3-D Point Cloud Object Detection Based on Supervoxel Neighborhood With Hough Forest Framework

Hanyun Wang, Cheng Wang, *Member, IEEE*, Huan Luo, Peng Li,
Yiping Chen, and Jonathan Li, *Senior Member, IEEE*

***Abstract*—Object detection in three-dimensional (3-D) laser scanning point clouds of complex urban environment is a challenging problem. Existing methods are limited by their robustness to complex situations such as occlusion, overlap, and rotation or by their computational efficiency. This paper proposes a high computationally efficient method integrating supervoxel with Hough forest framework for detecting objects from 3-D laser scanning point clouds. First, a point cloud is over-segmented into spatially consistent supervoxels. Each supervoxel together with its first-order neighborhood is grouped into one local patch. All the local patches are described by both structure and reflectance features, and then used in the training stage for learning a random forest classifier as well as the detection stage to vote for the possible location of the object center. Second, local reference frame and circular voting strategies are introduced to achieve the invariance to the azimuth rotation of objects. Finally, objects are detected at the peak points in 3-D Hough voting space. The performance of our proposed method is evaluated on real-world point cloud data collected by the up-to-date mobile laser scanning system. Experimental results demonstrate that our proposed method outperforms state-of-the-art 3-D object detection methods with high computational efficiency.**

***Index Terms*—Hough forest, local reference frame (LRF), mobile laser scanning (MLS), object detection, point clouds, supervoxel neighborhood.**

I. Introduction

With the development of mobile laser scanning (MLS) systems [1]–[3] in recent years, the time of data collection is tremendously reduced for subsequent use in urban planning, maintenance functions, emergency response preparation, virtual tourism, and multimedia entertainment. As an important approach for urban scene analysis, automatic extraction of important urban scene structures such as road signs, lamp posts, and cars from three-dimensional (3-D) point cloud data has become a highly attractive research topic. The major challenges in detecting objects from laser scanning point clouds are huge point cloud data volumes, intraclass shape variation, overlap between neighboring objects, point-density variation, orientation variation, and incompleteness of object caused by occlusion.

Most of existing object detection work is based on prior knowledge of the specified object categories [4]–[16] or based on object global shape description [17]–[19]. These methods either are difficult to extend from the specific object categories to more generic object categories or require the completeness of objects, which is hard to ensure in complex real-world urban scenes.

Hough forest [20], [21] is a successful object detection framework based on a partial object description model—implicit shape model (ISM). ISM [22] is essentially a codebook of local part appearances. Similar features are grouped into the same codebook entry, and each entry contains the same type of local structure. Based on ISM, each local part of an object votes the possible location of the object center. This attribute makes ISM robust to occlusion and overlap, which commonly exist in complex urban environments. Hough forest has been successfully used for object detection in two-dimensional (2-D) images [20], [21] and 3-D shapes [23]–[26].

However, Hough forest lacks the ability to detect objects with arbitrary orientation [21]. Because objects, such as cars and pedestrians, are typically upright in normal images, lacking the ability to detect objects with arbitrary orientation is not an issue in normal image applications [20]. But in real-world 3-D scenes, identical categorical objects are commonly placed in various azimuth orientations. Therefore, for object detection in point clouds of complex urban environment, a critical requirement is the invariance to azimuth rotation. The detection performance of Hough forest also heavily depends on the distinctiveness of the object's local parts. Existing part-based object detection methods either cannot control the size of each part [18], [27] or does not consider the interior shape structures [26]. Moreover, detecting objects from MLS point clouds requires high computational efficiency.

In this paper, we propose a method with high computational efficiency to detect objects from real-world 3-D laser

Manuscript received August 29, 2014; revised October 22, 2014; accepted January 14, 2015. This work was supported by the National Science Foundation of China under Project 61371144. *(Corresponding author: C. Wang.)*

H. Wang, P. Li, and Y. Chen are with the School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China.

C. Wang and H. Luo are with the Fujian Key Laboratory of Sensing and Computing For Smart Cities, Department of Computer Science, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: cwang@xmu.edu.cn).

J. Li is with the Fujian Key Laboratory of Sensing and Computing For Smart Cities, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China and also with the GeoSpatial Technology and Remote Sensing Laboratory, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2015.2394803

scanning point clouds. Supervoxel, which groups the 3-D points into perceptually meaningful clusters with high efficiency, is incorporated into the Hough forest framework to accelerate the 3-D local patch extraction. The point cloud is first segmented into supervoxels through the voxel cloud connectivity segmentation (VCCS) algorithm [28]. Each supervoxel, together with its first-order neighborhood, is grouped into one local patch. All the local patches are described by both structure and reflectance features, and then used in the training stage for learning a random forest classifier as well as the detection stage to vote for the possible location of the object center. To cope with the azimuth rotation of objects, we use both the local reference frame (LRF) [29]–[31] and the circular voting [26] in the Hough voting stage. The LRF-based voting strategy is the 3-D counterpart of gradient-based voting strategy, which is used for 2-D remote sensing images [21]. By defining LRFs for the 3-D local parts of ISM, we estimate the rotation transformation between matched pairs of asymmetrical training and testing parts. Then, we align the offset vector, according to the rotation transformation, to correctly vote the object center. If one of the matched pair of parts is symmetrical, a circular voting strategy is introduced by rotating the offset vector, and then all locations with certain distances in the horizontal plane and in the vertical direction are voted. Experimental results demonstrate the robustness and efficiency of our proposed algorithm on complex urban environment point clouds acquired by an MLS system.

II. Related Work

A. Existing Object Detection Methods

Most of existing object detection methods can be divided into two classes. The first class of approaches is based on prior knowledge of the specified object categories. Jaakkola *et al.* [4] extracted curbstones and road markings, such as zebra crossings, from MLS data based on intensity and height information, and modeled the pavement as a triangulated irregular network. Guan *et al.* [5] extracted road markings by interpolating 3-D points into geo-referenced intensity images. Yu *et al.* [6] detected buildings from airborne laser scanning points based on a marked point process algorithm. Brenner [7] developed a rule-based pole extraction method for intelligent driving guidance and assistance. Lehtomäki *et al.* [8] detected pole-like objects based on scan lines that are not applicable to unorganized point clouds. Elhinney *et al.* [9] presented a road edge detection method, where the cross section is modeled as a 2-D cubic spline, and the road edges are extracted by detecting peaks and troughs from that fitted spline. For extracting buildings and trees, Yang *et al.* [10] proposed a method to generate feature imagery from point clouds. The generated feature image provides an alternative solution for extracting road markings [11] and facade footprints [12]. Yang *et al.* [13] proposed a marked point process-based method to extract building outlines from airborne laser scanning point clouds. Pu *et al.* [14] proposed a method to recognize basic structures, like poles near roads, according to their geometric attributes. Becker [15] proposed a model-driven method to extract windows based on formal grammar. Friedman and Stamos [16] detected repeated structures, such as windows, based on Fourier analysis. Some research, specifically designed for road extraction, has also been proposed [32]–[35]. Nevertheless, the use of prior knowledge makes it difficult to extend these methods to more generic object categories.

The second class of approaches is based on an object detection framework for generic categories. The methods belonging to this class can also be classified into methods based on global and partial shape description. Global shape-based methods require accurate segmentation in advance, whereas partial shape-based methods are insensitive to segmentation. Aijazi *et al.* [17] proposed a point cloud classification method in urban environments. The point cloud is first segmented by the super-voxel segmentation method. All super-voxels are then merged into segmented objects. The classification is performed based on the geometric descriptors extracted from these segments. Ning *et al.* [18] segmented the scene into clusters using a surface growing algorithm [36] and represented the clusters through primitive shape elements. Golovinskiy *et al.* [19] proposed a framework for recognizing small objects from 3-D laser scanning point clouds in urban environments. Their framework is divided into four steps: location, segmentation, characterization, and classification. The completeness of objects, which is hard to ensure because of occlusion and overlap between neighboring objects in real-world scenes, restricts the performance of these global shape-based methods. Although some point cloud segmentation algorithms have been proposed [27], [37], [38], accurate segmentation in a complex environment is still an unsettled problem. A typical part-based 3-D object detection method is proposed in [26]. Their method is based on Hough forest object detection framework, and the effectiveness is demonstrated through 3-D laser scanning point clouds of real-world urban scenes.

B. Studies on 3-D Local Patch Extraction

One type of patch extraction method [18], [27] represents the objects through primitive shape elements and then merges the elements based on topological connectivity. This type of method cannot control the size of each element and, thus, is not robust to occlusion when used for detection. Another type of method [26] represents the whole point cloud through octree partition [39], and extracts local parts according to the leaves of octree. The number of local parts is related to the size of whole point cloud and the size of octree leaves. However, this type of method does not consider the interior shape structures. Supervoxel, an analogue of superpixel [40]–[42] that is widely used in 2-D image applications, groups the 3-D points into perceptually meaningful clusters with high efficiency. Points within each supervoxel must have similar features and be spatially connective. To speed up processing, supervoxels, instead of the original points, are usually treated as the basic processing units, which is very suitable for laser scanning point cloud applications. VCCS is a novel supervoxel algorithm [28]. Points within each supervoxel have similar feature appearances, such as normal and fast point feature histogram (FPFH) [43]. A point cloud is segmented into individual supervoxels of similar size according to the constraint that each supervoxel cannot flow

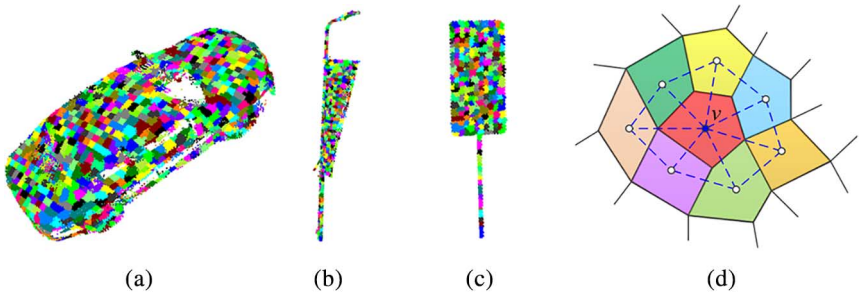


Fig. 1. Point clouds over-segmentation examples for: (a) a car; (b) a street lamp; and (c) a traffic sign based on VCCS. In (d), all the supervoxels (represented by colored surfaces) within the first-order neighborhood of a supervoxel (red surface) constitute a local patch. The white and blue points are the centers of supervoxels, and an adjacent graph is constructed by the connected dashed blue lines.

across the object boundaries. However, this constraint degrades the distinctiveness of the supervoxels.

III. Test MLS Data

Our proposed method was evaluated on three datasets containing three different categorical objects: street lamp, car, and traffic sign. All of these datasets were collected by the RIEGL VMX-450 system [1] (400 lines per second, 1.1 million measurements per second, and 8 mm accuracy) in Xiamen, China.

A. Street Lamps

The dataset, used to evaluate the performance of street lamp detection, covers about 188 150 m^2 and contains about 480 million points. The intersection point of lamp pole and lamp header is considered as the object center. Within the test scene, containing 183 street lamps with various azimuth rotations, only 159 street lamps were completely segmented from the scene. Due to their overlap with other objects in the scene, the other 24 street lamps failed to be segmented.

B. Traffic Signs

The dataset, used to evaluate traffic sign detection, covers a distance of about 10 km along the surveyed road and contains about 24 million points selected from the raw point clouds. This dataset contains 73 traffic signs with various azimuth rotations and sizes, of which 38 traffic signs were completely segmented from the scene and 35 traffic signs failed to be segmented.

C. Cars

The dataset, used to evaluate car detection, covers about 8700 m^2 and contains about 61 million points. The dataset contains 134 cars with various azimuth rotations, of which 125 cars were successfully segmented from the scene and the other nine cars were failed to be segmented. Moreover, nearly half of the cars are seriously occluded during scanning.

IV. Object Detection With Supervoxel Neighborhood-Based Hough Forest

In this section, we first introduce our local patch extraction strategy based on supervoxel neighborhood. Then, we introduce the Laplace–Beltrami scale space (LBSS) theory [44], [45] to demonstrate our strategy's superior distinctiveness. We introduce the LRF definition and the appearance of each local patch in Sections IV-C and IV-D, respectively. The training and detection procedure of supervoxel neighborhood-based Hough forest is introduced in Section IV-E.

A. 3-D Local Patch Extraction Based on Supervoxel Neighborhood

In this paper, we define the 3-D local patch as a cluster that contains a supervoxel and its neighborhood. Given a point cloud, we start the extraction of 3-D local patches by over-segmenting the point cloud into supervoxels through the VCCS algorithm. Fig. 1(a)–(c) shows three examples of over-segmented supervoxels on the point clouds of real-world objects using the VCCS algorithm with the voxel resolution 0.05 m and seed resolution 0.1 m. The voxel resolution is used to construct the voxel-cloud space, and the seed resolution is used to select the initial seed points of supervoxel. Then, we construct an adjacency graph G for all supervoxels. The vertices V of the graph G are composed of supervoxel centers, and the edges exist only between directly neighboring supervoxels. For a supervoxel centered at v, we define the nth-order neighborhood $N_n(v)$ as follows:

$$N_n(v) = \{v_i \,|d(v, v_i) \leq n, \quad v_i \in V\} \tag{1}$$

where the distance $d(v, v_i)$ is defined as the minimum number of edges between two vertices v and v_i. For a supervoxel centered at v, all the adjacent supervoxels within the distance n constitute a local patch, and the center v is treated as the center of the local patch. From now on, we denote a local patch centered at v by the neighborhood $N_n(v)$. For example, $N_0(v)$ denotes that the supervoxel itself is treated as a local patch, and $N_1(v)$ denotes that the supervoxel together with its first-order neighborhood is treated as a local patch. Fig. 1(d) illustrates the construction of a local patch $N_1(v)$.

B. LBSS Theory

By comparing our approach with directly treating supervoxels as local patches, we demonstrate the effectiveness of our local patch extraction approach through the measure of distinctiveness. We use the LBSS theory [44], [45], which has been

proposed to detect interest regions in 3-D unorganized point clouds, to define the distinctiveness of a local patch. LBSS measures the distinctiveness of a point at different scales and selects the maximum along the scale dimension of a suitable saliency function.

In this paper, the distinctiveness of a local patch $N_n(v)$ is defined as follows:

$$\rho(p,t) = \frac{2\left\|p - A(p,t)\right\|}{t} e^{-\frac{2\|p-A(p,t)\|}{t}} \tag{2}$$

where $A(p,t)$ is an operator that can be interpreted as the displacement of a point along its normal $n(p)$ by a quantity proportional to the mean curvature, $C_H(p)$, as given in (3)

$$A(p,t) \approx p + C_H(p)n(p)t^2 = p + \frac{t^2}{2}\Delta_M p \tag{3}$$

where Δ_M is the Laplace-Beltrami operator, p denotes the keypoint of the local patch, and t is the current scale attached to p. For the local patch $N_n(v)$, we treat, as the keypoint p, the point in $N_0(v)$ that is nearest to the centroid of $N_0(v)$, and define the scale t as the largest Euclidean distance from the points in $N_n(v)$ to the keypoint p. The scale t is formulated as follows:

$$t = \max\{d_e(p, v_i), \quad v_i \in N_n(v)\} \tag{4}$$

where $d_e(p, v_i)$ is the Euclidean distance from the keypoint p to the point v_i in $N_n(v)$. The evaluation results demonstrate that the distinctiveness of individual supervoxel is obviously lower than supervoxel neighborhood, and we treat the first-order supervoxel neighborhoods as local patches for better time performance. Detail discussion is shown in Section VI-A.

C. Definition of LRF

To constrain the Hough voting in the detection stage, we define a LRF for each local patch. LRF, a full 3-D coordinate system, was first proposed to construct rotation-invariant feature descriptors [29]–[31] and also has been used to acquire the transformation between query and database objects [30], [31]. LRF essentially relies on the neighboring shape structure of a specific point. The three axes of the LRF are determined by performing eigenvalue decomposition on the scatter matrix of all points lying on a local surface. Usually, to improve the robustness of the LRF to occlusion and clutter, the scatter matrix is weighted by the distance. As described in [30], the sign of each axis is disambiguated by aligning the direction to the majority of the point scatters. However, the LRF is only stable on asymmetrical local surfaces. Thus, in this paper, the LRF for a symmetrical local surface is defined as a zero matrix.

Let $\{\lambda_1, \lambda_2, \lambda_3\}$ denote the eigenvalues of the scatter matrix in descending order of magnitude, and let

$$\varepsilon = \frac{\lambda_2}{\lambda_1} \tag{5}$$

denote the ratio between the second and the first largest eigenvalues. Then, $\varepsilon = 1$ for symmetrical surfaces, and $\varepsilon < 1$ for asymmetrical surfaces. To cater for noise, we choose a threshold $0 < \varepsilon_0 < 1$ to proceed with the definition of LRF. Thus, for each local patch, we define LRF as follows:

$$LRF = \begin{cases} \{\tilde{v}_{\mathbf{1}}, \tilde{v}_{\mathbf{3}} \times \tilde{v}_{\mathbf{1}}, \tilde{v}_{\mathbf{3}}\}^T, & \text{if } \varepsilon < \varepsilon_{\mathbf{0}} \\ \mathbf{0}, & \text{if } \varepsilon > \varepsilon_{\mathbf{0}} \end{cases} \tag{6}$$

where $\tilde{v}_{\mathbf{1}}$ and $\tilde{v}_{\mathbf{3}}$ are two unique unambiguous orthogonal eigenvectors corresponding to the maximal and minimal eigenvalues. In our experiments, the threshold ε_0 is set to be 0.9.

D. Appearance of 3-D Local Patch

The local patch is described by both structure and reflectance features. The structure features are as follows: spectral features [46], eigenvalues of the covariance matrix, 3-D invariant moments, and FPFH [43]. The spectral features describe the local patch topology by assigning a saliency describing the degree of scatter, linearity, and planarity. By defining $\lambda_1 \geq \lambda_2 \geq \lambda_3$ to be the eigenvalues of the scatter matrix M defined over a local patch, the saliency of scatter, linearity, and planarity are measured by $\{\sigma_s = \lambda_3, \sigma_l = \lambda_1 - \lambda_2, \sigma_p = \lambda_2 - \lambda_3\}$. The eigenvalues of the covariance matrix describe the extent of local patch spanning along three directions. The 3-D moments are measures of the spatial distribution of a mass of 3-D points. The 3-D invariant moments are moments invariant to translation and rotation transformation. FPFH are pose-invariant features, which describe the local surface properties of points using combinations of their k nearest neighbors. Besides spatial information, laser scanning systems also capture reflectance information of objects. The reflectance feature is the median of reflectance intensities. Works based on both spatial and spectral information have been discussed [47]–[50]. In addition, we also use other features, such as the height of the local patch center relative to the lowest point in the point cloud and the occupied area of the local patch in the horizontal plane.

The appearance of each local patch is composed of four components

$$P = (I, F, l, d) \tag{7}$$

where I is the feature description described above, F is the LRF of the local patch, l is the class label with one for positive samples and zero for negative samples, and d is the offset vector, which goes from the object center to the patch center. Negative samples have a pseudo offset, i.e., $d = 0$.

E. Object Detection With Supervoxel Neighborhood-Based Hough Forest

The Hough forest algorithm is divided into two stages: training and detection. In our proposed method, the training stage starts with over-segmenting the point cloud into individual supervoxels through the VCCS method. Each supervoxel, together with its first-order neighborhood, constitutes a local patch. The appearance of the ith local patch P_i is composed of four components: $\{P_i = (I_i, F_i, l_i, d_i)\}$. Based on these local appearances, the optimal parameters of the split function on

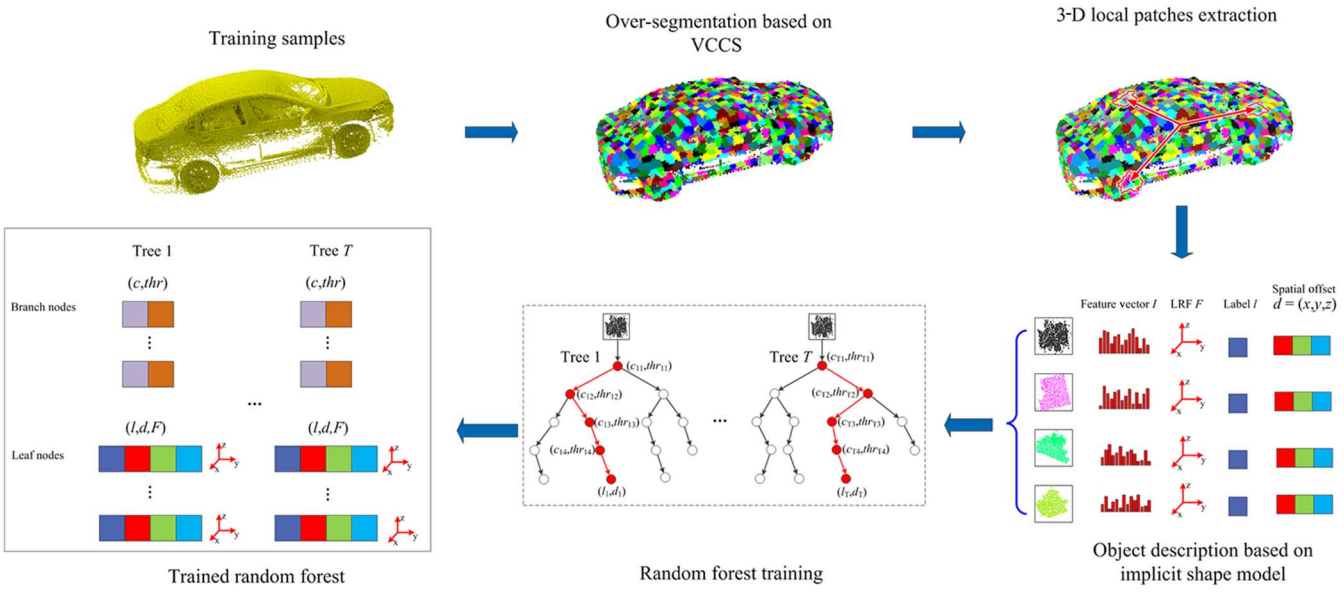


Fig. 2. Training procedure of supervoxel-based Hough forest algorithm. T is the number of trees in the random forest.

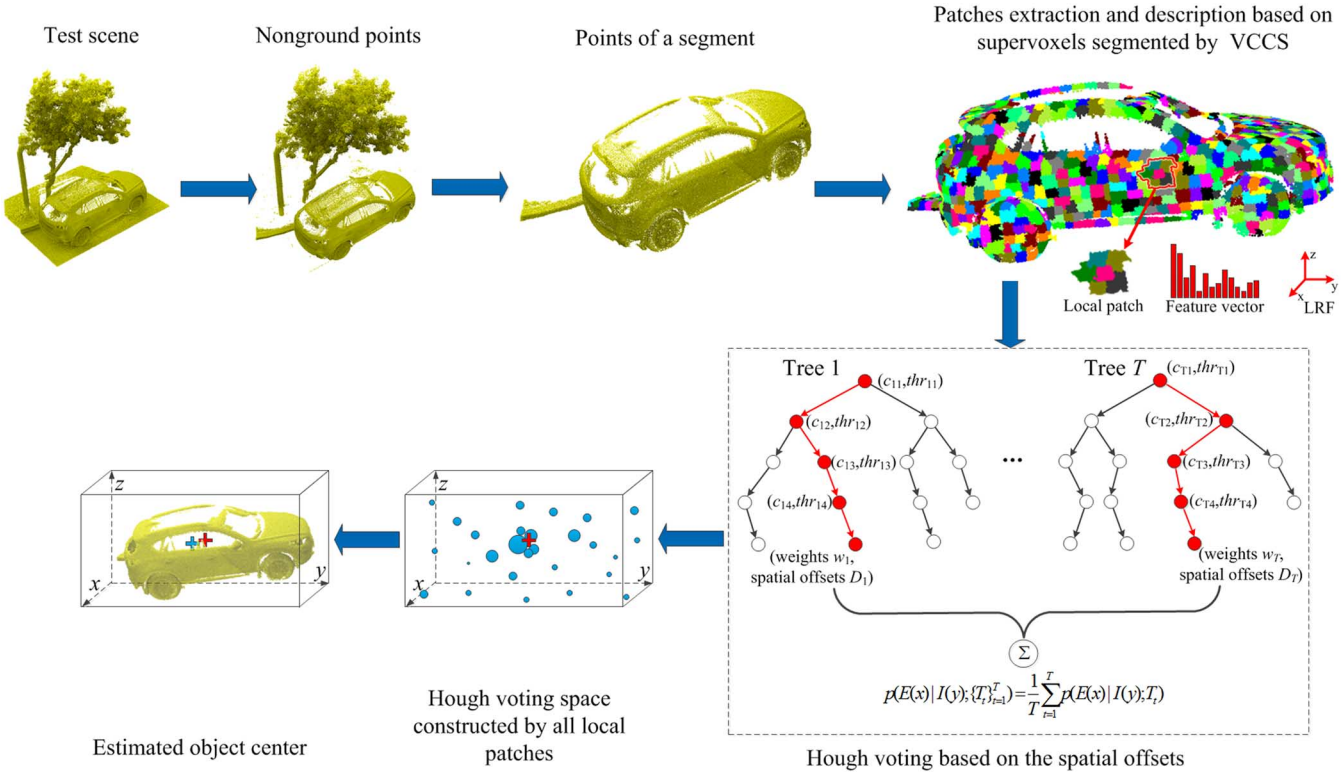


Fig. 3. Detection procedure of supervoxel-based Hough forest algorithm. The red cross represents the real object center, and the light blue cross represents the estimated object center.

each branch node are determined [20]. To meet the requirement of rotation invariance, instead of using the appearances at two selected different positions within a local patch as described in [20], we use the entire local patch's appearance to learn the split function. Afterward, according to the split function, the training patches that reach a branch node are split into two subsets. The aforementioned splitting step is repeated until the depth of the node reaches a maximum or the number of samples is smaller than a given threshold. Each branch node of the trained trees stores the selected feature channel and the corresponding feature threshold. Each leaf node stores the proportion, the offset vectors, and the LRF of the positive training patches that reach this node in the training stage. Fig. 2 shows the complete training procedure of our algorithm.

Fig. 3 shows the complete object detection procedure. First, the ground points are removed from the test scene [51]. By gradually increasing the window size of the filter and using elevation difference thresholds, the point clouds of nonground

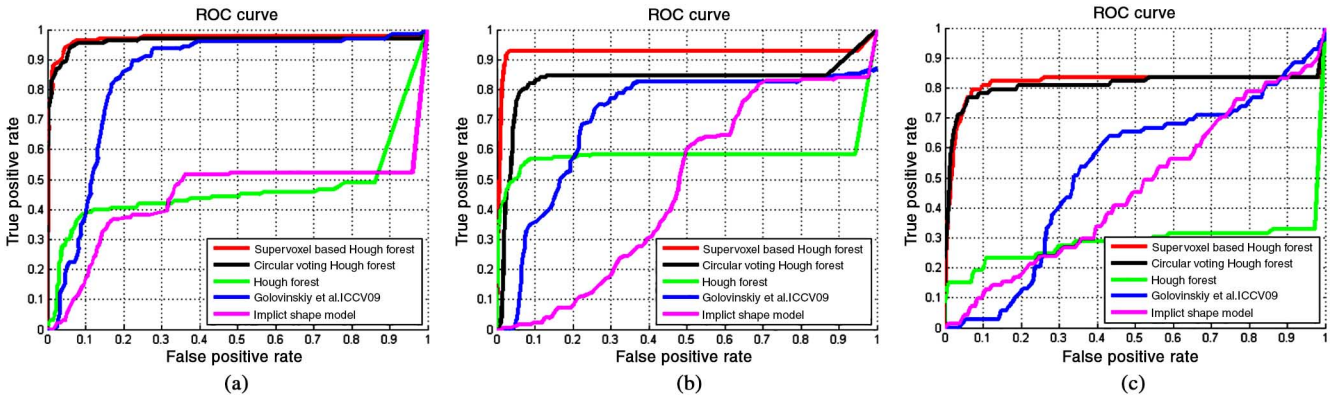


Fig. 4. Proposed method (red curve) is compared with the original Hough forest (green curve) method, Golovinskiy's method (blue curve), our previous method (black curve), and implicit shape model-based method (magenta curve) on three different categorical objects. Comparison of three methods for: (a) car; (b) street lamps; and (c) traffic signs detection.

objects such as vehicles, vegetation, and buildings are preserved, and ground points are removed. Next, a segmentation method is used to partition the scene into individual segments. The segmentation approach is based on a nearest neighborhood graph [19], and the segmentation error is minimized by the min-cut algorithm. The local patches are extracted and described, based on the method used in the training stage, and then passed downward through the trained trees to a leaf node in each tree according to the information stored in the branch nodes. The spatial offsets stored in the leaf nodes are used to cast votes to the object center. Finally, all votes create a 3-D Hough voting space, and the object center is determined by a traditional nonmaximum suppression process.

However, because of rotations that exist between the training samples and the test samples, the spatial offset vectors stored in the leaf nodes cannot be directly used to vote for the object center. In urban environments, the objects of interest such as cars, traffic signs, and street lamps usually have azimuth rotations. To achieve the invariance to the azimuth rotation, we use different voting strategies in the Hough voting stage. If both of the matched training and testing local patches are asymmetrical, the rotation transformation is estimated by aligning the LRF of the training patch to the LRF of the matched testing patch. Specifically, given a patch v_s and the corresponding LRF F_s centered at p_s in the test sample, we pass v_s down through the trained trees and reach one leaf node containing a patch v_m with offset vector d_m and LRF F_m. The rotation transformation between v_s and v_m is estimated by

$$R = F_s^T F_m. \tag{8}$$

The correct offset vector d_s starting from the object center to the patch center p_s for the patch v_s is estimated as

$$d_s = d_m R. \tag{9}$$

Then, the object center is estimated as

$$o = p_s - d_s = p_s - d_m(F_s^T F_m). \tag{10}$$

If one of the matched training and testing local patches is symmetrical, the estimated rotation matrix R equals to zero. To achieve invariance to the azimuth rotation, we adopt a circular voting strategy. We rotate the offset vector for all orientations in the azimuth direction. All positions with a certain distance d_h to the local patch center p in the horizontal plane and d_z to p in the vertical direction are potential positions of the object center. By using the circular voting approach, we achieve rotation invariance in the azimuth direction. For a patch centered at $\{p_x, p_y, p_z\}$, the object center is estimated by

$$\begin{cases} o_x = p_x + d_h \cos(\theta) \\ o_y = p_y + d_h \sin(\theta) \\ o_z = p_z - d_z \end{cases}. \tag{11}$$

The circular voting approach is computationally efficient because the operation of rotating offset vectors is implemented by defining a discrete lookup table.

V. Experimental Results and Comparison

Our proposed method was evaluated on three datasets described in Section III. All the category-specific random forests are trained based on the samples with various azimuth rotations. To evaluate the performance of our object detection algorithm, we manually labeled the target objects in all training and testing point clouds as the ground truth. A detection is marked as a true positive only if the estimated center falls into certain distance thresholds relative to the labeled object center in both horizontal and vertical directions. Each target object matches only one detection. When there are multiple detections for an object, only the one closest to the labeled center is labeled as true detection and the others are labeled as false positives. The detection performance is shown by the ROC curve.

We compared our algorithm with the original Hough forest, our previous work [26], the method proposed in [19], and the method based on ISM [25]. The performance of different methods including ours is shown in Fig. 4. As seen in Fig. 4, our method significantly outperforms the state-of-the-art. Golovinskiy's method [19] is based on an object's global shape features, and consequently the performance seriously

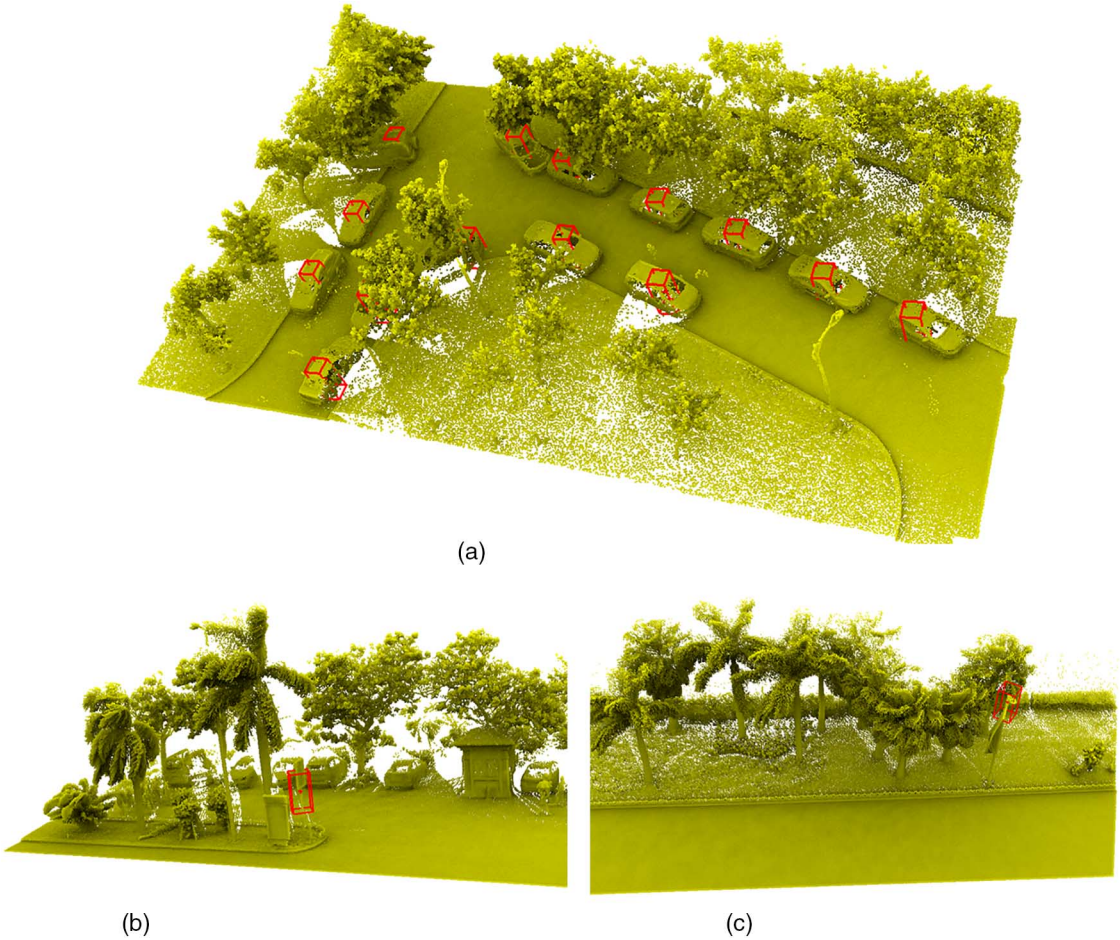


Fig. 5. Results of the proposed detection algorithm. Red 3-D bounding boxes represent the true correct detection results. (a) Car detection result. (b) Traffic sign detection result. (c) Street lamp detection result.

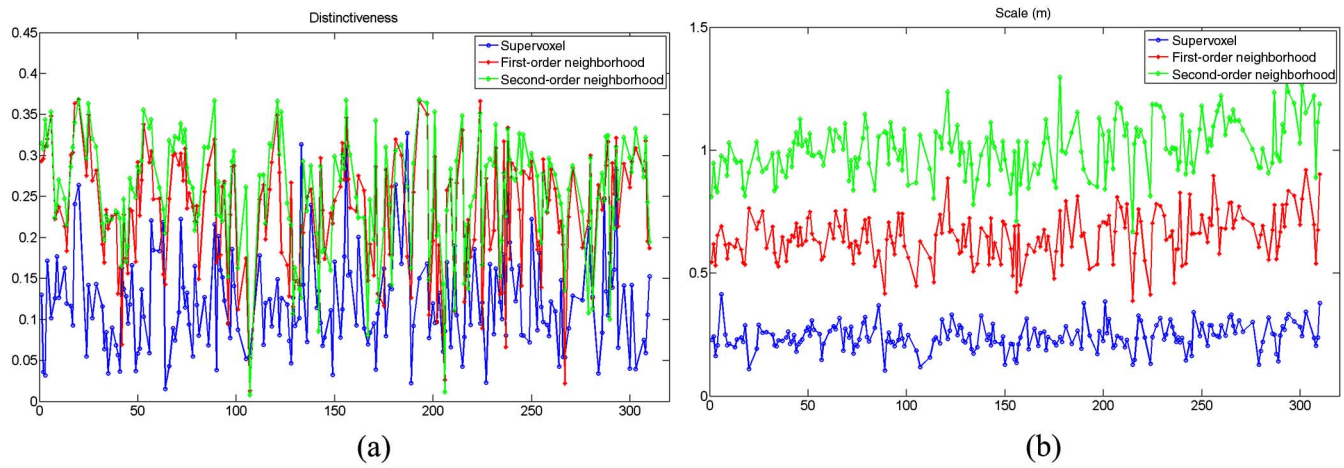


Fig. 6. Distinctiveness evaluation based on LBSS theory. (a) Shows the distinctiveness of the local patches extracted from different neighborhood sizes, and the Y-axis represents the distinctiveness of the local patch; (b) shows the corresponding scales, and the Y-axis represents the size of the local patch. In both figures, the X-axis represents the labels of the supervoxels.

depends on the completeness of the objects. Therefore, [19] is vulnerable to scenes where objects cannot be segmented from the background due to the overlap between neighboring objects or where objects are seriously occluded when scanning. On the contrary, our method is based on object part appearance and thus is robust to occlusion and overlap. The method based on ISM [25] treats the centroid as the object center in the training stage, which is not stable for the point density variation. In [25], the rotation invariance is achieved by aligning the offset vector according to the normal of the

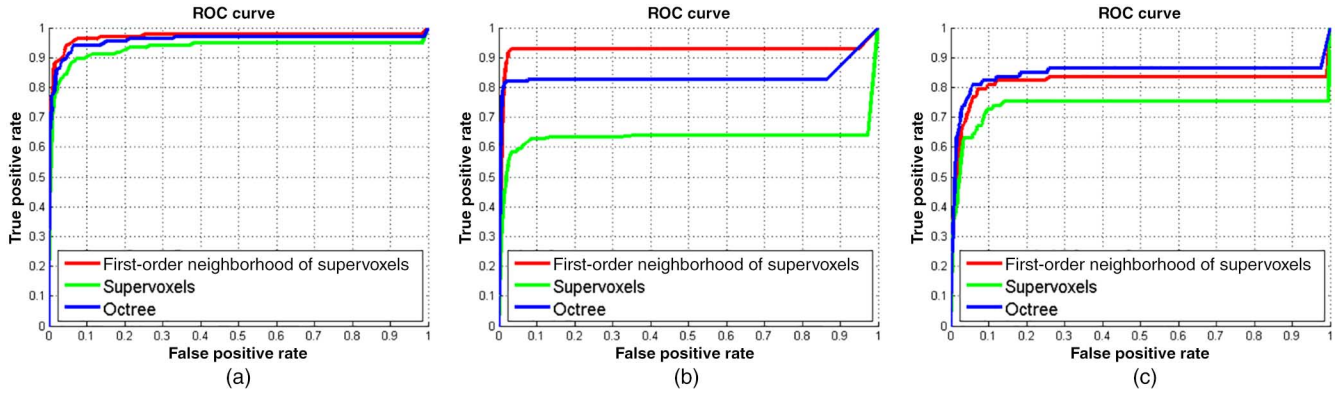


Fig. 7. Detection performance on three object categories: (a) car; (b) street lamp; and (c) traffic sign using different local patch extraction approaches.

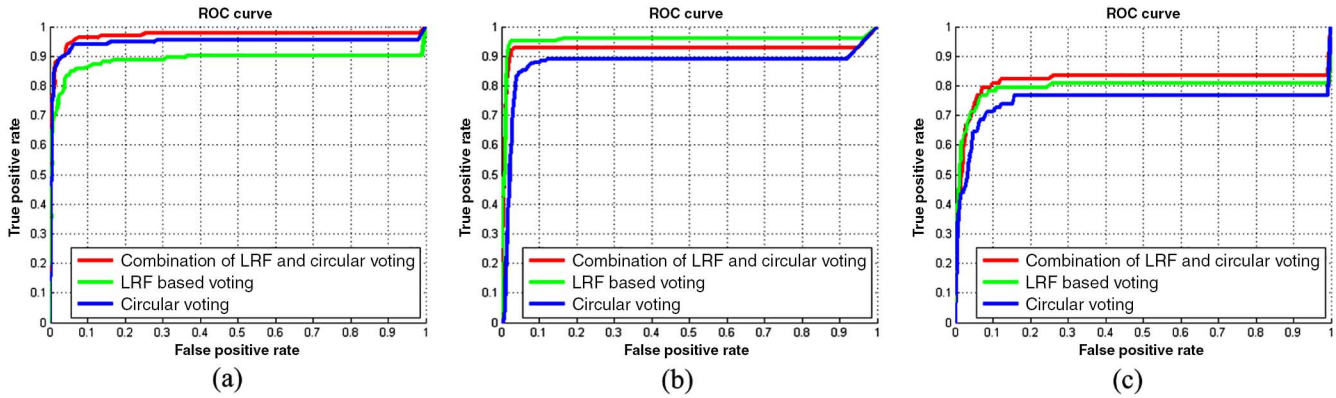


Fig. 8. Performances of the detection for: (a) car; (b) street lamp; and (c) traffic sign using combination of LRF and circular voting, LRF-based voting, and circular voting strategies.

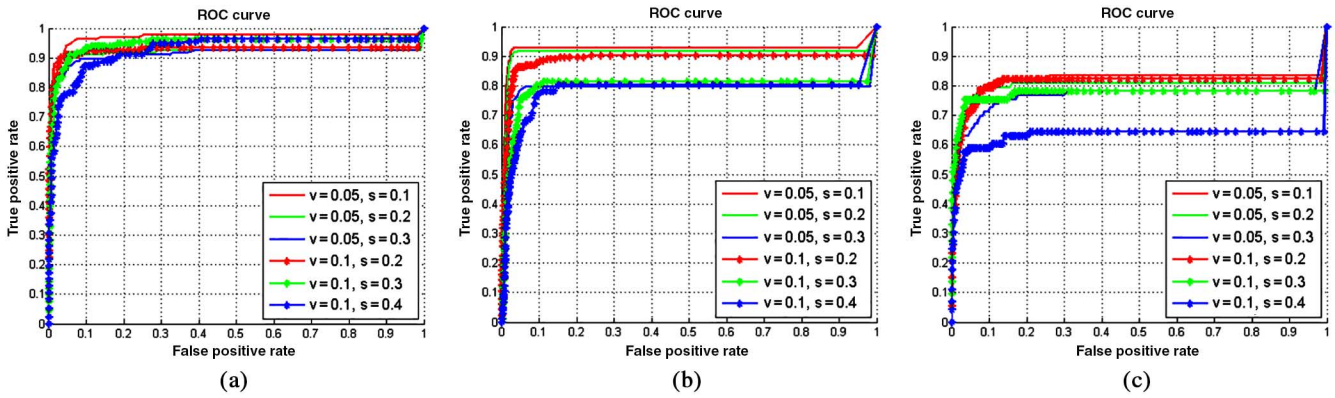


Fig. 9. Performances of the detection for: (a) car; (b) street lamp; and (c) traffic sign under different supervoxel resolutions.

keypoints. However, the normal is easily disturbed by the noise, occlusion, and overlap. Moreover, the LRF used in our method is constrained by both the maximal and minimal principal components of local covariance matrix, whereas the normal is constrained by only the minimal principal component of local covariance matrix. Compared to our previous work [26], the proposed method also achieves better performance. From the results, we conclude that our method has the ability to cope with overlap, occlusion, and rotation in cluttered real-world scenes. Fig. 5 shows the detection results for three different categorical objects.

VI. Discussion

In this section, we first evaluate the local patch distinctiveness under different sizes of supervoxel neighborhood based on the LBSS theory. The sensitivities of the detection performance under different local patch extraction strategies and different

Hough voting strategies are evaluated in Sections VI-B and VI-C. The sensitivity of our method to supervoxel resolutions is analyzed in Section VI-D. Finally, we analyze the feature importance and time performance of our method in Sections VI-E and VI-F, respectively.

A. Local Patch Distinctiveness Evaluation Based on LBSS theory

We evaluate the distinctiveness of our method with experiments conducted on the point clouds shown in Fig. 1(a). The voxel resolution is set to be 0.1 m, and the seed resolution is set to be 0.3 m. In Fig. 6(a), the distinctiveness evaluation results are shown when n is equal to 0, 1, and 2. The corresponding scales are shown in Fig. 6(b). From Fig. 6(a), we observe that the distinctiveness of individual supervoxel (blue curve) is obviously lower than the supervoxel neighborhood (red and green curves). Distinctiveness is lowered because a supervoxel cannot flow across an object's boundaries, whereas a supervoxel neighborhood can cross an object's boundaries. Comparing the red curve to the green curve in Fig. 6(a), we observe that increasing neighborhood size does not obviously improve distinctiveness. Thus, in this paper, we consider only the first-order supervoxel neighborhood for extracting local patches for better time performance.

B. Effectiveness of the Proposed Local Patch Extraction Approach

To demonstrate the effectiveness of our proposed local patch extraction approach, we compared it with the method that directly treats a supervoxel as a local patch and the method based on octree [26]. All the experiments were performed with the same Hough voting method, a combination of LRF and circular voting. From Fig. 7, we observe that the proposed local patch extraction approach (red curves) outperforms the approach that directly treats supervoxels as local patches (green curves) on all three categorical objects. The reason for this superior performance is that a supervoxel neighborhood can cross the object boundaries, whereas a supervoxel cannot flow across the object boundaries. Using the supervoxel neighborhood as a local patch, we maintain the dominant local structures and substantially improve the discriminant ability. Compared to the octree-based local patch extraction approach (blue curves), we observe that our supervoxel neighborhood-based approach achieves state-of-the-art detection performance.

C. Sensitivity Analysis of the Hough Voting Strategy

We also conducted experiments to test the proposed method under different Hough voting strategies. Specifically, we tested our method under the LRF-based voting, circular voting, and combination of LRF and circular voting strategies while keeping all other conditions the same. The LRF-based voting strategy achieves rotation invariance according to the actual rotation transformations, which rely on the LRFs associated with each pair of matched training and testing asymmetrical local patches. The circular voting strategy, because it also votes some irrelevant locations, achieves rotation invariance at the cost of a low false positive rate. From Fig. 8, we observe that the performance of the combination of LRF and circular voting strategy outperforms the other voting strategies on car and traffic sign categories. In Fig. 8(b), the performance of the LRF-based voting is slightly better than the combination of the LRF and circular voting strategy for the street lamp category. The reason for the better performance is because most of the extracted local patches of street lamps are asymmetrical, and the votes are mainly based on the LRF-based voting. Considering the comprehensive performance and the generality of the method, we conclude that the combination of LRF and circular voting has the best detection performance.

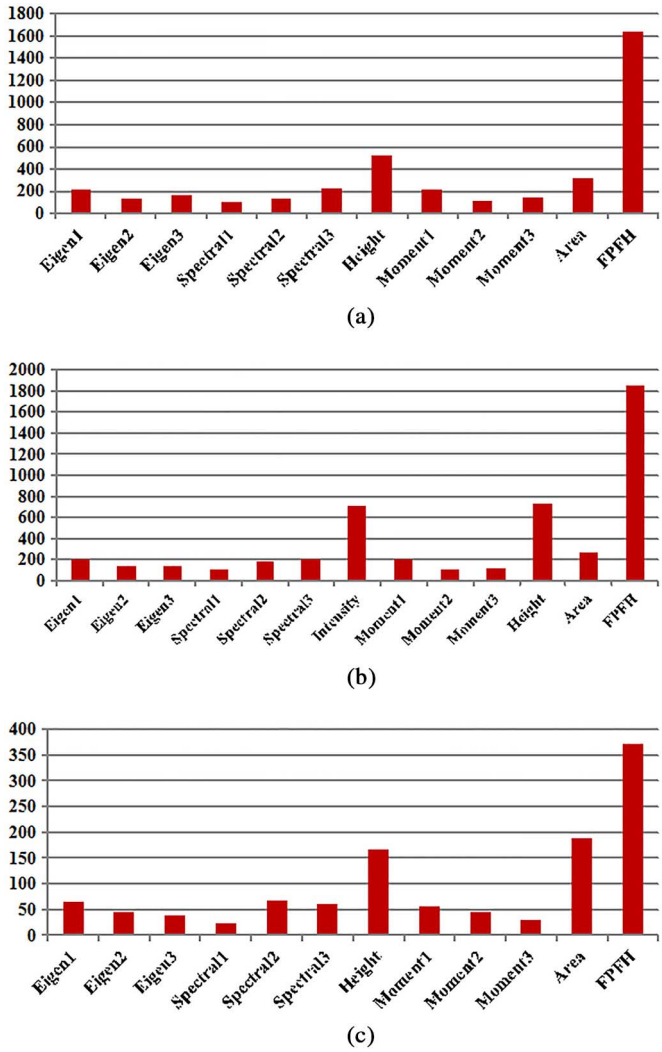


Fig. 10. Feature analysis for: (a) car; (b) street lamp; and (c) traffic sign detection.

D. Sensitivity Analysis of Supervoxel Resolutions

The sensitivity of our proposed method to the supervoxel resolutions is analyzed in this section. We conducted six groups of experiments with different voxel resolution and seed resolution for each object category. Fig. 9 shows the detection performance under different supervoxel resolutions for car, street

TABLE I
RUNNING TIME OF TRAINING AND DETECTION

	Training points (mil.)	Training time (s)	Detection points (mil.)	Ground points filtering time (s)	Nonground points segmentation time (s)	Detection time (s)
Street lamp	0.84	540	480	2022	432	1188
Car	1.26	720	61	720	194	3636
Traffic sign	0.14	61	24	690	144	670

lamp, and traffic sign. From Fig. 9, we observe that when voxel resolution is set to be 0.05 m and seed resolution is set to be 0.1 m, our method achieves best performance for all the three categorical objects. The reason is that the robustness of our method to occlusion and overlap is degraded when increasing the size of local patch, although the distinctiveness of local patch is promoted. Thus, in all of our experiments, the voxel resolution is set to be 0.05 m and the seed resolution is set to be 0.1 m.

E. Feature Analysis

To evaluate which feature is important for object detection from 3-D laser scanning point clouds, we counted the number of times that the feature is selected to segment the feature space in the training stage. According to the splitting principle of random forest classifier, only one feature channel is selected optimally from the feature vector at each segmentation step. Fig. 10 shows the feature analysis histograms for the three categorical objects. The horizontal axis represents the features, and the vertical axis represents the number of times the corresponding feature is chosen for training the random forest classifier. Because of color variation, when detecting cars and traffic signs, the reflectance intensity is not used. From Fig. 10, we observe that all the features have made their contributions. As seen from Fig. 10(a) and (c), the height, area, and FPFH play important roles for distinguishing cars and traffic signs from other objects. For street lamp detection, intensity, height, area, and FPFH play important roles. In conclusion, the feature analysis result is consistent with our knowledge of distinguishing objects from others.

F. Time Performance

All experiments were conducted on a machine with an Intel Core i3 3.3 GHz processor and a 16-GB RAM. The running time of training and detection on these three categorical objects is presented in Table I. For the whole detection stage, we list the time spending on its three main stages: ground points filtering, nonground points segmentation, and object detection. We also compared the object detection time of our proposed method with our previous work [26]. From Table II, we observe that, by extracting local patches based on supervoxels, our method achieves higher computational efficiency. The reason for the higher computational efficiency of supervoxel-based method is that we treated each supervoxel as the basic processing unit, whereas in our previous work, we treat each leaf of the octree as the basic processing unit.

TABLE II
DETECTION TIME COMPARISON

	Supervoxel based Hough forest (s)	Circular voting Hough forest (s)
Car	**3636**	8028
Street lamp	**1188**	4032
Traffic sign	**670**	1332

VII. CONCLUSION

In this paper, we proposed a novel supervoxel neighborhood-based method for object detection from laser scanning point clouds of complex urban environments. The main contributions of this paper include: firstly, we incorporate the supervoxel into the Hough forest object detection framework to accelerate the computational efficiency; secondly, we propose to treat a supervoxel neighborhood as a 3-D local patch, which is proved to have a higher distinctiveness than an individual supervoxel; and finally, we solve the limits of Hough forest framework for dealing with rotated objects through the combination of LRF and circular voting. Our method was tested on three different categorical real-world objects with high computational efficiency and achieves higher performances compared with state-of-the-art 3-D object detection methods. Moreover, the evaluation and comparison results, under different local patch extraction strategies and different Hough voting strategies, also experimentally verify the superiority of our method. Overall, our method achieves improvements with high computational efficiency over the existing 3-D object detection methods in real-world 3-D laser scanning point clouds.

The limitation of the proposed method is mainly caused by the difficulty of extracting supervoxels from point clouds of nonsolid-surface structures such as tree canopy. Thus, our method is not suitable for detecting objects such as trees with nonsolid-surface structures.

REFERENCES

[1] Riegl. *Riegl VMX450* [Online]. Available: http://www.riegl.com/nc/products/mobile-scanning/produktdetail/product/scannersystem/10/, accessed on 2014.
[2] Trimble. *Trimble MX8* [Online]. Available: http://www.trimble.com/imaging/Trimble-MX8.aspx, accessed on 2014.
[3] Optech. *Optech Lynx* [Online]. Available: http://www.optech.com/index.php/products/mobile-survey/, accessed on 2014.
[4] A. Jaakkola, J. Hyyppä, H. Hyyppä, and A. Kukko, "Retrieval algorithms for road surface modelling using laser-based mobile mapping," *Sensors*, vol. 8, pp. 5238–5249, 2008.

[5] H. Guan *et al.*, "Using mobile laser scanning data for automated extraction of road markings," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 93–107, 2014.

[6] Y. Yu, J. Li, H. Guan, C. Wang, and J. Yu, "Automated detection of road manhole and sewer well covers from mobile LiDAR point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 9, pp. 1549–1553, Sep. 2014.

[7] C. Brenner, "Extraction of features from mobile laser scanning data for future driver assistance systems," *Advances in GIScience*, 2009, pp. 25–42.

[8] A. J. M. Lehtomäki, J. Hyyppä, A. Kukko, and H. Kaartinen, "Detection of vertical pole-like objects in a road environment using vehicle-based laser scanning data," *Remote Sens.*, vol. 2, pp. 641–664, 2010.

[9] C. Mc Elhinney, P. Kumar, C. Cahalane, and T. McCarthy, "Initial results from European Road Safety Inspection (EURSI) mobile mapping project," in *Proc. ISPRS Commission V Tech. Symp.*, 2010, pp. 440–445.

[10] B. Yang, Z. Wei, Q. Li, and J. Li, "Automated extraction of street-scene objects from mobile lidar point clouds," *Int. J. Remote Sens.*, vol. 33, pp. 5839–5861, 2012.

[11] B. Yang, L. Fang, Q. Li, and J. Li, "Automated extraction of road markings from mobile lidar point clouds," *Photogramm. Eng. Remote Sens.*, vol. 78, pp. 331–338, 2012.

[12] B. Yang, Z. Wei, and J. Li, "Semi-automated building facade footprint extraction from mobile lidar point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 766–770, Jul. 2013.

[13] B. Yang, W. Xu, and Z. Dong, "Automated extraction of building outlines from airborne laser scanning point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1399–1403, Nov. 2013.

[14] S. Pu, M. Rutzinger, G. Vosselman, and S. O. Elberink, "Recognizing basic structures from mobile laser scanning data for road inventory studies," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, pp. S28–S39, 2011.

[15] S. Becker, "Generation and application of rules for quality dependent façade reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, pp. 640–653, 2009.

[16] S. Friedman and I. Stamos, "Online detection of repeated structures in point clouds of urban scenes for compression and registration," *Int. J. Comput. Vision*, vol. 102, pp. 112–128, 2013.

[17] A. K. Aijazi, P. Checchin, and L. Trassoudaine, "Segmentation based classification of 3D urban point clouds: A super-voxel based approach with evaluation," *Remote Sens.*, vol. 5, pp. 1624–1650, 2013.

[18] X. Ning, Y. Wang, and X. Zhang, "Object shape classification and scene shape representation for three-dimensional laser scanned outdoor data," *Opt. Eng.*, vol. 52, p. 024301, 2013.

[19] A. Golovinskiy, V. G. Kim, and T. Funkhouser, "Shape-based recognition of 3D point clouds in urban environments," in *Proc. IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 2154–2161.

[20] J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 1022–1029.

[21] Z. Lei, T. Fang, H. Huo, and D. Li, "Rotation-invariant object detection of remotely sensed images based on texton forest and Hough voting," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1206–1217, Apr. 2012.

[22] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vision*, vol. 77, pp. 259–289, 2008.

[23] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool, "Hough transform and 3D SURF for robust three dimensional classification," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 589–602.

[24] J. Knopp, M. Prasad, and L. V. Gool, "Scene cut: Class-specific object detection and segmentation in 3D scenes," in *Proc. Int. Conf. 3D Imag. Model. Process. Vis. Transmiss. (3DIMPVT'11)*, 2011, pp. 180–187.

[25] A. Velizhev, R. Shapovalov, and K. Schindler, "Implicit shape models for object detection in 3D point clouds," in *Proc. Int. Soc. Photogramm. Remote Sens. Congr.*, 2012, pp. 179–184.

[26] H. Wang *et al.*, "Object detection in terrestrial laser scanning point clouds based on Hough forest," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1807–1811, Oct. 2014.

[27] B. Yang and Z. Dong, "A shape-based segmentation method for mobile laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 81, pp. 19–30, 2013.

[28] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, "Voxel cloud connectivity segmentation-supervoxels for point clouds," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 2027–2034.

[29] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 356–369.

[30] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. Wan, "Rotational projection statistics for 3D local surface description and object recognition," *Int. J. Comput. Vision*, vol. 105, pp. 63–86, 2013.

[31] A. Mian, M. Bennamoun, and R. Owens, "On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes," *Int. J. Comput. Vision*, vol. 89, pp. 348–361, 2010.

[32] A. Boyko and T. Funkhouser, "Extracting roads from dense point clouds in large scale urban environment," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, pp. S2–S12, 2011.

[33] B. Yang, L. Fang, and J. Li, "Semi-automated extraction and delineation of 3D roads of street scene from mobile laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 79, pp. 80–93, 2013.

[34] J. Lam, K. Kusevic, R. Mrstik, P. Harrap, and M. Greenspan, "Urban scene extraction from mobile ground based LiDAR data," in *Proc. Int. Symp. 3D Data Process. Vis. Transmiss. (3DPVT)*, 2011, pp. 478–486.

[35] H. Wang *et al.*, "Automatic road extraction from mobile laser scanning data," in *Proc. Int. Conf. Comput. Vision Remote Sens. (CVRS'12)*, 2012, pp. 136–139.

[36] X. Ning, X. Zhang, and Y. Wang, "Automatic architecture model generation based on object hierarchy," in *Proc. ACM SIGGRAPH ASIA Sketches*, 2010, p. 39.

[37] A. G. T. Funkhouser, "Min-cut based segmentation of point clouds," in *Proc. Int. Conf. Workshop Comput. Vision*, 2009, pp. 39–46.

[38] M. Vieira and K. Shimada, "Surface mesh segmentation and smooth surface extraction through region growing," *Comput. Aided Geom. Des.*, vol. 22, pp. 771–792, 2005.

[39] W. T. Wong, F. Y. Shih, and T. F. Su, "Thinning algorithms based on quadtree and octree representations," *Inf. Sci.*, vol. 176, pp. 1379–1394, May 2006.

[40] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[41] L. Ming-Yu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constraint," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 99–112, Jan. 2014.

[42] J. Wang and X. Wang, "VCells: Simple and efficient superpixels using edge-weighted centroidal Voronoi tessellations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1241–1247, Jun. 2012.

[43] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA'09)*, 2009, pp. 3212–3217.

[44] R. Unnikrishnan and M. Hebert, "Multi-scale interest regions from unorganized point clouds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. Workshops (CVPRW'08)*, 2008, pp. 1–8.

[45] F. Tombari, S. Salti, and L. Di Stefano, "Performance evaluation of 3D keypoint detectors," *Int. J. Comput. Vision*, vol. 102, pp. 198–220, 2013.

[46] D. Munoz, J. A. Bagnell, N. Vandapel, and M. Hebert, "Contextual classification with functional max-margin Markov networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 975–982.

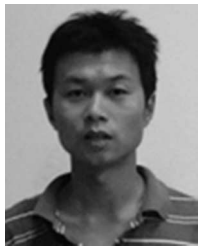
[47] A. Antonarakis, K. S. Richards, and J. Brasington, "Object-based land cover classification using airborne LiDAR," *Remote Sens. Environ.*, vol. 112, pp. 2988–2998, 2008.

[48] C. Hug, "Combined use of laser scanner geometry and reflectance data to identify surface objects," in *Proc. OEEPE Workshop '3-D City Models,'* Bonn, Germany, 1996, pp. 9–11.

[49] G. Chust, I. Galparsoro, Á. Borja, J. Franco, and A. Uriarte, "Coastal and estuarine habitat mapping, using LIDAR height and intensity and multi-spectral imagery," *Estuarine Coastal Shelf Sci.*, vol. 78, pp. 633–643, 2008.

[50] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Using the Dempster–Shafer method for the fusion of LIDAR data and multi-spectral images for building detection," *Inf. Fusion*, vol. 6, pp. 283–300, 2005.

[51] K. Zhang *et al.*, "A progressive morphological filter for removing non-ground measurements from airborne LIDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 4, pp. 872–882, Jun. 2003.



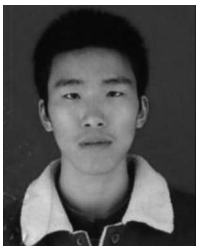
Hanyun Wang received the Master's degree in information and communication engineering from National University of Defense Technology, Changsha, China, in 2010. He is pursuing the Ph.D. degree at the School of Electronic Science and Engineering, National University of Defense Technology.

His research interests include computer vision, machine learning, pattern recognition, mobile laser scanning, and point cloud processing.

Cheng Wang (M'12) received the Ph.D. degree in information communication engineering from National University of Defense Technology, Changsha, China, in 2002.

He is a Professor and Vice Dean with the School of Information Science and Engineering, Xiamen University, Xiamen, China. He has co-authored more than 80 papers. His research interests include remote sensing image processing, mobile laser scanning data analysis, and multisensor fusion.

Dr. Wang is the Co-Chair of the ISPRS WG I/3, Council Member of the Chinese Society of Image and Graphics (CSIG), and Member of SPIE and IEEE GRSS.



Huan Luo received the B.S. degree in computer science from Nanchang University, Nanchang, China, in 2009. He is currently pursuing the Ph.D. degree at the Department of Computer Science, Xiamen University, Xiamen, China.

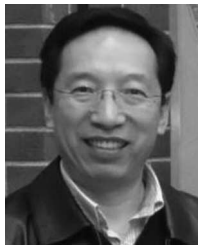
His research interests include computer vision, machine learning, mobile laser scanning, and semantic segmentation in 3-D point clouds.

Peng Li received the Ph.D. degree in electrical engineering from National University of Defense Technology, Changsha, China, in 2012.

His research interests include computer vision, remote sensing, data mining, and big data application.

Yiping Chen received the Ph.D. degree in information communication engineering from National University of Defense Technology, Changsha, China, in 2011.

She is an Assistant Professor with National University of Defense Technology, and a Postdoctoral Fellow with the School of Information Science and Engineering, Xiamen University, Xiamen, China. From 2007 to 2011, she was an Assistant Researcher with Chinese University of Hong Kong, Hong Kong, China. Her research interests include image processing, mobile laser scanning data analysis, and 3-D point cloud object detection.



Jonathan Li (M'00–SM'11) received the Ph.D. degree in geomatics engineering from the University of Cape Town, Cape Town, South Africa, in 2000.

He is currently with the Fujian Key Laboratory of Sensing and Computing For Smart Cities, School of Information Science and Engineering, Xiamen University, Xiamen, China. He is also heading the Laboratory for GeoSpatial Technology and Remote Sensing, Faculty of Environment, University of Waterloo, Waterloo, ON, Canada, where he is a Professor and an Elected Member of the University Senate. He has co-authored more than 200 publications, over 60 of which were published in refereed journals. His research interests include information extraction from earth observation images and 3-D surface reconstruction from mobile laser scanning point clouds.

Dr. Li is the Chair of the Inter-Commission Working Group I/Va on Mobile Scanning and Imaging Systems of the International Society for Photogrammetry and Remote Sensing from 2012 to 2016, the Vice Chair of the Commission on Hydrography of the International Federation of Surveyors from 2011 to 2014, and the Vice Chair of the Commission on Mapping from Remote Sensor Imagery of the International Cartographic Association from 2011 to 2015.