2011

# Pedestrian sensing using time-of-flight range camera

Xue Wei
*University of Wollongong*, xw158@uow.edu.au

Son Lam Phung
*University of Wollongong*, phung@uow.edu.au

Abdesselam Bouzerdoum
*University of Wollongong*, bouzer@uow.edu.au

## Publication Details

X. Wei, S. Phung & A. Bouzerdoum, "Pedestrian sensing using time-of-flight range camera," in IEEE CVPR 2011: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 43-48. Original conference information available here

# Pedestrian sensing using time-of-flight range camera

**Abstract**
This paper presents a new approach to detect pedestrians using a time-of-flight range camera, for applications in car safety and assistive navigation of the visually impaired. Using 3-D range images not only enables fast and accurate object segmentation and but also provides useful information such as distances to the pedestrians and their probabilities of collision with the user. In the proposed approach, a 3-D range image is first segmented using a modified local variation algorithm. Three state-of-the-art feature extractors (GIST, SIFT, and HOG) are then used to find shape features for each segmented object. Finally, the SVM is applied to classify objects into pedestrian or non-pedestrian. Evaluated on an image data set acquired using a time-of-flight camera, the proposed approach achieves a classification rate of 95.0%.

**Keywords**
era2015

**Disciplines**
Physical Sciences and Mathematics

**Publication Details**
X. Wei, S. Phung & A. Bouzerdoum, "Pedestrian sensing using time-of-flight range camera," in IEEE CVPR 2011: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 43-48. Original conference information available here

# Pedestrian Sensing Using Time-of-Flight Range Camera

Xue Wei, Son Lam Phung, and Abdesselam Bouzerdoum
School of Electrical, Computer and Telecommunication Engineering
University of Wollongong, Australia
xw158@uowmail.edu.au, phung@uow.edu.au, bouzer@uow.edu.au

## Abstract

*This paper presents a new approach to detect pedestrians using a time-of-flight range camera, for applications in car safety and assistive navigation of the visually impaired. Using 3-D range images not only enables fast and accurate object segmentation and but also provides useful information such as distances to the pedestrians and their probabilities of collision with the user. In the proposed approach, a 3-D range image is first segmented using a modified local-variation algorithm. Three state-of-the-art feature extractors (GIST, SIFT, and HOG) are then used to find shape features for each segmented object. Finally, the SVM is applied to classify objects into pedestrian or non-pedestrian. Evaluated on an image data set acquired using a time-of-flight camera, the proposed approach achieves a classification rate of* $95.0\%$.

## 1. Introduction

Detecting pedestrians in a given scene has applications in road safety and autonomous vehicles [8], assistive navigation for the blind [12], and surveillance [2]. Papageorgiou and Poggio at MIT presented a vision system that is used in Daimler-Chrysler Urban Traffic Assistant to detect pedestrians [15]. Haritaoglu and Flickner at IBM developed an intelligent billboard that counts the number of people in front of it [9]. Collins *et al*. at CMU created a multi-camera surveillance system that detects and tracks people over a wide area [2].

Existing approaches to pedestrian detection rely mostly on two-dimensional (2-D) cameras [5, 20]. Their major limitation is that other positioning or distance sensors are required to determine the three-dimensional (3-D) position of the target. In this paper, we propose a novel pedestrian detection approach that uses a 3-D range camera. This approach not only enables fast and accurate object segmentation, but also provides an estimation of distances and speeds of the pedestrians and their probabilities of collision with the user.

The paper is organized as follows: Section 2 reviews systems for 3-D image acquisition and algorithms for processing range images. Section 3 presents the proposed pedestrian sensing system, including range image segmentation, feature extraction, and pedestrian versus non-pedestrian classification. Section 4 analyzes the performance of the proposed method and Section 5 gives the concluding remarks.

## 2. Related work

In this section, we review the major aspects in processing range images: 3-D image acquisition, image segmentation and classification.

### 2.1. 3-D image acquisition

Three-dimensional information of an object can be acquired using passive or active approaches. The passive approach (stereo camera) uses two simultaneous cameras to capture the existing light in the environment. The active approach illuminates the object with light and analyses the reflected signal. Major types of active 3-D sensors are triangulation-based laser cameras, fringe projection-based cameras, and time-of-flight cameras.

- *Triangulation-based laser* cameras use reflected laser light and a color sensor to compute the distance between a projected laser point and the collection lens. Their depth of field (DOF) is between $25$mm to $200$mm, which is not suitable for pedestrian detection.
- *Fringe-based projection* cameras are high-density devices for 3-D surface measurement. They use a projector to illuminate the target with fringe patterns, which are then captured with one or more color cameras, located at fixed viewpoints. The target object needs to be located near the system, as the DOF is only between $50$mm and $400$mm [17].
- *Time-of-flight* (TOF) cameras generate distance images, where the value of each pixel is its distance to the camera. The distance is measured according to

the time taken for light to travel from the illumination source to the object and back to the receiver. The camera consists of a LED or laser diode source to lighten the scene, an optical filter to gather the reflected light, an image sensor to generate the pixel distance, and driver electronics to control the system. The DOF ranges from 5m to as high as 1000m.

Compared with other sensors, the TOF 3-D cameras have the advantages in a large depth-of-view and a high acquisition speed [17]. Furthermore, recent 3-D cameras can operate in both indoor and outdoor environments by using background light suppression. In this research, we use a 3-D TOF camera called MESA SwissRanger.

### 2.2. Range image segmentation and classification

Numerous approaches have been proposed for color image segmentation. For range images, there are fewer methods, and they focus mainly on finding planar surfaces or regular curved surfaces. The principle of these methods is to divide the image into closed regions with similar surface functions. Note that the major challenges in processing range images are that they typically have low resolutions and contain high noise.

Chandrasekaran *et al*. [1] introduced a dynamic neural network to segment range images. Their method identifies the crease edge pixels, but it focuses on segmenting only eight basic surface types. Xiao and Han [21] considered image segmentation as a combinatorial optimization problem and developed an image segmentation method based on Markov random field. This method represents image pixels in each region by a fixed polynomial function. For segmenting a complex scene, Feng *et al*. [7] presented a jump-diffusion method where objects are not limited to polyhedral shapes.

Algorithms have also been proposed to classify objects in 3-D range images. For example, for *Smart Airbag* systems, Devarakota *et al*. [4] used 3-D images to classify vehicle occupants as adults or children, leaning forward or backward. The classification is evaluated on several classifiers, including linear-regression, Bayes quadratic, Gaussian mixture, and polynomial classifiers. A pedestrian recognition system based on depth and intensity images was proposed by Rapus *et al*. [18]. Combining the features of depth and intensity, the pedestrian recognition with low camera resolution is advisable.

## 3. Proposed approach

The proposed approach for detecting pedestrians in a scene consists of three main stages: range image segmentation, feature extraction, and pedestrian versus non-pedestrian classification (see Fig. 1).

### 3.1. 3-D segmentation

We propose an image segmentation approach based on local variation (LV). This concept was originally proposed by Felzenszwalb and Huttenlocher for processing color or gray-scale images [6]. The local variation is a graph-based segmentation method. It merges two components $C_1$ and $C_2$ if the external variation $Ext(C_1, C_2)$ is small relative to both internal variations:

$$Ext(C_1, C_2) \leq \min[Int(C_1) + \tau(C_1), \\ Int(C_2) + \tau(C_2)]. \quad (1)$$

Here, $Int(C_1)$ and $Int(C_2)$ denote the internal variation of component $C_1$ and $C_2$, respectively. The term $\tau$ is a function of the component size, $\tau(C) = k/|C|$, where $k$ is a constant.

Consider an image with an undirected graph $H = (V, E)$, where each pixel has a corresponding vertex $v_i \in V$ and edge $e_i \in E$. The local variation algorithm finds a segmentation, described by an edge set $F$, using the steps as follows [6].

1. Sort $E$ in the image according to a non-decreasing edge weight. Let $\pi$ be the sorted set, $\pi = \{e_1, e_2, ..., e_k\}$. The edge weight is difference between two pixels.

2. Start with $F_0 = \{\emptyset\}$.

3. Repeat Step 4 for each edge $e_q$ where $q = 1, 2, ..., k$.

4. Construct $F_q$ from $F_{q-1}$. Suppose that edge $e_q$ connects vertex $v_i$ and $v_j$. If there is no path from $v_i$ to $v_j$ in $F_{q-1}$ and edge weight of $e_q$ is small compared to the internal variation of the components containing $v_i$ and $v_j$, then $F_q = \{F_{q-1}, e_q\}$. Otherwise, $F_q = F_{q-1}$.

The local variation algorithm has two main problems when applied to range images. Variations between regions are distance. First, it defines the external variation as the smallest difference between two components, and therefore is easily affected by noise. In outdoor range images, noise is a significant factor because of saturated background lighting. Second, an object in a range image may be over-segmented into two or more regions if it is partially occluded by another object.

To address the first problem, we apply a $3 \times 3$ median filter on the range image to remove noise and stabilise the distance values filtered image. Other filter sizes such as $5 \times 5$ or $7 \times 7$ also work. In addition, the confidence map generated by the 3-D camera is threshold to discard pixels with unreliable distance values. The confidence map has integer grades between 0 and 7, where 0 is for the most unreliable measure and 7 if for the most reliable measure.


Figure 1. Block diagram of the proposed pedestrian sensing approach from range images.

To address the second problem, we propose grouping the over-segmented regions by finding regions with similar normal vectors and distance values. For each segmented region, all 3-D pixels in the region are identified $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N\}$, where $\mathbf{p}_i = (x_i, y_i, z_i)$. The Delaunay triangulation algorithm is then applied to group the pixels into triangular surfaces (or meshes) [16]. For each surface $k$, where $k = 1, 2, ..., K$, the planar equation $a_k\, x + b_k\, y + c_k\, z + d_k = 0$ is found and its normal vector $\mathbf{n}_k = (a_k, b_k, c_k)$ is calculated. The average normal vector for the region is defined as

$$\mathbf{n}_a = \frac{1}{K} \sum_{k=1}^{K} \mathbf{n}_k. \tag{2}$$

The average distance for the region is also calculated as

$$d_a = \frac{1}{N} \sum_{i=1}^{N} d_i, \tag{3}$$

where $d_i$ is the distance or depth of pixel $i$.

Two regions $u$ and $v$ are merged if they have similar average normal vectors and similar average distances:

$$|\mathbf{n}_{a,u} - \mathbf{n}_{a,v}| \leq \tau_n \text{ AND } |d_{a,u} - d_{a,v}| \leq \tau_d, \tag{4}$$

where $\tau_n$ and $\tau_d$ are two thresholds.

### 3.2. Feature extractors

Many feature extraction methods have been proposed for general color or gray-scale images. In this paper, we investigate the suitability of three state-of-the-art feature extractors for classification of pedestrian versus non-pedestrian in 3-D range images. They are: the scale invariant feature transform (SIFT), the histogram of oriented gradient (HOG), and the holistic shape feature based on the spatial envelope (commonly known as GIST).

- *Scale invariance feature transform* (SIFT): The SIFT, proposed by Lowe [11], extracts image features that are invariant to image scale, rotation, changing illumination, and 3-D projection. It involves four main steps. First, the difference-of-Gaussian filter is applied to identify interest points that are invariant to scale and orientation. Second, the key points with high stability are selected from the outputs of the first step. Third, one or more orientations are assigned to each key point. Fourth, the local shape distortion and illumination changes are removed from the selected key points. For clear images, the SIFT algorithm performs well, but for blurred image, extraction of edge features is less effective.

- *Histogram of oriented gradient* (HOG): The HOG algorithm was originally developed for human detection in gray-scale images [3]. Its basic idea is that object shape and appearance can be characterized by the local distribution of intensity gradients or edge directions. The HOG algorithm extracts features by computing normalized local histograms of image gradients in a dense grid. An input image is first normalized by power-law equalization. Then, the image gradient is computed. Next, histograms for multiple orientations are computed for each cell. The cell can be rectangular or radial, and each pixel in the cell contributes a weighted score to a histogram. Finally, the cell histograms are normalized and grouped in blocks to form HOG descriptor. HOG descriptors have been applied also for detection of cars, buses, and bicycles in gray-scale images.

- *Holistic shape feature based on the spatial envelope* (GIST): The GIST, proposed by Oliva and Torralba [14], is a low-dimensional, holistic descriptor of the scene. First, for pre-processing an input image is normalized to reduce noise. Next, a set of Gabor filters are applied to find spectral and coarsely localized information. Finally, GIST features are extracted in several blocks. GIST descriptors are effective at processing blur images.

### 3.3. Feature classification

For each segmented region, the extracted features need to be classified as pedestrian or non-pedestrian. Numerous classifiers exist, and the support vector machine (SVM) is chosen to use in the proposed system. SVM is a widely used classifer in learning and pattern recognition. It was first proposed by Vapnik in 1995 based on the theories of Vapnik Chervonenks dimension and the stuctural risk minimization (SRM) [19]. SVM has several advantages: (i) it maximizes the margins between the two classes and hence improve the generalisation error; (ii) it works well when the number of training samples is limited; (iii) it can produce non-linear decision surfaces by projecting samples to

a high-dimensional space; this is known as the kernel approach.

Several kernels can be used with the SVM classifiers, for example linear, polynomial, and radial basis function (RBF). In this paper, we use the RBF kernel, which is defined as

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}, \quad (5)$$

where $\gamma$ is a positive scalar. The RBF kernel has been demonstrated to work well in numerous practical applications. This non-linear kernel also has fewer parameters compared to the polynomial kernel.

The SVM classifier is trained to differentiate pedestrian and non-pedestrian patterns. For road safety and assitive navigation of the blind applications, our approach can be extended to recognize other traffic objects such as cars and bikes.

## 4. Experiments and results

In this section, we analyze the performance of the proposed system for pedestrian detection on a set of 3-D range images.

### 4.1. Experiment data

The data set was acquired using a TOF camera produced by MESA Imaging, model SwissRanger 4000 [13]. The frame rate is 30 frames per second, the frame size is $144 \times 176$ pixel, and the pixel value is 16 bits. The data set was taken in indoor and outdoor environments, at different days, lighting conditions, and scenes. Example images are shown in Fig. 2. For each pixel, the camera produces five outputs: x, y, and z coordinates; amplitude; and the confidence score. The confidence score indicates the probability that the distance measurement is accurate. The radial distance for the pixel is calculated as

$$d(i,j) = \sqrt{x(i,j)^2 + y(i,j)^2 + z(i,j)^2}, \quad (6)$$

where $x(i,j)$, $y(i,j)$, and $z(i,j)$ are the x, y, and z Cartesian coordinates, respectively.

### 4.2. Results of range image segmentation

The proposed algorithm for range image segmentation was evaluated using a framework developed by Hoover *et al*. [10]. We prepared the ground truth by manually segmenting the test images. In the evaluation framework, an output region is considered *correctly-segmented* with a tolerance rate $\tau$ if the ratio $u/\max(\mathrm{g}, \mathrm{o})$ is greater than or equal to $\tau$. Here, $u$ is the area of the overlap between the ground-truth region $g$ and the output region $o$.

The segmentation rate of an algorithm is defined as

$$E = C/T, \quad (7)$$

Table 1. Segmentation rates of the LV algorithm and the proposed algorithm for tolerance rate $\tau = 0.5$.

| *Test Image* | *LV Algorithm* (%) | *Proposed Algorithm* (%) |
|---|---|---|
| image_01 | 80.0 | 89.8 |
| image_02 | 88.0 | 88.0 |
| image_03 | 91.0 | 93.5 |
| image_04 | 81.8 | 95.2 |
| image_05 | 82.0 | 82.8 |
| image_06 | 75.2 | 86.4 |
| image_07 | 82.4 | 88.0 |
| image_08 | 86.0 | 90.3 |
| image_09 | 86.5 | 86.6 |
| image_10 | 88.7 | 88.0 |
| Average rate | 84.2 | 88.9 |

where $C$ is the number of correctly-segmented regions and $T$ is the total number of segmented regions produced by a segmentation algorithm. Note that an output region is considered *under-segmented* if it consists of multiple ground truth regions. An output region is considered *over-segmented* if it is smaller than the corresponding ground truth region.

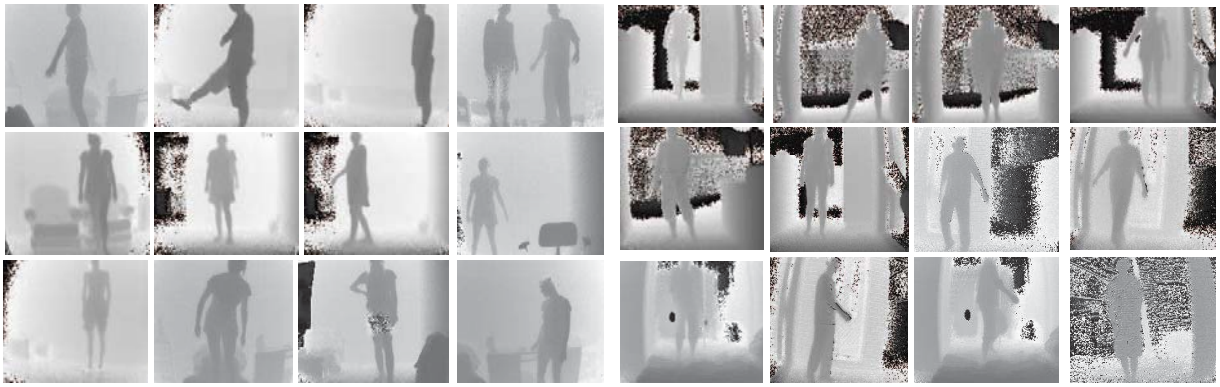
The segmentation rates were computed using 10 test images, for the local variation algorithm and the proposed algorithm (see Table 1). The proposed algorithm has a segmentation rate of 88.9%, whereas the local variation has a segmentation rate of 84.2%.

The segmentation ouputs of the LV and proposed algorithms on a test image are shown in Fig. 3. The differences are clearer when viewed from the online color images.

- With the LV algorithm (Fig. 3b), the pedestrian object is over-segmented into several regions, such as legs. The same background objects such as the road surface are also partitioned into several regions.
- With the proposed algorithm, segmentation noise is significantly reduced. However, some background regions are still over-segmented (Fig. 3c). After the proposed region merging step, over segmentation is reduced (Fig. 3d). This leads to improved performance in the classification stage.

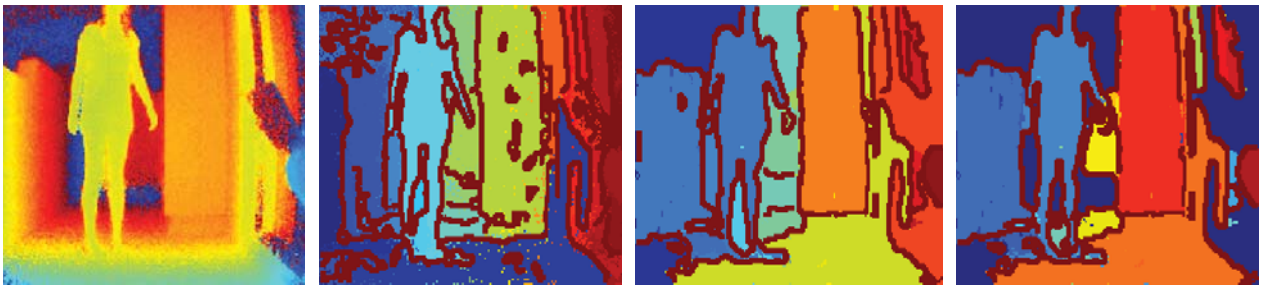
### 4.3. Comparison of feature extractors

For each segmented region, different feature extractors are applied. The feature extractors are GIST, SIFT and HOG. A GIST feature vector has 512 elements, where SIFT and HOG feature vectors each has 1000 elements. Each feature vector is processed by a trained SVM classifier. The experiment was performed on a set of 400 pedestrian patterns and 400 non-pedestrian patterns were used in the experiment. We used 60% of the data for training and 40% for test.



(a) indoor images (b) outdoor images

Figure 2. Samples from the data set of 3-D range images.



(a) (b) (c) (d)

Figure 3. Visual result of segmentation: (a) Input range image, (b) Output of the LV algorithm, (c) Output of the proposed algorithm before region merging, (d) Output of the proposed algorithm after region merging. See online color figure.

The classification rates of the three feature extractors are shown in Table 2.

- The SIFT method has the lowest classification rate among the three methods. The SIFT is suitable for recognizing objects based on texture. However, in range images object texture is not so visible unless there is a significant change in distance.

- The GIST method has the highest classification rate of $95.0\%$. A possible reason is that the GIST method extracts global shape information and ignores details in the inner regions of the object. For range images, shape and contour features are the most dominant.

- The HOG method has a good classification rate ($93.8\%$), but not as high as that of the GIST method. Combining GIST and HOG features may improve classification accuracy.

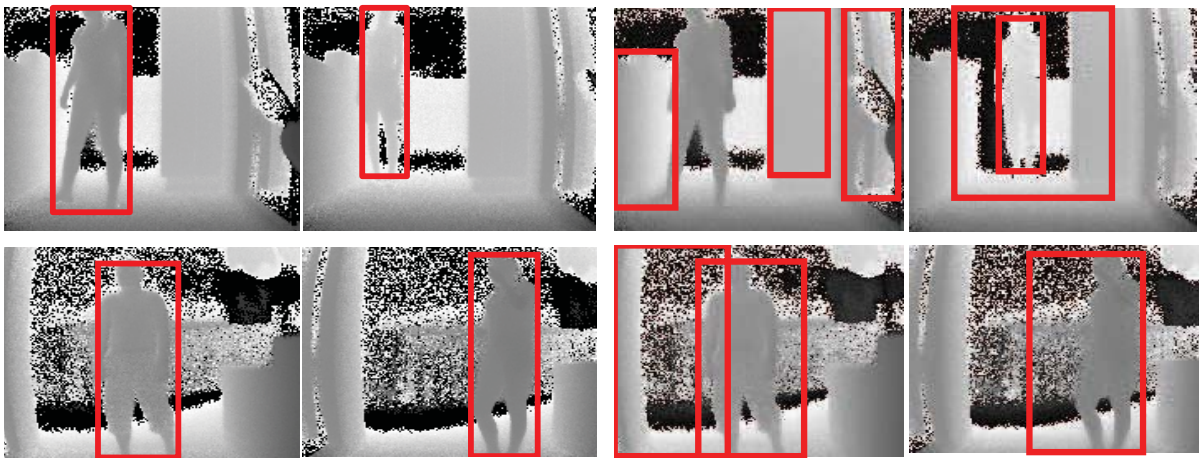
Figure 4 shows example ouputs of the proposed system for pedestrian sensing, using the GIST and HOG features. GIST method has fewer false detects compared to the HOG method. Our approach presented here can be extended to detect other traffic obstacles such as cars and bikes.

| *Method* | *Classification Rate* (%) |
|---|---|
| GIST | 95.0 |
| HOG | 93.8 |
| SIFT | 88.8 |

Table 2. Performance on pedestrian versus non-pedestrian classification of three feature extraction methods.

## 5. Conclusion

In this paper, we have presented a new approach for detecting pedestrians from time-of-flight range images. Using range images leads to more efficient segmentation of objects and provides distance and collision probability that are useful for assistive navigation. However, recognizing objects in range images poses significant challenges because range cameras typically have lower resolutions and higher noise, compared to color cameras. We proposed an image segmentation method that reduces over segmentation by processing surface normal vectors and distance data. In this paper, we have also analyzed three state-of-the-art feature extractors: HOG, SIFT and GIST. The GIST feature is found to be more effective in classifying pedestrian versus non-pedestrian in range images.



(a) using GIST features (b) using HOG features

Figure 4. Visual result of pedestrian detection from 3-D range images.

## References


[1] V. Chandrasekaran, M. Palaniswami, and T. M. Caelli. Range image segmentation by dynamic neural network architecture. *Pattern Recognition*, 29(2):315–329, 1996.

[2] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893 vol. 1, 2005.

[4] P. R. Devarakota, M. Castillo-Franco, R. Ginhoux, B. Mirbach, and B. Ottersten. Occupant classification using range images. *IEEE Transactions on Vehicular Technology*, 56(4):1983–1993, 2007.

[5] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Image segmentation using local variation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 98–104, 1998.

[7] H. Feng, T. Zhuowen, and Z. Song-Chun. Range image segmentation by an effective jump-diffusion method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1138–1153, 2004.

[8] T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):413–430, 2007.

[9] I. Haritaoglu and M. Flickner. Attentive billboards. In *International Conference on Image Analysis and Processing*, pages 162–167, 2001.

[10] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher. An experimental comparison of range image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):673–689, 1996.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] P. B. L. Meijer. An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121, 1992.

[13] MESA Imaging. SwissRanger 3-D TOF camera, 2011. http://www.mesa-imaging.ch/index.php.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[15] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *International Conference on Image Processing*, volume 4, pages 35–39 vol.4, 1999.

[16] P. Per-Olof and G. Strang. A simple mesh generator in Matlab. *SIAM Review*, 46(2):329–345, 2004.

[17] PolyWorks. 3-D metrology hardware review, 2010. http://www.innovmetric.com.

[18] M. Rapus, S. Munder, G. Baratoff, and J. Denzler. Pedestrian recognition using combined low-resolution depth and intensity images. In *IEEE Intelligent Vehicles Symposium*, pages 632–636, 2008.

[19] V. N. Vapnik. *The Statistical Learning Theory*. Springer, 1995.

[20] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[21] W. Xiao and W. Han. Markov random field modeled range image segmentation. *Pattern Recognition Letters*, 25(3):367–375, 2004.