Object Detection in Terrestrial Laser Scanning Point Clouds Based on Hough Forest

Hanyun Wang, *Student Member, IEEE*, Cheng Wang, *Member, IEEE*, Huan Luo, Peng Li, Ming Cheng, Chenglu Wen, and Jonathan Li, *Senior Member, IEEE*

***Abstract*—This letter presents a novel rotation-invariant method for object detection from terrestrial 3-D laser scanning point clouds acquired in complex urban environments. We utilize the Implicit Shape Model to describe object categories, and extend the Hough Forest framework for object detection in 3-D point clouds. A 3-D local patch is described by structure and reflectance features and then mapped to the probabilistic vote about the possible location of the object center. Objects are detected at the peak points in the 3-D Hough voting space. To deal with the arbitrary azimuths of objects in real world, circular voting strategy is introduced by rotating the offset vector. To deal with the interference of adjacent objects, distance weighted voting is proposed. Large-scale real-world point cloud data collected by terrestrial mobile laser scanning systems are used to evaluate the performance. Experimental results demonstrate that the proposed method outperforms the state-of-the-art 3-D object detection methods.**

***Index Terms*—Hough forest, implicit shape model (ISM), object detection, point clouds, terrestrial laser scanning (TLS).**

I. Introduction

RECENT advances in terrestrial laser scanning (TLS) provide abilities to quickly collect 3-D point clouds with high density and high accuracy over large areas. Such detailed point clouds enable us to detect not only the common large structures (e.g., road and building), but also the small objects (e.g., street lamp, tree, and car) in cluttered scenes. However, class-specific object detection from cluttered laser scanning point clouds is an essential but challenging task. Object detection from point clouds is limited by the following factors: intra-class shape variation, incompleteness of object caused by occlusion, overlapping between neighboring objects, point-density variance, and orientation variance.

Nowadays, much work [1]–[6] has been proposed on extracting objects, such as buildings, doors, mailboxes, street lamps, and trees, from 3-D laser scanning point clouds in cluttered urban environments. However, most of the existing work is based on prior knowledge or invariant feature descriptors of the specified object categories, and is difficult to extend from the specific object categories to more generic object categories. The emerging demands on automatic extraction of a large number of object categories require more robust and easy-to-expand object detection methods. Aleksey *et al.* [7] proposed an extensible framework for recognizing small objects in large-scale 3-D laser scanning point clouds in urban environments. However, the performance of this method is restricted by the performance of segmentation, and the segmentation errors, i.e., under-segmentation and over-segmentation will contaminate the performance of this method. Although some point cloud segmentation algorithms are proposed in [8], [9], accurate segmentation in complex environment is still an unsettled problem.

Hough forest was proposed as a promising discriminative approach based on implicit shape model (ISM) to detect objects in cluttered images [10]. Hough forest combines the generalized Hough voting framework with the random forest classifier. Through ISM instead of explicit codebook, Hough forest describes objects and learns a direct mapping between the local appearance and its Hough vote [10]–[13]. ISM is essentially a codebook for object category and only depends on object's local patches. This attribute makes ISM robust to occlusion and overlapping which commonly exist in complex urban environment.

However, Hough forest cannot effectively handle the rotations of objects [14]. This is not an issue in the normal image applications [10] because objects such as cars and pedestrians are typically in an upright direction in the images. But in real-world 3-D scenes, the same categorical objects are commonly placed in varying azimuth directions. Therefore, for object detection in 3-D laser scanning point clouds of complex urban environment, a critical requirement is rotation invariance in azimuth direction. Many papers have discussed rotation invariance. In [14] rotation invariance was achieved by incorporating rotation into split function and rotating the offset vector according to the dominant gradient orientation. However, the dominant gradient orientations are difficult to be calculated in unorganized 3-D point clouds.

This letter presents a novel approach to object detection from 3-D point clouds based on Hough forest. The state-of-the-art Hough forest object detection framework [10] for 2-D images is extended to deal with 3-D point clouds. Compared to imagery, TLS points are in real world coordinates thus there is no effect from scale factors. In the training stage, the 3-D local patches are extracted from labeled samples and described by

Manuscript received November 22, 2013; revised February 12, 2014; accepted March 1, 2014. This work was supported by the National Science Foundation of China (Project No. 61371144). *(Corresponding author: C. Wang.)*

H. Wang is with the National University of Defense Technology, School of Electronic Science and Eigineering, Changsha 410073, Hunan, China.

C. Wang is with the School of Information Science and Engineering, Xiamen University, Computer Science, Xiamen 361005, China.

H. Luo and M. Cheng are with the School of Information Science and Engineering, Xiamen University, Computer Science, Xiamen 361005, China.

P. Li is with the National University of Defense Technology, School of Electronic Science and Engineering, Changsha 410073, China.

C. Wen is with the School of Information Science and Engineering, Xiamen University, Cognitive Science, Xiamen 361005, China.

J. Li is with the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Ministry of Education), 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.

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/LGRS.2014.2309965

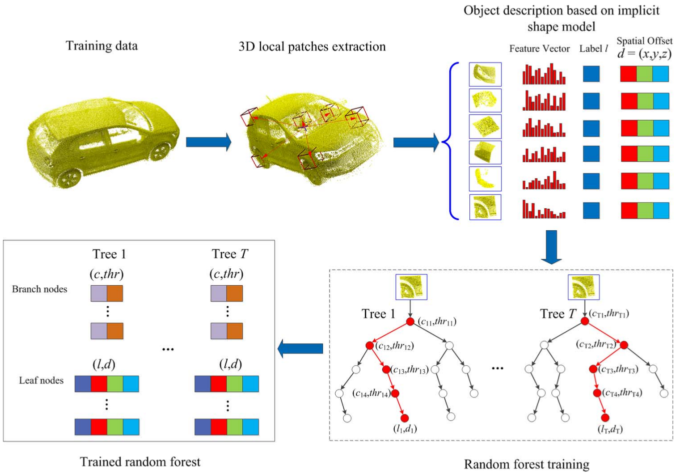


Fig. 1. Training procedure of the proposed algorithm.

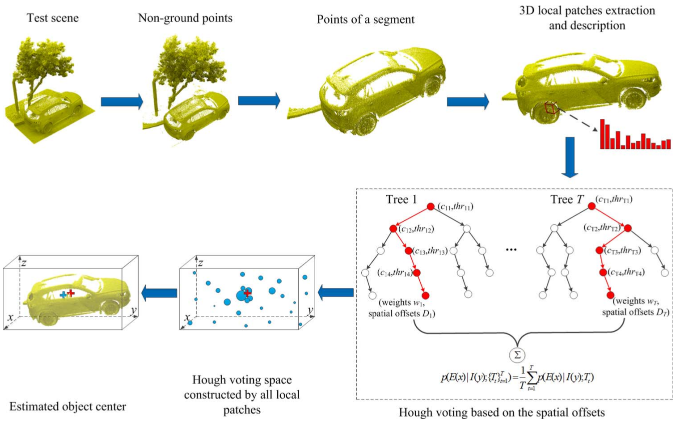

Test scene
Non-ground points
Points of a segment
3D local patches extraction and description
Tree 1
Tree T
(c_{11},thr_{11})
(c_{12},thr_{12})
(c_{13},thr_{13})
(c_{14},thr_{14})
(weights w_1, spatial offsets D_1)
(c_{T1},thr_{T1})
(c_{T2},thr_{T2})
(c_{T3},thr_{T3})
(c_{T4},thr_{T4})
(weights w_T, spatial offsets D_T)
Σ
$p(E(x)|I(y);\{T_t\}_{t=1}^{T})=\frac{1}{T}\sum_{t=1}^{T}p(E(x)|I(y);T_t)$
Hough voting based on the spatial offsets
Hough voting space constructed by all local patches
Estimated object center
x
y
z

Fig. 2. Detection procedure of the proposed algorithm. The red cross represents the real object center, and the light blue cross represents the estimated object center.

rotation-invariant features, then Hough forest is learned base on a rotation-invariant split function; in the testing stage, the extracted local patches are used to establish a Hough space voting accumulation and the object are detected at the voting peaks. The feature description, the Hough forest learning, and the Hough voting strategy steps are adapted to meet the requirement of rotation invariance for terrestrial mobile laser scanning data processing. The experimental results demonstrate the robustness and effectiveness of the proposed algorithm on large-scale mobile laser scanning point clouds, especially the robustness to the cluttered situation of occlusion and overlapping between neighboring objects. Compared to existing methods, our method is more robust and easily extended to detect different categorical objects.

The rest of this letter is organized as follows: Section II describes the proposed method. Section III presents extensive experimental results and evaluates the performance of the proposed method. Finally, Section IV states the concluding remarks.

II. Method

In the next few sections, we first introduce an overview of the proposed algorithm in Section II-A. Section II-B details 3-D local patch extraction and feature description. Circular voting and distance weighted voting are introduced in Section II-C and D, respectively.

A. Overview of the Algorithm

Our algorithm is divided into two stages: the training and the detection. As shown in Fig. 1, the training procedure starts with densely extracting and describing a set of 3-D local patches (to be described in Section II-B) from training samples. Each patch appearance P_i is composed of three components $\{P_i = (I_i, c_i, d_i)\}$, where I_i is the feature description, c_i is the class label with 1 for positive samples and 0 for negative samples, and d_i is the offset vector which starts from the object center to the 3-D local patch center. Negative samples have a pseudo offset, i.e., $d_i = 0$. Based on these local patch appearances, the optimal parameters of the split function on each branch node are determined [10].To meet the requirement of rotation invariance, instead of using only the appearances at two selected different positions within a local patch, we use the entire local patch's appearance to learn the split function. Afterwards, according to the split function, the training patches reaching 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, and each leaf node stores the proportion and the offset vectors of the positive training patches reaching this node in the training stage.

The complete object detection procedure is shown in Fig. 2. First, the ground points are removed from the test scene [15]. Then, a segmentation method is used to partition the scene into individual segments [7]. The 3-D local patches of each segment are extracted based on the method used in the training stage (to be described in Section II-B) and described by the same features as those of the training patches. Each 3-D local patch is then passed through the trained trees downwards to a leaf node in each tree according to the information stored in the branch nodes. Next, the spatial offsets stored in the leaf nodes are used to cast votes to the object center. Finally, all votes create a Hough voting space, and the object center is determined by a traditional non-maximum suppression process.

B. Three-Dimensional Local Patch Extraction and Feature Description

We define a 3-D local patch as a 3-D box with fixed size. The octree partitioning is applied to the entire point cloud for extracting the 3-D local patches. The 3-D space of the point cloud data is recursively subdivided into eight octants until each octant node is of the given size. Each 3-D local patch that contains 27 (3 by 3 by 3) leaf nodes is centered at one randomly selected non-empty leaf node of the constructed octree, as shown in Fig. 3. Fig. 3(a) shows the constructed octree with a resolution of 0.1 m for a sample. Fig. 3(b) shows one of the extracted 3-D local patches with the size of 0.3 m and Fig. 3(c) shows a close-up view of the local patch. The size of extracted local patches is three times of the size of octree leaf nodes. As a result, the local patches are densely extracted and overlapped with the neighboring local patches. Through treating all points in the local patch as a unit, rather than the individual points,

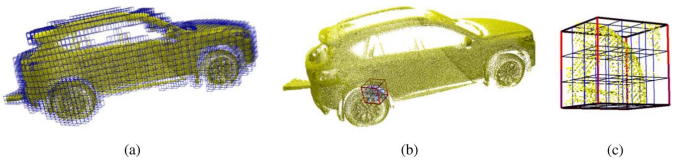


Fig. 3. Three-dimensional local patch extraction based on octree. (a) shows the constructed octree for an unorganized point cloud; (b) shows an extracted 3-D local patch; and (c) shows the close-up view of the extracted patch in (b).

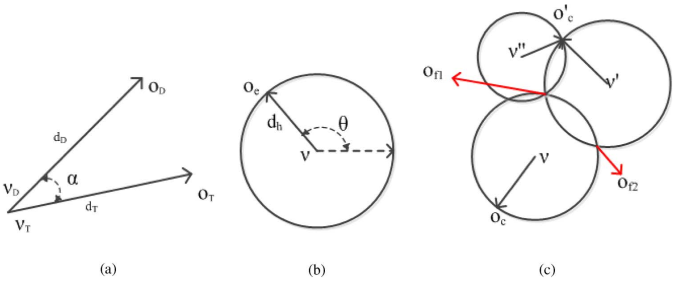


Fig. 4. Circular voting. (a) Illustration of direct voting for a rotated object; (b) illustration of circular voting; and (c) false peaks caused by circular voting for adjacent objects.

makes the local patch robust to noise and outliers. In the training stage, a fixed number of 3-D local patches are extracted. In the detection stage, each non-empty leaf node of the octree serves as the center of 3-D local patch.

The local patch is described by both the structure and reflectance features, such as spectral features [16], eigenvalues of covariance matrix, 3-D invariant moments, fast point feature histograms (FPFH) [17], and median of reflectance intensities. 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.

C. Circular Voting

A key component of our method is a generalized Hough voting procedure. To detect 3-D objects in real world scenes, the voting step needs to be rotation-invariant. The voting is based on the offset vectors learned in the training stage. If the offset vector is directly used in Hough voting, the votes for rotated objects will be cast to some wrong positions because each offset vector only represents a candidate object center with a certain orientation, which may not be consistent with the object of interest in the test scene. Fig. 4(a) illustrates the direct voting for an object which is rotated by an angle of α along the counter-clockwise direction in the horizontal plane from training samples. In Fig. 4(a), v_T is a local patch which originates from the training sample centered at o_T, and v_D is a local patch which originates from the object of interest centered at o_D in the test scene. The two corresponding offset vectors of these two local patches are d_T and d_D. We assume that these two patches are the same except that they are originated from two rotated samples. If the offset vector d_T is directly used to vote for object center, the voted center position is at a wrong position o_T rather than the true position o_D.

In urban environments, the objects of interest such as trees, cars, traffic signs, and street lamps, are usually rotated only in the azimuth direction. Thus, in this letter we only consider rotations in the azimuth direction in the entire 3-D geometric transformation space. To handle rotation invariance, the offset vector is rotated for all orientations in the azimuth direction, which essentially defines a circular voting field, as shown in Fig. 4(b). All positions with a certain distance d_h to v in the horizontal plane and d_z to v in Z direction are potential positions of the object center. By using the circular voting, we can achieve rotation invariance in the azimuth direction. For a 3-D local patch v, the object center is estimated by

$$\begin{cases} o_x = v_x + d_h \cos(\theta) \\ o_y = v_y + d_h \sin(\theta) \\ o_z = v_z - d_z \end{cases} \tag{1}$$

$$d_h = \sqrt{d_x^2 + d_y^2}. \tag{2}$$

The voting process is computationally efficient because the rotation operation of offset vectors can be implemented through defining a discrete lookup table. However, the rotation invariance is obtained at the cost of a low false positive rate because some irrelevant locations are also voted by circular voting. After summing all votes, some meaningless locations may be peaks with high scores in the Hough voting space. This false voting phenomenon is especially obvious for the areas with concentrated objects, such as adjacent trees. In order to improve the detection accuracy and suppress false peaks, a distance weighted voting strategy is introduced, which will be discussed in the next subsection.

D. Distance Weighted Voting

As illustrated in Fig. 4(c), o_c and o'_c denote two adjacent object centers, respectively. v, v', and v'' represent three 3-D local patches centered at these two adjacent objects. v' and v'' originate from the same object, and v originates from the other one. Through circular voting, the right object center is voted. However, some false locations are also voted, such as o_{f1} and o_{f2}. In our method, a distance weighted voting strategy is proposed to improve the detection accuracy and suppress false peaks. For an offset vector d in the trained forest, the weight for the vote is defined as

$$w = \exp\left(-\frac{\left(d_x^2 + d_y^2\right)}{\sigma^2}\right) \tag{3}$$

where d_x and d_y are horizontal components of offset vector d in horizontal plane, σ is a smoothing parameter.

III. Experiment

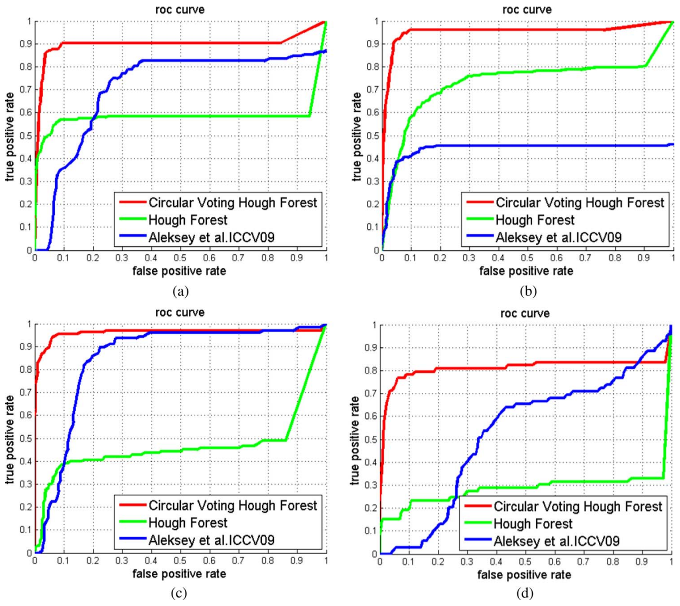
The proposed method was evaluated on three data sets containing four different categorical objects: street lamp, palm tree, car, and traffic sign. All of these data sets were collected by the RIEGL VMX-450 system (400 lines per second, 1.1 million measurements per second, and 8 mm accuracy) in Xiamen, China. We manually labeled the target objects in all training and testing point clouds as the ground truth. A detection to be marked as true positive must meet the condition that the estimated center falls into the certain distance thresholds in both horizontal and vertical directions relative to the labeled object center. Each target object can only match 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. In addition, our algorithm was compared with the original Hough forest and the method proposed in [7].

A. Comparison

The first data set covering about 188 150 m^2 and containing about 480 million points was used to evaluate the detection performances of street lamps and palm trees. Samples with various azimuths were selected to train lamp-specific forest through dense sampling patches. The intersection point of lamp pole and lamp header is considered as the object center. The test scene contains 183 street lamps and their azimuths vary from 0 to 360 degrees. 159 street lamps were completely segmented from the scene and the rest 24 street lamps were failed to be segmented because of the overlapping with other objects in the scene. For palm trees, the geometric center of palm tree is considered as the object center. The test scene contains 198 palm trees. 91 palm trees were segmented from the scene and the rest 107 palm trees were failed to be segmented because the palm trees in the scene are much dense and overlap with each other.

The second data set covering about 8700 m^2 and containing about 61 million points was used for the evaluation of car detection. Samples with various azimuths were selected to train car-specific forest through dense sampling patches. The data set contains 134 cars and their azimuths vary from 0 to 360 degrees. Moreover, nearly half of the cars are seriously occluded when scanning. The third data set containing about 24 million points selected from the raw point clouds was used for the evaluation of traffic sign detection. This data set covers a distance of about 10 km along the surveyed road. It contains 73 traffic signs with various azimuths and sizes. Thirty-eight traffic signs were completely segmented from the scene, and 35 traffic signs were failed to be segmented.

For all the four categorical objects, we have selected features from the following 28 feature channels: three spectral features, three eigenvalues of covariance matrix, three 3-D invariant moments, median of reflectance intensity, sixteen FPFH feature channels, height, and occupied area of a local patch in the horizontal plane. For street lamps, all the 28 features were used for both training and detection. For palm trees, the FPFH feature channels were not used because the FPFH is described based on the normal vector of each point in the local patch, however the normal vector is unstable on the leaves of the palm trees. For cars and traffic signs, the same feature channels as street lamps were selected except the median of reflectance intensity because that the variety of the colors of these two categories may cause large variety in the reflection intensity.



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

Fig. 5. The proposed method (red curve) is compared with original Hough forest (green curve) and the state-of-the-art method (blue curve) on four different categorical objects: (a) street lamp; (b) palm tree; (c) car; (d) traffic sign.

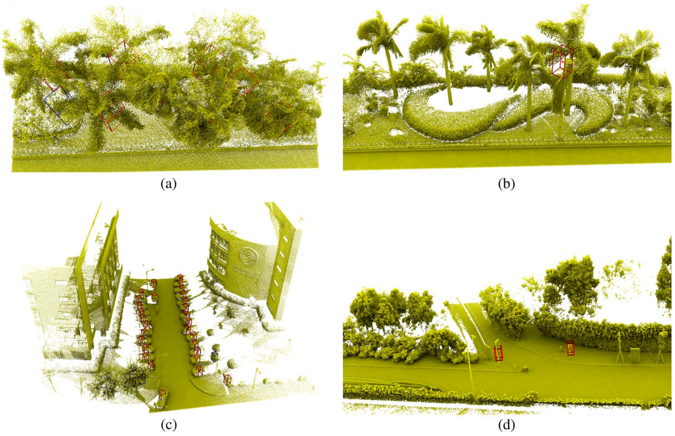

(a) (b) (c) (d)

Fig. 6. Results of the proposed detection algorithm. Red 3-D bounding boxes represent the true correct detection results, and blue 3-D bounding boxes represent the false positive detection results.

The performance of different methods including ours is shown in Fig. 5. As seen from Fig. 5, our method outperforms other methods considerably. The method in [7] is essentially based on an object's global shape features, and its detection performance is seriously dependent on the completeness of an object. As a result, for the test scenes where objects cannot be segmented from the background because of the overlapping between neighboring objects or objects occluded seriously when scanning, the performance of this method will degrade. On the contrary, our method is based on the object part appearance and can deal with occlusion and overlapping. Through the comparison results, we conclude that our method has the ability of dealing with overlapping, occlusion, and rotation in cluttered nature scenes. Fig. 6 shows the detection results for four different categorical objects. Fig. 6(a) shows the detection results of palm trees, Fig. 6(b) shows the detection results of street lamps, Fig. 6(c) shows the detection results of cars, and Fig. 6(d) shows the detection results of traffic signs. All experiments were finished on a machine with Intel Core i3 3.3 GHz processor and 16 GB RAM. The running time of training and detection on these four categorical objects is showed in Table I.

B. Sensitivity to Parameters

We also run experiments to evaluate the proposed method under different parameter settings. In particular, we test two

TABLE I
RUNNING TIME OF TRAINING AND DETECTION

	number of points in training dataset (mil.)	training time (s)	number of points in detection dataset (mil.)	ground points filtering time (s)	non-ground points segmentation time (s)	detection time (s)
Palm tree	0.67	468	480	2016	432	5616
Street lamp	0.84	432	480	2016	432	4032
Car	1.26	252	61	720	180	8028
Traffic sign	0.14	72	24	684	144	1332

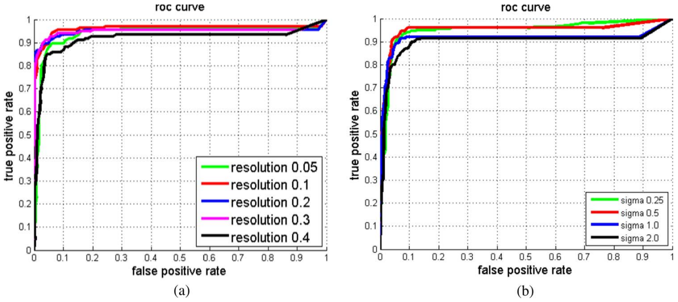


Fig. 7. (a) Sensitivity of octree resolution for car detection. (b) Sensitivity of distance weighted parameter σ for palm tree detection.

parameters: resolution of octree and distance weighted smoothing parameter σ. The resolution of octree decides the size of local patch. Fig. 7(a) shows the detection performance under different octree resolutions for car detection. From the Fig. 7(a), we can observe that setting the resolution of octree to 0.1–0.3 m achieves better performance than other values for car detection.

This distance weighted strategy pays more attention to the patches closer to the object center, compared to the further ones. Although the distance weighted technique can reject false detections caused by circular voting, it lessens the contributions of long range patches. Sometimes this may cause degradation of detection performance, especially for such objects, whose long range patches play important roles in distinguishing from others. To evaluate the sensitivity to the parameter σ, several experiments have been carried out with different values of σ for palm trees detection because of their serious overlapping. From Fig. 7(b), we can observe that the detection performance decreases when σ is too big or too small. This is because a small σ decreases the importance of patches on voting for object center while a large σ increases false detections.

IV. CONCLUSION

In this letter, we proposed a novel rotation-invariant method for object detection from TLS point clouds of complex urban environments. The main contributions of this letter include: firstly, the extension of the Hough forest method from 2-D images to 3-D point clouds, and secondly, the detection of objects with rotations in azimuth direction through a novel distance weighted circular voting strategy. The comparative experiments showed that our 3-D object detection method is robust to overlapping, occlusion and rotation, and is easily extended to various object categories. Tested on four different categorical real-world objects, our method achieves better performances compared to the original Hough forest and the state-of-the-art 3-D object detection method. The experimental results demonstrate the robustness and effectiveness of ISM on category-level object description and the Hough forest framework on object detection in 3-D laser scanning point clouds of complex urban environments.

REFERENCES

[1] 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, no. 3, pp. 641–664, 2010.

[2] F. Monnier, B. Vallet, and B. Soheilian, "Trees detection from laser point clouds acquired in dense urban areas by a mobile mapping system," in *Proc. ISPRS Ann. Photogramm., Remote Sens. Spatial, Inf. Sci., I-3*, 2012, pp. 245–250.

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

[4] 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.

[5] P. M. a. K. D. Alexander Patterson, IV, "Object detection from large-scale 3D datasets using bottom-up and top-down descriptors," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 553–566.

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

[7] V. G. K. Aleksey Golovinskiy and T. Funkhouser, "Shape-based recognition of 3D point clouds in urban environments," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[8] A. G. T. Funkhouser, "Min-cut based segmentation of point clouds," in *Proc. Int. Conf. Workshop Comput. Vis.*, 2009, pp. 1–8.

[9] 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.

[10] J. Gall, Lempitsky, and Victor, "Class-specific hough forests for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1022–1029.

[11] B. Leibe, Leonardis, Aleš, Schiele, and Bernt, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 259–289, 2008.

[12] J. Knopp, Prasad, Mukta, Gool, and Luc Van, "Scene cut: Class-specific object detection and segmentation in 3D scenes," in *Proc. IEEE Conf. 3DIMPVT*, 2011, pp. 180–187.

[13] R. S. Alexander Velizhev and K. Schindler, "Implicit shape models for object detection in 3D point clouds," in *Proc. ISPRS Congr.*, 2012, pp. 1–6.

[14] T. F. Zhen Lei, 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.

[15] Y. Zhou, Y. Yu, G. Lu, and S. Du, "Super-segments based classification of 3D urban street scenes," *Int J. Adv. Robot. Syst.*, vol. 9, pp. 1–8, 2012.

[16] D. Munoz, Bagnell, J. Andrew, Vandapel, Nicolas, Hebert, and Martial, "Contextual classification with functional max-margin markov networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 975–982.

[17] R. B. Rusu, Blodow, Nico, Beetz, and Michael, "Fast point feature histograms (fpfh) for 3d registration," in *Proc. IEEE Conf. Robot. Autom.*, 2009, pp. 3212–3217.