TOLDI: An effective and robust approach for 3D local shape description

Article *in* Pattern Recognition · November 2016

DOI: 10.1016/j.patcog.2016.11.019

CITATIONS
33

READS
916

4 authors:

Jiaqi Yang
Huazhong University of Science and Technology
39 PUBLICATIONS **225** CITATIONS
SEE PROFILE

Qian Zhang
Hubei University
20 PUBLICATIONS **288** CITATIONS
SEE PROFILE

Yang Xiao
Huazhong University of Science and Technology
104 PUBLICATIONS **860** CITATIONS
SEE PROFILE

Zhi-Guo Cao
Huazhong University of Science and Technology
214 PUBLICATIONS **1,450** CITATIONS
SEE PROFILE

TOLDI: An effective and robust approach for 3D local shape description

Jiaqi Yang, Qian Zhang, Yang Xiao, Zhiguo Cao*

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

Abstract

Feature description for the 3D local shape in the presence of noise, varying mesh resolutions, clutter and occlusion is a quite challenging task in 3D computer vision. This paper tackles the problem by proposing a new local reference frame (LRF) together with a novel triple orthogonal local depth images (TOLDI) representation, forming the TOLDI method for local shape description. Compared with previous methods, TOLDI manages to perform efficient, distinctive and robust description for the 3D local surface simultaneously under various feature matching contexts. The proposed LRF differs from many prior ones in its calculation of the z-axis and x-axis, the z-axis is calculated using the normal of the keypoint and the x-axis is computed by aggregating the weighted projection vectors of the radius neighbors. TOLDI feature descriptors are then obtained by concatenating three local depth images (LDI) captured from three orthogonal view planes in the LRF into feature vectors. The performance of our TOLDI approach is rigorously evaluated on several public datasets, which contain three major surface matching scenarios, namely shape retrieval, object recognition and 3D registration. Experimental results and comparisons with the state-of-the-arts validate the effectiveness, robustness, high efficiency, and overall superiority of our method. Our method is also applied to aligning 3D object and indoor scene point clouds obtained by different devices (i.e., LiDAR and Kinect), the accurate outcomes further confirm the effectiveness of our method.

Keywords: Local reference frame, Local feature descriptor, Shape retrieval, Object recognition, 3D registration

1. Introduction

Local shape description is a fundamental and critical problem in 3D computer vision areas. Compared with global shape descriptors, local shape descriptors hold many pleasurable peculiarities including being robust to clutter, occlusion, and missing regions [1]. These merits have facilitated many real-word applications related to local descriptor-based surface matching such as 3D registration [2], shape retrieval [3], human face recognition [4], object recognition [5], to name a few. With the development of numerous low-cost 3D acquisition systems, e.g., Microsoft Kinect and Intel RealSense, 3D data including point clouds, polygon meshes and depth images is now readily accessible, which further highlights the significance of investigation on vision tasks grounded on local shape description.

The concept of local shape description is given as encoding the geometric information contained in the local surface into a feature vector representation [6]. It is worthy noting that the scope of this paper concentrates on local shape description for rigid objects.

*Corresponding author

Email addresses: jqyang@hust.edu.cn (Jiaqi Yang), hangfanzq@163.com (Qian Zhang), Yang_Xiao@hust.edu.cn (Yang Xiao), zgcao@hust.edu.cn (Zhiguo Cao)

Essentially, local shape descriptors should be invariant to objects' pose, i.e., rigid transformation. Moreover, in real applications such as surface registration and object recognition, local shape descriptors are amenable to resist the impacts of noise, varying mesh resolutions (point densities), clutter, occlusion and missing regions. Above mentioned nuisances make it quite challenging to design a local shape descriptor with overall good performance. In the literature, many attempts have been made to cope with these difficulties. Examples include point signatures [7], spin images [8], fast point feature histograms (FPFH) [9], signature of histograms of orientations (SHOT) [6] and rotational projection statistics (RoPS) [10] (a comprehensive survey is available in [1]). These descriptors can be categorized into two classes: with or without local reference frame (LRF). For descriptors without LRF, they mainly take the statistics of local geometric features as feature representations, e.g., spin images and FPFH. Since the local spatial information is discarded, descriptors without LRF usually exhibit limited descriptiveness [10]. Conversely, LRF-based descriptors first build an LRF for the local surface and then characterize the local geometric and spatial information with respect to the LRF. Examples include point signatures and RoPS. LRF is a local coordinate system established in the local surface which on one hand makes the local descriptor invariant to rigid transformation, and on the other provides fully spatial information for local shape description [10]. A recent quantitative evaluation for local shape descriptors [11] shows that LRF-based descriptors neatly outperform those descriptors without LRF in most public datasets.

Among LRF-based descriptors, LRF and feature representation are their two major concerns. The reasons are: 1) the effectiveness and robustness of LRF-based descriptors are bounded up with their associated LRFs [6], and 2) the feature representation of a descriptor would directly affect the discriminative power of a descriptor [12]. At present, typical LRF proposals include the methods proposed by Mian et al. [13], Tombari et al. [6], Petrelli et al. [14], and etc. They are either based on covariance analysis (CA) or point spatial distributions (PSD). Unfortunately, most CA-based methods suffer from sign ambiguity [6], and PSD-based methods exhibit weak robustness to high levels of noise and variations of mesh resolutions [15]. Feature representation focuses on exploring an effective and robust manner to encode the information contained in the local surface. Examples include 2D/1D representations of point density [5, 10], deviation angle between normals [16, 9], local depth [12, 17], and their combination [18]. However, many of them suffer from either unilateral information encoding (e.g., snapshots [12]) and/or loss of information going from 3D to 2D/1D representations (e.g., RoPS), and therefore deteriorate the descriptiveness of the final constructed descriptors.

Motivated by these considerations, we propose a novel method named triple orthogonal local depth images (TOLDI) that comprises a new LRF proposal and TOLDI descriptor for effective, robust and efficient local shape description. To construct the LRF, we resort to the normal of the keypoint for the calculation of its z-axis using a subset of the radius neighbors, and the vector sum of all projection vectors of the radius neighbors to calculate the x-axis. Weighting strategies are adopted for each projection vector to achieve a balanced robustness to noise, varying mesh resolutions, clutter, and occlusion. The other axis, i.e., the y-axis, is computed directly via cross-product. As for feature representation, the distances from each radius neighbor to a virtual view plane, or referred to as local depth, are adopted to represent the local shape geometry from one view. Such representation has several advantages, e.g., preserving the essential geometric information, and being highly efficient for calculation [12]. To tackle the problem of loss of information caused by clutter and occlusion (including self-occlusion), three orthogonal view planes in the LRF are selected for comprehensive shape description. By concatenating these LDIs into a 1D feature vector, the TOLDI descriptor is generated. Different from many previous local shape description methods, which require complex preprocessings for the initial data such as triangulation, our TOLDI method is directly performed on the initial point clouds. We conduct a set of experiments on three public datasets, which respectively comprise shape retrieval, shape registration and object recognition scenarios, to comprehensively evaluate our pro-

posals. The experimental results, together with comparisons with several state-of-the-art methods, show coherently that the proposed method achieves the best overall performance against the exiting methods. The major contributions of this paper are summarized as follows:

1) A new LRF is proposed for the 3D local surface, which exhibits high repeatability under the impacts of noise, varying mesh resolutions, clutter and occlusion. It can be also seamlessly grafted to other LRF-based descriptors to promote their feature matching performance.
2) A novel local shape descriptor called TOLDI is proposed to achieve a satisfactory and balanced performance in terms of descriptiveness, robustness, and time efficiency.

The rest of this paper is structured as follows. Section 2 presents a brief review of related work on LRF construction and local shape descriptors. Section 3 gives a detailed description for the proposed TOLDI method. Section 4 presents the experimental evaluation of our method and several state-of-the-art methods on three standard datasets. The conclusions and future work are drew in section 5.

2. Related work

Various methods have been proposed in the literature for 3D local shape description. Considering that the proposed TOLDI approach falls in the family of LRF-based methods, we first review the existing LRF proposals because LRF is a critical component for LRF-based methods [6, 10], and then briefly describe the existing local shape descriptors.

2.1. LRF methods

As mentioned in section 1, we categorize the existing LRF methods into CA-based and PSD-based methods. As for CA-based methods, Novatnack et al. [19] defined the normal of the keypoint as the z-axis, they then calculated a covariance matrix for the neighboring points of the keypoint. The eigenvector associated with the largest eigenvalue of the covariance matrix was first projected on the tangent plane of the surface, the projected vector was then normalized as the x-axis. The third axis was given by $z \times x$. Mian et al. [13] defined the unit vectors of the LRF via performing covariance analysis on the radius neighbors of the keypoint, the three eigenvectors of the covariance matrix were defined as the unit vectors of the LRF. The problem of sign ambiguity has not been solved yet in these CA-based methods. Later on, Tombari et al. [6] employed a weighted covariance matrix for the calculation of LRF, where the sign of each axis was also unambiguous. This method has been proven to be quite robust to noise, its major limitation is the sensitivity to the variance of mesh resolutions. Recently, Guo et al. [10] proposed a novel technique for the construction of LRF by applying two weighting strategies to the covariance matrices of each triangle in the local surface. Different from other CA-based methods, the method in [10] employed every single triangle in the local surface for covariance analysis, and achieved significant improvement in terms of robustness.

For PSD-based LRFs, Chua and Jarvis [7] placed a sphere centered at the keypoint, and obtained a contour at the intersection region of the surface. The point with the largest signed projection distance to the tangent plane of the keypoint was selected to calculate the x-axis. The z-axis was directly along the normal of the keypoint, and the y axis was computed via cross-product. Petrelli et al. [14] picked a small subset of the radius neighbors for the calculation of the z-axis to achieve robustness to occlusion. The normals of the points lying in the border region of the local surface were then calculated, the point with largest deviation angle between its normal and the z-axis was selected for the calculation of x-axis. Later in [15], they improved the repeatability of the LRF by using the point with the largest local depth instead of deviation angle between normals. However, noise of large scale would have an obvious impact on the calculation of the x-axis in [14, 15].

2.2. Local shape descriptors

Numerous local shape descriptors have been proposed in the literature. Some are designed for deformable shapes, while the others deal with rigid shapes. The former category of descriptors, namely the intrinsic local descriptors, pursue the ability of being invariant to non-rigid transformations. Among

descriptors fall in this category, Sun et al. [20] proposed the heat kernel signature (HKS) using the Laplace-Beltrami operator on surfaces as well as its derived spatial embeddings. HKS is invariant to isometries and robust to small non-isometric deformations. Subsequently, Aubry et al. [21] proposed the wave kernel signature (WKS) by employing a quantum mechanical method to describe multi-scale details, which is shown to be more descriptive than HKS. In addition to these hand-crafted descriptors, some learned ones have been recently proposed. Litman and Bronstein [22] proposed a learned descriptor called the optimal spectral descriptor (OSD) by applying a bank of filters, which were constructed by a learning approach, to the shape's geometric features at different frequencies. Boscaini et al. [23] proposed the anisotropic diffusion descriptors (ADD) lately. ADD tackles the isotropic problem that exists in many prior approaches by using anisotropic diffusion on meshes and point clouds, where the directed local kernels were learned in a task-specific manner.

Most related to our method are the works from the other category (i.e., the rigid category) of descriptors, which are either LRF-based or not. For descriptors without LRF, Johnson and Hebert [8] proposed the spin image descriptor for 3D surface matching. They first specified the normal of the query point as the reference axis and projected the neighbors defined by the support angle onto a 2D coordinate, then a gray image was generated to describe the 2D point distribution using a 2D array accumulator. The spin image is one of the most cited feature descriptor, while it suffers limited descriptiveness and susceptibility varying mesh resolutions. Rusu et al. [24] proposed a novel method of characterizing the local geometry by using point feature histograms (PFH). PFH has great discriminative power but appears to be quite time-consuming. To address this problem, they later used the simplified point feature histogram (SPFH) of neighbors to construct the fast point feature histograms (FPFH) descriptor [9], which turns out to be distinctive and computationally efficient. Albarell et al. [25] proposed a low-dimensional descriptor called the surface hashes embedded in their isometry-enforcing game theory based surface alignment scheme. The surface hashes descriptor was created by computing local geometric properties such as normal deviation, integral volume or their combination at multiple scales. Recently, Yang et al. [18] proposed a local feature statistics histograms (LFSH) by integrating multiple faint correlated statistical information into a feature vector. LFSH is faster than most existing descriptors, while it suffers from limited descriptiveness [18].

For descriptors with LRF, Stein and Medioni [26] proposed a splash descriptor by encoding the relationship (angular distance) between the query point and its geodesic neighbors, the relationship was then stored using a 3D vector which was finally transformed to curvatures and torsion angles. Frome et al. [27] extended the 2D shape context method [28] to a 3D shape context (3DSC) descriptor, they divided the radius neighbors of the keypoint into several bins and counted the weighted number of points in each bin. Malassiotis et al. [12] proposed the "snapshots" descriptor by taking "snapshots" of the local surface using a virtual camera oriented perpendicularly to the surface. The snapshots descriptor first highlighted the discriminative power of local depth feature, whereas it still exhibits limited descriptiveness due to that only the information of a single view is contained. Zaharescu et al. [29] calculated and projected the gradient vectors of local vertices onto the three orthonormal planes of an LRF. Each plane was then divided into four polar subregions and represented by an eight-element histogram, and the concatenation of these histograms finally constituted a MeshHOG descriptor. Tombari et al. [6] proposed the signature of histograms of orientations (SHOT) descriptor by first dividing the spherical neighborhood space into 3D volumes. Then, a local histogram for each volume was generated by accumulating the number of points according to the deviation angles between the normal at the keypoint and the ones at the radius neighbors. All local histograms were finally concatenated, forming the SHOT descriptor. Despite SHOT's high descriptiveness, it is sensible to varying mesh resolutions [6]. Recently, Guo et al. [10] proposed the rotational projection statistics (RoPS) descriptor by calculating a set of statistics from several point density maps after rotation. The RoPS descriptor first encoded the local shape geome-

try from multiple perspectives and achieved the state-of-the-art performance in terms of feature matching [1]. Several limitations of the RoPS descriptor include high time consumption and susceptibility to the nonuniformity of points. Similar to the view-based mechanism in RoPS, Guo et al. [30, 31] proposed a Tri-Spin-Image (TriSI) through integrating and compressing three spin image signatures, which were generated from the three orthogonal coordinate axes of an LRF, into a feature vector. TriSI is time-consuming, though, it possesses stronger robustness to clutter and occlusion than RoPS [31].

3. TOLDI-based local shape description

This section presents the details of our proposed TOLDI technique for local shape description. Specifically, we first introduce a repeatable and robust LRF for the local surface. It is based on calculating the normal of the keypoint, and the vector sum of the projection vectors of its radius neighbors. Then, we present a TOLDI feature representation (Fig. 2) based on the proposed LRF by performing virtual view planes selection, neighboring points projection, local depth feature calculation and sub-features concatenation. Finally, the key parameters of TOLDI are quantitatively analyzed.

3.1. A Novel LRF Proposal

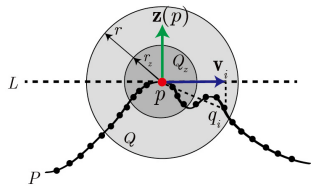


Figure 1: Illustration of the proposed LRF in 2D. The red point, green arrow and blue arrow denote the keypoint in the local surface, the z-axis and the projection vector of an exemplar radius neighbor of the keypoint, respectively. See text for the representations of variables.

An illustration of the proposed LRF is shown in Fig. 1. Given a keypoint p in point cloud P, the LRF at p can be represented as:

$$\mathbf{L}_p = \{\mathbf{x}(p), \mathbf{x}(p) \times \mathbf{z}(p), \mathbf{z}(p)\}, \tag{1}$$

where $\mathbf{x}(p)$ and $\mathbf{z}(p)$ are the x-axis and z-axis of $\mathbf{L}_p$, and the y-axis is computed via cross-product. Note that variables shown in bold font represent vectors, and $\times$ between vectors denotes cross-product operation. The calculation of LRF therefore contains two steps: estimation of z-axis and x-axis.

As for z-axis, an intuitive choice is to directly use the normal of the keypoint as the z-axis. Actually, such method has been confirmed to be very repeatable in [14]. We therefore use a similar way to calculate the z-axis. However, determining the point set required for the estimation of normal is critical [32]. That is, the distribution and count of points used for normal calculation would effect the final outcomes. To achieve robustness to clutter and occlusion, we only employ a small subset of points in the local surface [14, 15]. To attain robustness to varying mesh resolutions, radius neighbors are used to calculate such subset, in opposite to the popular k nearest neighbors. The detailed calculation of z-axis is as follows. First, we place a sphere of support radius r at p. The points (except p) inside the sphere are defined as the radius neighbors of p. All these radius neighbors constitute a local surface $Q = \{q_1, q_2, \ldots, q_k\}$. Next, we employ a subset of Q for the calculation of z-axis. Specifically, the points in Q with their distances to the keypoint p being smaller than a threshold r_z, are selected to form the subset $Q_z = \{q_1^z, q_2^z, \ldots, q_s^z\}$. In this paper, r_z is set to $\frac{r}{3}$ as suggested in [15]. At last, a covariance analysis [33] is performed on Q_z as:

$$Cov\left(Q_z\right) = \begin{bmatrix} q_1^z - \overline{q^z} \\ \cdots \\ q_s^z - \overline{q^z} \end{bmatrix}^T \cdot \begin{bmatrix} q_1^z - \overline{q^z} \\ \cdots \\ q_s^z - \overline{q^z} \end{bmatrix}, \tag{2}$$

where s is the size of Q_z, and $\overline{q^z}$ is the centroid of Q_z. The eigenvector $\mathbf{n}(p)$ corresponding to the minimum eigenvalue of $Cov\left(Q_z\right)$ is computed as the normal of p. Owing to that the sign of $\mathbf{n}(p)$ is not defined [6], we disambiguate the sign and calculate the z-axis as:

$$\mathbf{z}(p) = \begin{cases} \mathbf{n}(p), & \text{if } \mathbf{n}(p) \cdot \sum_{i=1}^{k} \mathbf{q}_i\mathbf{p} \geq 0 \\ -\mathbf{n}(p), & \text{otherwise} \end{cases}, \tag{3}$$

where k is the number of radius neighbors, $\cdot$ between vectors denotes dot-product, and $\mathbf{q}_i\mathbf{p}$ represents the vector between q_i and p.

Once the z-axis is determined, the next step is to calculate the x-axis. We denote the tangent plane of p with respect to $\mathbf{z}(p)$ as L, as shown in Fig. 1. The task is then to find a canonical direction in L. Since many surfaces may exhibit flat or symmetry geometry, the estimation of x-axis turns out to be more challenging than the z-axis. To tackle this problem, we first project all the radius neighbors of p on plane L, and calculate a projection vector for each neighboring point q_i as:

$$\mathbf{v}_i = \mathbf{pq}_i - (\mathbf{pq}_i \cdot \mathbf{z}(p)) \cdot \mathbf{z}(p). \tag{4}$$

With these vectors on plane L, we can either define a salient function for each projection vector to choose a representative vector as the x-axis (after normalization), or integrate all these vectors into a single vector. We choose the latter one, because employing all the points for the calculation of x-axis yields better robustness [6, 10]. Specifically, the vector sum of all the projection vectors in L is used for the estimation of x-axis:

$$\mathbf{x}(p) = \sum_{i=1}^{k} w_{i1} w_{i2} \mathbf{v}_i \Bigg/ \left\| \sum_{i=1}^{k} w_{i1} w_{i2} \mathbf{v}_i \right\|, \tag{5}$$

where k represents the number of radius neighbors of keypoint p, and $\|\cdot\|$ denotes L2 norm.

In Eqs. 5, w_{i1} is a weight that related to the distance from q_i to p, that is:

$$w_{i1} = (r - \|p - q_i\|)^2, \tag{6}$$

w_{i2} is a weight that related to the projection distance from q_i to L, that is:

$$w_{i2} = (\mathbf{pq}_i \cdot \mathbf{z}(p))^2, \tag{7}$$

At last, the y-axis is calculated via the cross-product between x-axis and z-axis.

We remark that, the first weight w_{i1} is designed to improve the robustness of the LRF to clutter, occlusion and incomplete border regions [6, 10]. As a result, distant radius neighbors contribute less to the x-axis. The second weight w_{i2} is set to make the points which have larger projection distance contribute more to the x-axis, since such distance feature is a distinctive cue and can provide high repeatability on flat regions [15].

3.2. TOLDI descriptor

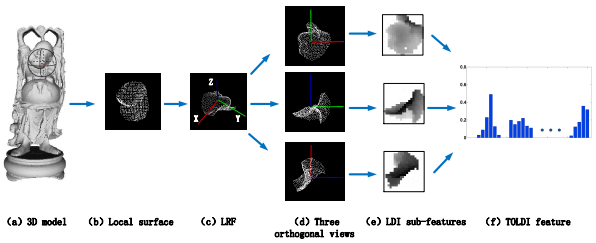


Figure 2: An illustration of the TOLDI feature descriptor. The red point in (a) represents a keypoint in the 3D model and the points inside the surrounding sphere of the keypoint constitute a local surface.

Once the LRF is constructed for the local surface Q, we are left with the task of encoding the spatial and geometric information contained in the local surface. As shown in Fig. 2, we first transform the local surface Q with respect to the LRF in order to achieve invariance to rigid transformation. The rotated surface is denoted by $Q' = \{q'_1, q'_2, \ldots, q'_k\}$. Then, we need to seek for an appropriate manner for feature representation.

In the literature, there are many effective approaches to encode the geometric information of a local shape, e.g., deviation angles between normals [32, 9], the point density of 2D projected points [8, 10] and the local depth [12]. In this paper, we adopt the local depth feature for local shape description for two reasons. First, the local depth feature, sometimes referred as signed projection distance, preserves the majority of the essential information of a shape [12]. In contrast, the limitation of the other methods is loss of information going from a 3D to 2D/1D representation of the local shape. Second, the local depth feature is quite efficient for calculation. Usually, the calculation of local depth feature requires view planes selection and 3D-to-2D projection. A typical example is the "snapshots" descriptor [12], which captures a local depth image from a view point that is perpendicular to the xy plane of their defined LRF. However, due to clutter and occlusion, a single view is not always sufficient for fully employing the contained geometric information in the local surface, as illustrated in Fig. 3. Therefore, we define several virtual view

planes in the LRF in order to take a comprehensive encoding for the local surface. Specifically, three orthogonal view planes that are respectively parallel to the xy, yz and xz planes of the LRF with a distance of r are chose. Put in other words, the equations of the three view planes deployed in the LRF system are defined as $z-r=0$, $x-r=0$ and $y-r=0$, respectively. All our discussed operations are performed in the LRF system, therefore, the views only exist in the LRF system. Here, employing three orthogonal views could provide us with complementary and relatively irredundant information as suggested in the TriSI method [30, 31]. Using the same multi-view mechanism as the TriSI method, though, our TOLDI method differs from the TriSI method from at least two aspects. As highlighted in [1], LRF construction and feature represention are two cirtical parts for a LRF-based descriptor. First, TOLDI is fed with our proposed LRF that based on point spatial distribution rather than classical covariance analysis used in TriSI's LRF. Second, in opposite to the spin image representation employed in TriSI, which has been proven to hold limited descriptiveness and robustness [12, 6], TOLDI uses local depth feature to form LDI representation for feature description, in order to achieve stronger discriminative power and robustness (as demonstrated in section 4.3).

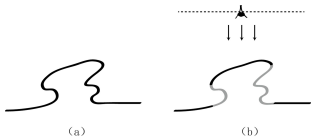


Figure 3: Illustration of loss of information caused by occlusion from a single view. (a) A shape in 2D. (b) The captured (shown in dark) and hidden (shown in gray) information from one view.

After the view planes are determined, we can calculate the local depth features f_i^{xy}, f_i^{yz} and f_i^{xz} of each radius neighbors in Q' with respect to xy plane, yz plane and xz plane as:

$$\begin{cases} f_i^{xy} = r - {q'}_i.z \\ f_i^{yz} = r - {q'}_i.x \\ f_i^{xz} = r - {q'}_i.y \end{cases}, \tag{8}$$

where ${q'}_i.x$, ${q'}_i.y$ and ${q'}_i.z$ represent the x-value, y-value and z-value of ${q'}_i$, respectively. The values of the three features are in the range of $[0, 2r]$ and they are further normalized to $[0, 1]$. Then, we respectively project the points in Q' on the three view plane and generate an image I of size $w \times w$ for each view plane using 2D array accumulators. The pixel value is defined as the smallest local depth feature value among the points falling in the pixel. It is based on imitating that the human observes a 3D object from one side, where the occluded regions are invisible. Note that some pixels may contain no points during projection, we assign a large value for them to distinguish such "holes". Finally, three images I_{xy}, I_{yz}, I_{xz} can be derived from the local surface. In order to integrate these sub-features efficiently, we directly use concatenation operation to combine them into a 1D histogram as in [6, 18], forming the TOLDI descriptor with $3 \times w \times w$ bins.

In the following, we describe three major intrinsic properties including invariance to rigid transformations, stability and high computational efficiency inherent to our TOLDI descriptor through theoretical analysis:

Invariance to rigid transformations: The TOLDI descriptor is calculated within a spherical shaped form, and relies on a unique and unambiguous LRF. Thus, the geometric primitives (i.e., the local depth feature used in this paper), are computed on transformed local points with respect to the LRF. Endowed with the intrinsic property of the defined LRF, these computed geometric primitives are therefore invariant to rigid transformations, satisfying the theorem for 3D invariant descriptors in [34]. To aggregate these invariant features, LDI "signature" (belongs to the signature category of 3D local feature descriptors [6, 1]), which preserves the local spatial information of point primitives, is used to generate the final TOLDI descriptor to resist the global Euclidean transformations of the 3D object (as verified in section 4.4).

Stability: The stability of TOLDI descriptor on one hand relies on the robustness of its LRF, and on the other grounds on its feature representations. First, the proposed LRF takes a small subset of radius neighbors to calculate the

z-axis to resist clutter and occlusion [14]. The x-axis is the vector sum of a set of projection vectors of the radius neighbors, containing two weights to strike a balanced robustness to noise and varying mesh resolutions (as verified in section 4.2). Second, sparse partition is performed on the LDI and a single point with local maximum depth value is selected to calculated the value of a bin in the LDI. Thus, noise is of little chance to cause an obvious change of bin values in the LDI. Similarly, effect of variation in point densities is rather faint because only one point, in spite of the number of points in a bin, is required to calculate the bin value. These designing principles definitely boost the stability of TOLDI against a variety of nuisances (as verified in section 4.3).

High computational efficiency: The theoretical computational complexity of the TOLDI descriptor for a given local surface with k points is $O(3 \cdot k) \rightarrow O(k)$, where 3 means that three local depth features with respect to different reference planes need to be computed for a local point. Despite the low computational complexity of TOLDI, computing the local depth feature per time is quite fast [12]. It is therefore expected that the TOLDI descriptor possesses high computational efficiency (as verified in section 4.3.3).

3.3. Analysis of TOLDI parameters

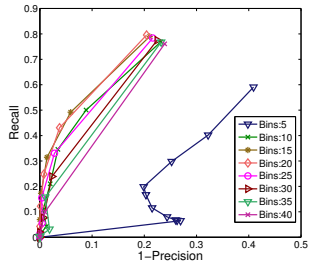


Figure 4: Effect of varying the partition bins w on the TOLDI descriptor.

There are very few parameters in our TOLDI method, where the dimensionality of the TOLDI descriptor is simply controlled by the number of partition bins w. This parameter has two major effects on the TOLDI descriptor. First, it determines the dimensionality of the TOLDI descriptor. Large values of w would cause more time consumption during feature matching and make the descriptor less compact. Second, it affects the descriptiveness and robustness of the TOLDI descriptor. More partition bins would bring a more distinctive TOLDI feature. Meanwhile, it would also deduce the robustness of the descriptor to varying mesh resolutions and the nonuniformity of points. Thus, it is a tradeoff between feature's descriptiveness and robustness. Notably, the support radius r for a local feature descriptor is also an arguably important parameter, since it determines the scale of a feature descriptor [27, 6]. However, the parameter r is closely related to a specific application context. For instance, in 3D registration and object recognition areas, a moderate size of r is necessary because large values of r would increase the descriptors's sensitivity to missing border regions, clutter and occlusion, whereas small values of r would make a descriptor be less distinctive [14, 10]. While in retrieval scenario, a relatively large r would be beneficial because more discriminative information would be included without bringing additional nuisances. In these regards, we set r as 15 mr (mr denotes mesh resolution) as suggested in [10, 35] to strike a balance among a descriptor's descriptiveness, robustness and time efficiency. We thereupon focus on the analysis of parameter w.

In order to determine a reasonable w for TOLDI, we test the performance of TOLDI descriptor against different w on the Bologna Retrieval dataset (see section 4.1.1). Specifically, we downsample the scenes to $\frac{1}{4}$ of their original mesh resolution and add Gaussian noise with a standard deviation of 0.3 mr (mr denotes mesh resolution) to them, so as to create a combination of nuisances in this dataset. The RPCs (see section 4.1.2 for a detailed introduction) under various parameter settings are thereafter calculated for analysis as shown in Fig. 4.

From the result, we find that the performance of the TOLDI descriptor improves as w increases from 5 to 20, and it gradually drops when w is larger than 20. Specifically, the performance with $w = 5$ is signif-

icantly inferior to the others. It is because that the descriptor would lose many details about the local shape when w is too small. Based on above observations and discussions, we set $w = 20$ in this paper.

4. Experiments

In this section, the performance of our TOLDI approach (i.e., the LRF and TOLDI descriptor) is tested on three standard datasets, i.e., the Bologna retrieval (BR) dataset [36], the UWA object recognition (UWAOR) dataset [37, 13], and the UWA 3D modeling (UWA3M) dataset [38]. Our method is also compared with several state-of-the-art methods for rigorous evaluation.

4.1. Experimental setup

Before evaluation, the implementation details of the experiments, including the description of datasets, the adopted evaluation criteria and the parameter settings of the considered methods are reported. All the experiments in this paper are implemented on a PC with an Intel Core i3-2120 3.3GHz CPU and 8GB of RAM.

4.1.1. Datasets

The experimental datasets incorporate three public datasets that are related to different application scenarios, i.e., the BR [36] dataset for shape retrieval, the UWAOR dataset [37, 13] for 3D object recognition, and the UWA3M dataset [38] for partial 3D view matching. The variety of feature matching datasets definitely helps us to justify the performance of our method with a comprehensive manner.

Specifically, there are 6 models and 18 synthetic scenes in the BR datasets. The 6 models are noise free ones taken from the Stanford 3D Scanning Repository [39], which are scanned by a Cyberware 3030 MS scanner. The scenes in the BR dataset are the rotated and noisy copies of the models with three increasing levels of noise (i.e., 0.1 mr, 0.3 mr and 0.5 mr Gaussian noise). Note that, we create another seven sets of scenes with respectively two levels of noise (i.e., 0.2 mr and 0.4 mr Gaussian noise) and five levels of mesh decimation ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ of original mesh resolution), so as to enrich the nuisances contained in this dataset. Afterwards, the BR dataset includes 6 models and 60 scenes in total. The UWAOR [37, 13] dataset contains 5 models and 50 real scenes, where the scenes were generated by randomly placing four or five models together in a scene and scanned from one view using a Minolta Vivid 910 scanner. Clutter and occlusion are the major challenges in this dataset. The UWA3M dataset involves 22, 16, 16, and 21 2.5D view scans respectively from four objects. These 2.5D views are scanned by a Minolta Vivid 910 scanner. Due to that a single viewpoint can not capture the structure of an entire 3D model, feature description and matching in this dataset are confronted with the nuisances of missing regions, holes and self-occlusion. Note that all the experiments are performed between two point clouds, hereinafter referred as a model and scene point clouds. Fig. 5 shows two exemplar models and scenes in each dataset.

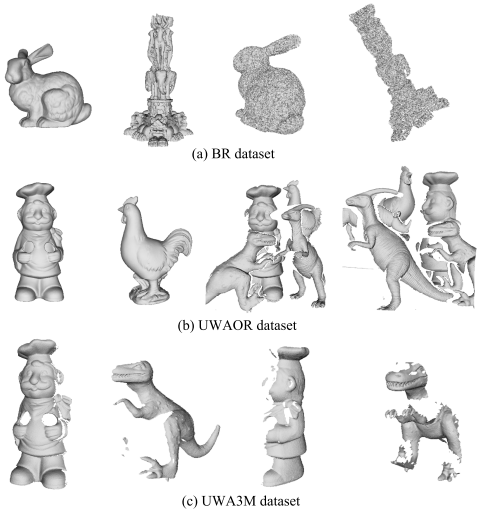


Figure 5: Two exemplar models and scenes (shown from left to right) respectively taken from the BR, UWAOR, and UWA3M datasets.

4.1.2. Evaluation criteria

The performance of the proposed LRF and TOLDI descriptor is quantitatively assessed using the popular *MeanCos* [6, 14] and recall vs. 1-precision curve

(RPC) [6, 11], respectively. Specifically, they are calculated as follows.

The $MeanCos$ criterion measures the mean angular error of the corresponding axes between the two LRFs, and is calculated as:

$$MeanCos(\mathbf{L}_i^m, \mathbf{L}_j^s) = \frac{Cos'(X)_{i,j} + Cos(Z)_{i,j}}{2}, \quad (9)$$

where $\mathbf{L}_i^m$ and $\mathbf{L}_j^s$ represent two corresponding LRFs between the model and scene, $Cos(Z)$ is the cosine of the angle between the z-axis of the $\mathbf{L}_i^m$ and the transformed $\mathbf{L}_i^s$ obtained using ground truth transformation, and $Cos(X)$ coincides with the x-axis angular error after aligning the z-axes of $\mathbf{L}_i^m$ and the $\mathbf{L}_i^s$. Note that, the third axis (y-axis) can always be computed from the other two axes, and therefore it is not necessary to involve it in $MeanCos$ calculation [14]. The readers may also refer to [14] for a detailed description of the $MeanCos$ criterion. In order to obtain an aggregated result for a dataset, for each model-scene pair in this dataset, 1000 points are first randomly selected from each model, and the corresponding points in the scene are extracted using the ground truth transformation. Then, LRFs are computed for these selected points in the model and scene. At last, the average $MeanCos$ of the $MeanCos$ values of all the corresponding LRFs between any model-scene pair in a dataset is used as the final result. Ideally, the $MeanCos$ of two well coincided LRFs equals to 1. Remarkably, the calculation of $MeanCos$ only requires the calculation of LRFs for the keypoints, and no feature descriptors are computed in this process.

The RPC is calculated as follows. Given a model, a scene and the ground truth transformation, a model feature is matched against all scene features to find the closest and the second closest features. If the ratio between the smallest and the second smallest distances is smaller than a threshold, the model feature and the closest scene feature would be considered as a match. Then, a match would be further defined as a correct one if the distance between its associated points is sufficiently small (i.e., being smaller than half of the support radius of the descriptor in this paper), otherwise it is judged as a false one. By varying the threshold, a curve would be generated. Here, the recall is defined as:

$$\text{recall} = \frac{\text{the number of correct matches}}{\text{total number of corresponding features}}. \quad (10)$$

The 1-precision is defined as:

$$1 - \text{precision} = \frac{\text{the number of false matches}}{\text{total number of matches}}. \quad (11)$$

If the feature obtains both high recall and precision, the RPC would fall in the top left of the plot. In our experiments, 1000 points are randomly selected from each model as the keypoints, and their corresponding points are extracted from the scene using the ground truth transformation as suggested in [10, 31]. The ground truth transformations of the BR and UWAOR are given by the publishers. For the UWA3M dataset, the ground truth transformation of any two partially overlapped views is obtained through first manually aligning the two views and then refining using the ICP algorithm [40].

Both evaluation criteria are calculated between two surfaces to be matched, i.e., a model and a scene, and take effect only if the two surfaces have overlapped regions. All the model-scene pairs in the BR and UWAOR satisfy this rule owing to their retrieval and model-based object recognition contexts, while in the case of 3D registration in the UWA3M dataset, not every two views in an object share overlaps. In this case, we only choose the view pairs with an overlap ratio (i.e., the ratio between the number of corresponding points and the minimum number of the point counts of the two view pairs [38]) being larger than 10% for experiments, and 496 valid pairs are eventually screened out from the UWA3M dataset.

4.1.3. Parameter setting

Our method is also compared to several existing methods for a thorough evaluation. In particular, the proposed LRF is compared with four recent methods proposed by Mian et al. [13], Tomabari et al. [6], Petrelli et al. [15] and Guo et al. [10]. The support radii of all these LRFs are kept the same as 15 mr for a fair comparison. As for the proposed TOLDI descriptor, it is compared with six the-state-of-art descriptors, including the spin image [5], snapshots [12],

Table 1: Parameter settings for seven feature descriptors, where mr denotes mesh resolution.

	Support Radius (mr)	Dimensionality	Length
Spin image	15	15×15	225
Snapshots	15	40×40	1600
FPFH	15	3×11	33
SHOT	15	$8 \times 2 \times 2 \times 10$	320
RoPS	15	$3 \times 3 \times 3 \times 5$	135
TriSI	15	29×1	29
TOLDI	15	$3 \times 20 \times 20$	1200

FPFH [9], SHOT [6], RoPS [10] and TriSI [30]. Note that the spin image descriptor is one of the most cited one in the area of local shape description, and the RoPS descriptor is a recent proposed descriptor that exhibits superior feature matching performance in many standard datasets [1]. The parameter settings of the six feature descriptors are shown in Table 1. All these parameter settings, unless otherwise specified, are used for all the experiments in this paper.

4.2. Performance evaluation of the proposed LRF

The proposed LRF is comprehensively evaluated on the three experimental datasets from several aspects, i.e., its repeatability performance, generalization ability and time efficiency. The repeatability term is the major concern for an LRF [14, 10], the generalization ability aims at validating the benefits of our LRF to other LRF-based descriptors, and the time efficiency term is also an important factor for an LRF proposal.

4.2.1. Repeatability performance

Following the implementation details provided in section 4.1.2, we obtain the *MeanCos* results for each method (see section 4.1.3) on three experimental datasets as shown in Fig. 6.

In the BR dataset (the results correspond to Fig. 6(a)(b)), different levels of Gaussian noise and mesh decimation are injected so as to separately assess the repeatability performance of an LRF with respect to noise and mesh decimation. As witnessed by Fig. 6(a), our LRF achieves the best performance in stability and repeatability aspects with respect to all levels of noise, followed by the methods proposed

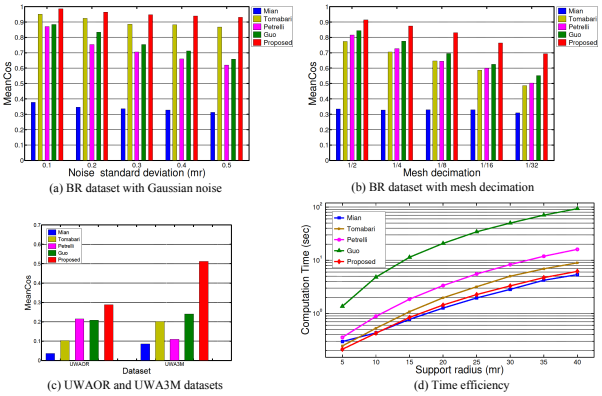


Figure 6: Repeatability and time efficiency performance of five LRF methods on the BR, UWAOR, and UWA3M datasets. The Y axis of the figure in (d) is logarithmic axis for best view.

by Tombari et al., Guo et al., Petrelli et al. and Mian et al. Note that the method proposed by Mian et al. performs significantly worse than the others, because it only partially solves the sign ambiguity problem of the three axes (i.e., only the z-axis is unambiguous). In contrast, the methods with sign disambiguity technique, e.g., proposed by Tombari et al. and Guo et al., achieve much better performance. Besides, we can see that the method proposed by Petrelli et al., which employs a single point for the determination of the x-axis, shows strong sensitivity to high levels of noise. From the results in Fig. 6(b), one can make two major observations. First, the proposed LRF neatly outperforms the others at all levels of mesh resolutions. Specifically, the proposed LRF surpasses the others by a large margin at low levels of mesh resolutions. Second, the method proposed by Guo et al., in turn, outperforms the method proposed by Tomabri et al. with respect to varying mesh resolutions, in opposite to the results in Fig. 6(a). It is because that the method proposed by Guo et al. includes an additional weight for each triangle in the local mesh to achieve invariance to variation of mesh resolutions. However, it is still sensitive to high levels of mesh decimation.

As for the results on the UWAOR and UWA3M datasets (Fig. 6(c)), a significant deterioration of

MeanCos performance can be found for all the tested LRF methods. It is because that clutter and occlusion exist in the UWAOR dataset, and missing border regions as well as self-occlusion are included in the UWA3M dataset, which make the two datasets far more challenging than the BR dataset. In such challenging cases, our LRF still achieves the best performance on both datasets, and the gap appears to be more obvious on the UWA3M dataset. The second best one is the LRF proposed by Guo et al., which is somewhat not surprising since it is originally proposed for object recognition [10], and also demonstrated to be effective in 3D registration scenario [35]. The common trait of our LRF and the one proposed by Guo et al. is that weighting strategies are designed for each triangle or point to resist the impacts of clutter, occlusion, and missing border regions.

There are at least two factors can explain the strong robustness of our LRF. First, all points are employed for the calculation of the critical x-axis to achieve robustness to noise and mesh resolution variation. Second, weighting strategies are applied to each point in order to reduce the impacts brought by clutter, occlusion, and missing border regions.

4.2.2. Generalization ability

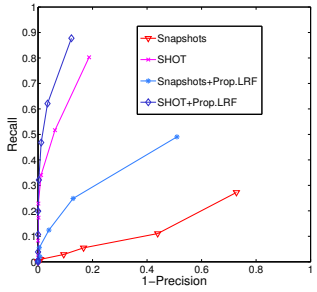


Figure 7: The impact of the proposed LRF on the snapshots [12] and SHOT [6] descriptors in terms of RPC.

To further evaluate the effectiveness of the proposed LRF, we replace the LRFs in two LRF-based descriptors, i.e., the snapshots [12] and SHOT [6] descriptors, with the proposed LRF, and then compare their feature matching performance. Note that the LRF of the snapshots descriptor is the one proposed by Mian et al. [13], and that in SHOT is the one proposed by Tombari et al. [6]. This experiment is conducted on a tuned BR dataset, where the scenes are first simplified to $\frac{1}{4}$ of their original mesh resolution and then added with Gaussian noise with a standard deviation of 0.3 mr. The RPC results of the two original and two modified descriptors are presented in Fig. 7.

It is clear that both the snapshots and SHOT descriptors based on our LRF outperform their original versions. In particular, the snapshots descriptor gains a significant improvement with the proposed LRF. This result has demonstrated two conclusions. One is the universality of our LRF for other LRF-based descriptors. The other is that a more robust and repeatable LRF can boost the feature matching performance of a descriptor without changing its feature representation.

4.2.3. Time efficiency

For each LRF method, we collect the timing statistics of computing the LRFs for the randomly selected points in the models of the BR dataset to test its time efficiency performance. Owing to that the computational time for LRF construction is related to the number of points in the local surface, we calculate the time costs for generating the LRFs with respect to varying support radii for a thorough evaluation. The results are reported in Fig. 6(d).

It can be observed that, the proposed LRF is the most efficient one for calculation when the support radius r is within 10 mr, and then surpassed by the method proposed by Mian et al. when r becomes larger. Note that the repeatability performance of our LRF far surpasses the one proposed by Mian et al. The high time efficiency of our LRF is due to that only a fraction of the local surface vertices is employed for the calculation of z-axis, and computing the x-axis requires add operation of vectors merely. Obviously, the method proposed by Guo et al. is significantly inferior to the other methods in terms of time efficiency. It can be explained that multiple covariance matrices are calculated since they performed covariance analysis for each triangle in the local surface. On the contrary, the methods, e.g., proposed

by Tombari et al. and Mian et al., which calculate a single covariance matrix for the local surface produce far less time cost. Besides, the method proposed by Petrelli et al. is the second time-consuming one. It is because that locating the point with the largest signed projection distance in the local surface requires sort operation.

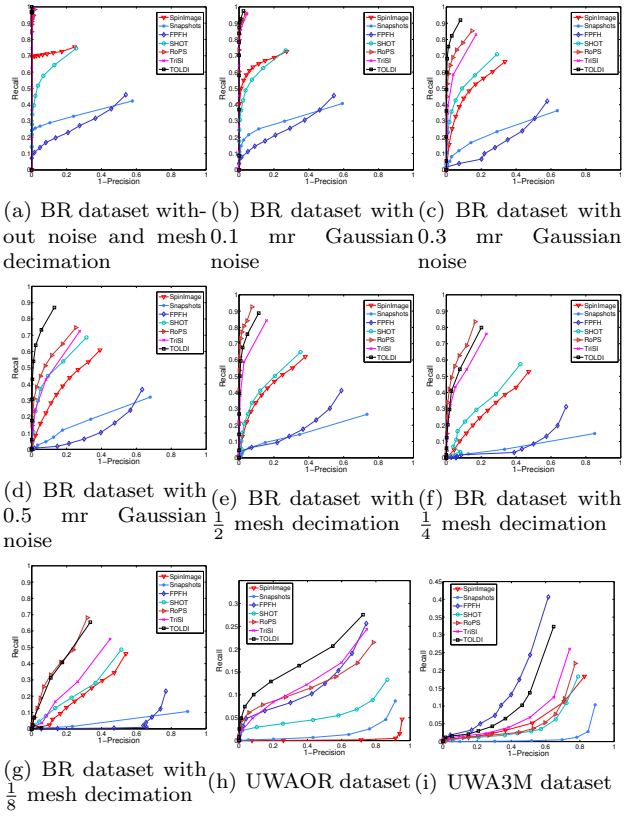
4.3. Performance evaluation of the TOLDI descriptor

The performance of the proposed TOLDI descriptor is also tested on the three experimental datasets using the RPC criterion (see section 4.1.2). In addition to feature matching performance, the time efficiency test of our TOLDI descriptor is also given because some time-crucial applications, such as robotics and mobile phones, have strict limits on a descriptor's computational efficiency.

4.3.1. Feature matching performance

For each descriptor listed in Table 1, we follow the steps described in section 4.1.2 to generate a RPC on a given dataset. The RPC results of our TOLDI descriptor and the other six compared descriptors on the three datasets are presented in Fig. 8.

Regarding aspect of robustness to noise (Fig. 8(a)-(d)), our TOLDI descriptor achieves the best performance on the BR dataset with or without noise, followed by the RoPS and TriSI descriptors. It also can be seen that the margin between our TOLDI and RoPS descriptor gets larger at high levels of noise (i.e., 0.3 mr and 0.5 mr Gaussian noise). The spin image and SHOT descriptors perform comparable with respect to different levels of noise, while the snapshots and FPFH descriptors are significantly inferior to others. It is because that the LRF of snapshots suffers from sign ambiguity and the normal characteristics (i.e., deviation angle between normals) encoded in FPFH is susceptible to noise. With the impact of varying mesh resolutions (Fig. 8(e)-(g)), the proposed TOLDI descriptor behaves the second best under low levels of mesh decimation, which is slightly surpassed by the RoPS descriptor. However, when the mesh decimation reaches $\frac{1}{8}$, our TOLDI descriptor achieves comparable performance with the



(a) BR dataset without noise and mesh decimation (b) BR dataset with 0.1 mr Gaussian noise (c) BR dataset with 0.3 mr Gaussian noise

(d) BR dataset with 0.5 mr Gaussian noise (e) BR dataset with $\frac{1}{2}$ mesh decimation (f) BR dataset with $\frac{1}{4}$ mesh decimation

(g) BR dataset with $\frac{1}{8}$ mesh decimation (h) UWAOR dataset (i) UWA3M dataset

Figure 8: RPC performance evaluation of seven feature descriptors on the BR, UWAOR, and UWA3M datasets.

RoPS descriptor. These observations clearly demonstrate the strong robustness of our TOLDI descriptor against noise and mesh resolution variation. In terms of robustness to clutter and occlusion, as shown in Fig. 8(h), it is necessary to rank the performance of these tested descriptors again. Because the FPFH descriptor, which performs poorly on the BR dataset, shows comparable performance with the RoPS and TriSI descriptors on the UWAOR dataset, and the performance of the spin image descriptor in turn downgrades significantly in object recognition scenario. Nevertheless, the proposed TOLDI descriptor again achieves the best performance on the UWAOR dataset. Eventually, the feature matching performance on the UWA3M dataset, which provides partial view matching scenario, is shown in Fig. 8(i). It can be witnessed that the FPFH and the proposed TOLDI descriptors outperform the other descriptors by a large margin on this dataset. Specifically, our TOLDI is slightly inferior to the FPFH descriptor. It can be inferred that different types of nuisances bring various impacts on feature descriptors. In general, our TOLDI descriptor achieves the best overall robustness to these tested nuisances confirmed by across-dataset experiments.

The high distinctiveness and strong robustness of our TOLDI descriptor can be explained from at least three aspects. First, feature representation in our TOLDI descriptor is based on a repeatable and robust (as demonstrated in section 4.2) LRF, which therefore boosts the robustness of its derived descriptor. Second, local depth feature is adopted to encode the geometric information contained in the local surface, so as to avoid the loss of information going from 3D to 2D/1D representations in opposite to the spin image and RoPS descriptors. Third, the information of three orthogonal views is encoded to achieve a comprehensive information characterization for the local surface, whereas those descriptors (e.g., the snapshots) which only describe the information of a single view would suffer from loss of information caused by clutter and occlusion.

Table 2: The mean and standard deviation (shown in the bottom right brackets) of AUC_{rp} values of feature descriptors with bootstrap sampled keypoints [41] on the experimental datasets. GN, MD and mr respectively represent Gaussian noise, mesh decimation and mesh resolution. The top 2 results for each dataset are shown in bold face.

	Spin image	Snapshots	FPFH	SHOT	RoPS	TriSI	TOLDI
BR	0.745(±3.84E−3)	0.376(±3.06E−3)	0.358(±3.22E−3)	0.713(±1.49E−3)	0.983(±1.33E−3)	**0.987**(±9.00E−4)	**0.999**(±4.00E−4)
BR+0.1 mr GN	0.704(±2.98E−3)	0.347(±2.59E−3)	0.344(±2.94E−3)	0.696(±1.89E−3)	**0.959**(±1.55E−3)	0.952(±1.36E−3)	**0.974**(±1.35E−3)
BR+0.3 mr GN	0.602(±4.62E−3)	0.281(±4.29E−3)	0.270(±3.83E−3)	0.658(±2.21E−3)	**0.834**(±2.83E−3)	0.797(±1.51E−3)	**0.911**(±2.53E−3)
BR+0.5 mr GN	0.518(±2.17E−3)	0.223(±4.15E−3)	0.199(±2.92E−3)	0.625(±2.35E−3)	**0.699**(±3.88E−3)	0.660(±2.13E−3)	**0.850**(±3.06E−3)
BR+$\frac{1}{2}$ MD	0.545(±3.89E−3)	0.183(±4.22E−3)	0.266(±6.45E−3)	0.575(±3.85E−3)	**0.917**(±2.62E−3)	0.814(±3.70E−3)	**0.872**(±2.03E−3)
BR+$\frac{1}{4}$ MD	0.409(±4.04E−3)	0.084(±2.91E−3)	0.144(±3.67E−3)	0.464(±3.12E−3)	**0.797**(±2.87E−3)	0.706(±5.28E−3)	**0.751**(±5.20E−3)
BR+$\frac{1}{8}$ MD	0.314(±5.07E−3)	0.053(±2.58E−3)	0.068(±2.14E−3)	0.349(±2.53E−3)	**0.586**(±2.08E−3)	0.419(±2.60E−3)	**0.561**(±4.24E−3)
UWAOR	0.004(±7.64E−3)	0.021(±2.97E−3)	0.148(±3.34E−3)	0.063(±6.99E−3)	0.138(±4.70E−3)	**0.153**(±6.73E−3)	**0.197**(±6.83E−3)
UWA3M	0.077(±1.36E−3)	0.017(±5.88E−3)	**0.233**(±4.98E−3)	0.067(±4.45E−3)	0.082(±2.57E−3)	0.114(±6.91E−3)	**0.171**(±4.44E−3)

4.3.2. Performance using bootstrap sampled keypoints

The results in section 4.3.1 reflect the feature matching performance of the tested descriptors using random sampled keypoints, following the standard pipeline in [6, 10]. To test the stability of the results under various keypoint samples, bootstrap sampling strategy [41] is considered for further evaluation. That is, the random sampling procedure of keypoints is repeated for several rounds. Afterwards, the fluctuations of the performance of the tested descriptors are calculated. In particular, the mean and standard deviation of the area under RPC curve (denoted by $\mathrm{AUC}_{\mathrm{rp}}$) values of each descriptor are computed. Here, $\mathrm{AUC}_{\mathrm{rp}}$ is a simple and aggregated metric to measure how an algorithm performs over the whole precision-recall space [42]. The results with bootstrap samples of keypoints for all the experimental datasets after 10 rounds are reported in Table 2.

In general, the proposed TOLDI descriptor behaves quite stable under bootstrap sampled keypoints. This is also the case of many other descriptors such as RoPS and TriSI. The performance ranking of all the tested descriptors is consistent with that shown in Fig. 8. To be specific, our TOLDI descriptor still achieves the best overall (ranks either the first or the second for all datasets) feature matching performance with bootstrap sampled keypoints. The result further reveals the distinctiveness and robustness of our TOLDI descriptor when using various keypoint samples. It is an important trait due to that various 3D keypoint detecting strategies such as ISS [43] and MeshDOG [29] would result to different locations of keypoints on 3D point cloud [36]. Preserving high discriminative power under various detected keypoints is necessary for a well-designed surface descriptor.

4.3.3. Time efficiency

We test the time efficiency of these feature descriptors as follows. First, we randomly select 1000 points from each model in the BR dataset. Second, for each feature descriptor, we collect the total time costs of feature extraction for these points with respect to different support radii r. Here, r increases from 5 mr to 40 mr with an interval of 5 mr. The aim is to examine the time efficiency of feature descriptors against varying numbers of points in the local surface, since the major factor that would affect the computational time is the number of points in the local surface [11]. All the descriptors are implemented in the point cloud library (PCL) [44]. The resulting time statistics are reported in Fig. 9.

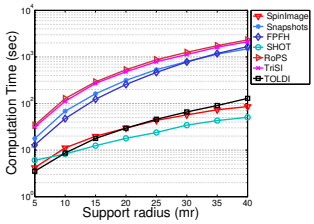


Figure 9: Computational time of feature descriptors with respect to varying support radius. The Y axis is shown logarithmically for clarity.

We can see that our TOLDI descriptor is the most efficient one when r is less than 10 mr, and then ranks the second and the third when r is in the range of $[10, 20]$ mr and $[20, 40]$ mr, respectively. It can be explained that when more points are contained in the local surface, the time consumption of finding the point with the minimum local depth value for each bin in the TOLDI feature vector would increase dramatically. Overall, the RoPS and TriSI descriptors are two most time-consuming one, and their main time consumptions are produced by the time-consuming LRF construction process (as demonstrated in section 4.2). It should be remarked that both RoPS and TriSI descriptors exhibit comparable feature matching performance to our TOLDI descriptor in many cases, yet, they are about one order of magnitude slower compared to our TOLDI descriptor. FPFH is designed for efficient local shape encoding, however, its time consumption increases rapidly as the support radius gets larger. From the results, we can conclude that the spin image, SHOT and TOLDI descriptors are the three most efficient ones among the six descriptors. Particularly, our TOLDI descriptor outperforms the spin image and SHOT descriptors by a large margin in terms of descriptiveness and robustness.

4.4. Application to 3D matching

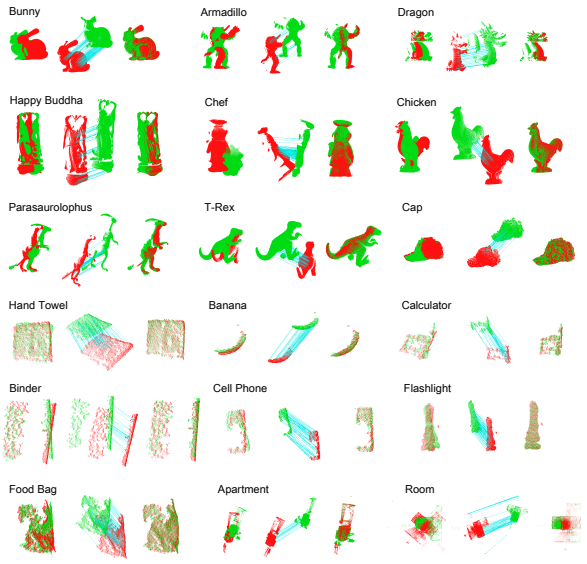


Figure 10: Visual results of 18 pairs of 3D object and scene point clouds with various resolutions using the proposed TOLDI method. From left to right for each point cloud pair, the initial position, point-to-point correspondences, and final registration result of the input point cloud pair. For clarity, the two point clouds to be aligned are respectively shown in red and green, and the correspondences are represented by cyan lines.

To further validate the effectiveness of the proposed TOLDI approach, we employ it to align point cloud scans that are captured from different views. Specifically, the general registration process coincides with our previous proposed local feature-based 3D registration pipeline [18], where the local feature description task is accomplished using our TOLDI method. In this experiment, a variety of real word scanned point cloud data is used for assessment, including the "Bunny" "Armadillo", "Dragon" and "Happy Buddha" data taken from the Stanford Repository [39], the "Chef", "Chicken", "Parasaurolophus", and "T-Rex" data in the UWA3M dataset [38], the "Cap", "Hand Towel", "Banana", "Calculator", "Binder", "Cell Phone", "Flashlight" and "Food Bag" data in the Washington RGB-D object dataset [45], the "Room" data obtained from the PCL [44] website, and the "Apartment" data provided in [46]. The challenges in these sets of point cloud data are many fold. First, they consist of both object (e.g., the Bunny and Banana) and indoor scene data (e.g., the Room and Apartment), which exhibit different geometric properties. Second, these point clouds are acquired by different devices including the LiDAR scanner (e.g., the Room and Chicken) and Kinect (e.g., the Banana and Cap), creating different levels of real noise for 3D matching. Third, both high and low resolution point clouds, such as the Room data with hundreds of thousands of points and the Calculator data with a few hundreds of points, are involved. The registration results are presented in Fig. 10.

One can make several observations from the results. First, all the tested point cloud scans have been automatically and accurately aligned based on our TOLDI method. Second, for both high-quality and low-quality data, sufficient consistent correspondences are established between two point clouds to be aligned, which well validates the robustness of our TOLDI descriptor. Third, even for the indoor scene data that exhibit poor shape geometry, our TOLDI method manages to generate correct correspondences for these indoor scene point clouds. It clearly confirms the high descriptiveness of the proposed TOLDI descriptor.

5. Conclusions and future work

In this paper, we presented a novel TOLDI approach to achieve high descriptiveness, strong robustness and high time efficiency for 3D local shape description. It contains 1) a repeatable and stable LRF proposal, and 2) a distinctive and robust TOLDI feature descriptor.

We constructed the LRF by calculating the normal of the keypoint and the weighted projection vectors of all the radius neighbors of the keypoint. Our technique differs from the existing methods in at least two aspects. First, compared with these CA-based methods [13, 10], which generally calculate the LRF based on the eigenvectors of the covariance matrix computed for the local surface, our method obviates the sign ambiguity problem of the eigenvectors.

Second, compared with those prior PSD-based methods [14, 15], the proposed method on one hand utilizes all points for the calculation of the critical x-axis, and on the other assign weights to them to achieve a balanced robustness to noise, varying mesh resolutions, clutter, and occlusion. In addition, the proposed LRF is universal for other LRF-based descriptors.

Feature representation was then performed for the local surface with three LDI signatures, which were generated from three orthogonal views in the LRF. Both spatial and geometric information was encoded in the TOLDI descriptor in a comprehensive manner. The main characteristics of the proposed TOLDI descriptor are concluded as follows. First, the TOLDI descriptor is associated with a repeatable and robust LRF, and it is invariant to rigid transformation. Second, TOLDI is highly informative because it captures the rich spatial and geometric information of the local surface from multiple views. Besides, differing from most exiting methods that encode shape geometry from the derivative of the local surface, as in [8, 10, 31], our encoding technique is directly performed on 3D points to achieve a minimal loss of information. Third, the TOLDI feature can be extracted right on the initial scanned point clouds, and no complex preprocessings (such as triangulation) are demanded. In contrast, some descriptors including MeshHOG [29] and RoPS [10] are calculated on the mesh representation of 3D models.

In order to evaluate our method, we performed a set of experiments and comparisons on the BR, UWAOR and UWA3M datasets which are respectively related to shape retrieval, object recognition and 3D registration applications. The outcomes conclude that our LRF is highly robust and repeatable against a variety of nuisances, and the proposed TOLDI descriptor exhibits overall superiority in terms of descriptiveness and robustness when compared to the state-of-the-arts. In addition, both the proposed LRF and TOLDI feature are computational efficient.

In the future, there are two interesting directions for our further research. One is about the improvement of the TOLDI descriptor. In this paper, we directly take all the pixel values in one LDI to describe the local depth information, creating a relatively high-dimensional descriptor. We look forward to devising a more compact representation for the LDI feature, as in [10]. In addition, many low-cost instruments have been developed recently, e.g., the Microsoft Kinect device, stereo sensors and structure from motion systems, which are able to capture the texture of a 3D object as well. Integrating RGB information to the TOLDI descriptor would be beneficial when the 3D models exhibit poor geometric features but rich photometric cues. The other one is to integrate the proposed LRF and TOLDI descriptor to specific application algorithms, e.g., 3D object recognition and surface registration.

Acknowledgment

The authors would like to acknowledge the Stanford 3D Scanning Repository, the University of Western Australia (UWA), the Autonomous Systems Lab (ASL), and the University of Washington for making their datasets available to us. We also thank the help of Dr. Alioscia and Dr. Guo for sharing the code of their proposals to us. This work is jointly supported by the National High Technology Research and Development Program of China (863 Program) under Grant 2015AA015904 and the China Postdoctoral Science Foundation under Grant 2014M562028.

References

[1] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, 3d object recognition in cluttered scenes with local surface features: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (11) (2014) 2270–2287.

[2] S. A. A. Shah, M. Bennamoun, F. Boussaid, A novel 3d vorticity based approach for automatic registration of low resolution range images, Pattern Recognition 48 (9) (2015) 2859–2871.

[3] M. Ovsjanikov, Q. Mérigot, F. Mémoli, L. Guibas, One point isometric matching with the heat kernel, in: Computer Graphics Forum, Vol. 29, Wiley Online Library, 2010, pp. 1555–1564.

[4] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, X. Zhou, A two-phase weighted collaborative representation for 3d partial face recognition with single sample, Pattern Recognition 52 (2016) 218–237.

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

[6] F. Tombari, S. Salti, L. Di Stefano, Unique signatures of histograms for local surface description, in: Proceedings of European Conference on Computer Vision, 2010, pp. 356–369.

[7] C. S. Chua, R. Jarvis, Point signatures: A new representation for 3d object recognition, International Journal of Computer Vision 25 (1) (1997) 63–85.

[8] A. E. Johnson, M. Hebert, Surface matching for object recognition in complex three-dimensional scenes, Image and Vision Computing 16 (9) (1998) 635–651.

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

[10] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, J. Wan, Rotational projection statistics for 3d local surface description and object recognition, International journal of computer vision 105 (1) (2013) 63–86.

[11] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, N. M. Kwok, A comprehensive performance evaluation of 3d local feature descriptors, International Journal of Computer Vision 116 (1) (2016) 66–89.

[12] S. Malassiotis, 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 29 (7) (2007) 1285–1290.

[13] A. Mian, M. Bennamoun, R. Owens, On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes, International Journal of Computer Vision 89 (2-3) (2010) 348–361.

[14] A. Petrelli, L. D. Stefano, On the repeatability of the local reference frame for partial shape matching, in: Proceedings of 2011 IEEE International Conference on Computer Vision, 2011, pp. 2244–2251.

[15] A. Petrelli, L. D. Stefano, A repeatable and efficient canonical reference for surface matching, in: Proceedings of 2012 IEEE 3D Imaging, Modeling, Processing, Visualization and Transmission, 2012, pp. 403–410.

[16] A. Flint, A. R. Dick, A. Van Den Hengel, Thrift: Local 3d structure recognition., in: Proceedings of the 9th Conference on Digital Image Computing: Techniques and Applications, Vol. 7, 2007, pp. 182–188.

[17] T. Masuda, Log-polar height maps for multiple range image registration, Computer Vision and Image Understanding 113 (11) (2009) 1158–1169.

[18] J. Yang, Z. Cao, Q. Zhang, A fast and robust local descriptor for 3d point cloud registration, Information Sciences 346 (2016) 163–179.

[19] J. Novatnack, K. Nishino, Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 440–453.

[20] J. Sun, M. Ovsjanikov, L. Guibas, A concise and provably informative multi-scale signature based on heat diffusion, in: Computer graphics forum, Vol. 28, Wiley Online Library, 2009, pp. 1383–1392.

[21] M. Aubry, U. Schlickewei, D. Cremers, The wave kernel signature: A quantum mechanical approach to shape analysis, in: Proceedings of the 2011 IEEE International Conference

on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 1626–1633.

[22] R. Litman, A. M. Bronstein, Learning spectral descriptors for deformable shape correspondence, IEEE transactions on pattern analysis and machine intelligence 36 (1) (2014) 171–180.

[23] D. Boscaini, J. Masci, E. Rodolà, M. M. Bronstein, D. Cremers, Anisotropic diffusion descriptors, in: Computer Graphics Forum, Vol. 35, Wiley Online Library, 2016, pp. 431–441.

[24] R. B. Rusu, Z. C. Marton, N. Blodow, M. Beetz, Persistent point feature histograms for 3d point clouds, in: Proceedings of the 10th International Conference on Intelligent Autonomous Systems, 2008, pp. 119–128.

[25] A. Albarelli, E. Rodolà, A. Torsello, Fast and accurate surface alignment through an isometry-enforcing game, Pattern Recognition 48 (7) (2015) 2209–2226.

[26] F. Stein, G. Medioni, Structural indexing: Efficient 3-d object recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 125–145.

[27] A. Frome, D. Huber, R. Kolluri, T. Bülow, J. Malik, Recognizing objects in range data using regional point descriptors, in: Proceedings of the European Conference on Computer Vision, 2004, pp. 224–237.

[28] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.

[29] A. Zaharescu, E. Boyer, K. Varanasi, R. Horaud, Surface feature detection and description with applications to mesh matching, in: Proceedings of 2009 IEEE Computer Vision and Pattern Recognition, 2009, pp. 373–380.

[30] Y. Guo, F. A. Sohel, M. Bennamoun, M. Lu, J. Wan, Trisi: A distinctive local surface descriptor for 3d modeling and object recognition., in: Proceedings of the 8th International Conference on Computer Graphics Theory and Applications, 2013, pp. 86–93.

[31] Y. Guo, F. Sohel, M. Bennamoun, et al., A novel local surface feature for 3d object recognition under clutter and occlusion, Information Sciences 293 (2015) 196–213.

[32] R. B. Rusu, N. Blodow, Z. C. Marton, M. Beetz, Aligning point cloud views using persistent feature histograms, in: Proceedings of the 21st IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008, pp. 3384–3391.

[33] M. Pauly, Point primitives for interactive modeling and processing of 3d geometry, Ph.D. thesis, Citeseer (2003).

[34] F. Ghorbel, Towards a unitary formulation for invariant image description: application to image coding, in: Annales des telecommunications, Vol. 53, Springer, 1998, pp. 242–260.

[35] Y. Guo, F. Sohel, M. Bennamoun, J. Wan, M. Lu, An accurate and robust range image registration algorithm for 3d object modeling, IEEE Transactions on Multimedia 16 (5) (2014) 1377–1390.

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

[37] A. S. Mian, M. Bennamoun, R. Owens, Three-dimensional model-based object recognition and segmentation in cluttered scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1584–1601.

[38] A. S. Mian, M. Bennamoun, R. A. Owens, A novel representation and feature matching algorithm for automatic pairwise registration of range images, International Journal of Computer Vision 66 (1) (2006) 19–40.

[39] B. Curless, M. Levoy, A volumetric method for building complex models from range images, in:

Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996, pp. 303–312.

[40] P. J. Besl, N. D. McKay, Method for registration of 3-d shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 239–256.

[41] B. Efron, R. J. Tibshirani, An introduction to the bootstrap, Chapman and Hall: Boca Raton, 1994.

[42] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd International Conference on Machine learning, ACM, 2006, pp. 233–240.

[43] Y. Zhong, Intrinsic shape signatures: A shape descriptor for 3d object recognition, in: Proceedings of the 12th International Conference on Computer Vision Workshops, IEEE, 2009, pp. 689–696.

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

[45] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view rgb-d object dataset, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2011, pp. 1817–1824.

[46] F. Pomerleau, M. Liu, F. Colas, R. Siegwart, Challenging data sets for point cloud registration algorithms, The International Journal of Robotics Research 31 (14) (2012) 1705–1711.