# Point Cloud Registration Using A Viewpoint Dictionary

David Avidar, David Malah, and Meir Barzohar
Department of Electrical Engineering, Technion, Haifa 32000, Israel

*Abstract* – **The use of 3D point clouds is currently of much interest. One of the cornerstones of 3D point cloud research and applications is point cloud registration. Given two point clouds, the goal of registration is aligning them in a common coordinate system. In particular, we seek in this work to align a sparse and noisy local point cloud, created from a single stereo pair of images, to a dense and large-scale global point cloud, representing an urban outdoors environment. The common approach of keypoint-based registration, tends to fail due to the sparsity and low quality of the stereo local cloud. We propose here a new approach. It consists of the creation of a dictionary of much smaller clouds using a grid of synthetic viewpoints over the dense global cloud. We then perform registration via an efficient dictionary search. Our approach shows promising results on data acquired in an urban environment.**

**Keywords – ICP, LiDAR, localization, 3D registration, SfM, sparse point cloud, stereo reconstruction, urban environment**

## I. INTRODUCTION

The process of point cloud registration is a crucial step in many applications involving 3D point clouds. A few examples of such applications are modeling of indoors or outdoors environments, robot navigation, obstacle avoidance, and landscape surveying. Given two (or more) 3D point clouds, the goal of point cloud registration is finding transformations which align the clouds in a common coordinate system. For rigid registration, these transformations include elements of rotation and translation (6DoF).

The registration challenge we face in this work involves a dense, large-scale point cloud, representing an outdoor environment (e.g., a neighborhood, a small town), and a sparse and noisy local cloud with limited range and field of view. The large-scale ("global") cloud may be created using Structure-from-Motion (SfM) or a LiDAR sensor, and the local cloud is typically created using stereo reconstruction. A challenging aspect of such registration is the substantially different properties of the global and local clouds. While the global cloud is dense and detailed, the local cloud may be very sparse and noisy. Moreover, the local cloud usually suffers from missing data, especially on smooth, uniform surfaces. Because of such differences, common approaches to point cloud registration, such as keypoint or plane-based registration, mostly fail. For example, high levels of noise degrade the unique localization of keypoints and cloud sparseness hinders plane-based registration. As a result, descriptor matching (and using the putative correspondences for finding an initial registration) becomes especially challenging.

Our proposed registration approach faces the previously mentioned challenges by utilizing the geometric information of the sparse local cloud as a whole. We use a synthetic grid of viewpoints over the global cloud in order to create from it a dictionary of much smaller clouds commensurate with the local cloud in terms of range and field of view. Given a sparse local cloud, we can now consider the registration problem as a dictionary search problem.

Keypoint-based registration involves establishing local correspondences between sparse keypoints in the point clouds, using feature descriptors. In recent years, several keypoint detectors for 3D surfaces have been proposed. A recent survey of keypoint detectors [1] has performed an evaluation of such detectors. Some keypoint detection methods are typically based on the eigenvalues of the covariance matrix that characterizes local point neighborhoods. A related way of characterizing the saliency of a point in the cloud is by its neighborhood Surface Variation [2], defined as the ratio between the smallest of the eigenvalues of the neighborhood covariance matrix and their sum. Thresholding of this ratio is used to select keypoints.

Once keypoints have been detected, a common next step is the computation of descriptors. A recent comparative study of 3D shape descriptors [3] has evaluated the performance of ten popular descriptors over several databases. In terms of time complexity, the best performing descriptor was FPFH (Fast Point Feature Histogram) [4] while RoPS (Rotational Projection Statistics) [5] was reported to have good results over several different types of datasets. Spin-Images [6] are another example of a popular, albeit less recent, 3D shape descriptor, which may be used to encode the geometry of a local 3D neighborhood as a 2D image.

Next, the keypoint descriptors may be used to establish correspondences between different point clouds. Then, a possible approach for finding an initial coarse registration, using the estimated correspondences, is by applying RANSAC (RANdom SAmple Consensus) [7]. Finally, an iterative refinement step may be applied using ICP (Iterative Closest Point) [8] or one of its variants [9]. Keypoint-based approaches for registration are often sensitive to noise, occlusion, and to point density differences. We have found them to be unreliable for registration of the stereo-reconstructed point clouds we have.

Another approach for point cloud registration involves matching geometric primitives, such as lines [10] or planes [11], detected in the scenes to be registered. These approaches are viable when all the scans are relatively of high quality, e.g., acquired using LiDAR. However, for stereo-reconstructed point clouds, which are very sparse and noisy, reliable detection of line or plane features is challenging.

978-1-5090-2152-9/16/$31.00 ©2016 IEEE

An approach which shares some similarities with ours was proposed for image-based localization given a Structure from Motion point cloud of a large-scale scene [12]. Their approach consisted of using synthetic viewpoints over a large-scale scene reconstructed using SfM, where the 3D points are characterized by SIFT descriptors [13]. They use synthetic viewpoints to efficiently index the 3D points with the goal of view (image) registration to the 3D scene. We, however, use synthetic viewpoints to transform the large-scale 3D scene into a dictionary of smaller point clouds with limited range and field of view. While their approach relies on the availability of image-based descriptors, our approach requires only the 3D point coordinates of the global and local clouds and allows use of other acquisition techniques such as LiDAR.

The rest of the paper is organized as follows. Section II describes a conventional keypoint-based registration. In Section III, we present our proposed approach: viewpoint dictionary based registration, in detail. Sections IV and V discuss our experimental setup and registration results, respectively. We summarize, draw conclusions, and outline directions for future work in Section VI.

## II. KEYPOINT-BASED REGISTRATION

In this section, we briefly outline the keypoint-based registration process that we used to attempt registration between a dense, large-scale, "global" cloud and a sparse and noisy "local" point cloud resulting from stereo reconstruction. An example of such clouds is shown in **Fig. 1**.

The first step is keypoint detection, used to identify a sparse set of "interesting" points in both global and local clouds. For this step, we used a Surface Variation based detector [2], mentioned above. Points on flat surfaces are characterized by small Surface Variation values while salient points, such as 3D corners, have larger values. Keypoints were selected as points with Surface Variation values that exceed a set threshold value.

Next, a Spin-Image descriptor [6], mentioned earlier, was computed for each of the identified keypoints.

Keypoint correspondences are established by computing the similarity between each Spin-Image of the local cloud and the Spin-Images of the global cloud.

The next step used is a RANSAC procedure [7], with the purpose of filtering out erroneous correspondences (outliers) and estimating an initial (coarse) registration between the local and global clouds. As a final step, ICP [8] is used to refine the registration.

## III. PROPOSED APPROACH: VIEWPOINT DICTIONARY BASED REGISTRATION

We now describe the steps of the proposed registration approach. We use a grid of synthetic viewpoints over the global cloud to transform it into a dictionary of smaller clouds and solve the registration problem via a dictionary search. A block diagram of the proposed method is shown in **Fig. 2**. If real-time localization is sought, i.e., finding the location from which the local cloud was acquired, in relation to the global cloud, the first two steps in the block diagram can be done offline. The third step may be done either offline or online, as will be discussed in Subsection C, and the remaining steps are done online, once a local cloud is received.

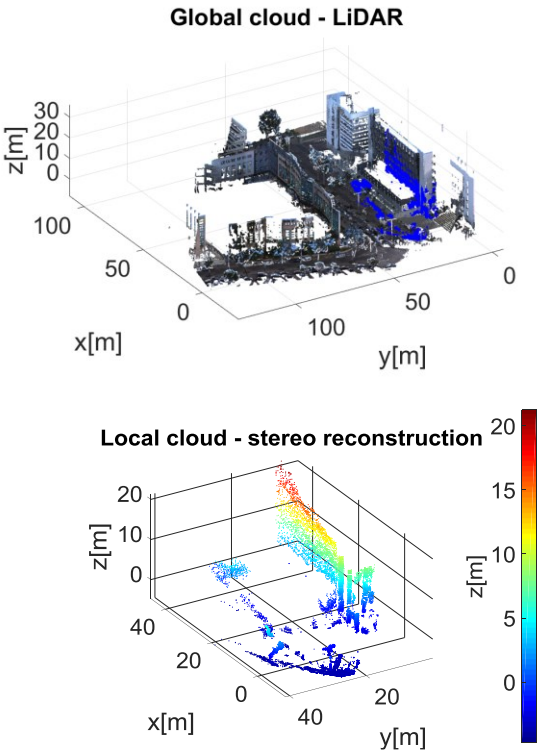


**Fig. 1** – Top: a LiDAR global cloud representing an urban environment. The pose of a stereo local cloud is shown in blue. Bottom: stereo local cloud example.

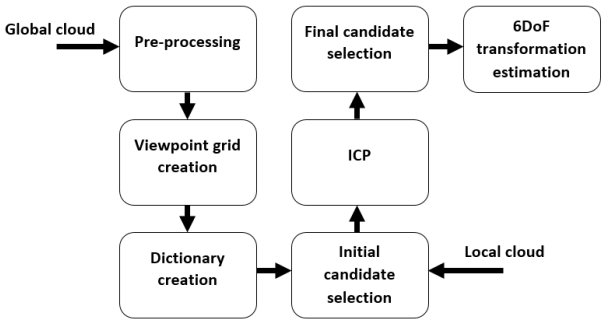


**Fig. 2** – Block diagram of the viewpoint dictionary based registration approach.

### *A. Pre-processing*

The pre-processing step involves subsampling of the global cloud. Typically, dense clouds of large-scale scenes, reconstructed using SfM or LiDAR, may contain millions or even tens of millions of points. Since the computational resources required to work with such clouds are currently unfeasible when aiming at real-time performance, we perform a random subsampling of the global cloud. In our experience, such subsampling is legitimate while the relevant underlying geometric information contained in the cloud (e.g., ground

surface, building walls etc.) is preserved. We note that more sophisticated subsampling strategies are possible [14]. Pre-processing may include also denoising of the global cloud data.

### *B. Viewpoint grid creation*

In this step, we create a grid of synthetic viewpoints over the global cloud. We aim for the synthetic viewpoints to resemble typical viewpoints of observers moving on foot or by vehicle throughout the area represented by the global cloud. In this step, we assume that the global cloud is given such that gravity is approximately pointing in the negative direction of the z-axis. Initially, we create a regular Cartesian grid in the x,y plane, over the global cloud. The distance between adjacent grid points is denoted as $d_{grid}$. For each grid point, we assign a z value equal to that of the global cloud point, whose projection on the x,y plane is closest. The closest point's normal vector is used as the grid point's normal vector. Normal vector estimation is done based on Principal Component Analysis of the local neighborhood [15].

To avoid synthetic viewpoints in unlikely positions, such as on rooftops, on walls, or in vegetation, two filtering steps are performed. Viewpoints whose height above the ground exceeds some threshold value are filtered out. The ground height may be estimated based on plane detection in a local neighborhood. In addition, viewpoints whose normal vector is substantially different than those of its neighboring viewpoints are rejected as well.

Since we want to place the synthetic viewpoints at some user-defined height above the ground, we move each remaining viewpoint in the direction of its assigned normal vector.

We note that the grid creation process we describe is not ideal but was chosen for its simplicity and because it was found to give reasonable results for typical urban global clouds. More advanced approaches, which are possible because this step is done offline, may involve a more robust ground surface detection.

### *C. Dictionary creation*

We use the synthetic viewpoint grid to compute a dictionary of small clouds ("dictionary clouds") which represent the 3D scenes that would be viewed by an observer from the synthetic viewpoints. The observer is assumed to have a limited viewing range and a restricted field of view - for example, a maximal range of $r_{max} = 60m$ and a $\sphericalangle FoV = 75^o$ horizontal angle of view ("yaw"), where the viewpoint normal defines the vertical direction $\hat{\boldsymbol{z}}_i = \boldsymbol{n}_i$ (see **Fig. 3**) where $i$ is the index of the corresponding viewpoint $\boldsymbol{pv}_i$. For each viewpoint, $\boldsymbol{pv}_i$, a number of viewing directions (e.g., $N_{dir} = 12$), denoted by $\hat{\boldsymbol{x}}_{ij}$ $(j \in 1,2,\dots,N_{dir})$, are selected in the plane perpendicular to $\hat{\boldsymbol{z}}_i$. Thus, a unique right-hand frame of reference is created for each dictionary cloud, using a cross product: $\hat{\boldsymbol{y}}_{ij} = \hat{\boldsymbol{z}}_i \times \hat{\boldsymbol{x}}_{ij}$. A vertical may be defined in addition to the horizontal one, but is not shown here for simplicity.

Using its unique frame of reference $(\hat{\boldsymbol{x}}_{ij}, \hat{\boldsymbol{y}}_{ij}, \hat{\boldsymbol{z}}_i)$, the angle of view $(\sphericalangle FoV)$, and the maximal range $(r_{max})$, each dictionary cloud is cropped from the global cloud.

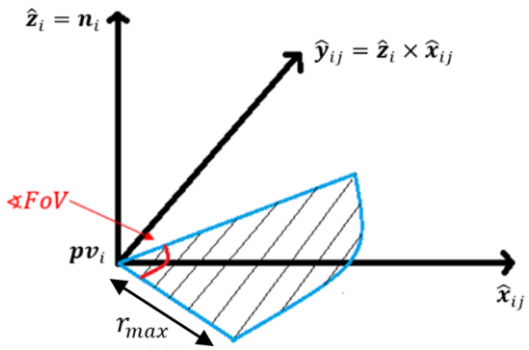


**Fig. 3** – A schematic drawing of the right-hand frame of reference defined for a dictionary cloud at a synthetic viewpoint $\boldsymbol{pv}_i$.

We then use a Hidden Point Removal (HPR) algorithm [16]. Intuitively, given a set of points and a viewpoint, the HPR algorithm identifies a subset of points that are visible from that viewpoint. Our motivation for doing so is that a stereo-reconstructed point cloud may only contain points which are visible from the viewpoint. Thus, keeping dictionary cloud points which will not be included in a stereo reconstruction will not assist in matching a local cloud to a dictionary cloud. Furthermore, to avoid including dictionary clouds that are not informative, we do not include those whose number of points is below a threshold number.

The dictionary clouds are stored in a uniform pose such that the viewpoint is located at the origin and the frame of reference $(\hat{\boldsymbol{x}}_{ij}, \hat{\boldsymbol{y}}_{ij}, \hat{\boldsymbol{z}}_i)$ is aligned with the standard Cartesian frame. This way, in the following step where we compare a given local cloud to the dictionary clouds, we avoid realigning the local cloud with each dictionary cloud in the matching process.

Dictionary creation may be performed either online or offline. When creating the dictionary offline (i.e., using a "static" dictionary), its storage may entail high memory requirements. This also depends on the size of the area represented by the global cloud, the number of synthetic viewpoints used and the number of viewing directions per viewpoint. In contrast, it is also possible to create the dictionary online (i.e., using a "dynamic" dictionary). Given an initial guess as to the location of the local cloud (e.g., GPS reading), dictionary clouds may be created in this case only for synthetic viewpoints in a relatively small surrounding area (e.g., within a radius of $30m$). This may significantly reduce memory requirements at the price of additional online computations. An advantage introduced by using a dynamic dictionary is the possibility of adapting the dictionary creation parameters (such as $N_{dir}$) while navigating. In addition, assuming a dense viewpoint grid is created offline (Subsection B), selection of viewpoints to be used for creating the dictionary may be adaptive as well.

### *D. Initial candidate selection*

Once we are given a local point cloud, which may be created using stereo reconstruction, we wish to select an initial set of candidate dictionary clouds that best resemble it. By doing so, we significantly reduce the computational load of the following ICP step (Subsection E). We first transform the local cloud to

the uniform pose described in the previous subsection. We assume that the local cloud viewpoint is known in its local reference frame and that its analog of the reference frame $(\hat{\boldsymbol{x}}_{ij}, \hat{\boldsymbol{y}}_{ij}, \hat{\boldsymbol{z}}_i)$ can be estimated by detecting the local ground plane (that defines $\hat{\boldsymbol{z}}_i$) and the center axis of the field of view (that defines $\hat{\boldsymbol{x}}_{ij}$).

Once the local cloud has been transformed to the uniform pose, we compute the RMSE over the nearest-neighbor distances between its points and the points of each dictionary cloud. A low RMSE value indicates a possibly matching dictionary cloud, while a high RMSE value indicates a different 3D scene and an unlikely match. However, lowest RMSE is not a perfect indicator of a match between a dictionary cloud and the local cloud due to the substantial noise in the latter. Thus, initial candidates for additional processing are selected as a fraction (e.g., 0.1) of the dictionary clouds whose RMSE scores are the lowest.

It should be noted that knowing the reference frame of a candidate dictionary cloud and the reference frame of the local cloud allows computing a potential coarse registration between the local and global clouds.

### *E. ICP and final candidate selection*

We use the ICP algorithm to refine the registration between the local cloud and the set of candidate dictionary clouds selected in the previous step. This allows our registration accuracy to be better than what is dictated by the selection of viewpoint grid spacing $(d_{grid})$ and the number of viewing direction per viewpoint $(N_{dir})$. The result of the ICP step is a refining transformation (rotation and translation) of the local cloud, which brings the RMSE between it and a dictionary cloud to a local minimum. The final dictionary cloud candidate is selected to be the one with lowest RMSE, following the ICP application, with respect to the local cloud as described in the previous section.

Given the transformations which brought the selected dictionary cloud and the local cloud to the uniform pose, and the refining transformation returned by the ICP algorithm, we can compute the equivalent 6DoF transformation which aligns the local and the global clouds.

## IV. EXPERIMENTAL SETUP

In our experiments we used a large-scale LiDAR cloud (SfM is possible as well) representing an urban scene, and two types of local clouds: simulated clouds and clouds reconstructed using stereo images that were acquired at the actual scene represented by the global cloud (see example in **Fig. 1**).

### *A. Simulated local clouds*

In order to simulate local clouds, the x,y coordinates of ten viewpoints in the global cloud are manually selected. Then, similarly to what is done during grid creation (Section III.B), we assign z values and normal vectors. The normal vectors are used as $\hat{\boldsymbol{z}}_i$, as mentioned before, and the $\hat{\boldsymbol{x}}_{ij}$ vectors are chosen arbitrarily in the plane perpendicular to $\hat{\boldsymbol{z}}_i$ such that a local reference frame is defined for each viewpoint. Using the reference frame, a local cloud can be cropped from the global cloud with respect to range ($r_{max} = 60m$) and field of view ($\sphericalangle FoV = 75^o$) limitations. Next, the cropped cloud is subsampled such that its point density is inversely proportional to the distance from the viewpoint. Spatial Gaussian noise is added with uniform distribution in orientation, but the noise standard deviation is proportional to the distance from the viewpoint. The maximal noise standard deviation, which applies to points in the maximal range of $r_{max} = 60m$, is denoted as $\sigma_{max}$. As will be shown in Section V, three different levels of noise were tested: $\sigma_{max} = 0m\ (no\ noise), 3m, 5m$. As a final step, the HPR [18] algorithm is applied.

### *B. Stereo local clouds*

Stereo reconstruction was used to create seven actual local clouds similar to the one shown in Fig. 1. A sequence of stereo image pairs was acquired from a single location in the scene while changing the viewing direction from pair to pair. The maximal range of the reconstructed stereo cloud was $r_{max} = 60m$ and the horizontal field of view was $\sphericalangle FoV = 75^o$.

## V. REGISTRATION RESULTS

The transformations we compute to perform the registration have six degrees of freedom - three for rotation and three for translation. We represent the error in localization as the Euclidean distance in meters between the actual (ground truth) local cloud's viewpoint and the estimated viewpoint resulting from the registration. The errors in rotation are represented using three intrinsic Euler angles, $R_{error} = R_z(\alpha)R_y(\beta)R_x(\gamma)$, where $R_{error}$ represents the rotation matrix from the true orientation to the estimated orientation, and $(\alpha, \beta, \gamma)$ represent yaw, pitch and roll respectively. $R_x(\gamma)$ thus represents a rotation by an angle $\gamma$ around the x-axis of the ground truth frame of reference. Note that since intrinsic rotation angles are used, the frame of reference rotates after each elemental rotation.

Registration results of the proposed viewpoint dictionary based approach are presented and compared below to the results of the keypoint-based approach presented in Section II. Both approaches are tested on both the simulated local clouds and the stereo local clouds.

### *A. Registration Results: Simulated local clouds*

The registration results of the proposed approach for the ten simulated local clouds, with $\sigma_{max} = 5m$, are summarized in **Table 1**. The average localization error was $1.01m$ with an STD of $0.79m$, averaged over the ten simulated local clouds. The maximal localization error was $2.28m$.

**Table 1** - Registration results of the proposed approach for the simulated local clouds, using a maximal noise STD of $\boldsymbol{\sigma_{max} = 5m}$.

| | Localization Error [m] | Yaw Absolute Error [deg] | Pitch Absolute Error [deg] | Roll Absolute Error [deg] |
|---|---|---|---|---|
| Mean over 10 clouds | 1.01 | 0.63 | 1.18 | 0.45 |
| STD over 10 clouds | 0.79 | 0.87 | 0.56 | 0.31 |

To compare the registration results of the proposed approach and the keypoint-based approach we count the number of simulated local clouds, for which the localization error is lower than $3m$. This is repeated for each level of noise (see **Table 2**). While for the keypoint-based approach the number of successful registrations substantially declines when larger noise is introduced, the viewpoint dictionary based approach shows better robustness to noise. We note that for the keypoint-based approach, when the localization error was larger than $3m$, it was mostly around several tens of meters due to failure to establish correct correspondences between the local and global clouds.

**Table 2** - comparison of registration results between keypoint-based registration and viewpoint dictionary based registration.

| Max. noise STD ($\sigma_{max}$) | | 0[m] | 2[m] | 5[m] |
|---|---|---|---|---|
| # Local clouds with localization error < 3m | Keypoints | 9/10 | 8/10 | 4/10 |
| | Viewpoint dictionary (proposed) | 10/10 | 10/10 | 10/10 |

### *B. Registration Results: Stereo local clouds*

Registration results of the viewpoint dictionary based approach for each of the seven stereo local clouds are shown in **Table 3**. As can be seen in that table, for five of the seven local clouds the localization error was less than $3m$ (with a mean of $1.5m$), while two of the seven local clouds were not successfully registered. In contrast, using the keypoint-based approach on the same data, none of the local clouds were successfully registered - all localization errors were as large as several tens of meters (the lowest error was $25m$). This demonstrates the advantage of the viewpoint dictionary based approach over the keypoint-based approach for noisy and sparse local clouds.

**Table 3** - Registration results of the proposed approach for the seven stereo local clouds.

| Local cloud # | Localization Error [m] | Yaw Error [deg] | Pitch Error [deg] | Roll Error [deg] |
|---|---|---|---|---|
| 1 | 2.96 | -6.35 | -3.40 | -6.00 |
| 2 | 74.94 | 135.83 | -3.93 | -2.75 |
| 3 | 2.41 | 4.18 | 2.87 | -2.85 |
| 4 | 0.48 | 0.94 | 1.90 | -1.31 |
| 5 | 0.51 | 2.23 | 1.59 | -3.29 |
| 6 | 1.16 | 3.13 | 0.45 | -2.02 |
| 7 | 12.10 | 51.65 | -0.19 | 3.87 |

## VI. CONCLUSION

In this paper, we have proposed a novel viewpoint dictionary based approach for the registration of a sparse and noisy stereo-reconstructed point cloud, and a dense large-scale urban point cloud. The proposed approach is compared to the common keypoint-based registration approach, and shown to achieve much better results in terms of registration accuracy. In future work, we intend to test the method's performance on larger datasets. In addition, we believe that the development of a dedicated descriptor for the dictionary and local clouds will benefit the method's performance in terms of accuracy, robustness and computational complexity.

## ACKNOWLEDGEMENTS

This work was supported by the Omek consortium, which is a program of the office of the chief scientist of the Israeli ministry of economy. The support by Nimrod Peleg and Yair Moshe of the Signal and Image Processing Lab (SIPL), is greatly appreciated. We would also like to thank Erez Nur, technical manager of Omek, for helpful discussions. Our research in Omek was done in collaboration with CEVA and Elbit Systems Land and C4I. We thank Elbit Systems for providing us with the point cloud data used in this paper.

## REFERENCES

[1] F. Tombari, S. Salti, L. Di Stefano, "Performance evaluation of 3D keypoint detectors," International Journal of Computer Vision., vol. 102(1-3), pp. 198-220, 2013.

[2] M. Pauly, M. Gross, L.P. Kobbelt, "Efficient simplification of point-sampled surfaces," In Proceedings of the conference on Visualization'02, pp. 163-170, IEEE Computer Society, 2002.

[3] Y. Guo et al., "A comprehensive performance evaluation of 3D local feature descriptors," International Journal of Computer Vision, vol. 116(1), pp. 66-89, 2016.

[4] R.B. Rusu, N. Blodow, M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," In Robotics and Automation, ICRA'09, IEEE International Conference, pp. 3212-3217, 2009.

[5] Y. Guo et al., "Rotational projection statistics for 3D local surface description and object recognition," International journal of Computer Vision, vol. 105(1), pp. 63-86, 2013.

[6] A.E. Johnson, M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21(5), pp. 433-449, 1999.

[7] Fischler, R.C, Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24(6), pp. 381-95, 1981.

[8] P.J. Besl, N.D. McKay, "Method for registration of 3-D shapes," In Robotics-DL tentative, pp. 586-606, International Society for Optics and Photonics, 1992.

[9] S. Rusinkiewicz, M. Levoy, "Efficient variants of the ICP algorithm," In Proc. 3rd International Conference on 3-D Digital Imaging and Modeling, pp. 145-152, 2001.

[10] W. Hansen, H. Gross, U. Thoennessen, "Line-based registration of terrestrial and airborne lidar data," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 37 Part B3a, pp. 161-166, 2008.

[11] T.A. Teo, S,H. Huang, "Surface-based registration of airborne and terrestrial mobile LiDAR point clouds," Remote Sensing, vol. 6(12), pp. 12686-12707, 2014.

[12] A. Irschara et al., "From structure-from-motion point clouds to fast location recognition," In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2599-2606, 2009.

[13] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of Computer Vision, vol. 60(2), pp. 91-110, 2004.

[14] H. Huang et al., "Consolidation of unorganized point clouds for surface reconstruction," In ACM transactions on graphics (TOG), vol. 28, No. 5, pp. 176-182, 2009.

[15] H. Hoppe et al., "Surface reconstruction from unorganized points," ACM, vol. 26, No. 2, pp. 71-78, 1992.

[16] S. Katz, A. Tal, R. Basri, "Direct visibility of point sets," ACM Transactions on Graphics (TOG), vol. 26(3), pp. 24-35, 2007.