Pole-Like Road Object Detection in Mobile LiDAR Data via Supervoxel and Bag-of-Contextual-Visual-Words Representation

Haiyan Guan, *Member, IEEE*, Yongtao Yu, Jonathan Li, *Senior Member, IEEE*, and Pengfei Liu

***Abstract*—This letter addresses the problem of detecting pole-like road objects (including light poles and traffic signposts) from mobile light detection and ranging (LiDAR) data for transportation-related applications. The method consists of two consecutive stages: training and pole-like object detection. At the training stage, a contextual visual vocabulary is created from the feature regions generated from a training data set by supervoxel segmentation. At the pole-like object detection stage, a bag-of-contextual-visual-words representation is generated for each semantic object segmented from mobile LiDAR data. The experimental results show that the proposed method achieves correctness, omission, and commission of 88.9%, 11.1%, and 2.8%, respectively, in detecting pole-like road objects. Computational complexity analysis demonstrates that our method provides a promising and effective solution to rapid and accurate detection of pole-like objects from large volumes of mobile LiDAR data.**

***Index Terms*—Bag-of-contextual-visual-words, detection, mobile light detection and ranging (LiDAR), pole-like objects, supervoxel segmentation.**

I. Introduction

MOBILE light detection and ranging (LiDAR) systems integrate laser scanner(s), a global navigation satellite system, an inertial measurement unit, a distance measurement indicator, and digital/video camera(s) [1], [2]. LiDAR systems have been used to acquire three-dimensional (3-D) geospatial data of roadways over a large area at a normal driving speed. Nowadays, high-density and high-accuracy LiDAR data are becoming a leading source for highway mapping [3], urban road distress assessment [4], [5], and road feature inventory [6], [7]. Pole-like road objects, including light poles and traffic signposts, located along roads/streets, are typical kinds of road infrastructure. For example, light poles provide illumination to pedestrians and vehicles at night for a clear visibility of the road environment. Traffic signposts, as a highly important transportation infrastructure, play a critical role in transportation, traffic safety, and route guidance. Thus, recently, detecting pole-like road objects (specifically light poles and traffic signposts) has attracted increased attention in the literature.

Based on eigenvalue analysis, principal component analysis (PCA) was a widely used method for detecting pole-like road objects from irregular point clouds [8], [9]. These methods detected linear pole-like structures by first constructing a covariance matrix for each point with its neighbors and then analyzing eigenvalues decomposed from the covariance matrix [10]. Eigenvalue-based PCA methods show high computational efficiency. However, other objects (particularly tree trunks) in a road scene might cause a considerable number of false alarms.

Shape and context features were also widely used in pole-like object detection methods. Shape features were characterized based on height, number, and types of attached part segments, whereas context features were computed based on surrounding distributions [8]. For example, a percentile-based method was developed in [6] with respect to the shape, height, and size of light poles. In [11], a pairwise 3-D shape context descriptor, which considers both local and global similarity measures, was proposed to detect light poles. In [7], a 3-D object matching framework was proposed for detecting light poles, with attachments, of varied shapes and sizes.

With the prior knowledge of pole-like objects in shape and size, by using grammar rules, a voxel structure was applied to mobile LiDAR data [12]. Through a 3-D neighborhood analysis of voxel representations, pole-like objects were detected. In [13]–[16], pole-like objects were extracted by analyzing scan lines, rather than raw point clouds. To improve computational efficiency, some studies convert 3-D point clouds into 2-D representations [17]. Point density in 2-D representations was also exploited to detect light poles [13].

Supervoxels, presented in [18], group 3-D points into perceptually meaningful clusters by using voxel cloud connectivity segmentation (VCCS). VCCS, an oversegmentation algorithm

Manuscript received September 17, 2015; revised December 21, 2015; accepted January 18, 2016. Date of publication March 2, 2016; date of current version March 23, 2016. This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20151524, in part by the National Natural Science Foundation of China under Grant 41501501, and in part by Tianjin Science and Technology Planning Project under Grant 15ZCZDSF00390.

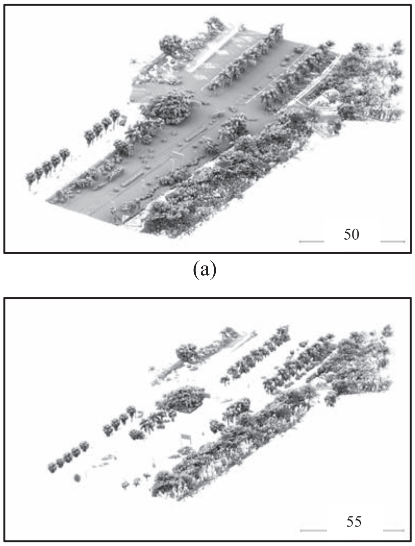
H. Guan is with the College of Geography and Remote Sensing, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: guanhy.nj@nuist.edu.cn).

Y. Yu is with the Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China (e-mail: allennessy.yu@gmail.com).

J. Li is with the Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, Xiamen 361005, China, and also with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@xmu.edu.cn; junli@uwaterloo.ca).

P. Liu is with the College of Urban & Environment Science, Tianjin Normal University, Tianjin 300387, China (e-mail: liupengfei0920@163.com).

Digital Object Identifier 10.1109/LGRS.2016.2521684



(a)

(b)

Fig. 1. Portion of the 11-km-long road containing the test data set. (a) Raw point cloud. (b) Filtered off-ground points by a voxel-based upward growing filtering.

for point clouds, performs computationally efficiently. Thus, in this letter, we supervoxelize a training data set to generate feature regions based on a first-order supervoxel neighborhood. Regarding generated feature regions, we analyze their spatial contextual information to generate a contextual visual vocabulary because spatial contextual information exhibits richer, more salient, distinctive representations than do only local feature regions. Then, by using the generated contextual visual vocabulary, semantic objects, segmented from mobile LiDAR data, are quantized to form a bag-of-contextual-visual-words representation to detect pole-like objects. Such a bag-of-contextual-visual-words representation has the advantage of effectively, saliently, and distinctively depicting an entire object, thereby providing a promising object-oriented detection solution.

II. Mobile LiDAR Data Set

In this letter, the survey area is within Xiamen Island (118°04′04″ E, 24°26′46″ N), a part of the City of Xiamen, China. A RIEGL VMX-450 system (see the RIEGL website for complete specifications) was used for this survey. In this survey, we collected data along Ring Road South at an average speed of 30–40 km/h. Ring Road South is a typical urban road area containing a plethora of road infrastructures (e.g., light poles and traffic signposts) and other objects (e.g., cars, buildings, and high bridges). The average point density in the survey is about 4082 points/m^2. From the collected data, we selected a test data set, containing about 1728 million points and covering a road segment of approximately 11 km. Fig. 1(a) shows a portion of the 11-km-long road segment.

To reduce the number of points to be processed, a voxel-based upward growing filtering method [7] is first performed to remove ground points from the test data set. Fig. 1(b) shows the filtered off-ground points.

III. Pole-Like Object Detection Framework

The method proposed for the detection of pole-like road objects from mobile LiDAR data includes the following two stages:

1) Training stage, which generates contextual visual vocabulary via supervoxel segmentation.
2) Detection stage, which detects pole-like objects, including light poles and traffic signposts, from the filtered off-ground points by a bag-of-contextual-visual-words representation.

A. Training Stage

The training stage aims to generate a contextual visual vocabulary for depicting the features of pole-like road objects. From the filtered off-ground points, to construct a contextual visual vocabulary, we select a group of 50 training data sets, each of which covers a road segment of about 50 m.

1) Generation of Feature Regions: To obtain salient and distinctive local geometric representations of pole-like road objects, the training data are first segmented into a set of supervoxels by using the VCCS algorithm [18]. Then, the feature region associated with a supervoxel is created by integrating this supervoxel and its first-order neighbors. During supervoxelization with the VCCS algorithm, two parameters, voxel resolution (S_r) and seed resolution (S_s), control the segmentation quality. S_r condenses a continuous point-cloud space to a voxel-cloud space, whereas S_s controls seed point selection to construct initial supervoxels. As demonstrated in [19], such a feature region generation strategy, by embedding first-order supervoxel neighbors, achieves higher saliencies and distinctiveness than does its counterpart, by using single supervoxels as feature regions.

2) Structural Description of Feature Regions: In this letter, we describe feature regions by a structural descriptor, integrating geometrical (F_g), orientation (F_o), and scale information (F_s).

Geometrical information (F_g) includes two components: a 3-D eigen-based geometrical feature descriptor (g_{eigen}) [20] and a 16-dimensional fast point feature histogram (FPFH) descriptor (g_{FPFH}) [21]. An FPFH descriptor can depict 3-D point clouds rapidly and saliently. For a feature region, g_{eigen} is derived from a covariance matrix $\mathbf{S}_{3\times3}$, constructed by all of the points in the feature region. The covariance matrix's three eigenvalues (λ_1, λ_2, and λ_3, $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$) are decomposed to calculate the three members of the eigen-based feature descriptor, $g_{\text{eigen}} = \{a_l, a_p, a_v\}$

$$a_l = \frac{\sqrt{\lambda_1} - \sqrt{\lambda_2}}{\sqrt{\lambda_1}} \quad a_p = \frac{\sqrt{\lambda_2} - \sqrt{\lambda_3}}{\sqrt{\lambda_1}} \quad a_v = \frac{\sqrt{\lambda_3}}{\sqrt{\lambda_1}} \tag{1}$$

where a_l, a_p, and a_v represent the linear, planar, and volumetric geometrical features, respectively [20].

Orientation information (F_o) denotes the orientation of a feature region and is represented by the eigenvector e_1 associated with the largest eigenvalue λ_1 of the covariance matrix $\mathbf{S}_{3\times3}$.

Scale information F_s is defined as the longest Euclidean distance between the centers of a supervoxel and its first-order neighbors in a feature region.

3) Generation of Contextual Feature Groups: Next, a contextual feature group is created for each feature region to obtain spatial contextual information. As stated in [22], the number of feature regions m plays an important role in constructing contextual feature groups. The greater the number of feature regions, the smaller the repeatability of the combination of feature regions. Moreover, because of increasing feature-to-feature matching orders, an increase of feature regions leads to high computational burdens for measuring spatial contextual similarities between two contextual feature groups. In this letter, m is empirically set to 3.

Accordingly, for a feature region r, its contextual feature group G is constructed by r and its two nearest neighboring feature regions. Within G, spatial and scale relationships ($R_O(G)$ and $R_S(G)$) are calculated by

$$R_O(G) = \sum_{i=1,j>i}^{m} \arccos\left(Fo_i^T \cdot Fo_j\right) \tag{2}$$

$$R_S(G) = \sum_{i=1,j>i}^{m} \log\left(1 + \frac{Fs_i}{Fs_j}\right) \tag{3}$$

where Fo_i and Fs_i are the orientation and scale information, respectively, for feature region r_i and Fo_j and Fs_j are the orientation and scale information, respectively, for feature region r_j.

To measure the spatial contextual similarity between two contextual feature groups, we use a discriminant distance metric. We define a possible match between two contextual feature groups as a matching order. Given that m is 3, the number of matching orders is 6. For each matching order, we calculate all of its spatial context weighted *Mahalanobis* distances between two contextual feature groups. Then, to obtain the best matching order, the discriminant distance metric is defined as a spatial contextual similarity. Let G_A and G_B denote two contextual feature groups, respectively. Then, their spatial contextual similarity $S(G_A, G_B)$ is defined by

$$S(G_A, G_B) = \max_{\psi\in\{0,\ldots,m!\}} \frac{1}{2}\left(S_O^{\psi}(G_A, G_B) + S_S^{\psi}(G_A, G_B)\right) \tag{4}$$

where $\psi \in \{0, \ldots, m!\}$ is a matching order and $S_O^{\psi}(G_A, G_B)$ and $S_S^{\psi}(G_A, G_B)$ are the orientation and scale similarities, respectively, under matching order ψ. The similarities are defined as

$$S_O^{\psi}(G_A, G_B) = \frac{\min\left(R_O^{\psi}(G_A), R_O^{\psi}(G_B)\right)}{\max\left(R_O^{\psi}(G_A), R_O^{\psi}(G_B)\right)} \tag{5}$$

$$S_S^{\psi}(G_A, G_B) = \frac{\min\left(R_S^{\psi}(G_A), R_S^{\psi}(G_B)\right)}{\max\left(R_S^{\psi}(G_A), R_S^{\psi}(G_B)\right)}. \tag{6}$$

The definitions of $R_O^{\psi}(G_A)$ and $R_S^{\psi}(G_A)$ are established in (2) and (3).

For each matching order, the spatial contextual similarity between contextual feature groups G_A and G_B is calculated based on (4). Let ψ^* denote the best matching order. A spatial context weighted *Mahalanobis* distance is defined by

$$C(G_A, G_B) = (1 - S(G_A, G_B)) \cdot \sum_{i=1}^{m} \left(F_g^i(G_A) - F_g^{\psi*(i)}(G_B)\right)^T \times \mathbf{A}^{-1}\left(F_g^i(G_A) - F_g^{\psi*(i)}(G_B)\right) \tag{7}$$

where $\mathbf{A} \in R^{19\times 19}$ is the covariance matrix over all geometrical features. $\psi^*(i)$ represents the best match between the feature region in G_B and feature region i ($i = 0$, 1, 2) within G_A. $F_g^i(G_A)$ is the geometrical information for feature region i within G_A. $F_g^{\psi*(i)}(G_B)$ is the geometrical information for the best matching feature region in G_B.

4) Generation of Contextual Visual Vocabulary: To generate a contextual visual vocabulary, we first vector-quantize the constructed contextual feature groups into a number of clusters C_i ($i = 0, 1, \ldots, N$, and N is the number of clusters) based on the spatial contextual similarity defined in (4). To reduce computational complexity, the contextual feature group G_i^*, with the shortest distances to the other contextual groups in C_i, is selected as the updated center.

To further improve computational efficiency, in practice, we store a group-to-group distance matrix for each cluster to rapidly update the cluster center. Once the distance matrix of a cluster is computed, the clustering operation in its corresponding subclusters is completed efficiently. After vector quantization, each cluster center is taken as a distinctive contextual visual word. Finally, such contextual visual words form a contextual visual vocabulary. Because the contextual visual vocabulary is generated using contextual feature groups rather than single feature regions, each word in the vocabulary preserves rich, salient, and distinctive spatial contextual information. In addition, a stop list analogy [23] is used to discard the most frequent contextual visual words that occur in almost all scenes.

B. Pole-Like Road Object Detection Stage

Based on the generated contextual visual vocabulary from 50 training data sets, a bag-of-contextual-visual-words is constructed to detect pole-like objects from the filtered off-ground points.

To generate bag-of-contextual-visual-words representations from the filtered off-ground points, individual objects are first obtained through the following steps: 1) a Euclidean distance clustering approach, with a clustering distance of d_c, is applied, followed by a voxel-based normalized cut (Ncut) segmentation method [7]. In [7], the Ncut method effectively segments connected, but not seriously overlapped, clusters into separated semantic objects. 2) Based on the generated contextual visual vocabulary from the training data, a set of contextual visual words is formed for each semantic object. 3) Finally, a bag-of-contextual-visual-words is represented for each semantic object by a standard "term frequency-inverse document frequency" weighting [23]. The bag-of-contextual-visual-words representation, in detail, is as follows: first, we define a semantic object as a document d and the number of words in a contextual visual vocabulary as D. Accordingly, each document is represented by a D-dimensional vector of weighted word frequencies

$$\Gamma_d = (t_1, t_2, \ldots, t_i, \ldots, t_D)^T \tag{8}$$

where t_i is the weighted word frequency of the ith word in the vocabulary and represented by

$$t_i = \frac{n_i^d}{\sum\limits_{j=1}^{D} n_j^d} \log \frac{N}{N_i} \tag{9}$$

where n_i^d is the number of occurrences of the ith word, N is the total number of documents in the database, and N_i is the number of documents containing word i. This weighting is a product of two terms: *word frequency* and *inverse document frequency*. In this way, the segmented semantic objects are represented by a bag-of-contextual-visual-words.

To detect pole-like objects from the segmented semantic objects, a clean and completely scanned pole-like sample is also selected, supervoxelized, characterized, and quantized to form bag-of-contextual-visual-words representation. Then, with the representation, we use the normalized histogram intersection distance metric [24] to measure the similarity between the pole-like sample and a selected semantic object. As a result, a series of similarity measures between the pole-like sample and the segmented semantic objects is calculated and then thresholded to obtain pole-like road objects of interest.

IV. Results and Discussion

A. Parameter Sensitivity Analysis

A test data set obtained over an 11-km-long road segment (Section II) was used to investigate the applicability of our method. The following parameters were used: d_c, S_r, S_s, m, and D. d_c was empirically set to 0.15 m. An S_r of 0.05 m and an S_s of 0.1 m were used to generate supervoxels from a set of separated semantic objects. The parameters, contextual visual vocabulary size D and contextual feature group size m, have a significant impact on the detection performance of pole-like objects. Thus, we designed two groups of experiments to investigate the sensitivity of the proposed algorithm to the selections of D and m.

Comparing the extracted pole-like objects with the manually interpreted ground truth, we quantitatively assessed the pole-like object detection results according to the following three measures: percentages of *correctness* (E_{crt}), *omission* (E_{omi}), and commission (E_{comi}) [4], [20]. E_{cpt} indicates the correctly detected objects, E_{omi} evaluates the number of missing pole-like objects, and E_{comi} evaluates a portion of nonpole-like objects being misclassified as pole-like objects.

In the first group, we held m constant and varied D from 90 000 to 140 000 in intervals of 10 000. As shown in Table I, the detection performance improves as the vocabulary size increases. This is because, the greater the number of contextual visual words in the vocabulary, the higher the degrees of distinction between different categories of objects. However, when vocabulary size exceeds 120 000, performance changes very slightly. In addition, an increase in the vocabulary size causes great computational burden at the vocabulary generation stage. Thus, to balance detection performance and computational complexity, we set the vocabulary size at $D = 120\,000$.

TABLE I
Parameter Sensitivity Analysis: Vocabulary Size D and Feature Group Size m

D(×10000)	**9**	**10**	**11**	**12**	**13**	**14**
E_{crt} (%)	85.4	87.5	88.1	89.8	89.9	90.0
E_{omi} (%)	14.6	12.5	11.9	10.2	10.1	10.0
E_{comi} (%)	3.4	3.5	3.0	3.0	2.9	2.9
m	**1**	**2**	**3**	**4**	**5**	**6**
E_{crt} (%)	83.3	85.5	87.1	87.0	86.1	80.9
E_{omi} (%)	16.7	14.5	12.9	13.0	13.9	19.1
E_{comi} (%)	7.2	6.8	5.0	5.1	5.5	7.0

In the second group, we kept $D = 120\,000$ and varied m from 1 to 6 in intervals of 1. $m = 1$ means that single feature regions (without spatial contextual information) are used to generate the contextual visual vocabulary. As shown in Table I, when $m \leq 4$, detection performance improves as the contextual feature group size increases. This is because, by considering spatial contextual information in feature regions, the quantized contextual visual words are more likely to obtain salient distinctive feature encodings, thereby able to differentiate objects in different categories. However, when $m > 4$, detection performance drops dramatically. If too many local feature regions are combined, the repetition of the combination decreases accordingly, leading to a decrease in detection performance. In addition, an increase of m slows down the generation of the contextual visual vocabulary. Therefore, to obtain promising detection performance, we set the contextual feature group size to 3.

B. Pole-Like Object Detection

To evaluate the performance of our proposed pole-like object detection algorithm, we applied it to the mobile LiDAR data set. After parameter sensitivity analysis, we set $D = 120\,000$ and $m = 3$. In this letter, two types of pole-like objects, light poles and traffic signposts, were tested. The clean and completely scanned light pole and traffic signpost samples were selected as query objects to generate bag-of-contextual-visual-words representations. Then, the query objects were compared with the segmented sematic objects from the filtered point clouds. We manually checked the accuracy assessment of pole-like object detection (Table II). The labeled ground truth is a total of 888 pole-like objects, including 647 light poles and 241 traffic signposts. For the whole data set, the correctness, omission, and commission attained 88.9%, 11.1%, and 2.8%, respectively. As shown in Table II, 48 out of 99 pole-like objects were really missed, and the left includes 44 light poles misclassified as traffic signposts and 7 traffic signposts misclassified as light poles. Due to serious incompleteness and serious overlapping with other objects, which cannot be effectively segmented, some pole-like objects failed to be detected. In addition, the misclassification was mainly caused by the high geometric similarities of other objects to the pole-like objects.

TABLE II
ACCURACY OF POLE-LIKE OBJECT DETECTION

	Pole type:	Light pole	Traffic signpost	Other	Total
Number of	Reference	647	241	0	888
Detected Pole-like objects	Correctness	601	188	0	789
	Omission	44	7	48	99
	Commission	0	0	23	23

E_{crt} = 789/888 = 88.9%; E_{omi} = 99/888= 11.1%; E_{comi} = 23/812= 2.8 %

C. Computational Complexity

The proposed algorithm was implemented using C++ and tested on an HP Z820 8-core-16-thread workstation. Accordingly, time complexity analysis was divided into two stages: the generation of contextual visual vocabulary and the detection of pole-like objects. The algorithm took approximately 39 min to generate the contextual visual vocabulary containing 120 000 contextual visual words and approximately 42 min to detect pole-like objects from the filtered off-ground points. To efficiently process the mobile LiDAR data, we segmented it into segments with about a road length of 50 m each. Accordingly, a multithread computing environment containing 16 parallel threads was adopted. Such a parallel computing strategy dramatically improves the computational efficiency and reduces the time complexity of the proposed algorithm.

V. CONCLUSION

In this letter, we have presented a novel pole-like object detection method by using supervoxel segmentation and bag-of-contextual-visual-words representations. The major tasks include feature training and pole-like object detection stages. The method was tested on a mobile LiDAR data set (collected over an 11-km-long road segment). Correctness, omission, and commission of 88.9%, 11.1%, and 2.8%, respectively, were achieved. Due to high similarities, the major errors occurred between light poles and traffic signposts. Computational efficiency analysis shows that the multithread computing strategy with 16 parallel threads contributes to the improvement of pole-like object detection from mobile LiDAR data.

REFERENCES

[1] S. Murray *et al.*, "Mobile mapping system for the automated detection and analysis of road delineation," *IET Intell. Transp. Syst.*, vol. 5, no. 4, pp. 221–230, Dec. 2011.

[2] M. Brogan, S. McLoughlin, and C. Deegan, "Assessment of stereo camera calibration techniques for a portable mobile mapping system," *IET Comput. Vis.*, vol. 7, no. 3, pp. 209–217, Jun. 2013.

[3] J. Gong, H. Zhou, C. Gordon, and M. Jalayer, "Mobile terrestrial laser scanning for highway inventory data collection," in *Proc. Int. Conf. Comput. Civil Eng.*, Clearwater Beach, FL, USA, 2012, pp. 17–20.

[4] H. Guan *et al.*, "Iterative tensor voting for pavement crack extraction using mobile laser scanning data," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1527–1537, Mar. 2015.

[5] Y. Yu, H. Guan, and Z. Ji, "Automated detection of urban road manhole covers using mobile laser scanning data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3258–3269, Dec. 2015.

[6] 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, no. 6, pp. S28–S39, Dec. 2011.

[7] H. Guan, Y. Yu, J. Zheng, J. Li, and Q. Zhang, "Deep learning based tree classification using mobile LiDAR data," *Remote Sens. Lett.*, vol. 6, no. 11, pp. 864–873, Sep. 2015.

[8] H. Yokoyama, H. Date, S. Kanai, and H. Takeda, "Detection and classification of pole-like objects from mobile laser scanning data of urban environments," *Int. J. CAD/CAM*, vol. 13, no. 2, pp. 31–40, 2013.

[9] H. Yokoyama, H. Date, S. Kanai, and H. Takeda, "Pole-like objects recognition from mobile laser scanning data using smoothing and principal component analysis," in *Proc. ISPRS Archives*, 2011, vol. 38-5/W12, pp. 1–6.

[10] S. I. El-Halawany and D. D. Lichti, "Detection of road poles from mobile terrestrial laser scanner point cloud," in *Proc. Int. Workshop Multi-Platform/Multi-Sensor Remote Sens. Mapping*, Xiamen, China, 2011, pp. 1–6.

[11] Y. Yu, J. Li, H. Guan, C. Wang, and J. Yu, "Semiautomated extraction of street light poles from mobile LiDAR point-clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1374–1386, Mar. 2015.

[12] C. Cabo, C. Ordoñez, S. García-Cortés, and J. Martínez, "An algorithm for automatic detection of pole-like street furniture objects from mobile laser scanning point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 47–56, Jan. 2014.

[13] Y. Chen, H. Zhao, and R. Shibasaki, "A mobile system combining laser scanners and cameras for urban spatial objects extraction," in *Proc. IEEE Conf. Mach. Learn. Cybern.*, Hong Kong, 2007, vol. 3, pp. 1729–1733.

[14] Y. Hu, X. Li, J. Xie, and L. Guo, "A novel approach to extracting street lamps from vehicle-borne laser data," in *Proc. IEEE Conf. Geoinf.*, Shanghai, China, 2011, pp. 1–6.

[15] M. Lehtomäki, A. Jaakkola, 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, Feb. 2010.

[16] A. Kukko, A. Jaakkola, M. Lehtomäki, H. Kaartinen, and Y. Chen, "Mobile mapping system and computing methods for modeling of road environment," in *Proc. Urban Remote Sens. Event*, Shanghai, China, 2009, pp. 1–6.

[17] S. I. El-Halawany and D. D. Lichti, "Detecting road poles from mobile terrestrial laser scanning data," *GISci. Remote Sens.*, vol. 50, no. 6, pp. 704–722, Dec. 2013.

[18] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter, "Voxel cloud connectivity segmentation—Supervoxels for point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Portland, OR, USA, 2013, pp. 2017–2034.

[19] H. Wang *et al.*, "3-D point cloud object detection based on supervoxel neighborhood with Hough forest framework," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1570–1581, Apr. 2015.

[20] B. Yang, Z. Dong, G. Zhao, and W. Dai, "Hierarchical extraction of urban objects from mobile laser scanning data," *ISPRS J. Photogramm. Remote Sens.*, vol. 99, pp. 45–57, Jan. 2015.

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

[22] S. Zhang *et al.*, "Building contextual visual vocabulary for large-scale image applications," in *Proc. Int. Conf. Multimedia*, Firenze, Italy, 2010, pp. 501–510.

[23] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 591–606, Apr. 2009.

[24] Y. Jiang, J. Meng, J. Yuan, and J. Luo, "Randomized spatial context for object search," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1748–1762, Jun. 2015.