J. Vis. Commun. Image R. 25 (2014) 74–85

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Object segmentation and classification using 3-D range camera

Xue Wei *, Son Lam Phung, Abdesselam Bouzerdoum

School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Australia

ARTICLE INFO

Article history:
Available online 13 April 2013

Keywords:
3-D range image segmentation
Range/intensity image processing
Feature extraction
Object classification
Assistive navigation
Pedestrian classification
Image segmentation evaluation
RGB-D image processing

ABSTRACT

This paper proposes a vision system using a 3-D range camera for scene segmentation and pedestrian classification. The system detects and segments objects in the foreground, measures their distances to the camera, and classifies them into pedestrians and non-pedestrian obstacles. Combining range and intensity images enables fast and accurate object segmentation, and provides useful navigation cues such as the range and type of nearby objects and the ground surface. In the proposed approach, a 3-D range image is segmented using histogram processing and mean-shift clustering. The ground surface is detected by estimating its normal vector in 3-D space. Fourier and GIST descriptors are then applied on each detected region to extract shape and texture features. Finally, support vector machines are used to classify objects; in this paper we focus on differentiating pedestrian and non-pedestrian regions. The performance of the proposed system is evaluated with two datasets. One dataset for object segmentation and pedestrian classification is acquired by us using a 3-D range camera; the other is a public RGB-D dataset for people detection. Experimental results show that the proposed system performs favorably compared to some existing segmentation and feature extraction approaches.

1. Introduction

Detecting and classifying objects in a 3-D scene plays an important role in assistive navigation for the blind [1], road safety [2,3], surveillance [4], and many other applications. In the traditional approach, visual recognition consists of image segmentation followed by classification [5]. Many image segmentation methods are based on low-level features such as color and texture. For example, Gould et al. proposed an object classification system based on multi-class image segmentation [6]. Their system labels pixels as background or foreground classes, and then classifies the foreground regions as cars, pedestrians or other. In an alternative approach, Leibe et al. suggested that the image segmentation and recognition are intertwined processes, and top-down knowledge from object recognition should guide the segmentation process [7]. Several top-down algorithms have been proposed to improve figure-ground segmentation of color images [7,5,8]. However, to avoid segmentation errors, several object detection algorithms without segmentation have been proposed, such as window scanning [9], local contour features [10], and implicit shape model [11].

With recent advances in 3-D cameras, range images have been used for object segmentation and recognition. Compared with color images, range images are less sensitive to changes in the environment illumination, object color or texture. Existing algorithms for range image segmentation focus mainly on segmenting planar surfaces or regular curved surfaces [12–17]. The principle of these algorithms is to divide the image into closed regions with similar surface functions. Harati et al. proposed an edge-based segmentation for range images [18]. In their algorithm, two bearing angle (BA) images for vertical and horizontal directions are calculated from range images, and the edges of BA images are detected using the Sobel operator. Segmentation is achieved by labeling the combined edge map. Coleman et al. proposed an edge detection method using Laplacian operators for irregular range images [19]. The improved Laplacian operators reduce noise in range images and achieve a higher segmentation rate than the traditional Laplacian operator. Markov random fields were also applied to range image segmentation. For example, Wang and Wang proposed a range image segmentation based on Bayes inference and Markov random field modeling, and used the surface function parameters to group distance pixels into planar regions [20]. Zhang et al. combined Markov random fields with graph cuts [21] to reduce over-segmentation for range images [22].

More recently, several algorithms have been proposed for object detection and classification in 3-D images [23–29]. Eunyoung and Medioni proposed a scalable framework for categorizing 3-D objects [23]. After range image segmentation, the objects are classified using an online learning system, which is based on a hierarchical structured model reported in [30]. Das et al. proposed an object detection and localization system based on both color and range images for robots [24]. Their method first removes saturation noise from range images, and then extracts the features of ob-

* Corresponding author. Fax: +61 2 42 21 3236.
E-mail addresses: xw158@uowmail.edu.au (X. Wei), phung@uow.edu.au (S.L. Phung), bouzer@uow.edu.au (A. Bouzerdoum).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.04.002

jects using the singular value decomposition filter and pLSA model [31]. The multi-level SVM classifier is used to categorize five objects. Fardi et al. presented a 3-D photonic mixer device for pedestrian detection [25]. This system first segments the image to create reliable detections of objects in the image plane, and then uses object distance and shape/motion features to detect the pedestrians. Devarakota et al. used range images to classify vehicle occupants (adults or children) and their actions (leaning forward or backward) [26]. Their method was evaluated with several classifiers, including linear-regression, Bayes quadratic, Gaussian mixture, and polynomial classifiers. Rapus et al. proposed a pedestrian recognition system based on depth and intensity images [27]. In their system, the ground plane from the depth image is extracted first; an AdaBoost head-shoulder detector is then used to classify pedestrians. Mozos et al. suggested a multipart-based people detection system from range data [28]. In their system, individual classifiers are trained to detect different body parts, and their outputs are combined to form the final detector. Spinello and Arras developed a people detection system based on RGB-D data [29]. They proposed using histogram of depths (HODs) to extract features from range images, and used histogram of oriented gradients (HOGs) to extract features from intensity images. The decisions from range and intensity images are fused to form the final classification. There are also several other approaches that have been proposed for pedestrian detection in 2-D intensity images [32–34].

This paper presents a scene segmentation and pedestrian classification system that relies on 3-D range images and 2-D intensity images to locate objects in the scene, determine their range and velocity, and classify them into pedestrian and non-pedestrian objects. Recently, we proposed a pedestrian detection system based on range images [35]. In this system, the local variation algorithm [36] is used to segment range images, and the GIST features [37] are used to classify pedestrian and non-pedestrian patterns. In this paper, we extend the work presented in [35], by proposing a new segmentation approach and new features for object classification, and conducting more comprehensive experimental evaluation, analysis and comparison with other state-of-the-art techniques.

The remainder of the paper is organized as follows. Section 2 presents the proposed object sensing system, including range image segmentation, feature extraction, and object classification. Section 3 analyzes the performance of the proposed segmentation and classification methods, and compares them with existing state-of-the-art techniques. Section 4 gives the concluding remarks.

2. Proposed segmentation and classification system

The proposed system for scene segmentation and pedestrian classification using a range camera is shown in Fig. 1. The SwissRanger SR4000 camera produces range images (x, y, z), an intensity image, and a confidence map (see Fig. 2). In the proposed system, the input range image (z) is first segmented by analyzing the depth histogram. Next, the mean-shift algorithm is applied to the 3-D range data (x, y, z) to reduce under-segmentation. Then, the GIST features and Fourier descriptors are extracted from range and intensity images for classification of the segmented regions as a pedestrian or non-pedestrian.

Range images are typically very noisy, especially when captured in an outdoor environment. To reduce the noise, several pre-processing steps are applied before the range data are fed to the object segmentation and classification system. Pixels with unreliable depth values are removed by thresholding the confidence map, and then removed by median filtering. Fig. 2f shows the result of pre-processing the image in Fig. 2a.

2.1. Scene segmentation based on depth histogram

After pre-processing, the range image is segmented into separate regions by analyzing the depth histogram. The image pixels are partitioned into distinct distance layers using a set of thresholds, which are determined adaptively as the local minima of the range image histogram. The edge pixels in the range image are excluded during histogram-thresholding because pixels on region boundaries have noisy depth measures.

Let $\mathbf{h} = \{h_1, h_2, \ldots, h_B\}$ be the depth image histogram with B bins, where h_j denotes the jth histogram bin. The first-order derivative is approximated as

$$d_j = h_j - h_{j-1}, \text{ for } j = 2, 3, \ldots, B. \tag{1}$$

A local minimum is detected at point j if the first-order derivative changes sign from negative to positive:

$$d_j \leqslant 0 \text{ and } d_{j+1} > 0. \tag{2}$$

After thresholding, connected component labeling is applied on each distance layer to form regions. An example of histogram-based segmentation is shown in Fig. 3. Histogram-based segmentation is very fast, but it relies only on thresholding the depth values, which may lead to under-segmentation. This problem will be addressed in the next stage by applying the mean-shift algorithm on the 3-D spatial points.

2.2. Mean-shift clustering of 3-D points

After histogram processing, the mean-shift algorithm is employed to analyze the 3-D points in each region. Consider a preliminary segmented region with n pixels. Let $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ denote the corresponding 3-D points, where $\mathbf{p}_i = (x_i, y_i, z_i)$. The mean-shift algorithm partitions the 3-D points into sub-regions via kernel density estimation, where each sub-region corresponds to a density center $\mathbf{c}_j$. Suppose that $\mathbf{c}_j$ is the current estimate of the center, an updated estimate is calculated as

$$\mathbf{m}(\mathbf{c}_j) = \frac{\sum_{i=1}^{n} K(\mathbf{c}_j - \mathbf{p}_i)\mathbf{p}_i}{\sum_{i=1}^{n} K(\mathbf{c}_j - \mathbf{p}_i)}, \tag{3}$$

where $K(\cdot)$ is a kernel function that controls the contribution of the sample $\mathbf{p}_i$ to the center. In our work, a flat kernel is used, which is defined as follows:

$$K(\mathbf{v}) = \begin{cases} 1, & \text{if } ||\mathbf{v}|| \leqslant r, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where r denotes the kernel radius. The difference between the updated center and the current center $\{\mathbf{m}(\mathbf{c}_j) - \mathbf{c}_j\}$ is called the mean-shift. Convergence to a center occurs when the mean-shift is smaller than a threshold $e^{-\lambda\sqrt{r}}$, where λ is a positive constant.

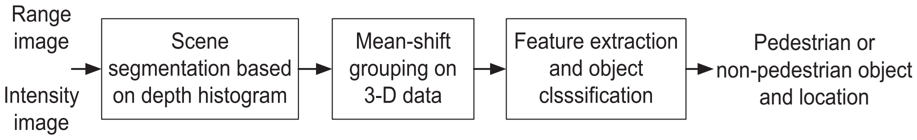
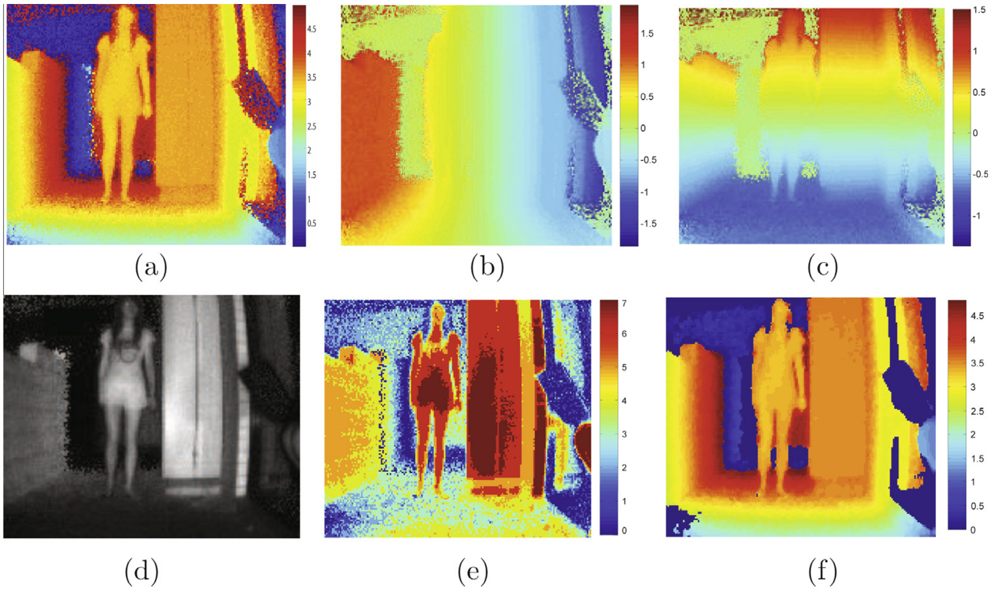

Range image
Intensity image
Scene segmentation based on depth histogram
Mean-shift grouping on 3-D data
Feature extraction and object clsssification
Pedestrian or non-pedestrian object and location

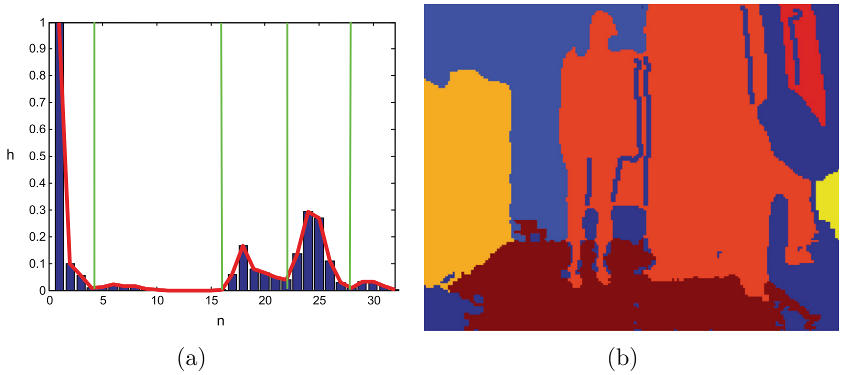
Fig. 1. Block diagram of the proposed scene segmentation and classification system.



(a) (b) (c)

(d) (e) (f)

Fig. 2. Example images captured by a 3-D camera SwissRanger SR4000: (a) range image (*z* dimension), (b) *x* dimension, (c) *y* dimension, (d) intensity image, (e) confidence map (a high score indicates a reliable distance measure) and (f) output of preprocessing stage (*z* dimension) – thresholding of the confidence map followed by median filtering.



(a) (b)

Fig. 3. Histogram-based segmentation: (a) histogram of the range image in Fig. 2f (the green lines are the detected local minima); (b) output of histogram-based segmentation. In this example, the number of histogram bins is $B = 32$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

All 3-D points located within a radius r from the density center are grouped into one sub-region.

To reduce computation cost and prevent over-segmentation, the radius r is selected adaptively according to the area and standard deviation of each region. Let A_i be the area of the ith preliminary segmented region, and $A_{\max}$ be the largest area. The bandwidth r_i for the ith segmented region is defined as

$$r_i = \frac{\tau_a A_i}{\sigma_i A_{\max}}, \tag{5}$$

where σ_i is the standard deviation of the depth pixels in region i, and τ_a is a positive parameter. Fig. 4 shows an example of 3-D mean-shift clustering, where a preliminary segmented region is partitioned into several sub-regions.

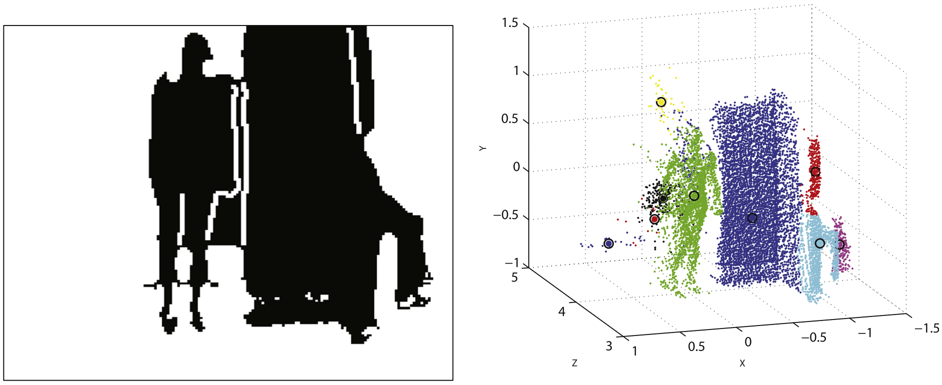
2.3. Ground segmentation and object localization

In range images, some object regions have the same depth values as the ground. To differentiate the ground region from other objects, we process the normal vector $\mathbf{u} = (u_x, u_y, u_z)$ of 3-D surfaces. Consider the set of 3-D points (x_i, y_i, z_i), where y_i is the vertical coordinate. The Delaunay triangulation is first applied on the (x_i, z_i) coordinates to generate horizontal triangulation surfaces. Let $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{p}_3$ be the three vertices of an output triangle. The normal vector of the triangulation surface is computed using the cross product:

$$\mathbf{u} = (\mathbf{p}_3 - \mathbf{p}_2) \times (\mathbf{p}_2 - \mathbf{p}_1). \tag{6}$$

The normal vector at a given point is estimated by averaging the normal vectors of all triangulation surfaces intersecting at that point. A 3-D point is considered as a ground pixel if the y-component of its unit normal vector exceeds a threshold τ_n. To increase the processing speed, the normal vectors are calculated only for the bottom-third region of the image, since the ground region is assumed to appear on the lower part of the image. The ground detection is applied to separate the objects from the ground and improve range image segmentation. Fig. 5 shows the ground detection for indoor and outdoor images. Note that for ground detection to be accurate, the 3-D range camera should be approximately parallel to the ground.

To finalize segmentation, over-segmented regions are recovered by merging adjacent regions that are separated by a weak edge.



(a) (b)

Fig. 4. Example of mean-shift clustering: (a) a preliminary segmented region; (b) output of mean-shift clustering. The circles ∘ denote the extracted density centers.

The common boundary of two adjacent regions $\mathbf{R}_a$ and $\mathbf{R}_b$ is removed if

$$\frac{L}{\min(L_a, L_b)} \leqslant \tau_l, \tag{7}$$

where L is the number of pixels along their common boundary, τ_l is a threshold, and L_a and L_b are the perimeter lengths of the two regions. Fig. 6 shows an example of edge merging.

After segmentation, the position $\hat{\mathbf{p}}_a = (\hat{x}_a, \hat{y}_a, \hat{z}_a)$ of the segmented object $\mathbf{R}_a$ is calculated as

$$\hat{\mathbf{p}}_a = \frac{1}{n_a}\sum_{\mathbf{p}_i \in \mathbf{R}_a} \mathbf{p}_i, \tag{8}$$

where n_a is the number of pixels in the region. The velocity of an object is estimated as the change in distance over time:

$$\hat{v}_a = ||\Delta\hat{\mathbf{p}}_a||/\Delta t, \tag{9}$$

where $\Delta\hat{\mathbf{p}}_a$ is the object displacement between two consecutive range images. Fig. 7 shows some segmented objects and their location in the scene.

2.4. Feature extraction

To classify the segmented regions, the Fourier and GIST features from range and intensity images are combined. Fourier features [38] represent the shape of the object, whereas, GIST features [37] model the dominant spatial structure of the object. These descriptors are introduced in the next two subsections.

2.4.1. Fourier descriptor

For each segmented region, the boundary of the object is extracted from the binary masks using morphological dilation. The noise edge pixels are removed by finding the largest connected component in the edge map, and tracing a close boundary of this component (clockwise direction, 8-connected labeling). The ith pixel on the boundary is represented as a complex number $p_i = x_i + jy_i$, where x_i and y_i are the horizontal and vertical coordinates of the pixel. The discrete Fourier transform (DFT) is applied to the 1-D complex signal p_i $(i = 0, 1, \ldots, N-1)$ to obtain the Fourier coefficients:

$$F_k = \frac{1}{N}\sum_{i=0}^{N-1} p_i e^{-j2\pi\frac{k\times i}{N}}, \quad k = 0, 1, \ldots, N-1. \tag{10}$$

The Fourier coefficients of low frequencies represent the general shape of the object, whereas the Fourier coefficients of high frequencies represent finer details about the object shape [38]. Let $F_{\max}$ be the maximum magnitude of all Fourier coefficients, $F_{\max} = \max(|F_k|), k = 1, 2, \ldots, N-1$. The DC Fourier coefficient F_0 is removed, and the remaining Fourier coefficients are normalized to form a Fourier descriptor as

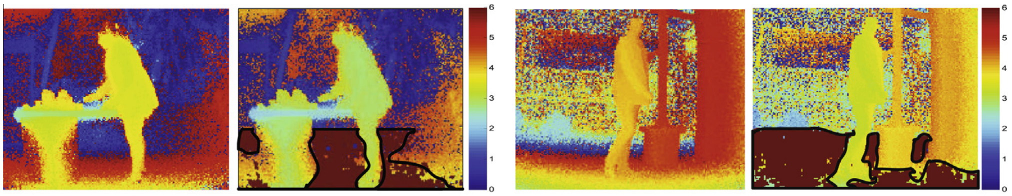
$$\mathbf{f} = \frac{[|F_1|, |F_2|, \ldots, |F_{N-1}|]^T}{F_{\max}}, \tag{11}$$

This Fourier descriptor has been shown to be relatively invariant to translation, scale, and rotation [38].

2.4.2. GIST descriptor

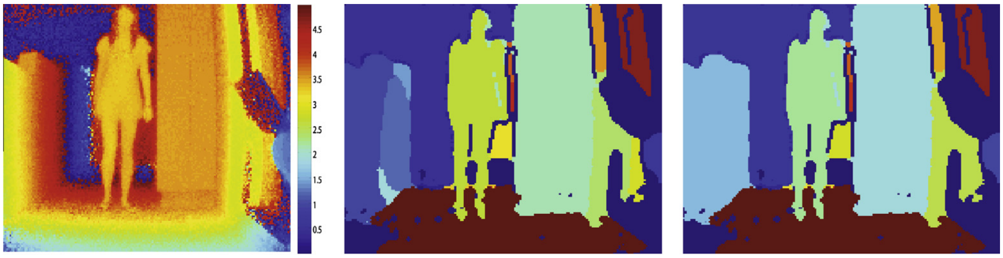
The GIST descriptor is a holistic and low-dimensional representation of images. It have been shown to be invariant to scale, orientation, and aspect ratio of objects [37]. For each segmented region, the GIST descriptor is applied to the corresponding rectangular regions in the range and intensity images.

Consider an image region $\mathbf{I}$ (range or intensity). First, the region is padded, whitened, and normalized to reduce the blocking artifact. Next, a set of multi-scale oriented Gabor filters are generated from one mother wavelet, through dilation and rotation. The im-



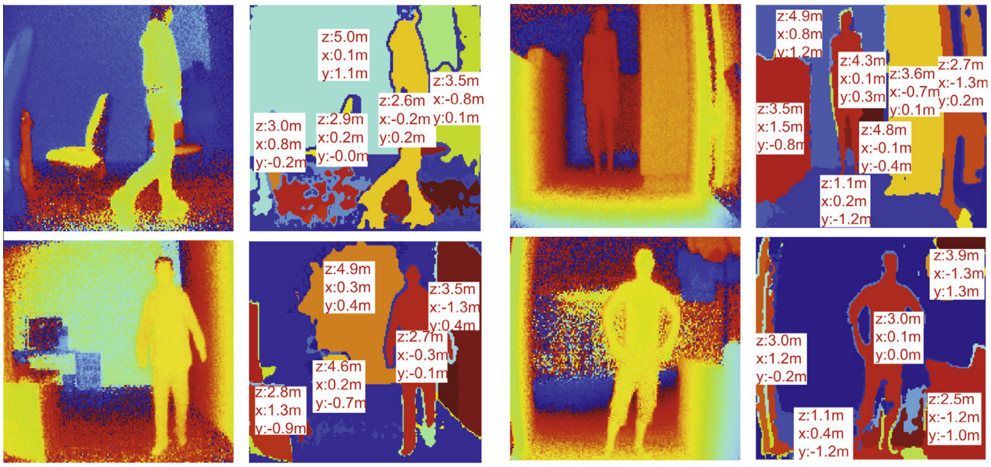
(a) indoor cases (b) outdoor cases

Fig. 5. Sample results of ground detection method. Left: input depth images; Right: outputs. The detected walkable regions are highlighted in dark red color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) (b) (c)

Fig. 6. An illustration of the proposed range image segmentation: (a) input range image, (b) after histogram thresholding and mean-shift segmentation and (c) after edge merging. In this example, the over-segmented surfaces on the right hand side of the person are recovered.



(a) indoor scenes (b) outdoor scenes

Fig. 7. Sample results of the proposed segmentation method. Left: input depth images. Right: outputs with calculated 3-D position in meter. See electronic color figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pulse response of a Gabor filter is the product of a harmonic and a Gaussian function:

$$g(x,y) = \cos(2\pi\frac{x'}{\lambda} + \Phi)\exp\left\{\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right\}, \tag{12}$$

where $x' = x\cos\theta + y\sin\theta, y' = -x\sin\theta + y\cos\theta$, θ is the rotation angle, Φ is the phase offset, λ is the wavelength of the sinusoidal component, σ is the parameter of the Gaussian function, and γ is the spatial aspect ratio of the harmonic function. The Gabor filters $\mathbf{g}_{k,\ell}$ with four scales $(k = 1, \ldots, 4)$ and eight orientations $(\theta_\ell = \frac{\pi\ell}{8}; \ell = 0, \ldots, 7)$ are applied to the region $\mathbf{I}$:

$$\mathbf{O}_{k,\ell} = \mathbf{I} * \mathbf{g}_{k,\ell}, \tag{13}$$

where $*$ is the convolution operator. The output of each filter is partitioned into 16 blocks. For each block, the average value is calculated:

$$\overline{O}_{k,\ell} = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|O_{k,\ell}(x,y)|. \tag{14}$$

Therefore, each filter produces 16 features. The features produced by all 4×8 Gabor filters are concatenated into a single feature vector.

2.5. Object classification

The proposed pedestrian detection system uses a feature vector consisting of 512 GIST features from the range region, 512 GIST features from the intensity region, and 100 Fourier descriptors from the object boundary. A support vector machine (SVM) classifier with the radial basis function (RBF) kernel is used to categorize pedestrian and non-pedestrian objects. The RBF kernel is given by

$$K(\mathbf{f}_i, \mathbf{f}_j) = \exp\{-\gamma||\mathbf{f}_i - \mathbf{f}_j||^2\}, \tag{15}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are two feature vectors, and γ is a positive scalar.

3. Experiments and results

The proposed segmentation and classification methods were evaluated on a dataset of range and intensity images captured by a MESA Imaging time-of-flight (TOF) camera. Section 3.1 describes the MESA image dataset and the performance measures used to assess the effectiveness of the proposed algorithms. Section 3.2 analyzes the effects of the various parameters on the performance of the algorithm, whereas Section 3.3 evaluates the effectiveness of each step in the proposed segmentation algorithm. Section 3.4 compares the proposed segmentation algorithm with existing algorithms. Section 3.5 evaluates the performance of the proposed pedestrian classification algorithm on two data sets: the MESA dataset and the RGB-D public dataset.

3.1. Experimental methods

A dataset was acquired using a TOF camera produced by MESA Imaging (model SwissRanger SR4000). For each pixel, the Swiss-

Ranger camera produces five outputs: (x, y, z) coordinates, intensity, and the confidence score. The camera operates at a speed of 30 frames per second, with a frame size of 144×176 pixels. The depth field of the camera extends from 0.5 m to 7.9 m. The images were taken for different indoor and outdoor scenes, on different days, and under various lighting conditions. Samples of the range and intensity image pairs are shown in Fig. 8.

Overall, 2000 images were extracted from a set of 50 videos, involving 20 different people performing different activities (walking, running, or standing). Each video sequence was subsampled, and only 20 frames with different background scenes were selected. A summary of the MESA dataset is given in Table 1. For segmentation, 220 images were segmented manually to generate the ground-truth. Twenty randomly selected images were used to determine suitable segmentation parameters; the remaining 200 range images were used for segmentation evaluation and comparison with other algorithms. For classification, a set of 2000 images (range and intensity) with labeled pedestrians was collected. Examples of the ground-truth for segmentation and pedestrian classification are shown in Fig. 9.

The segmentation performance was evaluated using several measures: correct segmentation rate, false segmentation rate, and weighted Jaccard coefficient. Consider an image with M machine segmented regions and G ground-truth regions. The Jaccard similarity coefficient between a machine segmented region R_m and a ground-truth region R_g is defined as

$$J(R_m, R_g) = \frac{|R_m \cap R_g|}{|R_m \cup R_g|}. \tag{16}$$

The segmentation quality was also measured by the weighted Jaccard coefficient J_w, which takes into account the region size:

$$J_w = \frac{\sum_{i=1}^{G} J_i \times A_i}{\sum_{i=1}^{G} A_i}, \tag{17}$$

Table 1
MESA range image dataset for segmentation and classification.

Image set	No. of images
Range images	1000
Intensity images	1000
Segmentation ground-truth	220
Classification ground-truth	2000

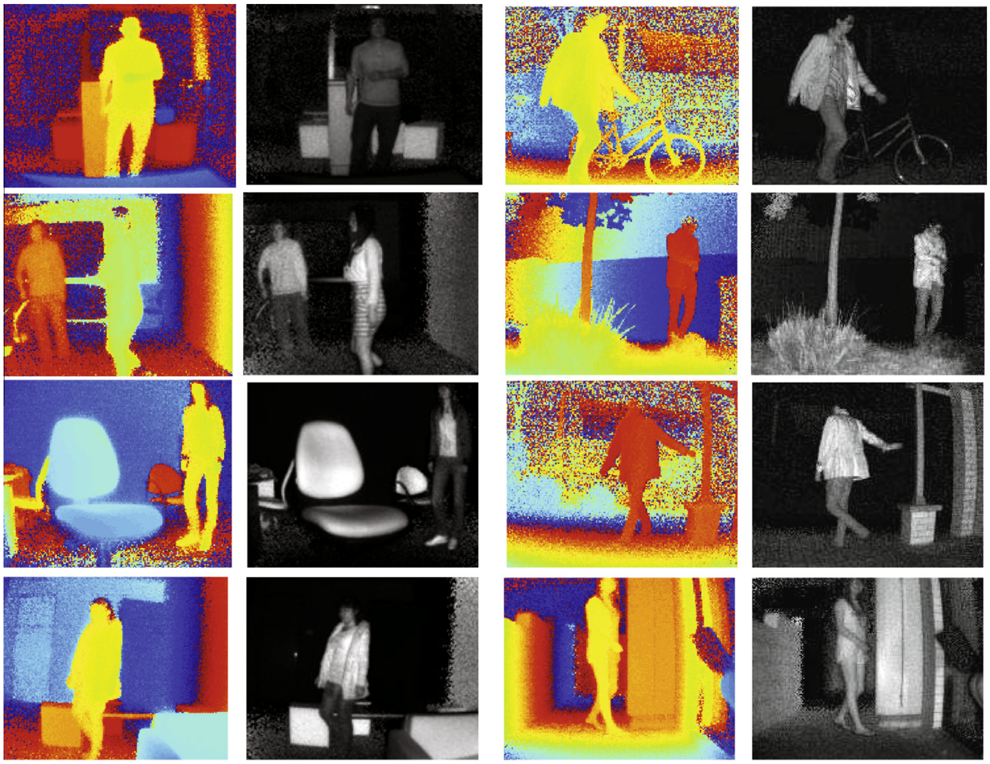
where A_i is the area of the ith ground-truth region.

In our experiments, region R_m was considered to be correctly segmented if the weighted Jaccard coefficient J_w is higher than 0.4. The correct segmentation rate P_c is defined as the percentage of ground-truth regions that are correctly segmented. The false segmentation rate P_f is defined as the percentage of machine-generated regions that are incorrectly segmented.

3.2. Selection of segmentation parameters

We first conducted experiments on a set of 20 training images to determine suitable parameters for the proposed segmentation algorithm. The segmentation accuracy was evaluated for different values of histogram size B, ranging from 16 to 128. Fig. 10 shows the segmentation outputs on a sample training image. For small B (e.g. $B = 16$), there is significant under-segmentation, whereas for large B ($B = 64$ or $B = 128$), there is significant over-segmentation. When $B = 32$, the segmentation output (Fig. 10d) resembles most the ground-truth (Fig. 10b). The weighted Jaccard coefficients for different histogram sizes, averaged over the 20 training images, are given in Table 2. The largest weighted Jaccard coefficient is 75.1%, obtained when $B = 32$. Therefore, a histogram size of 32 bins is selected for the remainder of the experiments.

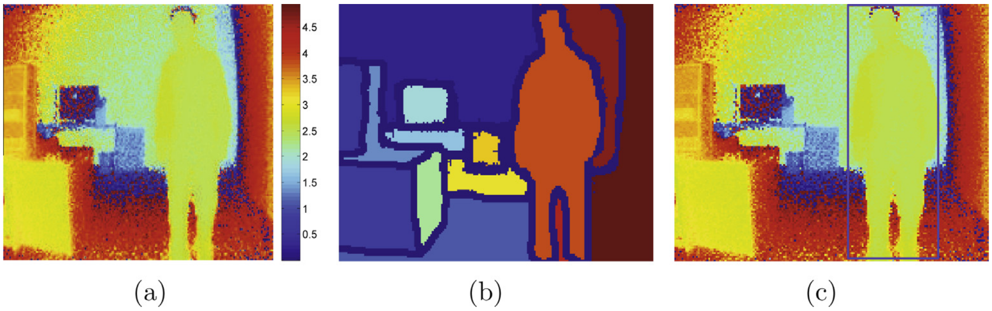
The segmentation performance on the 20 training images was also evaluated for different values of the segmentation parameters



(a) indoor scenes

(b) outdoor scenes

Fig. 8. Sample range and intensity images from the MESA dataset. Left columns: range images; Right columns: intensity images.



(a) (b) (c)

Fig. 9. Ground-truth data: (a) a range image; (b) segmentation ground-truth; (c) labeled pedestrian region.

τ_a, τ_n, and τ_l. Parameter τ_a is used in mean-shift clustering, τ_n is used in ground segmentation, and τ_l is used in post-processing (see Section 2). Fig. 11 shows the weighted Jaccard coefficient J_w and the correct segmentation rate P_c as functions of τ_a, τ_n, and τ_l. Based on these results, the following parameter values are selected for the proposed segmentation algorithm: $\tau_a = 0.09, \tau_n = 0.2$, and $\tau_l = 0.3$.

3.3. Analysis of the proposed segmentation algorithm

Experiments were conducted on the 200 test images to evaluate the effects of individual processing stages on the proposed segmentation algorithm. There are four main stages: pre-processing, histogram-based segmentation, mean-shift clustering, and post-processing. Table 3 presents the segmentation performance of different processing stages; the 95% confidence intervals are also shown. With only histogram thresholding, the correct segmentation rate is 58.9%. The median filter increases the correct segmentation rate to 63.2%. Applying ground detection, confidence map thresholding, and boundary removal increases the correct segmentation rate to 59.3%, 64.2%, and 66.6%, respectively. After all the pre-processing steps for the 3-D range images, the correct segmentation rate increases to 73.0%. Adding a post-processing step (edge-merging) increases the correct segmentation rate to 80.7%. By including mean-shift clustering, the final segmentation rate increases to 83.2%. These results indicate that combining four extra proposed processing stages improves the segmentation performance significantly.

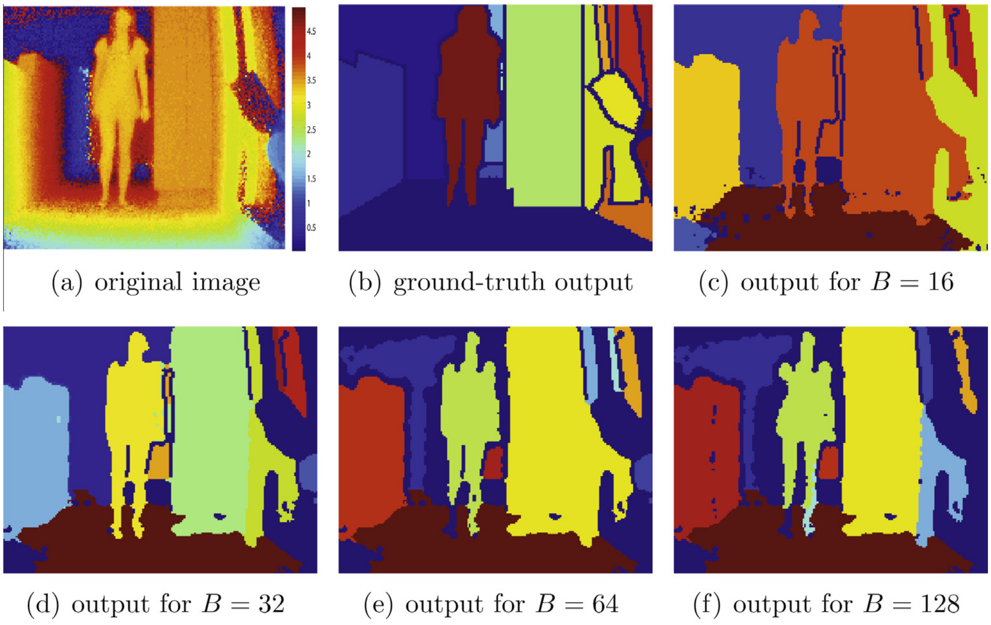
Table 2
The average weighted Jaccard coefficients for different histogram sizes B, on the 20 training images.

B	8	16	32	64	128
J_w (%)	67.7	72.6	75.1	73.9	70.6

In the proposed algorithm, mean-shift clustering was applied only to the 3-D data, i.e., $\mathbf{p}_i = (x_i, y_i, z_i)$. We also evaluated the segmentation performance when mean-shift clustering was applied to the combination of 3-D and intensity data, i.e., $\mathbf{p}_i = (x_i, y_x, z_i, I_i)$. For this approach, the final segmentation rate drops to 75.3%, false segmentation increases to 36.2%, and the weighted Jaccard coefficient drops to 66.7%. Furthermore, the average processing time increases to 2.0 s. Therefore, we can conclude that the combination of range and intensity data in mean-shift clustering does not improve segmentation performance. A possible reason is that adding intensity causes the objects to be over-segmented into several smaller regions.

3.4. Comparison with existing segmentation algorithms

Using the set of 200 test images, the proposed segmentation algorithm was compared with several state-of-the-art methods:



(a) original image (b) ground-truth output (c) output for $B = 16$

(d) output for $B = 32$ (e) output for $B = 64$ (f) output for $B = 128$

Fig. 10. Visual results of segmentation for different histogram size B.

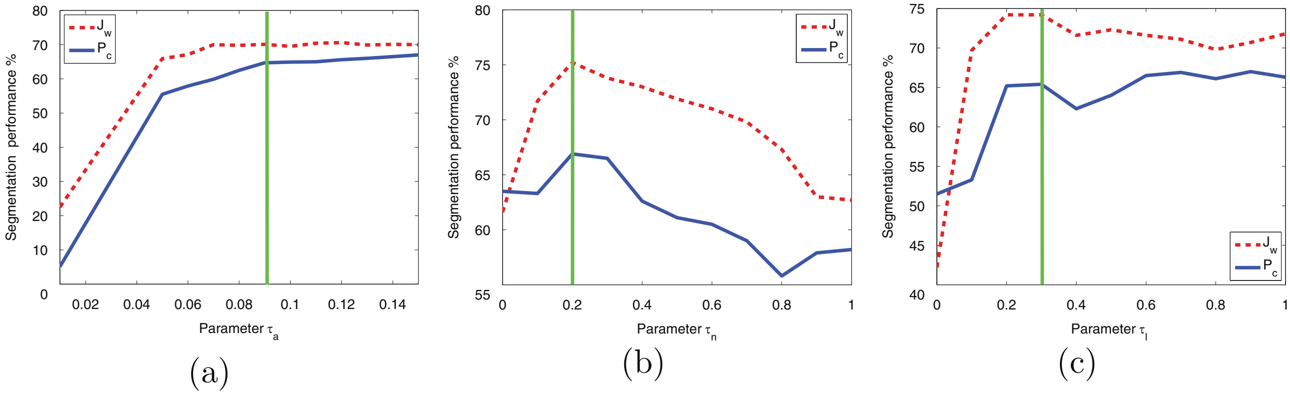


Fig. 11. Segmentation performances on a training set of 20 images for different values of segmentation parameters τ_a, τ_n, and τ_l. P_c is the correct segmentation rate, and J_w is the weighted Jaccard coefficient. The green lines indicate the selected parameter values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Segmentation performances with confidence intervals of the proposed method with different processing steps.

Algorithm	Correct segmentation rate (%)	False segmentation rate (%)	Weighted Jaccard (%)	Processing time (s)
Histogram thresholding only	58.9 ± 2.4	40.2 ± 2.4	54.0 ± 2.4	0.3
+ 2-D median filtering	63.2 ± 2.3	39.8 ± 2.4	54.4 ± 2.4	0.1
+ Ground detection	68.1 ± 2.2	36.8 ± 2.3	58.0 ± 2.4	0.4
+ Confidence map thresholding	59.3 ± 2.4	44.8 ± 2.4	54.0 ± 2.4	0.1
+ Boundary removal	66.6 ± 2.3	42.7 ± 2.4	58.5 ± 2.4	0.1
Histogram thresholding + all pre-processing steps	73.0 ± 2.1	36.9 ± 2.3	62.7 ± 2.3	1.0
+ Edge merging	80.7 ± 1.9	35.1 ± 2.3	66.2 ± 2.3	0.3
+ Mean-shift and edge merging	83.2 ± 1.8	34.9 ± 2.3	69.5 ± 2.2	1.8

Table 4
Segmentation performances with confidence intervals of six methods evaluated on the MESA dataset.

	Correct segmentation rate (%)	False segmentation rate (%)	Weighted Jaccard (%)	Processing time (s)
Markov random fields [43]	60.1 ± 2.4	36.8 ± 2.3	58.6 ± 2.4	30.0
Local variation [39]	63.0 ± 2.3	28.6 ± 2.2	64.5 ± 2.3	3.4
Graph cuts [21]	57.1 ± 2.4	39.2 ± 2.3	56.4 ± 2.4	0.5
K-means	56.9 ± 2.3	35.7 ± 2.4	60.3 ± 2.4	0.4
Bearing angles [18]	71.5 ± 2.2	32.2 ± 2.2	63.8 ± 2.3	0.1
Proposed method	83.2 ± 1.8	34.9 ± 2.3	69.5 ± 2.2	1.8

local variation [39], graph cuts [21], K-means, Markov random fields [40], and bearing angle (BA) with edge detection [18]. The same pre-processing and post-processing steps were applied to all methods. The same set of 20 training images was used to determine suitable parameters of each segmentation method. In the following, the implementation details of the various segmentation methods are presented.

Local variation (LV) is a graph-based segmentation method. It merges two components if the external variation is small relative to the internal variation. Our implementation of LV was based on the code of Felzenszwalb and Huttenlocher [39] and Su [41]. The three parameters in LV were smoothness parameter θ of Gaussian filter, minimum size of the segmentation component s, and radius of nearest neighborhood k. The parameters determined using the 20 training images are $\theta = 0.8$, $s = 300$ and $k = 10$.

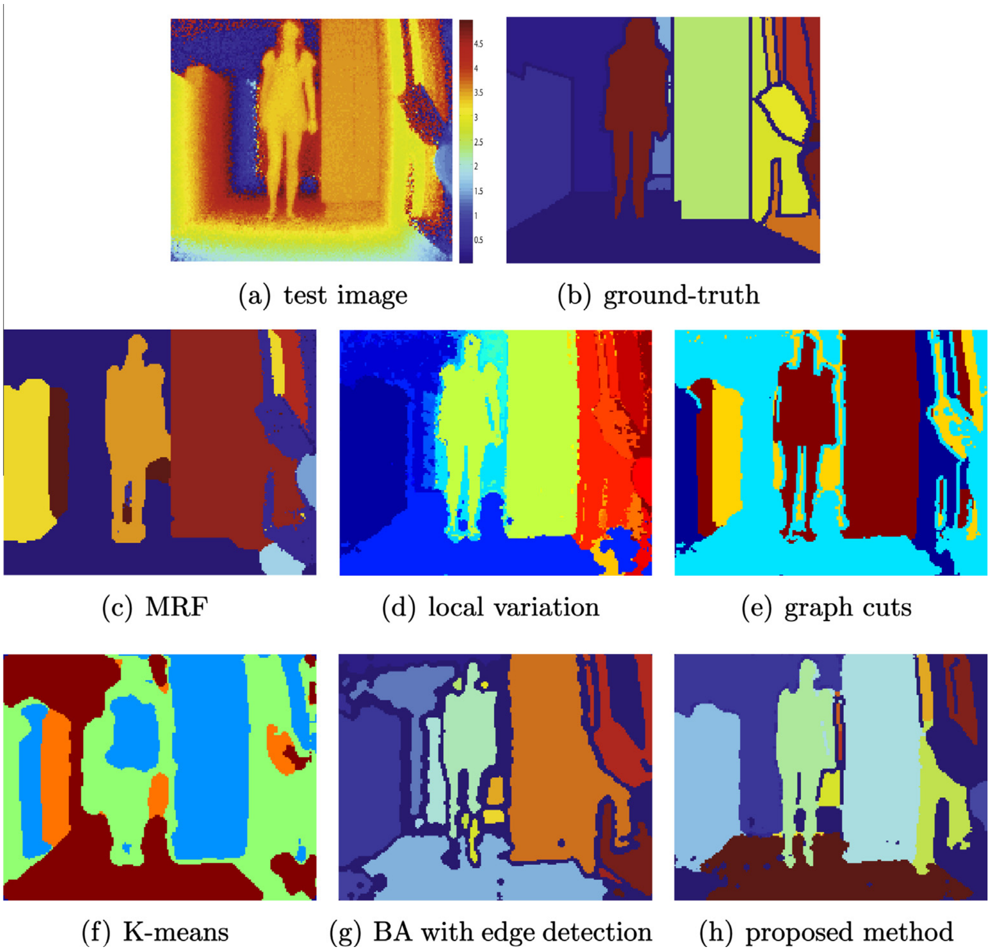
Graph cuts is a segmentation algorithm based on energy minimization. Boykov et al. proposed two graph cuts algorithms, namely swap and expansion [21]. The expansion algorithm with 10 clusters per image was used in our experiment. The implementation of graph cuts was based on the code of Bagon [42].

K-means is a cluster-based segmentation algorithm that is widely used in color image segmentation. Based on the training images, the cluster number K was set to 10 in our experiment.

Markov random fields have been used for segmenting 3-D images in [20,22,43]. In our experiment, the histogram-thresholding was used to provide an initial segmentation, which was then used as the input to the iterated conditional modes (ICMs) algorithm to estimate the MAP solution. The MRF method was implemented based on the MATLAB toolbox provided by Demirkaya et al. [43]. In our experiment, the number of classes was initialized to 15, the number of iteration was set to 10, and parameter β for the Gibbs energy function was set to 1.5.

Bearing angles with edge detection was proposed for indoor 3-D simultaneous localization and mapping (3D-SLAM) [18]. In this method, range images were first transformed to BA images in vertical and horizontal directions. The edges of the two BA images were detected by the Sobel edge detector, with threshold $\tau_{BA} = 0.3$. The final edge map was obtained using logical OR operator applied to the edge images. After edge detection, a median filter of size 3×3 was used to remove noise, and connected component labeling was applied to obtain the final segmentation.

Table 4 summarizes the segmentation results of different methods in terms of correct segmentation rate P_c, false segmentation rate P_f, weighted Jaccard coefficient J_w, and processing time. Compared to the other methods, the proposed method has a higher correct segmentation rate, a higher weighted Jaccard coefficient, and a



(a) test image (b) ground-truth

(c) MRF (d) local variation (e) graph cuts

(f) K-means (g) BA with edge detection (h) proposed method

Fig. 12. Range image segmentation using six different segmentation algorithms on a test image.

Table 5
Segmentation performance on MESA dataset according to the evaluation framework in [44] with $\tau_h = 0.6$.

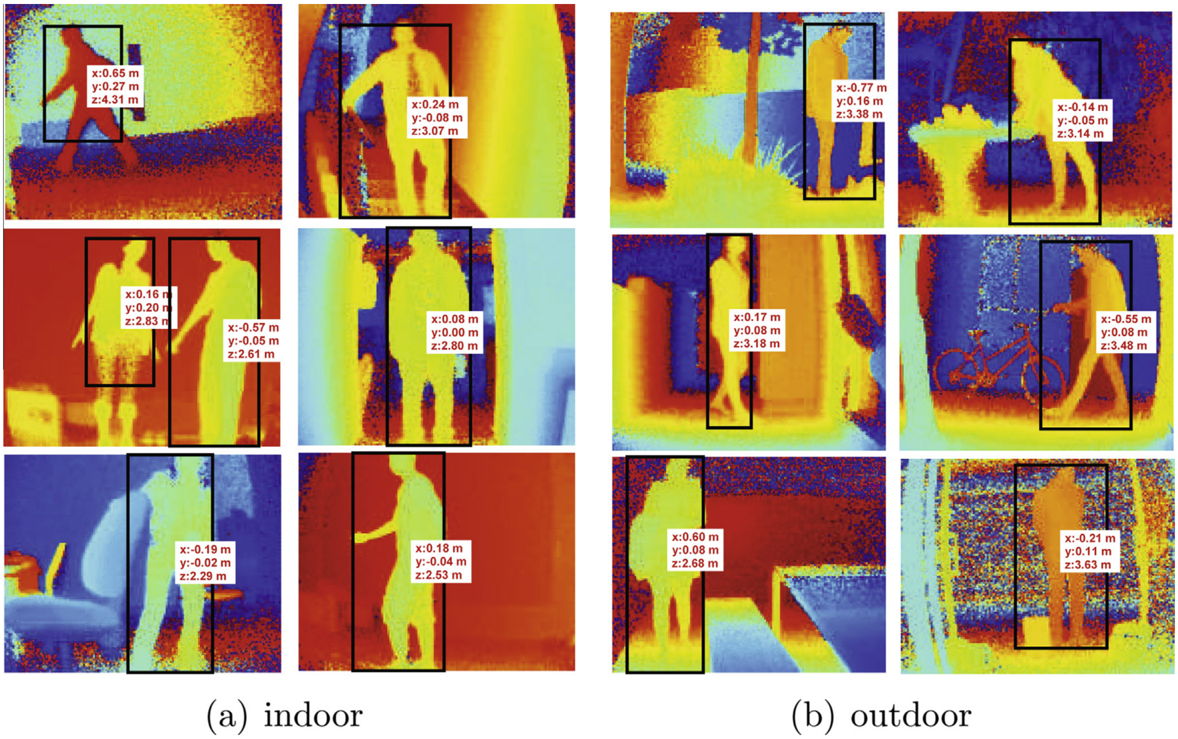
Algorithms	Correct segmentation rate (%)	Over segmentation rate (%)	Under segmentation rate (%)	Missed segmentation rate (%)	Noise segmentation rate (%)
Graph cut [21]	33.3 ± 2.3	33.7 ± 2.3	65.6 ± 2.3	14.2 ± 1.7	11.8 ± 1.6
K-means	34.2 ± 2.3	49.3 ± 2.4	47.1 ± 2.4	9.3 ± 1.4	25.8 ± 2.1
Local variation [39]	42.7 ± 2.4	46.9 ± 2.4	55.4 ± 2.4	5.5 ± 1.1	16.1 ± 1.8
Markov random fields [43]	38.0 ± 2.3	65.6 ± 2.3	38.2 ± 2.3	5.2 ± 1.1	22.9 ± 2.0
Bearing angles [18]	48.5 ± 2.4	24.0 ± 2.1	71.1 ± 2.2	15.4 ± 1.7	9.8 ± 1.4
Proposed method	63.0 ± 2.3	35.0 ± 2.3	61.6 ± 2.3	25.0 ± 2.1	3.4 ± 0.9

Table 6
Classification rates of pedestrian versus non-pedestrian classification on the MESA dataset.

Algorithms	CR for only range images (%)	CR for only intensity images (%)	CR for both range/intensity images (%)	Processing time (s)
HOG–HOD [9,29]	93.7 ± 0.7	88.6 ± 0.9	95.4 ± 0.6	0.06
SIFT–SPM [45,46]	93.2 ± 0.7	93.5 ± 0.7	96.2 ± 0.6	0.40
GIST [37,35]	97.5 ± 0.4	97.2 ± 0.5	98.2 ± 0.4	0.30
Proposed method	97.7 ± 0.4	97.2 ± 0.5	98.6 ± 0.3	0.32

lower false segmentation rate. The segmentation results of the six algorithms on a test image are shown in Fig. 12.

All segmentation algorithms were also compared based on the evaluation framework proposed by Hoover et al. [44]. In this evaluation framework, a machine-generated region R_m is considered *correctly-segmented* with a tolerance rate τ_h if the ratio $|R_m \cap R_g|/\max(|R_m|, |R_g|)$ is greater than or equal to τ_h. Here, $|R_m \cap R_g|$ is the area of the overlap between the machine-generated region R_m and the ground-truth region R_g. The correct segmentation rate of an algorithm is defined as the percentage of machine-generated regions that are correctly segmented. Note that a machine-generated region is considered *under-segmented* if it consists of multiple ground-truth regions. A machine-generated region is considered *over-segmented* if it is smaller than the corresponding ground-truth region. A region is considered as a *missed segmentation* when a segmenter fails to find a region which appears in the image (false negative). A region is considered a *noise segmentation* if the segmenter finds a region which does not exist in the ground-truth image (false positive).



(a) indoor

(b) outdoor

Fig. 13. Sample results of the proposed pedestrian classification. The red numbers are the positions in meters from the pedestrians. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Classification rates of pedestrian versus non-pedestrian classification on the RGB-D dataset.

Algorithms	Feature-level Fusion (%)	Decision-level Fusion (%)
HOG–HOD [9,29]	97.4 ± 0.5	95.0 ± 0.6
SIFT–SPM [45,46]	97.8 ± 0.3	97.4 ± 0.5
GIST [37,35]	98.6 ± 0.3	98.8 ± 0.3
Proposed method	98.4 ± 0.4	99.0 ± 0.3

Table 5 summarizes the segmentation results of different methods in terms of correct segmentation rate, over segmentation rate, under segmentation rate, and false segmentation rate. Compared with the other five algorithms, the proposed algorithm achieves the highest correct segmentation rate.

3.5. Pedestrian classification

In this section, the proposed pedestrian classification algorithm based on range and intensity images was evaluated on a dataset comprising 1000 pedestrian and 1000 non-pedestrian patterns. The background was varied to include both indoor and outdoor scenes. The Fourier and GIST features were extracted from range and intensity regions as described in Sections 2.4.1 and 2.4.2. The classification rate was measured using fivefold cross validation. For each validation fold, four subsets were used for training, and the remaining subset is used for testing. This was repeated 5 times (each time using a different test subset); the classification rate was the percentage of test patterns that are correctly classified over the five folds. The confidence interval of the classification rate was also computed. For comparison, three other state-of-the-art feature extractors (HOG–HOD [9,29], SIFT–SPM [45,46], and GIST [37,35]) were evaluated on the MESA dataset. We also evaluated the performances of feature vectors that are extracted separately from range images and from intensity images.

Table 6 shows the classification performances of different feature extractors. The HOG–HOD descriptor has a classification rate of 95.4%. The histogram of orientations does not perform well for low-resolution range and intensity images. In extracting the HOG–HOD features, all segmented regions are reshaped to a fixed aspect ratio, which increases the false classification rate. The SIFT–SPM descriptor has a classification rate of 96.2%. This descriptor is suitable for recognizing objects based on texture [45,46]. However, object texture is not very dominant in range images. Similarly to the HOG–HOD, the SIFT–SPM requires regions to be reshaped to a fixed aspect ratio. The GIST descriptor achieves a classification rate of 98.2%. It performs better than the HOG–HOD and SIFT–SPM across different image modalities (range only, intensity only,

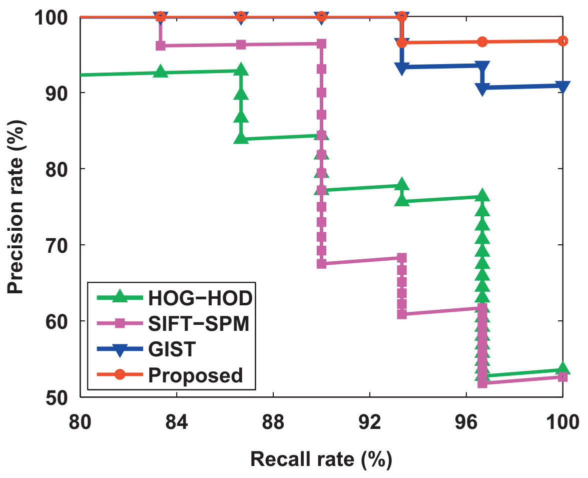


Fig. 14. Precision recall curves of HOG–HOD, SIFT–SPM, GIST, and the proposed method on the RGB-D dataset.

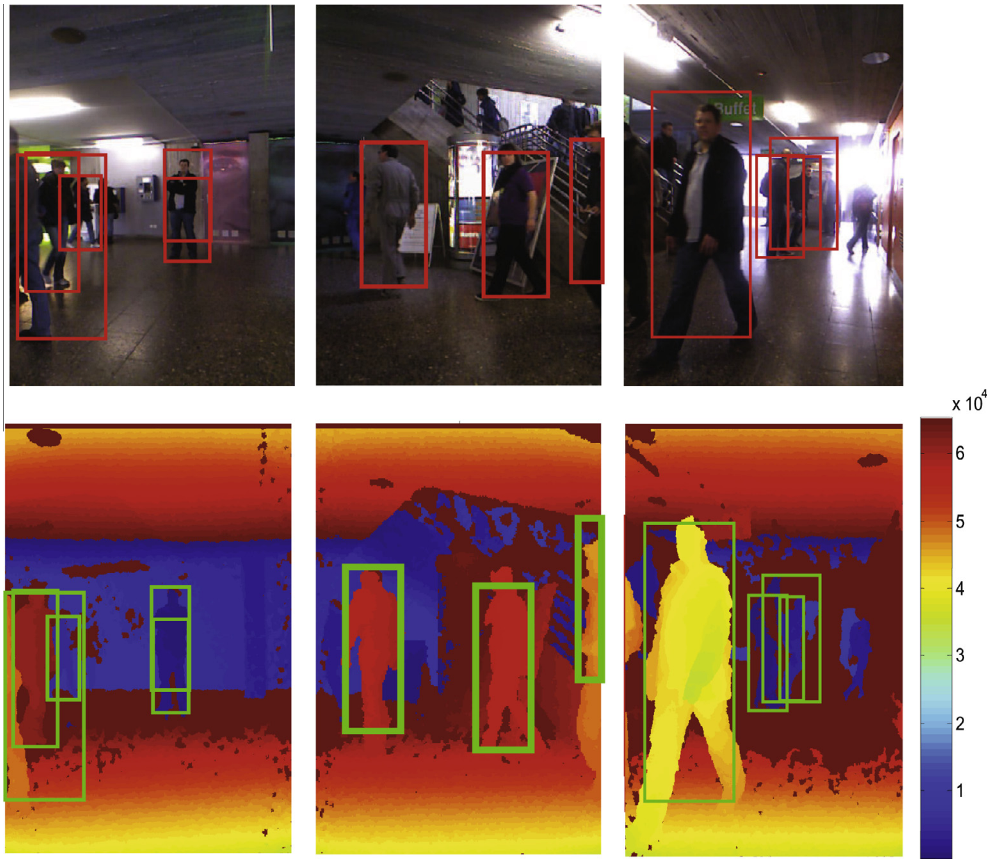


Fig. 15. Multiple-pedestrian detection using the proposed algorithm on the RGB-D dataset. Top row: color/intensity images. Bottom row: range images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

or combined range and intensity). The GIST descriptor is suitable for extracting global features of objects, especially from noisy low-resolution images. Furthermore, it is not sensitive to the aspect ratio of the object.

Finally, the proposed method, which combines GIST and Fourier descriptors, yields the highest classification rate of 98.6%. A possible reason is that the GIST method extracts global texture information in range and intensity images, whereas the Fourier descriptor enhances boundary features of the objects. Combining global and shape features improves the classification accuracy. Sample results of the pedestrian classification are shown in Fig. 13.

The proposed method was also compared with HOG–HOD, SIFT–SPM and GIST on the RGB-D people dataset [29]. This dataset was extracted from 3 movie sequences. It contains 1035 pedestrian and 1035 non-pedestrian patterns. The classification rate was measured using fivefold cross validation.

Table 7 shows the classification rates on RGB-D dataset. For each of the four methods (HOG–HOD, SIFT–SPM, GIST, and the proposed algorithm), two fusion approaches were evaluated: (i) fusion at the sensory level (feature-level); and (ii) fusion at the classifier level (decision-level). For feature-level fusion, a feature vector that combines all imaging modalities (range and intensity) is extracted, and a single SVM classifier is trained.

For decision-level fusion, one SVM classifier is trained based on the feature vector extracted from each imaging modality (range or intensity). The final classification is determined by fusing the decisions of several SVM classifiers. In this approach, the two SVM classifiers for range and intensity images are fused as follows [29]:

$$p = (1-k)p_r + kp_i, \tag{18}$$

where p is the resultant probability of detecting a pedestrian, and p_i and p_r are the probabilities of detection obtained from the intensity-image and range-image classifiers, respectively. The value of k in (18) is determined as $k = 1/(1+\theta^2)$, where $\theta = F_r/F_i$. Here, F_r and F_i are the false negative rate of the range-image classifier and the intensity-image classifier, respectively.

For HOG–HOD and SIFT–SPM features, using decision-level fusion does not improve the classification rate compared to feature-level fusion. For both the GIST descriptor and the proposed method, the decision-level fusion yields higher classification rates than the feature-level fusion. For the feature-level fusion approach, the GIST and the proposed method achieve higher classification rates than the HOG–HOD and SIFT–SPM methods. For the decision-level fusion approach, the proposed method has a higher classification rate than the HOG–HOD, SIFT–SPM, and GIST methods. The precision and recall curves of the compared algorithms in decision-level fusion are shown in Fig. 14. The equal error rates for HOG–HOD, SIFT–SPM, GIST, and the proposed method are 86.7%, 90.0%, 93.3%, and 96.7% respectively. Fig. 15 shows some pedestrian detection results obtained with the proposed method.

4. Conclusion

In this paper, a new approach for segmenting and classifying objects using a 3-D time-of-flight range camera was presented. An image segmentation approach was proposed that reduces segmentation errors by combining histogram processing, mean-shift clustering, edge merging, and ground normal vector thresholding. The segmented objects are classified into pedestrian and non-pedestrian patterns by combining Fourier descriptors and GIST features extracted from range and intensity images. The proposed segmentation and pedestrian classification algorithms were evaluated on 3-D range data, and compared with existing state-of-the-art methods. Experimental results show that the proposed methods are more effective in segmenting 3-D scenes and classifying pedestrians from non-pedestrian objects.

Acknowledgments

We thank the three anonymous reviewers for their constructive comments that helped improve this article.

References

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

[2] M. Bertozzi, A. Broggi, M. Cellario, A. Fascioli, P. Lombardi, M. Porta, Artificial vision in road vehicles, Proceedings of the IEEE 90 (2002) 1258–1271.

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

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

[5] M.E. Farmer, A.K. Jain, A wrapper-based approach to image segmentation and classification, IEEE Transactions on Image Processing 14 (2005) 2060–2072.

[6] S. Gould, T. Gao, D. Koller, Region-based segmentation and object detection, in: Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems, 2009, pp. 655–663.

[7] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, International Journal of Computer Vision 77 (2008) 259–289.

[8] A. Torralba, K.P. Murphy, W.T. Freeman, Contextual models for object detection using boosted random fields, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, 2005, pp. 1401–1408.

[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and, Pattern Recognition, 2005, pp. 886–893.

[10] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, Groups of adjacent contour segments for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 36–51.

[11] B. Leibe, A. Leonardis, B. Schiele, Combined object categorization and segmentation with an implicit shape model, in: Proceedings of the European Conference on Computer Vision Workshops, 2004, pp. 17–32.

[12] R. Hoffman, A.K. Jain, Segmentation and classification of range images, IEEE Transactions on Pattern Analysis and Machine Intelligence 9 (1987) 608–620.

[13] P.J. Besl, R.C. Jain, Segmentation through variable-order surface fitting, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (1988) 167–192.

[14] J. Mukherjee, P.P. Das, B.N. Chatterji, Segmentation of range images, Pattern Recognition 25 (1992) 1141–1156.

[15] G.M. Hegde, Y. Cang, A recursive planar feature extraction method for 3D range data segmentation, in: IEEE International Conference on Systems, Man, and, Cybernetics, 2011, pp. 3119–3124.

[16] G.P.M. Hegde, Y. Cang, Extraction of planar features from SwissRanger SR3000 range images by a clustering method using normalized cuts, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, pp. 4034–4039.

[17] K.M. Varadarajan, M. Vincze, Object part segmentation and classification in range images for grasping, in: International Conference on, Advanced Robotics, 2011, pp. 21–27.

[18] A. Harati, S. Gachter, R. Siegwart, Fast range image segmentation for indoor 3D-SLAM, in: IFAC Symposium on Intelligent Autonmous Vehicles, 2007.

[19] S.A. Coleman, B.W. Scotney, S. Suganthan, Edge detecting for range data using Laplacian operators, IEEE Transactions on Image Processing 19 (2010) 2814–2824.

[20] X. Wang, H. Wang, Markov random field modeled range image segmentation, Pattern Recognition Letters 25 (2004) 367–375.

[21] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 1222–1239.

[22] C. Zhang, L. Wang, R. Yang, Semantic segmentation of urban scenes using dense depth maps, in: Proceedings of the European Conference on Computer vision, 2010, pp. 708–721.

[23] K. Eunyoung, G. Medioni, Scalable object classification using range images, in: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2011, pp. 65–72.

[24] D. Das, Y. Kobayashi, Y. Kuno, Multiple object detection and localization using range and color images for service robots, in: ICCAS-SICE International Joint Conference, 2009, pp. 3485–3489.

[25] B. Fardi, J. Dousa, G. Wanielik, B. Elias, A. Barke, Obstacle detection and pedestrian recognition using a 3D PMD camera, in: IEEE Intelligent Vehicles, Symposium, 2006, pp. 225–230.

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

[27] M. Rapus, S. Munder, G. Baratoff, J. Denzler, Pedestrian recognition using combined low-resolution depth and intensity images, in: IEEE Intelligent Vehicles, Symposium, 2008, pp. 632–636.

[28] O.M. Mozos, R. Kurazume, T. Hasegawa, Multi-part people detection using 2D range data, International Journal of Social Robotics 2 (2010) 31–40.

[29] L. Spinello, K.O. Arras, People detection in RGB-D data, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011, pp. 3838–3843.

[30] E. Bart, I. Porteous, P. Perona, M. Welling, Unsupervised learning of visual taxonomies, in: IEEE Conference on Computer Vision and, Pattern Recognition, June, pp. 1–8.

[31] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 42 (2001) 177–196.

[32] M. Enzweiler, D. Gavrila, Monocular pedestrian detection: survey and experiments, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2008) 2179–2195.

[33] F.H.C. Tivive, A. Bouzerdoum, S.L. Phung, K.M. Iftekharuddin, Adaptive hierarchical architecture for visual recognition, Applied Optics 49 (2010) B1–B8.

[34] S.L. Phung, A. Bouzerdoum, Detecting people in images: an edge density approach, in: IEEE International Conference on Acoustics, Speech, and, Signal Processing, 2007, pp. 1229–1232.

[35] X. Wei, S.L. Phung, A. Bouzerdoum, Pedestrian sensing using time-of-flight range camera, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2011, pp. 43–48.

[36] P.F. Felzenszwalb, D.P. Huttenlocher, Image segmentation using local variation, in: IEEE Conference on Computer Vision and, Pattern Recognition, 1998, pp. 98–104.

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

[38] E. Persoon, K.-S. Fu, Shape discrimination using fourier descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (1986) 388–397.

[39] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 59 (2004) 167–181.

[40] O. Demirkaya, M.H. Asyali, M.M. Shoukri, Segmentation of CDNA microarray spots using markov random field modeling, Bioinformatics 21 (2005) 2994–3000.

[41] D. Su, Matlab code of efficient graph-based image segmentation, 2009. <http://www.mathworks.com.au/matlabcentral/fileexchange/25866-efficient-graph-based-image-segmentation>.

[42] S. Bagon, Matlab wrapper for graph cut, 2016. <http://vision.ucla.edu/brian/gcmex.html>.

[43] O. Demirkaya, M.H. Asyali, P. Sahoo, Image Processing with MATLAB: Applications in Medicine and Biology, CRC Press, 2008. <https://sites.google.com/site/odkaya/downloads>.

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

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

[46] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.