A 3D Descriptor based on Local Height Image

Tiecheng Sun, Shuaicheng Liu, Guanghui Liu, Shuyuan Zhu, Zhipeng Zhu

Institute of Image Processing
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

***Abstract*—This paper proposes a novel 3D local descriptor, which seeks a good balance between the efficiency and the accuracy. We use the *Local Reference Frame* (LRF) to estimate a robust coordinate system to describe the local 3D shape. A novel *Local Height Image* (LHI) is defined by projecting the 3D points in the support region onto the tangent plane of the basis point. The *Local Height Image Descriptor* (LHID) is then defined by calculating the averaged projection distances. We further smooth the LHID to resist various kinds of interferences. We setup several experiments to assess the performance of our descriptor by comparison with the state-of-the-art algorithms. The experimental results demonstrate the effectiveness of the proposed method, which not only achieves the high accuracy as well as the robustness, but also possesses low complexity for the efficiency.**

I. Introduction

The 3D local descriptor is one of the most fundamental and important research topic, which can be applied to many computer vision tasks such as point cloud registration [1], [2], [3], recognition [4], [5] and 3D reconstruction [6]. The 3D descriptor can be classified into two categories, the global descriptor and the local descriptor. Local descriptors have attracted greater attention due to their robustness against clutters and occlusions [7], [8], [9]. Local features are extracted based on the spatial neighborhood of the basis point. A good descriptor should be a descriptive representation of local 3D shapes. The article [9] gave a comprehensive survey of existing local surface feature used into 3D object recognition methods. And the article [10] evaluated ten popular local 3D descriptors from the aspects of the descriptiveness, the compactness, the robustness and the scalability.

Over the past decades, many works have been proposed regarding the 3D local descriptors. The Spin Image (SI) [7] is one of the classic local descriptor, whose support region is based on cylindrical coordinates established by the normal of the basis point. The cylinder is divided into several rings with different radius and heights. The indexes of bins formed the Spin Image. Frome *et al.* [8] proposed the 3D shape contexts (3DSC), which has spherical support region that being divided into bins by azimuth, elevation and radial dimension. The 3DSC has a good performance against noisy and cluttered data. These descriptors have a common characteristic that they have not tackled the uniqueness of the coordinate system. Later, Tombari *et al.* [11], [12] deployed a unique local Reference Frame (LRF) to improve the 3DSC performance which is named as Unique Shape Context (USC). The unique LRF was obtained by *EigenValue Decomposition* (EVD) of the covariance matrix. In order to disambiguate the sign of the eigenvectors, they reoriented the eigenvector so that its sign was coherent with the majority of the vectors. The USC yields an improved accuracy with less memory consumption [12]. They also proposed Signature of Histograms of orientation (SHOT) descriptor using LRF [11]. Guo *et al.* proposed a novel unique LRF [13]. The LRF used weighted scatter matrix to improve the robustness against varying mesh resolutions, occlusions and clutters. There were also some other LRFs based descriptors [14], [15].

High performance of a 3D descriptor is not only based on a unique and robust LRF but also a descriptive and exact 3D shape representation. Some descriptors are based on the histograms. For example, the Spin Image divided the cylinder to rings according to the radial and elevation coordinates [7]. The 3DSC [8], the SHOT [11] and the USC [12] divided the sphere into bins by azimuth, elevation, and radius. Guo *et al.* [13] rotated point cloud around three coordinate axis by a series of angles and projected point cloud to corresponding plane and counted the number of points that fall into the bin, yielding a descriptor named as RoPS. As for these methods, if the local point cloud is surface but not volume. There would be some empty bins if the size of bin is small. However, if the bin become bigger, the resolution of descriptor would be lower. With this dilemma, the division of volume is not the most effective way.

As such, Chua and Jarvis proposed Point Signatures [16]. They projected space curve gotten from intersection of a sphere with the surface to an approximate tangential plane. Points were characterized by signed distance from the point to its projected point and the angle about the normal and the reference direction. Novatnack *et al.* [14] mapped and encoded the local points to 2D domain using *exponential map* and extracted features in *geodesic polar coordinates*. Sometimes the spatial distribution of 3D points can be expressed by or embedded to a 2D image. Yang *et al.* divided the 3D lidar points into grids on the plane [17]. They defined the grids as geo-referenced feature image, whose intensity values reflect the spatial distribution of the 3D lidar points. There are also some other statistical properties applied to 3D descriptors. For instance, The PFH [18] and the FPFH [1] accumulated angles between pairs of points falling into the support region.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

However, they are sensitive to noise. MeshHOG computed histograms on three planes according to the gradients of the scalar function [15]. Castellani *et al.* utilized *Multicircular Hidden Markov Model* (MC-HMM) to analyze local geometric feature [19]. Bronstein *et al.* proposed the spectral shape distance according to generic diffusion distance, which used for nonrigid shape recognition [20]. In this paper, as aforementioned, we calculated the spatial distribution rather than the histogram, for a more realistic 3D shape representation.

Some of the methods, such as the USC [12], can achieve a good performance in terms of the accuracy. However, these methods are time consuming. Both the performance and the efficiency should be considered when designing the 3D descriptors. This is because more and more algorithms based on 3D descriptors are applied to real time applications such as self-driving. To this end, we propose a novel 3D descriptor *Local Height Image Descriptor*(LHID), which seeks a good tread-off between the efficiency and accuracy. With regards to the accuracy and the robustness, we use the LRF and collect the averaged projection heights. In terms of the complexity, only a few efficient calculations are involved, such as the 3D point projection, average calculation and EigenValue Decomposition. The experiments demonstrate the effectiveness of the proposed descriptor.

II. Our method

Local Reference Frame (LRF) has become a most common strategy in designing the 3D descriptors. Many famous descriptors have adopted the LRF for their descriptor extraction, such as the SHOT [11], the USC [12], and the RoPS [13]. Whereas, some traditional methods were based on the division of the 3D volumes. One issue is that this division often leads to empty bins when the resolution is high.

Our method belongs to the LRF, we first adopt the method of SHOT [11] to estimate a robust LRF. Then, we project the 3D surface points onto the tangent plane and divide the plane into mesh grids based on the LRF. Then, we record the average projection distances with respect to each mesh grid. Finally, the descriptor is generated by these averaged projection distances. In the following, we describe each step in detail.

A. Local Reference Frame (LRF) Estimation

We adopt the LRF in our local descriptor. Here, we begin by giving the definition of the LRF. First, for a given basis point p_b (an interest point), we collect the neighboring points by a radius R_b. In particular, in the 3D space, the searching space is a searching 3D sphere with the radius R_b and the center p_b, defined as the support region. With n points in the support region, the covariance matrix $C_{n\times n}$ is calculated as:

$$C_{n\times n}(p_b) = \frac{1}{\sum\limits_{i:d_i\leq R_b}(R_b - d_i)} \sum_{i:d_i\leq R_b} (R_b - d_i)\cdot(p_i - p_b)(p_i - p_b)^T \tag{1}$$

where $d_i = \|p_i - p_b\|$. Here, the p_b and p_i correspond to the 3D coordinates of the basis point and its neighboring points.

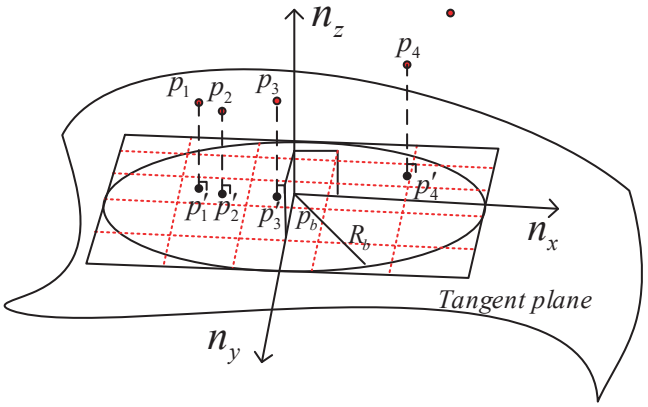


Fig. 1. LHID descriptor: *Local Height Image*(LHI) is established on LRF. The search region is a circle on the tangent plane. LHI is generated by the outer square of the circle. The pixel value of LHI is the averaged projection distance of the containing 3D points. Please refer to the text for the details.

Second, the matrix $C_{n\times n}$ is decomposed by EVD. The normal of the basis point is given by the eigenvector corresponding to the smallest eigenvalue, which is denoted as z axis with ambiguous signs. We form vectors by connecting all the p_i to p_b. The sign is then determined by the sign of the majority vectors. The x axis corresponds to the maximum eigenvalue, which is disambiguated by the same way. Finally, the y axis is the cross product of z and x. Please refer [11] for more details. The final LRF has three orthogonal vectors n_x, n_y, n_z, with n_z denotes the normal.

B. Local Height Image Generation

The tangent plane is uniquely determined by the basis point p_b and its normal n_z. The other two orthogonal components n_x, n_y are located within the plane. The points that fall into the support region are projected onto the plane along the normal vector. Equation 2 is the projection transformation.

$$p'_i = \begin{bmatrix} n_x^T \\ n_y^T \end{bmatrix} \cdot p_i, \tag{2}$$

where p'_i is the projected 2D point computed by p_i. Fig. 1 draws an illustration, in which we show the LRF (n_x, n_y, n_z), the tangent plane, the projection of the points and the circle projected from the searching sphere. Then, we define a square within the tangent plane as the outer bounding box of the circle. The length of the square is $2R_b$. We uniformly divide the square into mesh grids (Fig. 1, red lines). Next, we want to assign values to each of the grid. To achieve this, we collect all the projected points p'_i within a grid and calculate the mean projection distance $\bar{d}_{(i,j)}$. Here, the i, j index the mesh cell. The projection distance is indicated by the length of the dashed line in Fig. 1. The $\bar{d}_{(i,j)}$ is defined as:

$$\bar{d}_{(i,j)} = \frac{1}{k} \sum_{p'_i \in Mesh(i,j)} \left| n_z^T \cdot p_i \right|, \tag{3}$$

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

where p_i and p_i' denote the 3D coordinates of a point and its projected 2D version, respectively. k is the total number of points within a mesh cell $Mesh(i, j)$.

Now, let us consider the mesh cell as an image pixel, and the intensity of the pixel is the value of that cell. Then, we define the mesh as the *Local Height Image (LHI)*. In particular, we divide the mesh as $m \times m$, thus the resolution of the LHI image is also $m \times m$.

The value of the pixels is the averaged distance, which is robust to different sizes under noisy inputs. Meanwhile, the LHI generation is computationally cheap, only several dot product and average operations are involved.

C. Local Height Image Smoothing

To further improve the robustness against the leaking points, noises, and varying mesh resolutions. We smooth the LHI image with a gaussian filter.

$$H(u, v) = e^{-\frac{u^2}{2\sigma_x^2} - \frac{v^2}{2\sigma_y^2}}, \tag{4}$$

where we suppose that p_i' obeys the uniform distribution $U(-R_b, R_b)$ in the n_x and n_y directions, respectively. So we set $\sigma_x = \sigma_y = \frac{2R_b}{\sqrt{12}}$. The smooth not only increases the robustness, but also improves the performances in dealing with different point densities.

III. Experiments

In this paper we compare our method with several state-of-the-art 3D descriptors implemented in open source library *Point Cloud Lirbary* (PCL) [21] (version 1.8.1), including the Spin Image (SI) [7], the FPFH [1], the 3DSC [8], the USC [12], the SHOT [11] and the RoPS [5]. We use the common Bologna Dataset [22] to test the performances under different scenarios, including the noise free, the Gaussian noise, the varying mesh resolutions and the shot noise [10]. We also analyze the complexity of each descriptor.

The dataset consists of scenes. Each scene consists of several models. For each point in a scene, we can locate it in the model. In other words, we can establish the ground-truth correspondences between a scene and its containing models. Therefore, the matched correspondence can be compared with the ground-truth correspondence for the evaluation. Specifically, as proposed in [22], in each experiment, we randomly select 1000 key points in a model. For example, if a scene consists of 3 models, then we established 3000 ground-truth correspondences between the 3 models and the scene. Now, we want to find the correspondences by matching the descriptors. To do so, we extract descriptors of all the key points, 3000 in the scene, 3000 in the models, and matches them by comparing the descriptor distances. To find a match, for each scene point, we find the smallest and the second smallest distances in the models. Then, we compute the ratio between the two distances. A smaller ratio means a more reliable correspondence [22]. So a correspondence can only be established if the ratio is smaller than a threshold. The performance are reported by

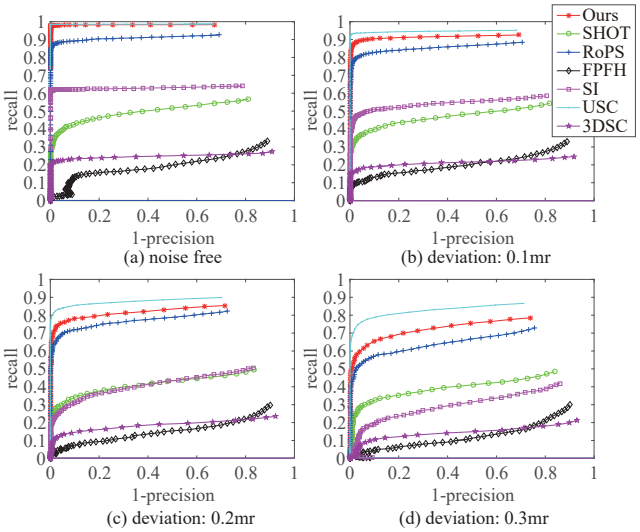


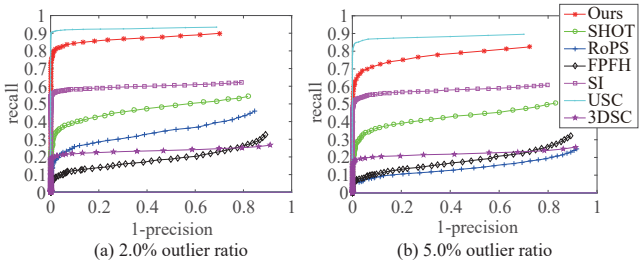
Fig. 2. Gaussian Noise Experiment: *Precision-Recall* curves.

Recall versus *1-Precision* curves (RPC) [23]. As proposed in [10], the recall is the ratio of the number of correct matches to the number of corresponding features. The precision is the ratio of the number of correct matches to the number of matches. RPC is obtained by increasing the threshold from 0 to 1. For the fairness, we use the same support radius and experiment environment. All the descriptors are implemented in C++ based on PCL. We use the same set random points. We use the randomized kd-tree algorithm in the *Fast Library for Approximate Nearest Neighbors* (FLANN) [24] to find two-nearest features. All parameters are set in units of point cloud density, sometimes also referred to as mesh resolution (mr). We run our experiment on a machine with i5-7500 and 8.00GB memory.

A. Gaussian Noise Experiment

We use Bologna dataset with 6 models ('Armadillo', 'Asian Dragon', 'Thai Statue', 'Bunny', 'Happy Buddha', 'Dragon') and 45 scenes taken from the *Stanford 3D Scanning Repository* [11], [12], [13]. In the experiment, we fixed search radius to 15mr and the other parameters were set to default in PCL. The parameters of the 3DSC and the USC were set according to [12]. The dimension of our descriptor was set to 400 (20×20). The performance of all methods was assessed by noise free data and Gaussian noise data with standard deviation of 0.1mr, 0.2 mr and 0.3 mr. Fig. 2 shows the RPC results. Our method and the USC have similar best performance to noise free data. As noise increases, the USC keeps the highest Recall. Our method and the RoPS are also robust to noise, ranking as second and third, respectively. The robustness of our descriptor is due to the fact that the pixel of LHI represents the average height. The FPFH is sensitive to noise due to its strong dependency on normals and angles.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE



(a) 2.0% outlier ratio (b) 5.0% outlier ratio

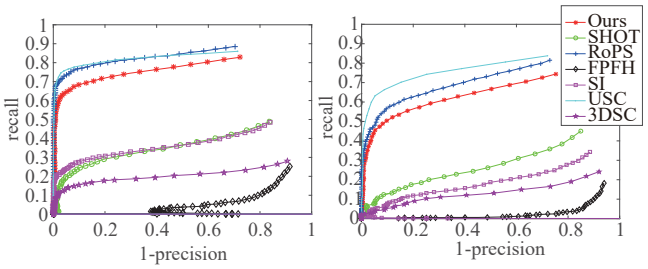
Fig. 3. Shot Noise Experiment: *Precision-Recall* curves.

B. Shot Noise Experiment

We also analyzed the performance on shot noise data [10]. We added shot noise with 2.0% and 5.0% outlier ratio to the scene. The amplitude was set to 6mr. Fig. 3 shows the results. It is clear that the USC has the best performance, followed by Ours, the SI, and the SHOT. The Recall rate of our descriptor can reach to 0.9 and 0.8 in the case of 2.0% and 5.0% outlier ratio, respectively. The robustness of our method to shot noise is due to the Gaussian filtering. The result is consistent with the conclusion in [10], that the RoPS and the FPFH are sensitive to shot noise. The USC has a better performance than the 3DSC due to the adoption of the LRF.

C. Cross Resolutions Experiment

We resampled the scene to $\frac{1}{2}$, $\frac{1}{4}$ of the original point density. We set the dimension of our descriptor to be 196 (14×14) and search radius to be 25mr. Fig. 4 reports the results. The experiment shows that the USC and the RoPS are very robust. Our method is close to their performance. In order to resist to different resolutions, our method tunes the size of LHI pixel. The Recall rate can reach to 0.8 and 0.7 in the case of $\frac{1}{2}$ decimation and $\frac{1}{4}$ decimation, respectively.



(a) 1/2 decimation (b) 1/4 decimation

Fig. 4. Cross Resolution Experiment: *Precision-Recall* curves.

D. Efficiency

We used all model points with different search radiuses to assess the efficiency of every descriptor. Notably, the input data was the raw point cloud without any additional information, such as the normals and the triangulations. Moreover, some methods may spend a period of fixed time, which is irrelevant to the descriptor extraction. For example, some methods require normal estimations such as the SI, the FPFH and the 3DSC while some other methods even need triangulations, such as the RoPS. Whereas, the USC and ours do not require either of these operations from the input data. Here, we refer this period of time as irrelevant time. As this irrelevant time is fixed per data for a method, for the fairness, in order to eliminate the impact of irrelevant time on the evaluation of complexity, for each data we extracted a large amount of point features to reduce the proportion of irrelevant time to the whole process.

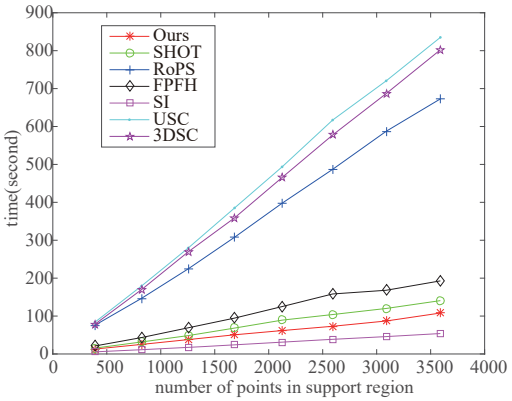


Fig. 5. Efficiency Experiment: Time spent versus number of points in the support region. The slope of the approximate line represents the complexity of different methods. The smaller slope corresponds to a lower complexity.

We compared the efficiency by varying the number of the points in the support region. We obtained the efficiency curve by changing the search radius. Fig. 5 reports the running time of different methods with respect to different number of points in the support region.

We can see that the amount of the running time is proportional to the number of points in the support region for all methods. The slope of the approximated line can reflect the efficiency of methods. The USC is the slowest, followed by the 3DSC and the RoPS. In contrast, the SpinImage (SI) is the most efficient method. Our method ranks at the second. However, our method is much more accurate than SpinImage as discussed in the previous sections. Our method keeps a relatively high efficiency due to its simple yet effective two-dimension grid division and omitting complex weights calculation. The experiments indicate that our method can achieve a good balance regarding the accuracy and the complexity.

IV. Conclusion

We have presented a novel 3D local feature descriptor, Local Height Image Descriptor (LHID), which can achieve a good performance while maintains the efficiency. The descriptor is based on LRF and LHI. The experimental results demonstrated that our method had strong robustness similar to the USC and the RoPS. With respect to the speed, our descriptor is faster several times than the most state-of-the-art algorithms. Our method achieves a good balance of accuracy and complexity, which is advantageous for the scenarios of mass point cloud data processing and could facilitate real-time applications.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

REFERENCES

[1] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *IEEE International Conference on Robotics and Automation*, 2009, pp. 3212–3217.

[2] Yulan Guo, Ferdous Sohel, Mohammed Bennamoun, Jianwei Wan, and Min Lu, "An accurate and robust range image registration algorithm for 3d object modeling," *IEEE Trans. on Multimedia*, vol. 16, no. 5, pp. 1377–1390, 2014.

[3] Hansung Kim and Adrian Hilton, "Influence of colour and feature geometry on multi-modal 3d point clouds data registration," in *International Conference on 3D Vision (3DV)*, 2014, vol. 1, pp. 202–209.

[4] Luıs A Alexandre, "3d descriptors for object and category recognition: a comparative evaluation," in *Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, vol. 1, p. 7.

[5] Yulan Guo, Ferdous Sohel, Mohammed Bennamoun, Min Lu, and Jianwei Wan, "Rotational projection statistics for 3d local surface description and object recognition," *International journal of computer vision*, vol. 105, no. 1, pp. 63–86, 2013.

[6] Alioscia Petrelli and Luigi Di Stefano, "On the repeatability of the local reference frame for partial shape matching," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2244–2251.

[7] Andrew E. Johnson and Martial Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999.

[8] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik, "Recognizing objects in range data using regional point descriptors," *European Conference on Computer Vision (ECCV)*, pp. 224–237, 2004.

[9] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, and Jianwei Wan, "3d object recognition in cluttered scenes with local surface features: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2270–2287, 2014.

[10] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, Jianwei Wan, and Ngai Ming Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 66–89, 2016.

[11] Federico Tombari, Samuele Salti, and Luigi Di Stefano, "Unique signatures of histograms for local surface description," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 356–369.

[12] Federico Tombari, Samuele Salti, and Luigi Di Stefano, "Unique shape context for 3d data description," in *Proceedings of the ACM workshop on 3D Object Retrieval*, 2010, pp. 57–62.

[13] Yulan Guo, Ferdous A Sohel, Mohammed Bennamoun, Jianwei Wan, and Min Lu, "Rops: A local feature descriptor for 3d rigid objects based on rotational projection statistics," in *International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 2013, pp. 1–6.

[14] John Novatnack and Ko Nishino, "Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images," *European Conference on Computer Vision (ECCV)*, pp. 440–453, 2008.

[15] Andrei Zaharescu, Edmond Boyer, and Radu Horaud, "Keypoints and local descriptors of scalar functions on 2d manifolds," *International Journal of Computer Vision*, vol. 100, no. 1, pp. 78–98, 2012.

[16] Chin Seng Chua and Ray Jarvis, "Point signatures: A new representation for 3d object recognition," *International Journal of Computer Vision*, vol. 25, no. 1, pp. 63–85, 1997.

[17] Bisheng Yang, Zheng Wei, Qingquan Li, and Jonathan Li, "Automated extraction of street-scene objects from mobile lidar point clouds," *International Journal of Remote Sensing*, vol. 33, no. 18, pp. 5839–5861, 2012.

[18] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz, "Aligning point cloud views using persistent feature histograms," in *IEEE International Conference on Intelligent Robots and Systems*, 2008, pp. 3384–3391.

[19] Umberto Castellani, Marco Cristani, and Vittorio Murino, "Statistical 3d shape analysis by local generative descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2555–2560, 2011.

[20] Michael M Bronstein and Alexander M Bronstein, "Shape recognition with spectral distances," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 1065–1071, 2011.

[21] Radu Bogdan Rusu and Steve Cousins, "3d is here: Point cloud library (pcl)," in *IEEE International Conference on Robotics and automation (ICRA)*, 2011, pp. 1–4.

[22] Samuele Salti, Federico Tombari, and Luigi Di Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014.

[23] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[24] Marius Muja and David G Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration.," *VISAPP (1)*, vol. 2, no. 331-340, pp. 2, 2009.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE