LOCAL VOXELIZED STRUCTURE FOR 3D LOCAL SHAPE DESCRIPTION: A BINARY REPRESENTATION

Siwen Quan, Jie Ma, Fangyu Hu, Bin Fang, Tao Ma*

National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Automation, Huazhong University of Science and Technology, P. R. China

siwenquan@hust.edu.cn, majie@mail.hust.edu.cn, husthoofy@hust.edu.cn, {lisben_cyan, whumatao}@163.com

ABSTRACT

This paper proposes a novel binary descriptor named local voxelized structure (LoVS) for 3D local shape description. Unlike many previous local shape descriptors relying on geometric attributes such as curvature and normals, LoVS simply uses point spatial locations to encode the local shape structure represented by point clouds into bit string. Specifically, LoVS is computed on a local cubic volume around the keypoint. The orientation of the cubic is determined by a local reference frame (LRF) to achieve rotation invariance. Then, the cubic is uniformly split into a set of voxels. A voxel is attached with label 1 if there are points inside, otherwise, it produces a 0 bit. All these labels therefore integrates into the LoVS descriptor. We evaluate our method on three public datasets. On each dataset, the LoVS descriptor outperforms all other descriptors tested.

Index Terms— Binary descriptor, 3D local shape, voxel

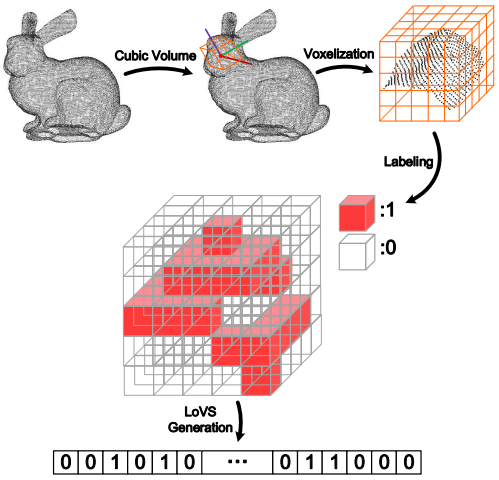


Fig. 1. An illustration of the LoVS binary descriptor.

1. INTRODUCTION

Shape description and matching are fundamental tasks in 3D computer vision. At present, an explosive growth of 3D sensors can be witnessed including many mobile devices such as Microsoft Kinect, Asus Xtion ProLive and Google Project Tango. These advances highlight the demand of crafting light-weight (binary) 3D local descriptors, especially in robotics and mobile phones.

3D local descriptors are feature vectors which represent the local shape geometry of a 3D object [1, 2]. A desired local shape descriptor should hold many peculiarities including being invariant to object's poses, distinctiveness, robustness to common nuisances. In the literature, numerous progresses can be found. *Spin images* [3] projects the points in a cylindrical volume on a 2D image through a plane that spins around the keypoint normal. *Fast point feature histograms* (FPFH) [4] calculates the statistical histograms of normal deviations between the keypoint and its neighbors. *Signature of histograms of orientations* (SHOT) [1] is the concatenation of normal deviation histograms which are calculated from numerous subspaces after space partition. *Rotational projection statistics* (RoPS) [5] encodes statistics from several 2D point density maps which are obtained via rotation and projection. Despite their effectiveness, they suffer from expensive footprint occupancy and poor matching efficiency.

A direct solution is arguably designing binary descriptors. In the 2D image domain, many binary feature descriptors [6, 7] were proposed whereas there are quite rare 3D binary ones. The very first 3D local binary proposal is the B-SHOT [8] descriptor, which transforms the SHOT descriptor with a binary converting algorithm. Compared with SHOT, B-SHOT requires 32 times lesser memory for its representation while being 6 times faster in feature descriptor matching. Unfortunately, owing to that B-SHOT is a direct quantization of SHOT, it therefore loses certain discriminative information and simultaneously inherits the demerit of SHOT, i.e., being sensitive to mesh resolution variation [5].

The corresponding author is Jie Ma (majie@mail.hust.edu.cn). This work is jointly supported by the Wisdom of Marine Science and Technology Foundation (Grant No. 2015HUST), and Shanghai Aerospace Science and Technology Foundation (Grant No. sast2016063).

Motivated by these considerations, we propose a new 3D local binary descriptor called local voxelized structure (LoVS) (Fig. 1). LoVS characterizes the local shape with a set of labeled voxels. In specific, we first establish a cubic volume at the keypoint. Then, we uniformly divide the cubic volume into several voxels and label a voxel as 0 or 1 by judging if there are points inside. The eventual generated LoVS bit string is the integration of these labels. The proposed LoVS differs from the existing methods from at least three aspects:

- LoVS represents the local shape geometry without extracting any complex geometric attributes, such as curvatures and normals (e.g., in FPFH [4] and SHOT [1]).
- Rather than the commonly used spherical volume [1, 5, 9], LoVS is computed in a cubic volume for uniform space partition.
- In contrast to the float-to-binary encoding method used in B-SHOT [8], the proposed LoVS descriptor instead performs 3D binary description from the intrinsic shape geometry perspective.

We finally justify our method on three popular datasets together with comparisons with the state-of-the-arts.

2. LOCAL VOXELIZED STRUCTURE

This section presents the proposed LoVS descriptor, which consists of cubic volume definition, voxelization and LoVS generation, as illustrated in Fig. 1.

2.1. Cubic Volume Definition

Our LoVS resorts to a local cubic volume with the purpose of splitting the neighboring space uniformly and efficiently. Considering that cubic volume is rarely used in the literature, we hereby introduce the relevant details. Let $\mathcal{P}$ be a point cloud, where p is a keypoint inside. The cubic volume should be orientated to attain rotation invariance, we therefore build a local reference frame (LRF) at the keypoint. In particular, we choose a recent mesh-independent LRF method [10] which is based on normal calculation and projected vector summarization (please refer to [10] for details). Assume that r is the support radius of the required spherical volume of the LRF $\mathcal{L}$, it is reasonable to first calculate a sphere-intersected surface using $\mathcal{Q}_{\mathcal{L}} = \{q_i : \|q_i - p\| \leq r, q_i \in \mathcal{P}\}$. Rather, we choose to calculate a larger-scaled one as:

$$\mathcal{Q}_{\mathcal{S}} = \{q_i : \|q_i - p\| \leq \sqrt{2}r, q_i \in \mathcal{P}\}. \quad (1)$$

The reasons are twofold. First, we could directly compute the desired local shape $\mathcal{Q}_{\mathcal{L}}$ for $\mathcal{L}$ by searching in $\mathcal{Q}_{\mathcal{S}}$, rather than performing greedy-search in point cloud $\mathcal{P}$. Second, due to that the cubic volume is also the inscribed cube of the spherical volume of $\mathcal{Q}_{\mathcal{S}}$, the cubic-intersected surface hence could be computed as:

$$\mathcal{Q}_{\mathcal{C}} = \{q_i : |q_i.x| \leq r, |q_i.y| \leq r, |q_i.z| \leq r, q_i \in \mathcal{Q}_{\mathcal{S}}{}'\}, \quad (2)$$

where $q_i.x$, $q_i.y$ and $q_i.z$ respectively denote the x, y and z values of q_i in the LRF coordinate system, $|\bullet|$ represents the absolute value of $\bullet$, and $\mathcal{Q}_{\mathcal{S}}{}'$ is the transformed $\mathcal{Q}_{\mathcal{S}}$ with respect to the reference frame $\mathcal{L}$ (to achieve rotation invariance). Here, since r also controls the size of our cubic-intersected surface $\mathcal{Q}_{\mathcal{C}}$, we name it as the support length of the local cubic, i.e., $\frac{1}{2}$ of the edge length of the cubic volume.

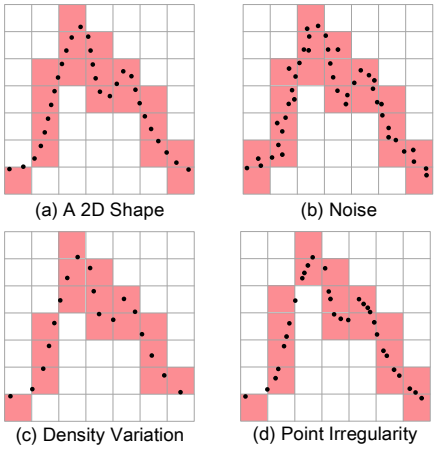


Fig. 2. Illustration of LoVS's robustness to (b) noise, (c) point density variation, and (d) point irregularity in 2D. For clarity, grids with points inside are marked with red.

2.2. Voxelization

Once the local cubic-intersected surface is calculated, we perform space partition in it to clarify the local structure of the object. The cubic volume provides us with the merit of fast yet uniform space division, we hence generate a set of voxels in it, also known as voxelization [11]. The basic principle of our LoVS is grounded on judging whether there are points inside a voxel. Accordingly, we first examine the corresponding voxel index of a given point q_i in the cubic-intersected surface $\mathcal{Q}_{\mathcal{C}}$. Assume that we have obtained $N_v = m^3$ voxels after voxelization, the voxel index of q_i is defined as:

$$voxel\_id(q_i) = \left\lfloor \frac{q_i.z + r}{l_{step}} \right\rfloor \times m^2 + \left\lfloor \frac{q_i.y + r}{l_{step}} \right\rfloor \times m + \left\lfloor \frac{q_i.x + r}{l_{step}} \right\rfloor, \quad (3)$$

where l_{step} is the division step, i.e., $\frac{2r}{m}$, during voxelization, and $\lfloor \bullet \rfloor$ denotes round-down operation. After all points in $\mathcal{Q}_{\mathcal{C}}$ are indexed, a voxel v_i would contain a point subset $\mathcal{Q}_{v_i}$. Further, we define a label $l(v_i)$ for voxel v_i with the help of $\mathcal{Q}_{v_i}$ as:

$$l(v_i) = \begin{cases} 1 & \text{if } |Q_{v_i}| > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

so as to achieve binary structure characterization (see Fig. 1).

2.3. LoVS Generation

Thus far, we have labeled each voxel in the cubic volume with 0 or 1. We then integrate these labels into a bit string, i.e., our final LoVS descriptor, as:

$$f_{\mathrm{LoVS}} = \{l(v_1), l(v_2), \cdots, l(v_{N_v})\}. \quad (5)$$

Admittedly, other feature integration techniques such as principle component analysis (PCA)-based and learning-based methods could also be employed to compute LoVS feature. Rather, we use concatenation currently considering low implementation complexity and efficiency. Note that we still analyze the effect of dimension reduction on current LoVS feature in Sec. 3.4.

Fig. 2 briefly illustrates our LoVS's robustness against common nuisances, including noise, point density variation and point irregularity. Clearly, with proper space divisions (i.e., the number of voxels N_v), the proposed LoVS could stay quite stable against these nuisances (as verified in Sec. 3.3).

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

We deploy our experiments on three public datasets, i.e., Bologna 3D Retrieval (B3R) [12], UWA 3D Object Recognition (UWAOR) [13, 14] and Queen's LiDAR (QuLD) [15] datasets. The B3R dataset contains 6 models and 18 scenes, the UWAOR dataset consists of 5 models and 50 scenes, and the QuLD dataset includes 5 models and 80 scenes. Note that we also create another 18 scenes in the B3R dataset with different levels of mesh decimation. The popular recall vs 1-precision curve (RPC) measure [1, 5] is used for quantitative evaluation (following the same calculation process as [1, 5]). Regarding comparative evaluation, the existing 3D binary one, i.e., B-SHOT [8], is considered. We also take several real-valued ones, i.e., spin image [3], snapshots [16], FPFH [4] and SHOT [1], into comparison. The parameter settings of all tested methods are reported in Table 1 ('mr' being mesh resolution). All the experiments are implemented in C++ with the help of point cloud library (PCL) [11].

3.2. Parameter Analysis of LoVS

The proposed LoVS has very few parameters, where m relating to the voxel number N_v (i.e., $N_v = m^3$) is the only relevant parameter. It decides the division extent of the cubic volume, affecting LoVS's descriptiveness and robustness. We therefore tune the parameter on a 'turning' dataset (i.e., the B3R dataset with 0.3mr Gaussian noise and $\frac{1}{4}$ mesh decimation). The results are shown in Fig. 3(a).

One can see that the RPC performance improves as m increases from 5 to 9, and then starts to deteriorate when m further increases. Accordingly, it suggests that using $m = 9$ could strike a well balance between descriptiveness and robustness. Therefore, $m = 9$ is used to generate the LoVS feature with 729 bits.

Table 1. Parameter settings for six feature descriptors.

	Support Radius/Length (mr)	Dimension	Storage (bit)
Spin image	15	225 (15 × 15)	225 × 8
Snapshots	15	1600 (40 × 40)	1600 × 8
FPFH	15	33 (3 × 11)	33 × 8
SHOT	15	352 (8 × 2 × 2 × 11)	352 × 8
B-SHOT	15	352 (8 × 2 × 2 × 11)	352 × 1
LoVS	15	729 (9 × 9 × 9)	729 × 1

Table 2. Effect of dimension reduction on LoVS on the B3R dataset (measured by $\mathrm{AUC}_{\mathrm{rp}}$), where GN and MD respectively represent Gaussian noise and mesh decimation.

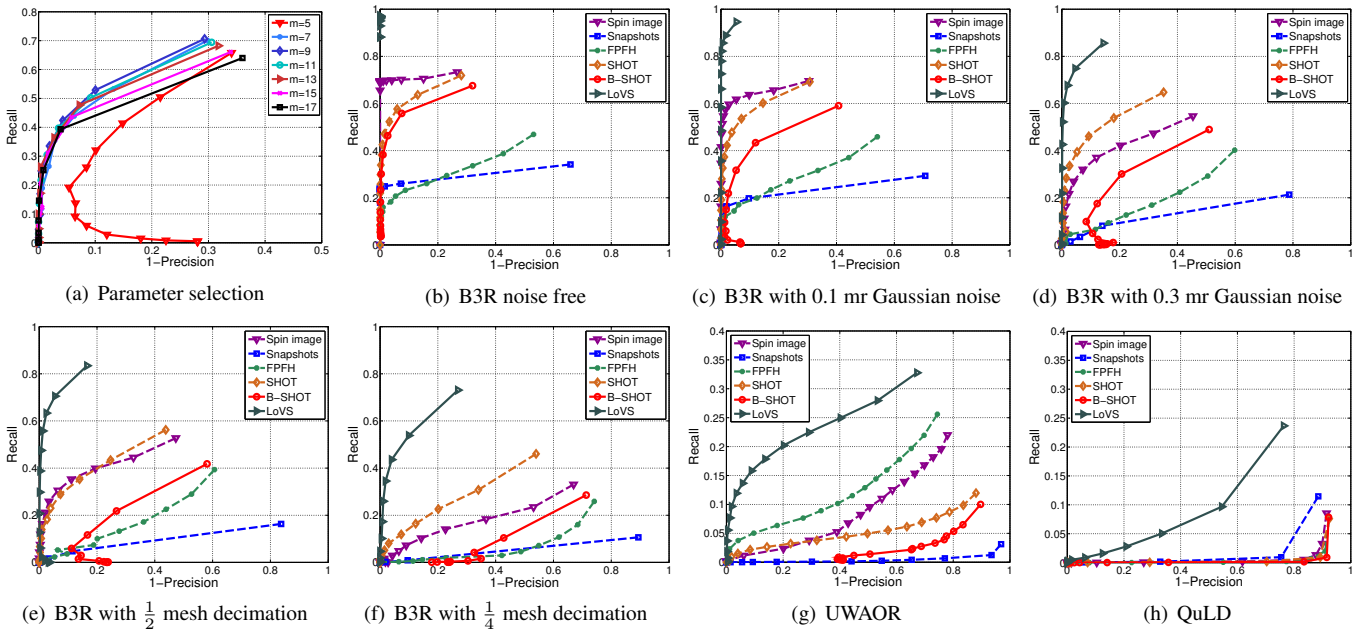
	B3R	B3R+0.1mr GN	B3R+0.3mr GN	B3R+$\frac{1}{2}$ MD	B3R+$\frac{1}{4}$ MD
LoVS	0.973	0.943	0.841	0.813	0.683
-20% dim.	0.972	0.941	0.838	0.810	0.673
-40% dim.	0.972	0.939	0.831	0.800	0.668
-60% dim.	0.971	0.933	0.818	0.793	0.642
-80% dim.	0.968	0.908	0.776	0.737	0.571

3.3. Feature Matching Performance

In terms of feature matching performance, we specifically choose the B3R, UWAOR and QuLD datasets for a rigorous examination. The aims include: testing the descriptiveness as well as robustness to noise and mesh decimations on the B3R dataset, judging the robustness to clutter and occlusion on the UWAOR dataset, and the robustness to the combination of clutter, occlusion and point irregularity on the QuLD dataset. It is worth noting that the term *point irregularity* has not been evaluated before. The RPC results of all tested descriptors are shown in Fig. 3.

Three major observation can be made from the results. **First**, for compared binary descriptor, i.e., B-SHOT, LoVS outperforms it by a large margin on all datasets. It is somehow not surprising because SHOT is also inferior to LoVS as well. As B-SHOT is a direct binary extension of SHOT, it loses certain distinctiveness, being surpassed by SHOT. **Second**, for compared real-valued descriptors, LoVS still ranks the first, while the second ranked one differs from datasets. For instance, SHOT ranks the second on the B3R dataset with mesh decimation, whereas FPFH is the second best one on the UWAOR dataset, reflecting challenges when across datasets. **Third**, the compared descriptors exhibit very poor performance on the QuLD dataset, owing to none of them take the resilience to point irregularity into their designing principles. By contrast, our LoVS attains an acceptable performance there. The time costs of computing 100 spin image, Snapshots, FPFH, SHOT, B-SHOT and LoVS descriptors with a 2.7GHz CPU are 0.21s, 2.34s, 0.25s, 0.20s, 0.22s, and 0.33s, respectively.

As before illustrated in Fig. 2, the overall superiority of



(a) Parameter selection (b) B3R noise free (c) B3R with 0.1 mr Gaussian noise (d) B3R with 0.3 mr Gaussian noise

(e) B3R with $\frac{1}{2}$ mesh decimation (f) B3R with $\frac{1}{4}$ mesh decimation (g) UWAOR (h) QuLD

Fig. 3. Parameter selection of LoVS and performance evaluation of feature descriptors on the B3R, UWAOR and QuLD datasets.

our LoVS descriptor could be explained from at least two aspects. **First**, LoVS is directly computed on the local shape, without second-order representations such as attribute histograms in SHOT, and 3D-to-2D projection in spin image. **Second**, LoVS is a coarse approximation of the shape, indicating well-preserved 3D structure and strong robustness.

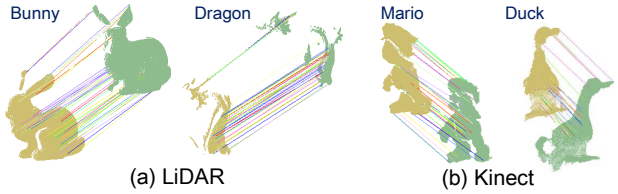
3.4. Effect of Dimension Reduction on LoVS

Our current LoVS is the direct concatenation of all voxel labels (see Sec. 2.3), it is efficient, though, would also cause redundancy. Here, we test the effect of dimension reduction on LoVS by uniformly sampling the voxel labels over the feature space. The experiment is performed on the B3R dataset. In order to measure the feature matching performance aggregately, the area under RPC ($\mathrm{AUC_{rp}}$) [17] is used. The $\mathrm{AUC_{rp}}$ results with respect to different sampling ratios are presented in Table 2.

The results suggest that LoVS holds strong discriminative power in high-quality case even with a compressing ratio of 80%, i.e., B3R. On low-quality datasets, e.g., B3R with $\frac{1}{4}$ mesh decimation, the performance meets a conspicuous decrease. Overall, it indicates that our LoVS could be further compressed without losing much descriptiveness.

3.5. Visual 3D Correspondences with LoVS

In addition to above quantitative results, we also present the visual feature matching results of LoVS on both LiDAR and Kinect data. In particular, the `Bunny` and `Dragon` data



(a) LiDAR (b) Kinect

Fig. 4. 3D correspondences via LoVS matching on both LiDAR and Kinect data (Figure best seen in color).

from the Stanford 3D Scanning Repository [18], and the `Mario` and `Duck` data from the Bologna Mesh Registration dataset [1] are considered. We use the standard registration module in PCL, where the feature extraction process is performed by LoVS, to register these view pairs. The results are shown in Fig. 4.

As witnessed by the figure, plenty of consistent correspondences are established between the point cloud views from both LiDAR and Kinect sensors, showing the effectiveness and robustness of LoVS in terms of shape registration.

4. CONCLUSIONS

In this paper, we proposed a novel 3D binary descriptor named LoVS. Across-dataset experiments together with state-of-the-art comparisons demonstrated the effectiveness and robustness of our method. Developing LoVS-based application algorithms would be our future work.

5. REFERENCES

[1] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 356–369. 1, 2, 3, 4

[2] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. 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. 1

[3] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, 1999. 1, 3

[4] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *Proceedings of IEEE International Conference on Robotics and Automation*, 2009, pp. 3212–3217. 1, 2, 3

[5] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. 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. 1, 2, 3

[6] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012. 1

[7] A. T. Tra, W. Lin, and A. Kot, "Dominant sift: A novel compact descriptor," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 1344–1348. 1

[8] S. M. Prakhya, B. Liu, and W. Lin, "B-shot: A binary feature descriptor for fast and efficient keypoint matching on 3d point clouds," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2015, pp. 1929–1934. 1, 2, 3

[9] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3384–3391. 2

[10] J. Yang, Q. Zhang, K. Xian, Y. Xiao, and Z. Cao, "Rotational contour signatures for robust local surface description," in *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2016, pp. 3598–3602. 2

[11] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2011, pp. 1–4. 2, 3

[12] F. Tombari, S. Salti, and L. Di Stefano, "Performance evaluation of 3d keypoint detectors," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 198–220, 2013. 3

[13] A. S. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1584–1601, 2006. 3

[14] A. Mian, M. Bennamoun, and R. Owens, "On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 348–361, 2010. 3

[15] B. Taati and M. Greenspan, "Local shape descriptor selection for object recognition in range data," *Computer Vision and Image Understanding*, vol. 115, no. 5, pp. 681–694, 2011. 3

[16] S. Malassiotis and M. G. Strintzis, "Snapshots: A novel local surface descriptor and matching algorithm for robust 3d surface alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1285–1290, 2007. 3

[17] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *International Journal of Computer Vision*, pp. 1–24, 2015. 4

[18] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1996, pp. 303–312. 4