidates are then applied to a Histogram of Oriented Gradients (HOG) feature descriptor to refine ROIs and Support Vector Machine (SVM) is used to classify them into pedestrian and non-pedestrian classes. The system is tested on a stereo based DAS dataset and results show that our system is able to detect pedestrians with different scales and illumination conditions and in presence of partial occlusion.

**Keywords** Advanced driver assistance system · ROI generation · Pedestrian detection · Stereo vision · Depth map

---

This work was supported by the Technology Development Program for Commercializing System Semiconductor funded By the Ministry of Trade, industry & Energy (MOTIE, Korea). (No. 10041126, Title: International Collaborative R&BD Project for System Semiconductor).

---

J. Kim (✉)  
Department of Electrical and Computer Engineering,  
Illinois Institute of Technology, Chicago, IL, USA  
e-mail: joohee@ece.iit.edu

M. Mesmakhosroshahi  
Multimedia Communications Laboratory,  
Illinois Institute of Technology, Chicago, IL, USA  
e-mail: mmesmakh@hawk.iit.edu

## 1 Introduction

In an effort to reduce driver errors in being the major cause of traffic accidents, there is a lot of research being conducted into the development of advanced driver assistance systems (ADAS). ADAS is a system aimed at helping the driver in tasks such as pedestrian and vehicle detection, traffic sign recognition, lane detection, etc. Such systems are classified into two categories based on the nature of the sensor used to acquire data. In the first category, active sensors such as lasers and radars are used that detect the distance of objects by measuring the travel time of the signal omitted by the sensor and reflected by the object. These sensors have high angular resolution and can detect long distance objects, however they have some drawbacks. A major drawback of active sensors is that colors and patterns cannot be detected, making object recognition very difficult. In the second category of ADAS systems, passive sensors and especially vision-based sensors are used that are relatively low cost and provide the possibility of object classification and texture and color detection. For this reason, passive sensors are preferred because image processing techniques can be used for interpreting and processing the information.

Driver assistance systems have several modules for traffic monitoring, driver's state monitoring, driver's communication and reasoning. Among these modules, traffic monitoring can be performed using vision based techniques. The goal of traffic monitoring module is to detect and track vehicles, people and road lanes and recognize traffic signs.

One of the main goals of driver assistance systems is to locate and track stationary and moving pedestrians in the path of the vehicle. Pedestrian detection is a challenging

task due to several complexities. These challenges can be summarized as follows:

- ADAS systems must be able to detect pedestrians in different environments including highways, rural and urban areas with cluttered backgrounds. Also, the system must be able to work properly in existence of occlusion and multiple pedestrians which is very common in urban scenarios.
- Environmental changes such as illumination and weather conditions can affect the performance of pedestrian detection.
- Pedestrians' shapes and postures are non-uniform. Large variety of people's appearance and cloths makes pedestrian detection and tracking a very challenging task.
- Unlike other human detection applications, in ADAS systems we have to deal with the camera motion too. Therefore, estimating the pedestrians' size and motion pattern is not easy from a moving platform.

Most of the pedestrian detection algorithms include two major steps: region of interest (ROI) generation and pedestrian classification. In some cases, a tracking and refinement step is also added to improve the detection rate. Figure 1 shows the general block diagram of pedestrian detection.

In this paper, we propose a pedestrian detection algorithm by combining the ROI generation using dense depth map and HOG feature descriptor in the color image to create a system that is invariant to scale and illumination and can handle occlusion and multiple objects.

The structure of the paper is as follows: In Section 2, we review previous work related to pedestrian detection. In Section 3, the proposed pedestrian detection algorithm is explained in detail. In Section 4, we present experimental results and in Section 5, the paper is concluded.

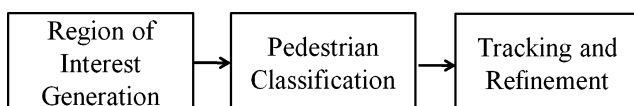
## 2 Related works

Several methods have been proposed for detecting humans with a static camera, however, pedestrian detection from a moving platform is a more complex task. Early pedestrian detection methods used a monocular camera mounted on a moving car and used techniques such as contour extraction and chamfer distance [1]. A literature survey on pedestrian detection for driver assistance systems is done in [2]. A part based SVM approach is presented in [3] for DAS systems. An algorithm for pedestrian detection in crowded

scenes is proposed by [4]. A pedestrian detection method using a moving camera is proposed in [5] which focuses on sudden detection of partially visible pedestrians in a three level approach. In [6], an on-board pedestrian detection and warning system is described using a monocular camera. In this algorithm, crosswalks are detected by image processing methods and optical flow is used for detecting moving objects on the crosswalk.

However, pedestrian detection and tracking performed by monocular cameras and only RGB data is not reliable. Recently, several pedestrian detection methods based on stereo vision have been proposed. In general, stereo vision based pedestrian detection improves the detection and tracking performance by exploiting depth information. Early stereo-based approaches for object detection using moving cameras were designed for robotic vehicles such as NASA's Mars rover [7] that uses 3-D optical flow for detection. In [8], a GPU-based system is proposed for pedestrian detection in mobile robots using stereo vision that uses edge properties to extract ROIs. In this system, HOG and SVM are used for classifying pedestrians.

In applications such as driving assistance systems, we have to deal also with the high speed of the car. In the past few years, a few approaches have been proposed for stereo-based pedestrian detection using a camera on a moving vehicle. One of the stereo based pedestrian detection method proposed in [9] uses dense depth for candidate window generation. In [10], a stereo-based preprocessing for human shape localization using  $v$ -disparity is proposed. In [11], a multi-cue pedestrian detection system is presented using SVM and Adaboost. A real time stereo-based system is proposed in [12] that uses ROI detection by projecting the stereo data into a polar-perspective map and classifies objects by computing geometric features. Another stereo-based method is proposed in [13] which uses Viola-Jones detection method [14] for ROI generation in left and right image and performs tracking based on correlation. The stereo-based system presented in [15] is a stereo modification of the system introduced in [1]. A method based on dense stereo is proposed in [16] that uses dense depth information for both detection and classification. A 2D-3D based pedestrian detection algorithm is presented in [17] that selects ROIs by estimating the road plane parameters. A Real AdaBoost and a combined set of Haar wavelets and edge orientation histograms are used in this algorithm for classification. In [18], dense stereo along with road and camera parameters are used for ROI generation.



**Fig. 1** Block diagram of pedestrian detection

## 3 Proposed algorithm

In this paper, we propose a new pedestrian detection algorithm which exploits both color and depth information for

stereo vision based DAS. As discussed above, lots of stereo based methods are available, however, a few of them use dense depth map in the detection process. In our proposed algorithm, we take advantage of the dense depth to present a novel method for ROI generation which is one of the most important steps in detecting pedestrians since missing any pedestrian in this step cannot be recovered.

In traditional methods, possible pedestrian locations were found by exhaustive search. However, exhaustive search based techniques incur huge computational complexity and have several drawbacks in detecting pedestrians with different scales, illumination conditions and in presence of occlusion. Also, because of the camera motion in DAS applications, scale of each object is changing frame by frame and it is very important to have a scale invariant ROI generation method to be able to detect objects in each frame. Motivated by these observations, we present an adaptive windowing method for ROI generation using dense depth maps. The detected ROIs are applied to a feature descriptor such as Histogram of Oriented Gradients for classification. Figure 2 shows an overview of our system.

### 3.1 ROI generation using dense depth maps

A depth map is an image in which the value of each pixel is related to the distance of that point from the camera in real world. Therefore, these depth values can be used to estimate the average size of the human at each distance. Stereo vision can be used to estimate depth map. Both sparse and dense depth maps have been used for pedestrian detection applications, however, dense depth map gives us more information. In our proposed method, we first divide the dense depth map into several layers and define rectangular windows with sizes varying for each depth layer based on the depth values. Then, by comparing the number of pixels that fall into the window with a predefined threshold, we decide whether the region contains a pedestrian or not. Our ROI generation method can be divided into four main steps of layering depth map, extracting skeleton for each depth layer, defining an adaptive window for each layer and sliding it

on the skeleton of the objects in each layer and refining the extracted regions.

#### 3.1.1 Layering depth map

At first step, a depth image is divided into several sub-images where each sub-image contains objects with the same depth values. Two methods are used for layering depth map. In the first method, depth histogram is obtained and pixel values that belong to each histogram bin constitutes a sub-image. Figure 3 shows an example of a depth histogram and sub-images extracted from a depth map.

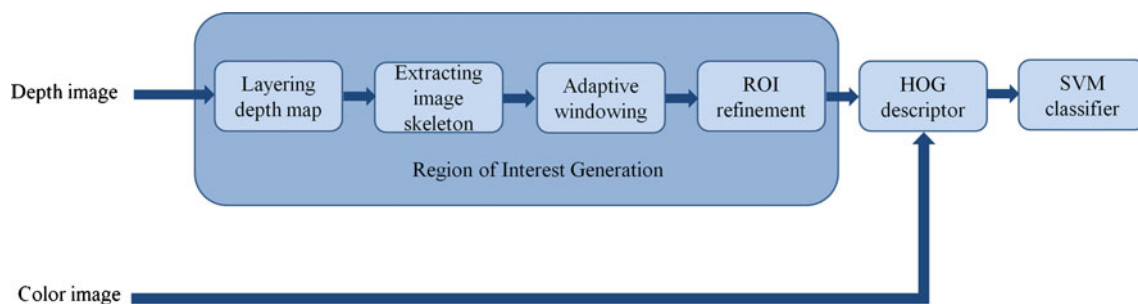
However, since depth maps are not perfect, we cannot expect to have the same depth values for each object. Therefore, by using depth histogram, pixels belonging to each object can fall into more than one sub-image. To solve this issue, we apply K-means clustering instead of image histogramming to the depth frame. K-means clustering is an unsupervised learning algorithm for grouping data into K classes where each observation belongs to the cluster with the nearest mean. K-means clustering can be applied to the depth image to segment it into  $k$  predefined clusters of  $(c_1, c_2, \dots, c_k)$  by solving Eq. 1 in an iterative process:

$$\operatorname{argmin} \sum_{i=1}^k \sum_{d_j \in c_i} \|d_i - \mu_i\|^2 \quad (1)$$

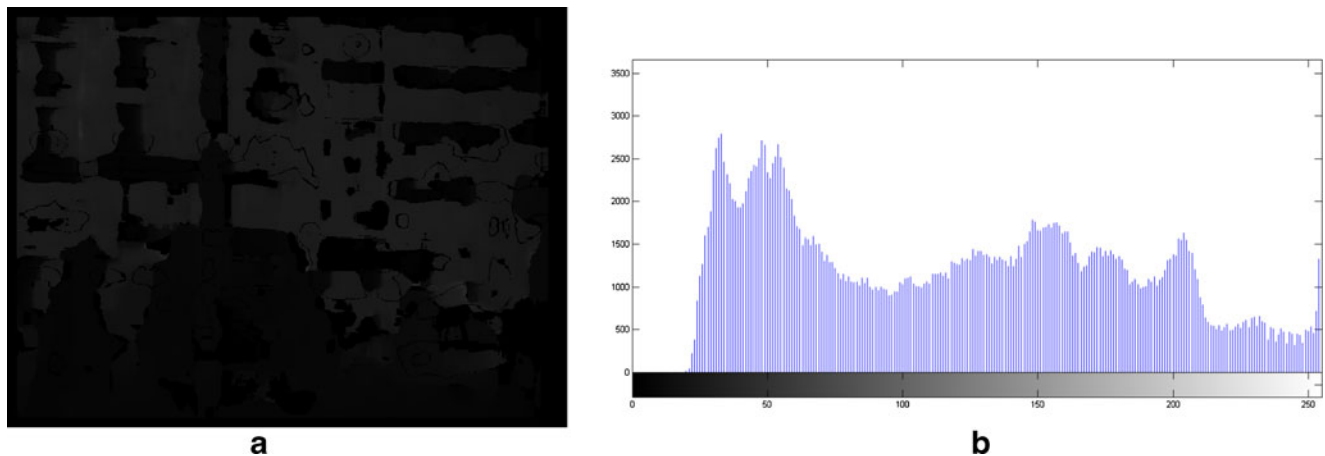
where  $(d_1, d_2, \dots, d_n)$  are depth values and  $\mu_k$  is the mean value of the group  $c_k$ . After clustering the depth map, pixel values that belong to each cluster, constitute a sub-image. Figure 4 shows a k-means clustered depth frame and a depth layer containing a pedestrian in both histogramming and k-means clustering methods.

#### 3.1.2 Extracting image skeleton

In exhaustive search based methods, all of the pixels are considered for searching but most of these pixels are not necessary for ROI decision. To reduce the complexity, we consider the pixels on the skeleton of the image only. For each sub-image, skeleton or medial axis is extracted and the



**Fig. 2** Overview of the proposed algorithm



**Fig. 3** **a** Depth image, **b** histogram of the depth image

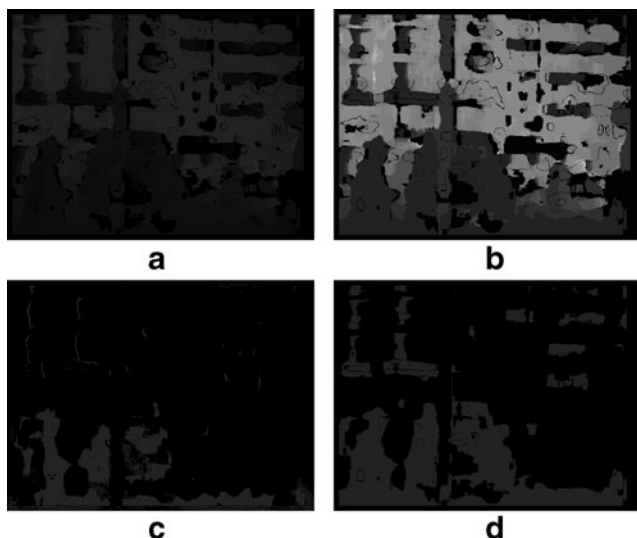
search window is slid on these pixels only. Image skeleton,  $S(X)$ , can be extracted using morphological operations given in Eqs. 2 and 3:

$$S(X) = \bigcup S_n(X) \tag{2}$$

$$S_n(X) = (X \ominus nB) - (X \ominus nB) \circ B, \tag{3}$$

where  $\ominus$  and  $\circ$  are morphological erosion and opening, respectively,  $n = 0, 1, \dots, N$ ,  $X$  is the input image,  $B$  is the structuring element and

$$nB = \underbrace{B \oplus \dots \oplus B}_{n \text{ times}} \tag{4}$$



**Fig. 4** **a** Depth frame, **b** k-means clustered depth map, **c** a depth layer containing pedestrian obtained by histogramming, and **d** a depth layer containing the same pedestrian obtained by k-means clustering

where  $\oplus$  is the morphological closing. Figure 5d shows the extracted skeletons for each sub-image.

### 3.1.3 Adaptive windowing

In ADAS systems, size of the pedestrian changes frame by frame because of the movement of the camera. Figure 6 shows a pedestrian in different distances from the camera.

Since the perceived size of an object is most strongly influenced by the object’s distance from the camera, we determine the detection window size for ROI search in each sub-image based on the depth value. Specifically, we first set the detection window size of the first depth layer which is closest to the camera as  $height/2 \times width/6$  considering the approximate shape of a human. Then, we adjust the height and width of the detection window in the n-th depth layer using Eqs. 5 and 6.

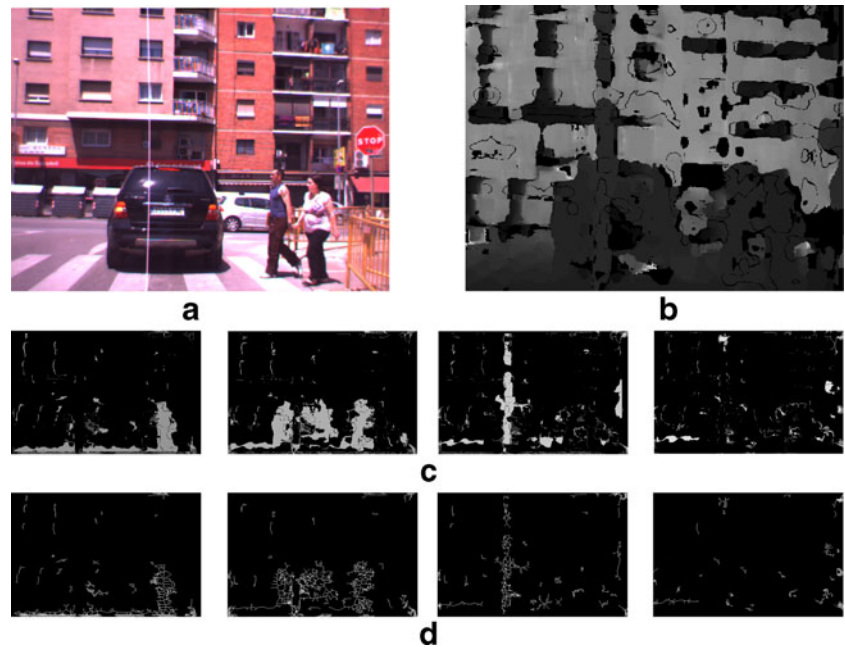
$$h_n = \frac{D - d_n}{D - d_1} \times h_1, \tag{5}$$

$$w_n = \frac{D - d_n}{D - d_1} \times w_1, \tag{6}$$

where  $d_1$ ,  $h_1$  and  $w_1$  are the depth value of the first layer and height and width at that layer,  $D$  is the maximum depth range (255 in our experiments), and  $d_n$  is the depth value in the n-th depth layer.

Then, we slide the detection window along the skeleton of the objects and calculate the amount of overlap between the detection window and a candidate region. If the amount of overlap is larger than  $1/3$ , the region is identified as a candidate region of interest. Figure 7 shows several examples of the detected ROIs.

**Fig. 5** Proposed ROI generation: **a** Color image, **b** depth map, **c** depth layers obtained using the depth histogram, and **d** skeleton of the objects in each depth layer



### 3.1.4 ROI refinement

Although searching for the regions of interest on the skeleton of depth layers reduces the number of search areas compared to the exhaustive search, there are still a large number of unwanted and repeated ROIs that must be removed. To decrease the number of ROIs, we first sorted the amount of overlaps between the detection window and the candidate regions of each layer in a descending order. Then, each region is considered as a region of interest if the amount of overlap is larger than  $1/3$  and the distance between the center of the detection window from the previously detected ROIs is larger than a certain threshold. We defined this threshold as  $T_0 \times h$  and  $T_0 \times w$  in vertical and horizontal directions respectively where  $T_0$  is between 0 and 1. The following algorithm shows the process of refining ROIs:

---

#### Algorithm 1 ROI refinement

---

**Data:** Amount of overlaps between the detection window and candidate regions in a descending order

**Result:** Choose ROIs

Retain the first candidate as ROI;

```

while Overlaps are larger than 1/3 detection window do
    find the center of the candidate region;
    if distance of the center from the previous ROIs are larger
      | than threshold then retain the region as ROI;
    else
      | discard the region;
    end
end

```

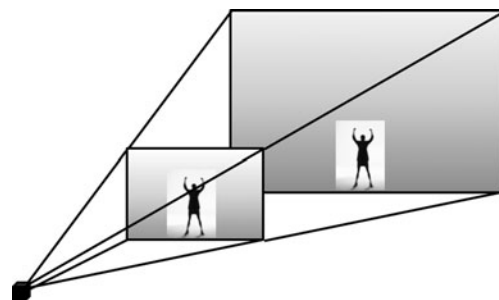
end

---

## 4 Experimental results

To test the performance of our proposed method, we used a stereo vision based DAS dataset provided by [17]. We used the classification and system packages of this dataset which provide both training and test sets and several video sequences. The test set contains 250 frames with  $640 \times 480$  pixels and range of 0 to 50m captured by a stereo vision camera from a moving vehicle in different illumination conditions and with different size and number of pedestrians. The training set contains 7649 negative and 1015 positive samples. Since the depth maps in the dataset have a poor quality and a limited range, we applied image inpainting and linear contrast stretching to improve the quality of the depth maps.

We performed several experiments to examine the performance of our ROI generation method and evaluate the performance of the extracted ROIs in pedestrian classification.



**Fig. 6** Pedestrian in different distances from the camera



**Fig. 7** An example of detected ROIs

#### 4.1 ROI generation

In the first experiment, we examined the performance of our ROI generation algorithm on 75 frames from the test set for various parameters defined for this module. Detecting pedestrians that are far from the camera is a very difficult task due to the small size of the pedestrians and poor quality of the depth maps. Therefore, we tested our method by thresholding the depth values and eliminating the regions where depth values are not available or they are larger than a fixed value. In the proposed method, two depth threshold values (specifically,  $T = 125$  and  $200$ ) are used for the experiments. Figure 8a and b show the detection rate and the number of extracted ROIs vs. the threshold used for ROI refinement.

According to the experimental results, we can observe that the detection rate improves when a smaller depth refinement threshold is used since pedestrians far from the camera (which are hard to detect) are discarded. In addition, we can also observe that the use of a larger depth refinement threshold decreases the computational complexity of the system

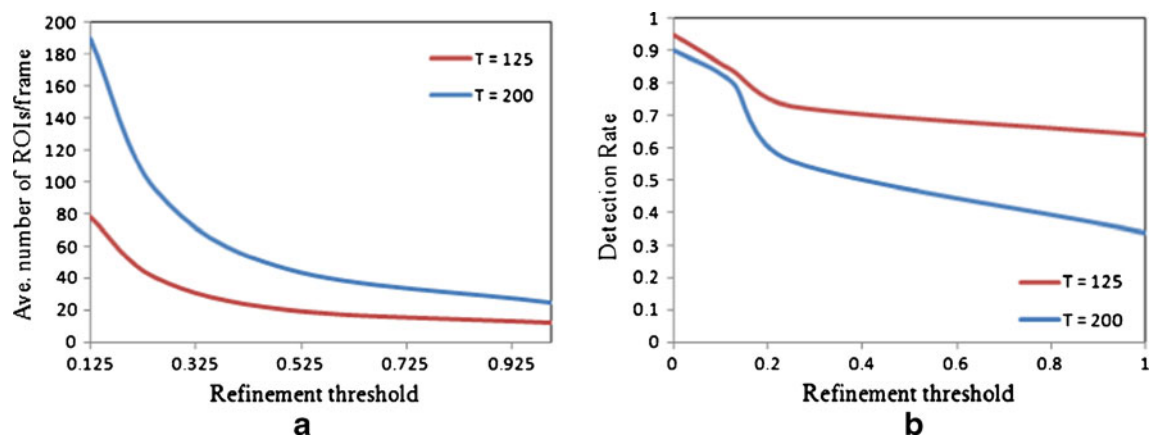
by reducing the number of ROIs but increases the number of missed pedestrians which is irrecoverable in the later stages.

#### 4.2 Pedestrian classification

In the second experiment, we evaluated the performance of the extracted ROIs in pedestrian classification. In the proposed system, Histograms of oriented gradients are computed for the extracted ROIs in both training and test sets and the performance is tested using support vector machine.

##### 4.2.1 Histogram of oriented gradients

Since ROIs contain both pedestrian and non-pedestrian regions, an object classification step is necessary to classify these regions. Among various silhouette based and appearance based methods, we use Histogram of Oriented Gradients (HOG) [19] for classifying ROIs since it shows the best performance in human detection. HOG is a feature descriptor inspired by Lowe's SIFT descriptor [20] and is



**Fig. 8** **a** Average number of extracted ROIs per frame vs. the refinement threshold, **b** detection rate vs. the refinement threshold

based on calculating orientation gradients in sub-blocks of a region. In HOG descriptor, the region is divided into smaller cells and orientation histograms are calculated for each cell. The histograms can be created by computing the orientation of gradients to form the histogram bins and its magnitude as a vote for the bin using Eqs. 7 and 8.

$$m(x, y) = \sqrt{L_x^2 + L_y^2} \quad (7)$$

$$\theta(x, y) = \arctan \frac{L_y}{L_x}, \quad (8)$$

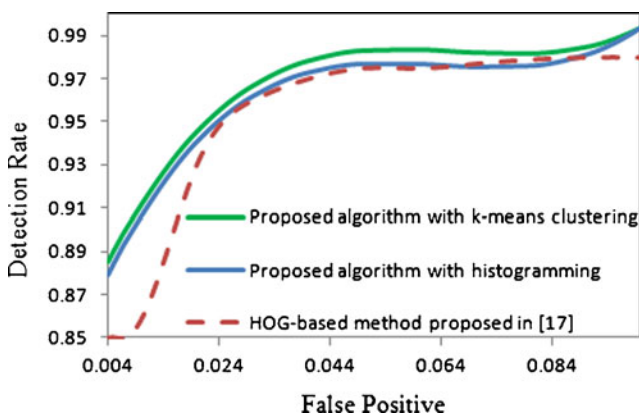
where  $m$  and  $\theta$  are magnitude and orientation at each pixel and  $L_x$  and  $L_y$  are gradients in  $x$  and  $y$  direction, respectively.

These cells are then grouped into larger regions called blocks and histograms of the blocks are normalized to improve illumination invariance. The normalized blocks are then concatenated to make the final feature vector.

#### 4.2.2 Support vector machine

We use a linear SVM to classify the feature vectors. To do this, SVM is first trained using 1000 pedestrian and 1000 non-pedestrian frames. Based on this support vectors, ROIs extracted from test frames are classified into pedestrian and non-pedestrian classes. ROIs extracted from the 250 frames of the test set using our ROI detection algorithm are then classified by this classifier. In the implementation of the ROI detector, the size of the initial bounding box is set to  $240 \times 107$  and the threshold for the filled area in the bounding box is set to  $1/3$ . For the HOG descriptor, the size of blocks and cells are set to  $3 \times 3$  and  $9 \times 9$ , respectively, and 9 bins are used for the histograms.

Figure 9 shows the ROC curves that compare the performance of the proposed method and the HOG based



**Fig. 9** ROC curve comparison between the proposed method and the HOG based method proposed in [17]

method proposed in [17]. Simulation results show that the proposed method outperforms the HOG based pedestrian detection method [17], especially when K-means clustering based depth map layering is used. The major reason for the improved performance of the proposed method is that the use of depth information makes ROI generation robust against illumination variations.

## 5 Conclusion

In this paper, a ROI generation method has been proposed for pedestrian detection in advanced driver assistance systems. The proposed method exploits both color and depth information obtained by a stereo camera system to improve the detection performance under illumination variations and in presence of partial occlusions. A depth-based adaptive windowing method has been proposed to detect pedestrians with different scales. A HOG descriptor and a linear SVM are used to classify the ROIs into pedestrian or non-pedestrian classes. Simulation results show that the proposed ROI generation method improves the performance of pedestrian detection in challenging environments.

As a future work, the ROI generation algorithm will be improved to perform better in detecting pedestrians in long distances and be more robust on occlusion and multiple objects detection in crowded scenes.

## References

- Gavrila DM (2000) Pedestrian detection from a moving vehicle. In: European conference on computer vision, pp 37–49
- Geronimo D, Lopez AM, Sappa AD, Graf T (2010) Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans Pattern Anal Mach Intell* 32(7):1239–1258
- Shashua A, Gdalyahu Y, Hayun G (2004) Pedestrian detection for driving assistance systems: single-frame classification and system level performance. *Proc Intell Veh Symp* 1–6
- Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. *IEEE Conf Comput Vis Pattern Recognit (CVPR'05)* 878–885
- Xu Y, Xu D, Lin S, Han TX, Cao X, Li X (2012) Detection of sudden pedestrian crossings for driving assistance systems. *IEEE Trans Syst Man Cybern Part B* 42(3):729–739
- Raksincharoensak P, Sakai Y, Shimizu I, Nagai M, Ulbricht D, Adomat R (2013) Development of an on-board pedestrian detection system using monocular camera for driver assistance applications. In: FISITA 2012 world automotive congress, pp 341–350
- Talukder A, Goldberg S, Matthies L, Ansar A (2003) Real-time detection of moving objects in a dynamic scene from moving robotic vehicles. In: International conference on intelligent robots and systems, pp 1380–1313
- Nam BD, Kang SI, Hong HK (2011) Pedestrian detection system based on stereo vision for mobile robot. In: *Frontiers of computer vision*, pp 1–7

9. Soga M, Kato T, Ohta M, Ninomiya Y (2005) Pedestrian detection with stereo vision. In: Proceeding of the IEEE international conference on data engineering
10. Broggi A, Fascioli A, Fedriga I, Tibaldi A, Rose MD (2003) Stereo-based preprocessing for human shape localization in unstructured environments. In: Proceeding of the IEEE intelligent vehicles symposium, pp 410–415
11. Gavrilla D, Munder S (2007) Multi-cue pedestrian detection and tracking from a moving vehicle. *Int J Comput Vis* 73(1):41–59
12. Bajracharya M, Moghaddam B, Howard A, Brennan S, Matthies L (2009) Results from a real-time stereo-based pedestrian detection system on a moving vehicle. In: International conference on robotics and automation (ICRA 2009)
13. Bhowmick B, Bhadra S, Sinharay A (2011) Stereo vision based pedestrians detection and distance measurement for automotive application. In: International conference on intelligent systems, modeling and simulation, pp 25–29
14. Viola P (2001) Rapid object detection using a boosted cascade of simple features. *Conf Comput Vis Pattern Recognit (CVPR 2001)* 1:511–518
15. Keller CG, Enzweiler M, Gavrila DM (2011) A New benchmark for stereo-based pedestrian detection. In: Intelligent vehicles symposium, pp 691–696
16. Keller CG, Enzweiler M, Rohrbach M, Llorca DF, Schnorr C, Gavrila DM (2011) The benefits of dense stereo for pedestrian detection. In: IEEE transactions on intelligent transportation systems, vol 12, pp 1096–1106
17. Geronimo D, Sappa AD, Ponsa D, Lopez AM (2010) 2D-3D based on-board pedestrian detection system. In: Computer vision and image understanding (special issue on intelligent vision systems), vol 114, pp 583–595
18. Keller CG, Llorca DF, Gavrila DM (2009) Dense stereo-based roi generation for pedestrian detection. In: Proceedings of the 31st DAGM symposium on pattern recognition, pp 81–90
19. Dalal N (2005) Histograms of oriented gradients for human detection. In: Computer vision and pattern recognition (CVPR 2005), vol 1, pp 886–893
20. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Intl J Comp Vision* 2:91–110



**Joohee Kim** received the B.S. and M.S. degrees in Electrical Engineering from Yonsei University, Seoul, Korea in 1991 and 1993, respectively. She received the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, in 2003. From 1993 to 1997, she was with Korea Telecom Research Laboratories as a Research Engineer. She joined Samsung Advanced Institute of Technology in 2003 as a

Senior Research Engineer and developed various video coding algorithms. From 2005 to 2008, she was an Assistant Professor in the Department of Information and Communication Engineering at Inha University in South Korea. She is currently an Assistant Professor in the Department of Electrical Engineering at the Illinois Institute of Technology. Her current research interests include image and video signal processing, multimedia communications, wireless video, 3D video coding, video analysis, and computer vision.



**Maral Mesmakhosroshahi** received the B.S. degree from Sharif University of Technology, Tehran, Iran in 2010 and received her M.S. degree in the Department of Electrical and Computer Engineering from the Illinois Institute of Technology, Chicago, IL in 2012. She is currently a PhD student in the Department of Electrical and Computer Engineering at the Illinois Institute of Technology. Her research

interests include image and video processing, computer vision and pattern recognition.