Int J Comput Vis (2012) 99:232–255
DOI 10.1007/s11263-012-0526-7

3D Geometric Scale Variability in Range Images: Features and Descriptors

Prabin Bariya · John Novatnack · Gabriel Schwartz · Ko Nishino

Received: 26 January 2011 / Accepted: 17 March 2012 / Published online: 4 April 2012

Abstract Despite their ubiquitous presence, little has been investigated about the scale variability—the relative variations in the spatial extents of local structures—of 3D geometric data. In this paper we present a comprehensive framework for exploiting this 3D geometric scale variability in range images that provides rich information for characterizing the overall geometry. We derive a sound scale-space representation, which we refer to as the geometric scale-space, that faithfully encodes the scale variability of the surface geometry, and derive novel detectors to extract prominent features and identify their natural scales. The result is a hierarchical set of features of different scales which we refer to as scale-dependent geometric features. We then derive novel local shape descriptors that represent the surface structures that give rise to those features by carving out and encoding the local surface that fall within the support regions of the features. This leads to scale-dependent or scale-invariant local shape descriptors that convey significant discriminative information of the object geometry. We demonstrate the effectiveness of geometric scale analysis on range images, and show that it enables novel applications, in particular, fully automatic registration of multiple objects from a mixed set of range images and 3D object recognition in highly cluttered range image scenes.

Keywords Range image · Scale variability · Scale-space · Geometric feature · Shape descriptor · Range image registration · 3D object recognition

1 Introduction

Real-world objects and scenes consist of geometric structures of varying scales. A scene may contain various objects of different dimensions and each individual object may consist of local structures of varying spatial extents. For instance, a forest comprises of a variety of trees of different heights and widths and each tree is made up of a trunk, branches, and leaves, whose spatial extents vary from meters to centimeters. At a finer scale, each individual leaf has geometric details such as a stem and veins of another order of magnitude smaller in size. This scale variation of local geometric structures often define the characteristic geometry of the object or scene. For instance, in a human face, both the tip of the nose and dimples are discriminative geometric features suitable for representing the underlying surface. The spatial extents of such geometric features, however, significantly differ from one another—they lie at entirely different scales. If extracted properly, this geometric scale variability is a source of significant additional information for accurately describing and discriminating the geometry of the object or scene.

This material is based in part upon work supported by the National Science Foundation under CAREER award IIS-0746717 and IIS-0803670. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

P. Bariya · J. Novatnack · G. Schwartz · K. Nishino (✉)
Department of Computer Science, Drexel University,
3141 Chestnut St., Philadelphia, PA 19104, USA
e-mail: kon@drexel.edu

P. Bariya
e-mail: pb339@drexel.edu

J. Novatnack
e-mail: jmn27@drexel.edu

G. Schwartz
e-mail: gbs25@drexel.edu

Geometric scale variability, however, has received little attention in the past: deemed as perturbations that need to be accounted for using multi-scale approaches which, for instance, exhaustively try a pre-determined set of spatial extents of the same spatial operator. In sharp contrast to such approaches that are inevitably tailored to the specific data and applications, our goal is to derive a comprehensive computational framework for *leveraging* the geometric scale variability as another source of information in general applications. In this paper, we focus on extracting scale-dependent 3D geometric features, a unified set of geometric features detected at their own intrinsic scales, and encoding scale-dependent/-invariant geometric structures around these features with local shape descriptors. To this end, we derive a novel representation of the surface geometry, analogous to the 2D image scale-space, that faithfully encodes and makes explicit the scale variability of the surface geometry at hand, which we refer to as the *geometric scale-space* (Novatnack and Nishino 2007, 2008).

We wish to emphasize that our focus is on deriving a sound foundation for analyzing and exploiting the 3D geometric scale-variability in range images, the main source of 3D geometric data in computer vision. To this end, our goal is to derive a canonical scale-space representation, feature detectors and local shape descriptors based upon the surface geometry captured in range images. This approach should not be mistaken with those of other feature detectors and shape descriptors that are tailored to represent 3D models, which usually have different goals such as 3D shape retrieval where the objective is to represent the entire shape compactly so that objects similar to a query object can be found.

We consider the normal field of a range image, which we refer to as the normal map, as the base representation of the surface geometry captured in the range image. We then compute the geometric scale-space by convolving the normal map with Gaussian kernels of increasing standard deviation, where the kernel is defined in terms of the geodesic distance. A rich set of scale-dependent features can then be extracted from the geometric scale-space. In particular, we derive a detector to extract geometric corners at different scales. In order to establish these detectors we carefully derive the first- and second-order partial derivatives of the normal map. We then derive an automatic scale selection method analogous to that of image scale-space theory to identify the natural scale of each feature and to unify all features into a single set. The result is a set of scale-dependent 3D geometric features that provide a rich and unique basis for representing the 3D geometry of the original data. We demonstrate the effectiveness of the geometric scale-space analysis and the resulting scale-dependent features by experimentally evaluating its localization accuracy, repeatability, and robustness against noise and sampling density variation.

We derive a novel scale-dependent local 3D shape descriptor which encodes the geometric information within the natural support region of each feature. This natural support region is defined as a geodesic disc with radius that is proportional to the estimated scale of the corner point. We carve out the local surface within the natural support region and map its normal field onto the tangent plane and interpolate in order to form a regular and dense description of the local surface. The set of scale-dependent local 3D shape descriptors collectively form a sparse hierarchical representation of the surface geometry.

Next, we show how this compact yet discriminative descriptor can be used to robustly match and align a set of range images with a consistent global scale, fully leveraging the hierarchy induced by the scale variation. We also show how we may define a local 3D shape descriptor that is invariant to the variation of the inherent local scale of the geometry, which can be used to register a set of range images with unknown or inconsistent global scales. We demonstrate the discriminative power encoded in the descriptors by using it to fully automatically register an unordered mixed set of range images corresponding to multiple objects—to automatically reconstruct multiple 3D models from a pile of range images. The results clearly demonstrate the power of leveraging 3D geometric scale variability in computer vision applications.

Finally, we show how the hierarchy induced by the scale variation can be exploited along with the highly discriminative local 3D shape descriptors to perform 3D object recognition in cluttered range image scenes. Unlike 3D shape retrieval, the goal for 3D object recognition is to correctly identify and localize 3D objects in scenes scanned from a single viewpoint that may contain multiple objects occluding each other. Thus, local shape descriptors are fundamental to the success of 3D object recognition as occlusion and clutter must be handled robustly, whereas global shape descriptors may suffice for 3D shape retrieval task. To this end, we utilize a tree-based matching scheme and introduce novel constraints based on the added geometric scale information that imposes a hierarchical coarse-to-fine structure to the tree-based matching and effectively culls the otherwise exponentially large search space of correspondences between model and scene features (Bariya and Nishino 2010). We further demonstrate the effectiveness of our approach of analyzing and encoding the scale-variability present in range images by performing recognition experiments on an extensive dataset of real as well as synthetic scenes with varying levels of occlusion and clutter. Preliminary results of the work reported in this paper have been published in Novatnack and Nishino (2007, 2008) and Bariya and Nishino (2010).

2 Related Work

Several methods have been proposed in the past that account for the geometric scale variability in 3D feature detection. Most of these methods are loosely based on the 2D scale-space theory—the analysis of scale variability in intensity images. In 2D scale-space theory, the space of images across different scales, the scale-space, is constructed by successively convolving the image with Gaussian kernels of increasing standard deviation (Koenderink 1984; Lindeberg 1994; Weickert et al. 1999; Witkin 1984). Rich visual features, including corners, edges, and blobs, can then be detected in this scale-space and their intrinsic scales can be identified (Lindeberg 1998).

Previous methods essentially apply the 2D scale-space theory to 3D geometric data by replacing pixel intensities with the 3D vertex coordinates of the mesh model. However, directly "smoothing" the 3D points, for instance with Gaussian kernels (Mokhtarian et al. 2001) or mean curvature flow (Schlattmann 2006), modifies the extrinsic geometry of the original model. This can lead to alterations of the global topology of the geometric data, in particular through fragmentation of the original model (Taubin 1995), which leads to an erroneous scale-space representation. Most past methods also use the Euclidean distance between 3D points as the distance metric in the operator for constructing the scale-space representation (Gelfand et al. 2005; Lalonde et al. 2005; Li and Guskov 2005; Pauly et al. 2006). This, however, can lead to the creation of erroneous features in the scale space, due to local topological changes within the support region of the operator.[1]

Several methods, mostly, for extracting scale-invariant or multi-resolution features or descriptors from range images based on smoothing 3D coordinates or curvature values of the vertices have been proposed in the past (Brady et al. 1985; Ponce and Brady 1985; Morita et al. 1992; Akagunduz and Ulusoy 2007; Li and Guskov 2007; Dinh and Kropac 2006). For instance, Li and Guskov (2007) detect multi-scale interest points for the purpose of object recognition by applying a smoothing operator directly to the 3D point and normal pairs; Gelfand et al. (2005) detect a set of geometric features at multiple scales by varying the radius of a volumetric surface descriptor; and Dinh and Kropac (2006) create a set of multi-resolution spin images (Johnson and Hebert 1999) by varying the bin size and the support size in a predetermined discrete range. Recently, Unnikrishnan and Hebert (2008) detected interest regions and their support sizes using a filtering operator that works in the input unorganized point cloud domain. In contrast, we analyze the surface geometry assuming that the connectivity of the 3D points is known, which can be estimated even for unorganized point clouds as we show in Sect. 7.4.2.

In addition, there are a wide variety of 3D shape descriptors that have been previously proposed (Stein and Medioni 1992; Chua and Jarvis 1997; Johnson and Hebert 1999; Sun and Abidi 2001; Frome et al. 2004; Skelly and Sclaroff 2007). Many of these suffer from the limitation that they are sensitive to the sampling density of the underlying geometry and the size of their support region cannot be canonically determined. Although these methods may achieve certain scale-invariance for the specific applications in mind, they are prone to topological errors induced by the lack of canonical scale analysis. Our fundamental belief is that the localization of the local structures to encode (feature detection) and the manner in which they are encoded (descriptor construction) should both incorporate the geometric scale variability and thus should be handled in an integrated framework such that the right amount of local structure is encoded at the most effective locations. In that regard, past methods do not fully exploit the rich discriminative information encoded in the scale-variability of local geometric structures that can in turn lead to novel computational methods for processing range images.

We demonstrate the effectiveness of our geometric scale analysis by using the resulting local scale-dependent/invariant descriptors for range image registration. Range image registration is a fundamental step of geometry processing in computer vision applications, for instance, to obtain 3D models from scanned data, to navigate based on 3D sensing, etc. In particular, we demonstrate exploiting the geometric scale variability to fully automatically register multiple 3D models from an unordered mixed set of range images of different objects—obtaining 3D models from a casually gathered pile of range images—that shall become a crucial capability given the increased use of raw 3D data in computer vision applications. This is in sharp contrast to previous work on fully automatic range image registration that assume that the given set of range images capture a single object or scene (Chen et al. 1999; Huber and Hebert 2003; Mian et al. 2004; Gelfand et al. 2005; Makadia et al. 2006; ter Haar and Veltkamp 2007).

We further demonstrate the effectiveness of our approach by utilizing the resulting scale-dependent/invariant local 3D shape descriptors for 3D object recognition, which falls under one of the fundamental goals of computer vision. Many of the previously proposed 3D shape descriptors that have been used for 3D object recognition such as *splash* (Stein and Medioni 1992), *point signatures* (Chua and Jarvis 1997), *spin images* (Johnson 1997), etc. suffer from any of a number of limitations such as sensitivity to the sampling rate and noise, low discriminating capability, robustness to occlusion and clutter and that the size of their support region cannot be canonically determined. Moreover, none of

[1] For instance, if two surfaces from different parts of the model lie close to each other, the support region of the scale-space operator will mistakenly include both surfaces.

the past approaches have explicitly explored the use of geometric scale-variability of local surface structures present in the data for 3D object recognition.

Following the publication of the preliminary results of our framework (Novatnack and Nishino 2007, 2008), several alternative constructions of the geometric scale-space have been proposed. For instance, Zou et al. (2009) represent the surface geometry with its Gaussian curvature field and directly smoothes it with Ricci flow to satisfy the causality assumption.[2] In contrast, our method represents the surface geometry with a regular and dense 2D embedding of the surface normal field and constructs the scale-space with successive Gaussian smoothing. This representation has the advantage of carrying directional information leading to rich descriptors, in addition to the fact that normals are less prone to noise since they are the first derivatives, enabling finer localization of features that are independent of the given range image resolution, and being simpler to implement that holds the potential of efficient implementation on the GPU that we leave as our future work. Most important, a 2D representation of the surface geometry and its scale-space is natural and suitable for range images, which are the main source of geometric data in computer vision applications, since they are already 2D projections of the 3D surface.

3 Geometric Scale-Space

Geometric structures that characterize the geometry captured in a range image reside on the surface. For this reason, we must construct a representation that faithfully encodes the scale variability on the surface, i.e., its surface geometry and not its embedding (point coordinates), and allows us to analyze geometric structures across different scales. We choose to represent the geometry of a given surface with its surface normals. By deriving and applying a scale-space operator that accounts for the geodesic distances on the surface, we build a scale-space of this surface normal field, which we refer to as the geometric scale-space.

3.1 2D Representation of the Surface Geometry

We assume that each point in a range image encodes a 3D coordinate of a surface point of the scene geometry $\mathbf{R}: D \to \mathbb{R}^3$, where D is a 2D domain in $\mathbb{R}^2$. We build the geometric scale-space of a range image by first constructing a normal map $\mathbf{N}$ on the same domain. In order to obtain the normal map $\mathbf{N}$, we first approximate the underlying surface by triangulating the range image and then compute a surface normal for each vertex.

We have chosen surface normals as our base representation due to the fact that they are less affected by noise as compared to higher-order derivative quantities such as curvature. Furthermore, they convey directional information of the surface geometry as opposed to scalar quantities such as mean or Gaussian curvature. Note that 3D coordinates cannot be used since they form the extrinsic geometry of the surface and altering their values will consequently alter the actual geometry of the surface.

In order to accurately construct a geometric scale-space of a range image that faithfully encodes the scale-variability of the underlying surface geometry, we define all operators in terms of the geodesic distance rather than the Euclidean distance. The geodesic distance between points in the range image can be directly approximated from the range image itself; given two points $\mathbf{u}, \mathbf{v} \in D$ we approximate the geodesic distance $d(\mathbf{u}, \mathbf{v})$ as

$$d(\mathbf{u}, \mathbf{v}) \approx \sum_{\mathbf{u}_i \in \mathcal{P}(\mathbf{u}, \mathbf{v}), \neq \mathbf{v}} \left\| \mathbf{R}(\mathbf{u}_i) - \mathbf{R}(\mathbf{u}_{i+1}) \right\|, \tag{1}$$

where $\mathcal{P}$ is a list of vertex points in the range image on the path between $\mathbf{u}$ and $\mathbf{v}$. If the path between $\mathbf{u}$ and $\mathbf{v}$ crosses an unsampled point in the range image then we define the geodesic distance as infinity. We also parse the range image and detect depth discontinuities by marking vertex points whose adjacent points lie further than a predetermined 3D distance and define the geodesic distance as infinity if the path crosses such points. Thus, we approximate the geodesic distance as the length of the line segments on the surface joining the two points of interest. Such an approximation is acceptable when the two points are not too far, which is true in our case as most of our analysis is local. Alternatively, other methods such as Fast Marching (Kimmel and Sethian 1998) could also be used.

3.2 Building the Geometric Scale-Space

Given the normal map $\mathbf{N}$, we construct the geometric scale-space of the range image that encodes the evolution of the surface normals on the range image as it is gradually smoothed. We define this as computing the (2-)harmonic flow of a harmonic map from $\mathbb{R}^2$ to $\mathbb{S}^2$ (Tang et al. 2000). The harmonic map is the minimizer of the harmonic energy,

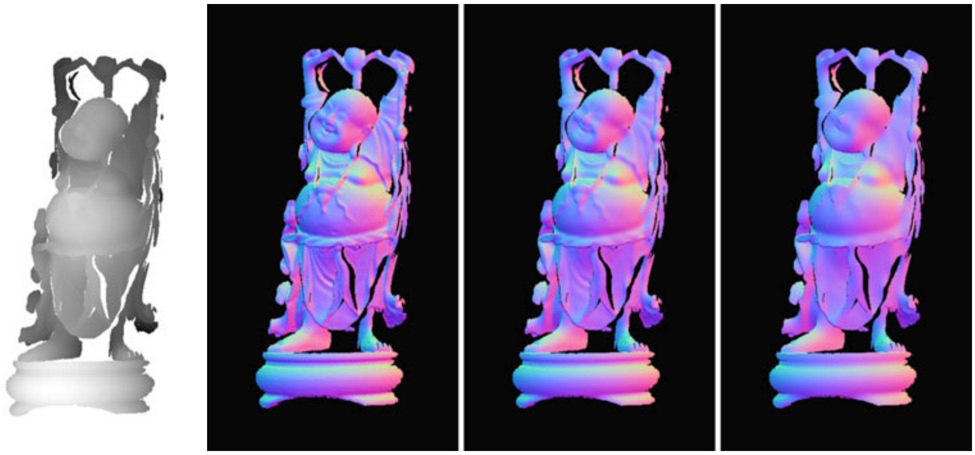
$$\min_{\mathbf{N}:\mathbb{R}^2 \to \mathbb{S}^2} \iint_D \|\nabla \mathbf{N}\|^2 \, ds \, dt, \tag{2}$$

and the harmonic flow corresponds to the gradient-descent flow of the Euler-Lagrange equation of the harmonic energy,

$$\frac{\partial N_i}{\partial t} = \Delta N_i + N_i \|\nabla \mathbf{N}\|^2 \quad (i = 1, 2, 3), \tag{3}$$

where N_i is the i-th component of $\mathbf{N}$ and t corresponds to the scale level in the geometric scale-space.

[2] Note that flow-based analysis of scalar fields on a 3D surface is not new and has been studied in the past, e.g., Kimmel (1997).



(a) Original (b) Scale $\sigma = 0$ (c) Scale $\sigma = 2$ (d) Scale $\sigma = 4$

Fig. 1 The geometric scale-space of a range image of a Buddha model (**a**). As the standard deviation increases from (**b**)–(**d**), finer details of the surface structures are smoothed away. The *red*, *green* and *blue* color channels, in (**b**)–(**d**), encode the direction of the surface normals in the x, y and z directions respectively (Color figure online)

The existence and uniqueness of the harmonic map for $\mathbb{R}^2 \to \mathbb{S}^2$ has been shown (Freire 1995; Struwe 1985). Thus we are able to construct a unique geometric scale-space based on the normal map.[3] However, it has been shown that the harmonic flow is only partially regular and can create singularities in finite time. This means that the geometric scale-space computed based on the normal map may not satisfy the causality assumption—"*any feature at a coarse level of resolution is required to possess a cause at a finer level of resolution*" (Koenderink 1984). However, the cases where the harmonic flow is known to blow up are when the initial data (original normal map in our case) is highly symmetric and at least C^1-continuous (Chang et al. 1992; Hardt 1991), which is very rare for real-world geometric data. Deriving the exact conditions that lead to non-causal geometric scale-space is a difficult problem which we leave as future work. For all the models in our experiments, we did not observe any singularities created in the computed geometric scale-space.

To construct a geometric scale-space, with a discrete set of scale levels, instead of iteratively computing the gradient-descent flow of (3), we convolve the normal map with a Gaussian kernel and renormalize the normals at each level.[4] As in 2D scale-space theory (Lindeberg 1994), the standard deviation of the Gaussian is monotonically increased from fine to coarse scale levels and naturally corresponds to the relative scale of the underlying geometric structure.

We use the geodesic distance as the distance metric to construct a geometric scale-space that encodes the surface geometry. Given a 2D isotropic Gaussian centered at a point $\mathbf{u} \in D$, we define the value of the *geodesic Gaussian kernel* at a point $\mathbf{v}$ as

$$g(\mathbf{v}; \mathbf{u}, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{d(\mathbf{u}, \mathbf{v})^2}{2\sigma^2}\right] \tag{4}$$

where $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is the geodesic distance between the 3D surface points $\boldsymbol{\phi}(\mathbf{u})$ and $\boldsymbol{\phi}(\mathbf{v})$.

Using this geodesic Gaussian kernel, we compute the normal at point $\mathbf{u}$ for scale level σ as

$$\mathbf{N}^\sigma(\mathbf{u}) = \frac{\sum_{\mathbf{v}\in\mathcal{W}} \mathbf{N}(\mathbf{v}) g(\mathbf{v}; \mathbf{u}, \sigma)}{\| \sum_{\mathbf{v}\in\mathcal{W}} \mathbf{N}(\mathbf{v}) g(\mathbf{v}; \mathbf{u}, \sigma) \|}, \tag{5}$$

where $\mathcal{W}$ is a set of points in a window centered at $\mathbf{u}$. The window size is also defined in terms of the geodesic distance and is set proportional to σ at each scale level. In our implementation, we grow the window from the center point while evaluating each point's geodesic distance from the center to correctly account for the points falling inside the window.

Figure 1 shows the geometric scale-space of a range image. A scale-space operator of increasing standard deviation is applied to the original range image (a), corresponding to discrete scale levels of its geometric scale-space (b)–(d). Finer surface structures, e.g., the wrinkles on the drape, are smoothed out early on while more prominent structures with coarser scales, such as the necklace, remain intact. The resulting geometric scale-spaces directly represent the inherent scale-variability of local geometric structures captured in range images and serves as rich basis for further scale-variability analysis of the underlying geometry.

4 Scale-Dependent Geometric Features

Given this geometric scale-space representation of a range image, we may now detect salient features in the geomet-

[3] On the other hand, the harmonic energy for $\mathbb{B}^3 \to \mathbb{S}^2$ has infinite number of solutions (Coron 1990) and hence a unique geometric scale-space cannot be constructed if the Euclidean distance is used.

[4] Observe that (3) can be seen as a diffusion equation with an additional term rooting from the unit vector constraint. The iterative computation of harmonic flow is usually computed by first computing the gradient-descent flow for the diffusion term and renormalizing the vectors at each step (Cohen et al. 1987).

ric scale-space that characterize the underlying 3D geometry across different scales. For this, we first derive the first- and second-order partial derivatives of the normal map $\mathbf{N}^\sigma$ of a range image. A novel corner detector can then be derived using these partial derivatives. We then devise an automatic scale-selection algorithm to identify the natural scale of each feature and unify the features detected at each scale into a single set of scale-dependent geometric features.

4.1 Derivatives of the Normal Map

We first derive the first-order partial derivatives of the 2D normal map in the horizontal (s) and vertical (t) directions. In the following, we describe them only for the horizontal (s) direction. The partial derivatives in the vertical direction (t) may be derived by simply replacing s with t.

At any point in the normal map of a range image, the horizontal direction corresponds to a unique direction on the tangential plane at the corresponding 3D point. The first-order derivative is thus the directional derivative of the normal along this specific direction in the tangential plane, known as the normal curvature. In the discrete domain D the normal curvature in the horizontal ($\mathcal{C}^s$) direction at a point $\mathbf{u} = (s,t)$ may be computed by numerical central angular differentiation:

$$\begin{aligned}\mathbf{N}_s(\mathbf{u}) &= \frac{\partial \mathbf{N}(\mathbf{u})}{\partial s} = \mathcal{C}^s(\mathbf{u}) \\ &\approx \frac{\frac{1}{2}\theta(\mathbf{u}_{-1},\mathbf{u}_{+1})}{L(\mathbf{u}_{-1},\mathbf{u}_{+1})} \\ &\approx \frac{\sin(\frac{1}{2}\theta(\mathbf{u}_{-1},\mathbf{u}_{+1}))}{L(\mathbf{u}_{-1},\mathbf{u}_{+1})},\end{aligned} \tag{6}$$

where $\mathbf{u}_{\pm 1} = (s \pm 1, t)$, $\theta(\mathbf{u}_{-1},\mathbf{u}_{+1})$ is the angle between the normal vectors $\mathbf{N}(\mathbf{u}_{-1})$ and $\mathbf{N}(\mathbf{u}_{+1})$, and $L(\mathbf{u}_{-1},\mathbf{u}_{+1})$ is the chord length between the 3D points $\boldsymbol{\phi}(\mathbf{u}_{-1})$ and $\boldsymbol{\phi}(\mathbf{u}_{+1})$. Note that the half-angle between the adjoining surface normals are assumed to be small such that its value can be approximated with its sinusoidal value. This approximation enables fast computation as

$$\sin\left(\frac{1}{2}\theta(\mathbf{u}_{-1},\mathbf{u}_{+1})\right) = \sqrt{\frac{1-\mathbf{N}(\mathbf{u}_{-1})\mathbf{N}(\mathbf{u}_{+1})}{2}}$$

from the half angle formula, but more important, it enables similar computation for the second order derivative as we show next. And because the normal curvature is a function of adjacent points in the 2D domain D the chord length L is simply the geodesic distance between these points. After applying the discrete geodesic distance in (4), we obtain

$$\mathbf{N}_s(\mathbf{u}) \approx \frac{\sin(\frac{1}{2}\theta(\mathbf{u}_{-1},\mathbf{u}_{+1}))}{d(\mathbf{u}_{-1},\mathbf{u}_{+1})}. \tag{7}$$

Note that because the angle between the two normal vectors is in the range $[0,\pi]$, the first-order derivative is nonnegative at both convex and concave surface points—it is unsigned.

The second-order derivative of the normal map can be derived as

$$\mathbf{N}_{ss}(\mathbf{u}) = \frac{\partial^2 \mathbf{N}(\mathbf{u})}{\partial s^2} = \frac{\partial C^s(\mathbf{u})}{\partial s}. \tag{8}$$

After applying the chain rule to (6) we obtain

$$\begin{aligned}\mathbf{N}_{ss}(\mathbf{u}) \approx{}& \frac{\partial \theta(\mathbf{u}_{-1},\mathbf{u}_{+1})}{\partial s}\frac{\cos(\frac{1}{2}\theta(\mathbf{u}_{-1},\mathbf{u}_{+1}))}{L(\mathbf{u}_{-1},\mathbf{u}_{+1})} \\ &- \frac{\partial L(\mathbf{u}_{-1},\mathbf{u}_{+1})}{\partial s}\frac{2\sin(\frac{1}{2}\theta(\mathbf{u}_{-1},\mathbf{u}_{+1}))}{L(\mathbf{u}_{-1},\mathbf{u}_{+1})^2}.\end{aligned}$$

If we assume that the sampling rate between every adjacent point in D is uniform, the derivative of the chord length L will be zero, and the second term vanishes. This assumption implies that in the local two-neighborhood, 3D distances to the adjacent surface points are approximately the same. This approximation would not hold for adjacent points with large depth variation, in which case we can compute the second term with an additional computational cost. After applying numerical central differentiation to θ and using the half angle formula, the second-order derivative reduces to

$$\begin{aligned}\mathbf{N}_{ss}(\mathbf{u}) \approx{}& \frac{\theta(\mathbf{u}_{-2},\mathbf{u}) - \theta(\mathbf{u}_{+2},\mathbf{u})}{d(\mathbf{u}_{-1},\mathbf{u}_{+1})} \\ &\times \frac{\sqrt{\frac{1}{2}(1+\mathbf{N}(\mathbf{u}_{-1})\cdot\mathbf{N}(\mathbf{u}_{+1}))}}{d(\mathbf{u}_{-1},\mathbf{u}_{+1})}.\end{aligned} \tag{9}$$

This form is particularly attractive as it enables us to compute the second-order derivative in terms of the original normal vectors, and the change in the local angle. The noise associated with higher-order derivatives is reduced as we have avoided an additional numerical differentiation of the first-order derivatives.

4.2 Corners

We wish to detect geometrically meaningful corners points that have high curvature isotropically or in at least two distinct tangential directions. The rich geometric information encoded in the normal maps enables accurate detection of these two types of 3D corners using a two-step geometric corner detector.

We begin by computing the Gram matrix $\mathcal{M}$ of first-order partial derivatives of the normal map $\mathbf{N}^\sigma$ at each point. The Gram matrix at a point $\mathbf{u}$ is defined as

$$\begin{aligned}&\mathcal{M}(\mathbf{u};\sigma,\tau) \\ &= \sum_{\mathbf{v}\in\mathcal{W}} \begin{bmatrix} \mathbf{N}_s^\sigma(\mathbf{v})^2 & \mathbf{N}_s^\sigma(\mathbf{v})\mathbf{N}_t^\sigma(\mathbf{v}) \\ \mathbf{N}_s^\sigma(\mathbf{v})\mathbf{N}_t^\sigma(\mathbf{v}) & \mathbf{N}_t^\sigma(\mathbf{v})^2 \end{bmatrix} g(\mathbf{v};\mathbf{u},\tau),\end{aligned} \tag{10}$$

where $\mathcal{W}$ is the local neighborhood around the point $\mathbf{u}$. In our implementation, we set $\mathcal{W}$ to include neighboring points with geodesic distances within 3σ times the median edge length. $\mathcal{M}$ has two parameters, one that determines the particular scale in the scale-space representation (σ), and one that determines the weighting of each point in the Gram matrix (τ). In our experiments, we set $\tau = \sigma/2$. The corner response at a point $\mathbf{u}$ is defined as the maximum eigenvalue of $\mathcal{M}$. However, due to the unsigned first-order derivative the resulting corner set will contain not only the aforementioned two desired types of geometric corners, but also points lying on 3D edges.

The second-order derivatives of the normal map can be used to prune the corners lying along the 3D edges. We first prune the corner points that are not centered on zero-crossings in both the horizontal and vertical directions. Next we keep only those points where the variance of the second-order partial derivatives in the neighborhood of $\mathbf{u}$ are within a constant factor of each other. The closer this constant factor is to 1, the greater the geometric variance of the selected corner points in both tangential directions.

4.3 Scale Selection

Once features are detected in each scale of the geometric scale-space, they can be unified into a single set. Although a feature may have a response at multiple scales, it intrinsically exists at the scale where the response of the feature detector is maximum. By determining this intrinsic scale for each feature, we obtain a comprehensive scale-dependent 3D geometric feature set.

In order to find the intrinsic scale of a feature we search for the maximum of the normalized feature response across a set of discrete scales, analogous to the 2D automatic scale selection method (Lindeberg 1998). The derivatives are normalized to account for a decrease in the derivative magnitude as the normal maps are increasingly blurred. We define the normalized first-order derivatives $\widetilde{\mathbf{N}}_s^\sigma$ and $\widetilde{\mathbf{N}}_t^\sigma$ as

$$\widetilde{\mathbf{N}}_s^\sigma = \sigma^\gamma \mathbf{N}_s^\sigma \quad \text{and} \quad \widetilde{\mathbf{N}}_t^\sigma = \sigma^\gamma \mathbf{N}_t^\sigma, \tag{11}$$

where γ is a free parameter that is set empirically.[5] The corresponding normalized second-order derivatives are defined as

$$\widetilde{\mathbf{N}}_{ss}^\sigma = \sigma^{2\gamma} \mathbf{N}_{ss}^\sigma \quad \text{and} \quad \widetilde{\mathbf{N}}_{tt}^\sigma = \sigma^{2\gamma} \mathbf{N}_{tt}^\sigma. \tag{12}$$

Figure 2 shows that the normalized first derivative magnitude, used for scale selection, achieves a single local maximum across a large range of scales.[6] Normalized feature responses are computed by substituting the normalized derivatives into the corner detector. The final scale-dependent geometric feature set is constructed by identifying the points in scale-space where the normalized feature response is maximized along the scale axis and locally in a spatial window.

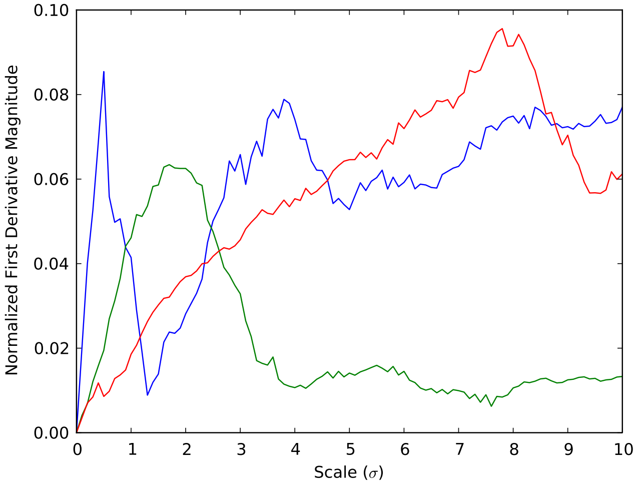


Fig. 2 Normalized first derivative magnitudes at three potential corner locations plotted for a range of scales. The magnitudes show peaks at the natural scales for each corner, and these peaks are maximal inside a large scale range

Figure 3 shows the set of scale-dependent corners detected on two range images of the Happy Buddha. Note that the corners are well dispersed across scales, and that there are a large number of corresponding corner points at the correct corresponding scales. The scale-dependent geometric features accurately encode the geometric scale-variability and can clearly be used as a unique representation of the underlying geometry. These suggest that we should be able to establish robust correspondences across range images once we establish local shape descriptors centered around these corner points, as we will derive in Sect. 5.

4.4 Robustness of Scale-Dependent Features

We experimentally evaluate the effectiveness of our framework for exploiting geometric scale-variability in range images in the form of scale-dependent features. We examine the accuracy by evaluating the repeatability and localization errors of the detected features across different range images of the same 3D object. In addition, we examine the robustness of the feature detection with different levels of noise and sampling rate of the underlying geometry by evaluating the repeatability and localization of the features along two dimensions, within each range image (intra) and across

[5] In our implementation, we use $\gamma = 1$ for the first derivatives and $\gamma = 2$ for the second.

[6] In our experiments, we use scales up to $\sigma = 5$. Also note that, as we use geodesic distance for building the scale-space and its derivatives, depth discontinuities will never contribute to a corner or descriptor.

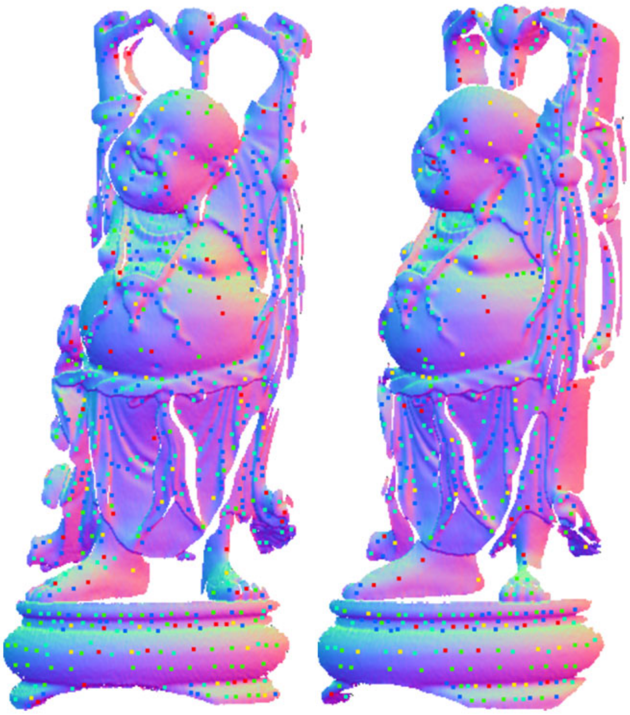


Fig. 3 Scale-dependent corners computed based on geometric scale-space analysis of two range images. The range images are depicted with their normal fields. The scale-dependent corners are colored according to their inherent scales. *Red*, *yellow*, *green*, *turquoise* and *blue* dots indicate the corners detected from the coarsest to the finest scales (Color figure online)

neighboring range image views (inter). The scale-dependent corners were detected on a total of 15 range images of the Happy Buddha model (see Fig. 9(a) for examples) and 12 range images of the Armadillo model. The average inter-repeatability of the corners for the Happy Buddha range images was 68.7 % and 66.49 % for the Armadillo range images. We first brought each pair of neighboring range images (adjacent views) into a common coordinate frame by aligning them using known transformation parameters. The inter-repeatability for such a pair of neighboring range images is then computed as the ratio of the number of corresponding corners found to the total number of corners in the overlapping region of the pair of range images. We count corresponding corners as those detected closest to each other within a pre-determined distance set as a factor of the median edge length between adjacent 3D points in the range images which was about 3 mm in these cases. The mean localization error for the Happy Buddha and Armadillo range images were 1.32 mm and 1.38 mm respectively. The median edge length for both set of range images was approximately 0.35 mm. The height of the Happy Buddha and Armadillo models are 175.8 mm and 197.4 mm respectively. The low localization errors together with the high repeatability indicates that the local geometric structures are reliably detected across varying scales and the resulting scale-dependent corners can likely provide powerful means to compute transformations between range images and 3D models.

4.4.1 Noisy Surface Geometry

We evaluate the robustness of the scale-dependent feature detection to noisy input geometric data by adding Gaussian noise with increasing standard deviation to (a) the surface normals and (b) the depth value of the 3D points in the range image. Figure 4(a) shows the scale-dependent corners detected on one range image of the Happy Buddha model without any added noise and as Gaussian noise is added to the surface normals with standard deviation of 0.02, 0.04, 0.06, 0.08 and 0.1. Although the corners detected at the finest scales are affected as the noise level increases, the corners detected at the coarser scales stay highly consistent.

We compute the intra-repeatability and localization error between the base (noiseless) range image and the noisy version of the same range image, as the percentage of scale-dependent corners detected on the base range image that have a corresponding corner on the noisy range image and the 3D Euclidean distance between corresponding pair of corners respectively. Inter-repeatability and localization error are computed between neighboring range images of the same model at the same noise level. Figure 5 shows the comparison of average repeatability and localization error, both intra and inter, with additive Gaussian noise to the surface normals for 12 range images of the Armadillo model, 15 range images each of the Happy Buddha and Dragon models as well as the overall set of these range images. Figure 6 shows the same comparison, for the same set of range images of the Armadillo, Happy Buddha and Dragon as well as the overall set of these range images, when Gaussian noise is instead added to the depth value of the 3D points of the range image. The standard deviation of the added Gaussian noise were chosen as a percentage of the average standard deviation of the depth values of the 3D points for each range image set. In this case, we added Gaussian noise with standard deviation corresponding to 0.1–0.5 % with increments of 0.1 %. For both cases, noisy surface normals and noisy depth values, the repeatability stays high for the same view (intra) and stays almost constant across different views (inter) despite the noise. The localization error also gracefully increases as the noise level increases. These results demonstrate the robustness of the scale-dependent feature detection and localization.

4.4.2 Varying Sampling Densities

We demonstrate the robustness of the scale-dependent corner detection to changes in surface sampling density by computing scale-dependent geometric corners. Figure 4(b)

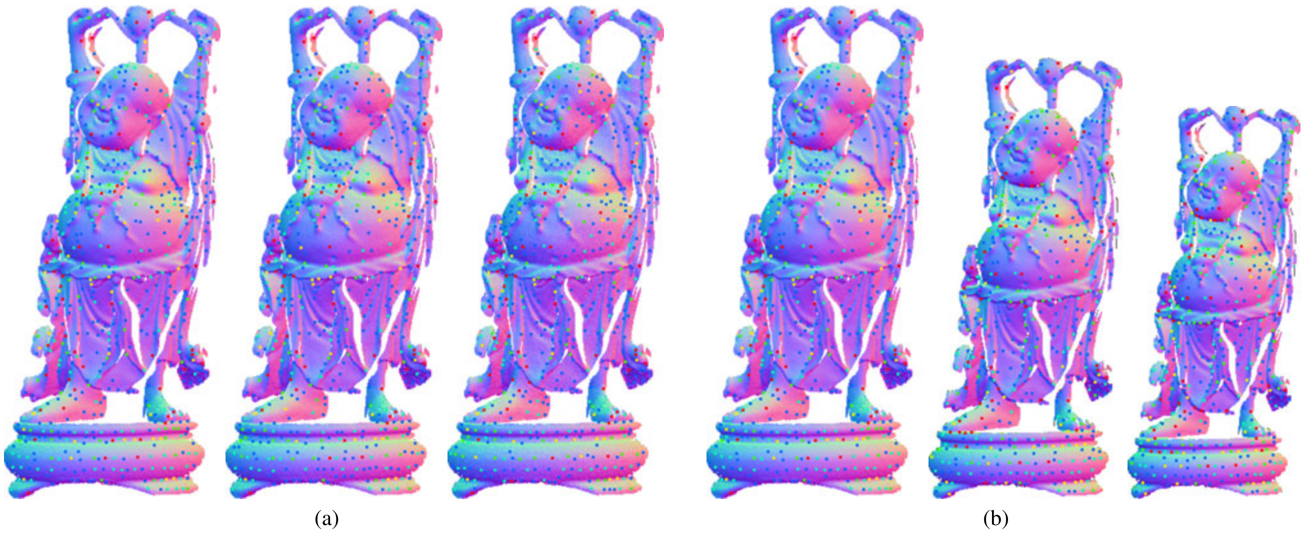


Fig. 4 (**a**) Scale-dependent geometric corner points detected in one view of the Happy Buddha model as noise is incrementally added to the surface normals of the range image. The standard deviation of the Gaussian noise range from 0 (*left most*) to 0.08 (*right most*) with 0.04 increments. The scale-dependent geometric corners are localized reliably despite the significant noise. *Red*, *yellow*, *green*, *turquoise* and *blue* dots indicate the corners detected from the coarsest to finest scales. (**b**) Scale-dependent corners detected for one view of the Happy Buddha model as the range image is subsampled. The original range image (*left most*) is subsampled to 80 % of its original number of points (*right most*) with 10 % decrements. Despite the significant reduction in the sampling of the surface geometry, coarser level corners are coherent. The differences in the size of the three range images shown here is to highlight the difference in the number of sampled points and do not represent any change in the model size (Color figure online)

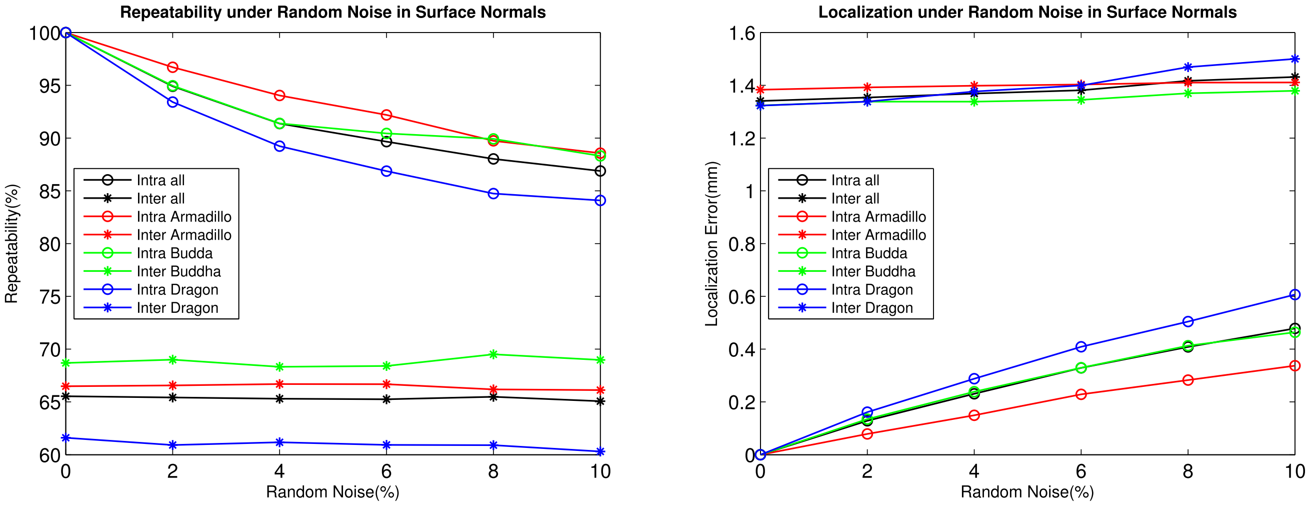


Fig. 5 Average intra and inter repeatability (*left*) and localization error (*right*) of scale-dependent geometric corners for 12 range images of the Armadillo model, 15 range images each of the Happy Buddha and Dragon models and the overall set of all these range images with varying degrees of additive Gaussian noise applied to the surface normals. The repeatability stay high for the same view and stays almost constant across different views. The localization error also increases gracefully as the noise level increases. These results demonstrate the robustness of the scale-dependent corner detection and localization

shows the scale-dependent corners detected on one range image of the Happy Buddha model (left most) and when the same range image is subsampled at increasing rates. The 3D coordinates for each subsampled range image point was obtained by inverse warping to the original range image and by using bilinear interpolation. As can be seen in Fig. 4(b), the corners detected at the coarser scales are highly consistent. Further analysis regarding the repeatability and localization error, both intra and inter, of the detected corners for Armadillo, Happy Buddha and Dragon range image sets as well as the overall set of all range images, is shown in Fig. 7. Similar to the case with noisy range images,

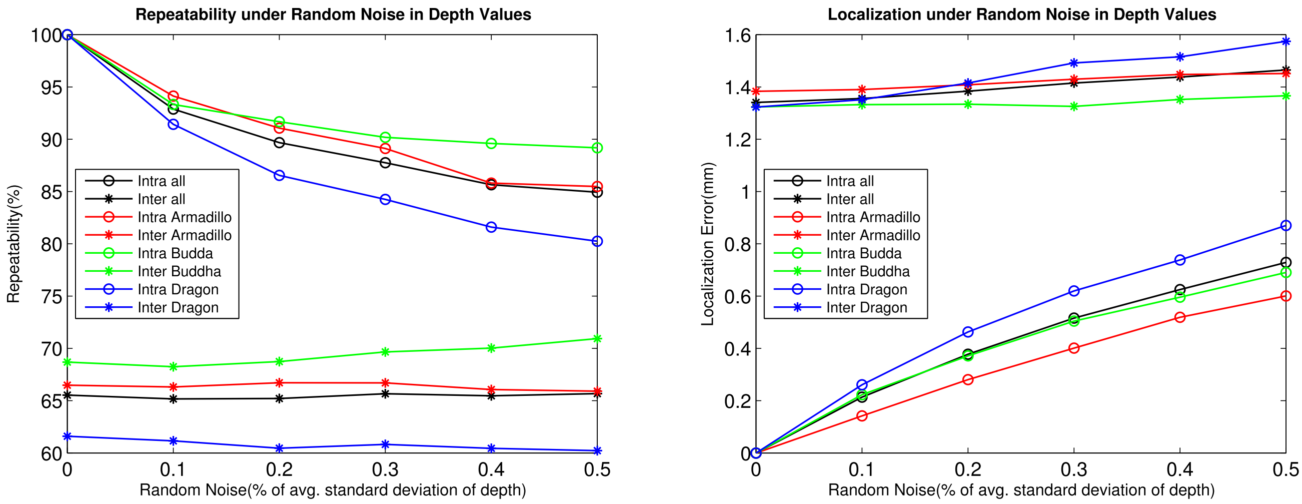


Fig. 6 Average intra and inter repeatability (*left*) and localization error (*right*) of scale-dependent geometric corners for the Armadillo, Happy Buddha and Dragon range image sets and the overall set of all these range images with varying degrees of Gaussian noise applied to the depth values of the 3D points

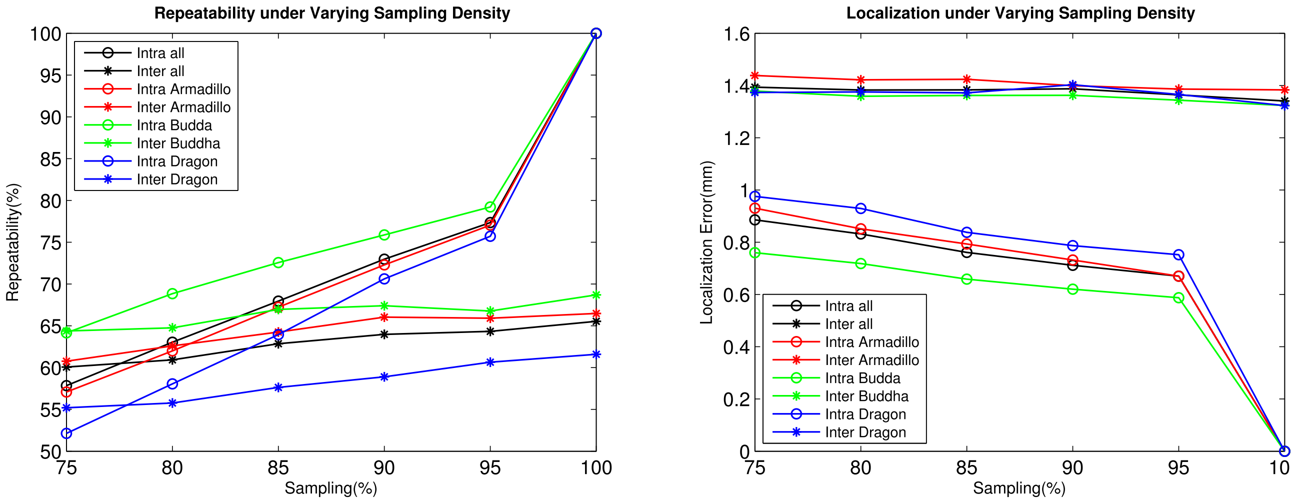


Fig. 7 Average intra and inter repeatability (*left*) and localization error (*right*) of scale-dependent geometric corners for 12 range images of the Armadillo model, 15 range images each of the Happy Buddha and Dragon models and the overall set of all these range images, under varying sampling densities. The repeatability and localization errors decrease gracefully as the subsampling is increased

intra-repeatability and localization error were computed between each range image and its sampled version, and inter-repeatability and localization error were computed between neighboring range images at the same sampling density. The results show that the scale-dependent corners can be detected and localized with accuracy that decreases gracefully with increasing subsampling. The drop-off in intra-repeatability from the original set of range images to their corresponding sampled versions can be attributed to a significant reduction in the number of corners detected at the finer scales of the sampled range images, due to the loss of fine scale geometry caused by the sampling and the interpolation. For example, for the view of the Happy Buddha shown in Fig. 4(b), the range image sampled at 95 % of the original contains approximately 9 % fewer corners and two-thirds of this loss is at the finest scale.

5 Scale-Dependent/Invariant Local Shape Descriptors

Once we detect scale-dependent features via geometric scale-space analysis, we can now carve out and encode the local region of the surface that characterizes the local geometric structure surrounding the feature, in particular, a representative corner point, in the form of a compact local shape descriptor. The associated inherent scale of each scale-dependent corner directly tells us the natural spatial extent (the support size) of the underlying local geometric structure. This information can then in turn be used to identify the size of the neighborhood of each corner that should be encoded in any local shape descriptor. For example, in the spin images descriptor, we may set the size of the neighborhood to be encoded in the descriptor based on the associated inherent scale of the scale-dependent corner point being represented. Here, we propose a novel local shape descriptor that carves out the region within the natural spatial extent of the feature and is insensitive to changes in the sampling rate. We will derive novel scale-dependent and scale-invariant local 3D shape descriptors, which retain the geometric scale-variability as a hierarchical representation or achieve scale-invariance, respectively. We focus on extracting such local shape descriptors to represent range images, in particular, for range image registration and 3D object recognition, which we present in the sections that follow.

5.1 Exponential Map

We construct both our scale-dependent and scale-invariant local 3D shape descriptors by mapping and encoding the local neighborhood of a scale-dependent corner to a 2D domain using the *exponential map*. The exponential map is a mapping from the tangent space of a surface point to the surface itself (Carmo 1976). Given a unit vector $\mathbf{w}$ lying on the tangent plane of a point $\mathbf{u}$, there is a unique geodesic Γ on the surface such that $\Gamma(0) = \mathbf{u}$ and $\Gamma'(0) = \mathbf{w}$. The exponential map takes a vector $\mathbf{w}$ on the tangent plane and maps it to the point on the geodesic curve at a distance of 1 from $\mathbf{u}$, or $\mathrm{Exp}(\mathbf{w}) = \Gamma(1)$. Following this, any point $\mathbf{v}$ on the surface in the local neighborhood of $\mathbf{u}$ can be mapped to $\mathbf{u}$'s tangent plane, often referred to as the Log map, by determining the unique geodesic between $\mathbf{u}$ and $\mathbf{v}$ and computing the geodesic distance and polar angle of the tangent to the geodesic at $\mathbf{u}$ in a predetermined coordinate frame $\{\mathbf{e}_1, \mathbf{e}_2\}$ on the tangent plane. This ordered pair is referred to as the *geodesic polar coordinates* of $\mathbf{v}$.

The exponential map has a number of properties that are attractive for constructing a 3D shape descriptor, most important, that it is a local operator. Although fold-overs[7] may occur if this neighborhood is too large, the local nature of the descriptors implies this will rarely happen. In practice we have observed fold-overs on an extremely small number of features, mostly near points of depth discontinuities in range images. Although the exponential map is not, in general, isometric, the geodesic distance of radial lines from the feature point are preserved. This ensures that corresponding scale-dependent corners will have mostly consistent shape descriptors among different views, e.g., different range images. In addition, because the exponential map is defined at the feature point, it does not rely on the boundary of the encoded neighborhood like harmonic images does (Zhang and Hebert 1999).

5.2 Scale-Dependent Descriptors

We construct a scale-dependent local 3D shape descriptor for a scale-dependent corner at $\mathbf{u}$ whose scale is σ by mapping each point $\mathbf{v}$ in the neighborhood of $\mathbf{u}$ to a 2D domain using the geodesic polar coordinates $\mathcal{G}$ defined as

$$\mathcal{G}(\mathbf{u}, \mathbf{v}) = \big(d(\mathbf{u}, \mathbf{v}), \theta_{\mathcal{T}}(\mathbf{u}, \mathbf{v})\big), \tag{13}$$

where again $d(\mathbf{u}, \mathbf{v})$ is the geodesic distance between $\mathbf{u}$ and $\mathbf{v}$ and $\theta_{\mathcal{T}}(\mathbf{u}, \mathbf{v})$ is the polar angle of the tangent of the geodesic between $\mathbf{u}$ and $\mathbf{v}$ defined relative to a fixed bases $\{\mathbf{e}_1, \mathbf{e}_2\}$. In practice we approximate this angle by orthographically projecting $\mathbf{v}$ onto the tangent plane of $\mathbf{u}$ and measuring the polar angle of the intersection point. We define the neighborhood of each feature that is encoded in the descriptor as points within a geodesic distance of $3\sigma_i$ times the median edge length, where σ_i is the intrinsic scale of the feature. The radius of the scale-dependent descriptor is also set proportional to the inherent scale of the scale-dependent corner σ to encode geometric information in the natural support region of each scale-dependent corner. In our implementation, we set the width and height of the scale-dependent descriptor to $30\sigma_i$ pixels.

After mapping each point in the local neighborhood of $\mathbf{u}$ to its tangent plane we are left with a sparse 2D representation of the local geometry around $\mathbf{u}$. We interpolate a geometric entity, the surface normals, encoded at each vertex to construct a dense and regular representation of the neighborhood of $\mathbf{u}$ at scale σ. We rely on the triangulation of the local neighborhood on the tangent plane, obtained from the triangulation of the range image, in order to aid in the interpolation of the surface normals on the tangent plane. Note that this makes the descriptor insensitive to resolution changes of the range images.

We also choose to encode the surface normals from the original range image, rotated such that the normal at the center point $\mathbf{u}$ points in the positive z direction. The resulting dense 2D descriptor is invariant up to a single rotation (the in-plane rotation on the tangent plane). We resolve this ambiguity by aligning the maximum principal curvature direction at $\mathbf{u}$ to the horizontal axis $\mathbf{e}_1$ in the geodesic polar coordinates, resulting in a rotation-invariant shape descriptor. We approximate this by first computing the eigenvectors of the covariance matrix of 3D points within the local neighborhood and by using the first eigenvector intersected with

[7] A fold-over can occur if the local neighborhood of the corner which gets encoded in the descriptor has multiple points at the same geodesic distance from the corner and in the same direction on the tangent plane.

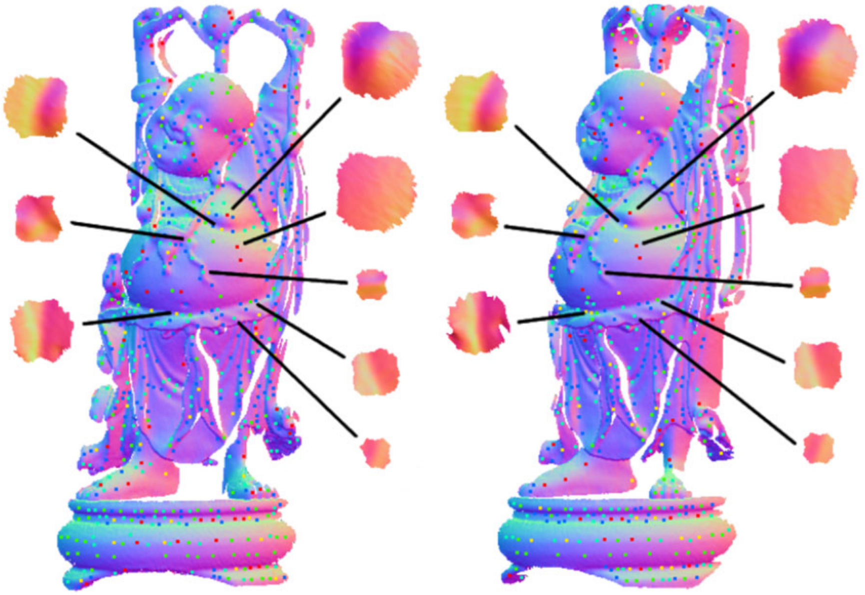


Fig. 8 Scale-dependent corners and scale-dependent local 3D shape descriptors computed based on geometric scale-space analysis of two range images. The range images are depicted with their normal fields. The scale-dependent corners are colored according to their inherent scales, with *red* and *blue* corresponding to the coarsest and finest scales, respectively. The scale-dependent local 3D shape descriptors capture local geometric information in the natural support regions of the scale-dependent features. Here, the colors in the descriptors represent the direction of the normals encoded in the descriptors (Color figure online)

the tangent plane as the positive horizontal axis. The sign of the eigenvector, however, is ambiguous and may be flipped (Bro et al. 2008). In order to remove any ambiguity as a result of this, we also constrain the first eigenvector to point in the same side of the tangent plane as the plane normal.

Once this local basis has been fixed, we re-express each point in terms of the normal coordinates, with the scale-dependent corner point $\mathbf{u}$ at the center of the descriptor. We refer to this dense 2D scale-dependent descriptor of the local 3D shape as $\mathbf{G}_{\mathbf{u}}^{\sigma}$ for a scale-dependent corner at $\mathbf{u}$ and with scale σ. Figure 8 shows subsets of scale-dependent local 3D shape descriptors computed at scale-dependent corners in two range images of the Happy Buddha.

5.3 Scale-Invariant Descriptors

The scale-dependent local 3D shape descriptors collectively provide a faithful sparse representation of the surface geometry in different range images when their global scales are the same or are known, e.g., when we know that the range images are captured with the same range finder. In order to enable comparison between range images that do not have the same global scale, we also derive a scale-invariant local 3D shape descriptor $\widehat{\mathbf{G}}_{\mathbf{u}}^{\sigma}$.

We may safely assume that the scales of local geometric structures relative to the global scale of a range image remains constant as the global scale of a range image is altered. Note that this assumption holds as long as the geometry captured in the range image is rigid and does not undergo any deformation, for instance, as it is captured with possibly different range sensors. We may then construct a set of scale-invariant local 3D shape descriptors by first building a set of scale-dependent local 3D shape descriptors and then normalizing each descriptor's size to a constant radius. In our implementation, we set the width and height of the scale-invariant descriptor to 50 pixels. Such a scale-invariant representation of the underlying geometric structures enables us to establish correspondences between a pair of range images even when the global scale is different and unknown.

5.4 Matching Descriptors

Since each descriptor is a dense 2D image of the surface normals in the local neighborhood, we may define the similarity of the local 3D shape descriptors as the normalized cross-correlation of surface normal fields using the angle differences,

$$\begin{aligned}\mathcal{S}\left(\mathbf{G}_{\mathbf{u}_1}^{\sigma}, \mathbf{G}_{\mathbf{u}_2}^{\sigma}\right) \\ = \frac{\pi}{2} - \frac{1}{|A \cap B|} \sum_{\mathbf{v} \in A \cap B} \arccos\left(\mathbf{G}_{\mathbf{u}_1}^{\sigma}(\mathbf{v}) \cdot \mathbf{G}_{\mathbf{u}_2}^{\sigma}(\mathbf{v})\right), \quad (14)\end{aligned}$$

where A and B are the set of points in the domain of $\mathbf{G}_{\mathbf{u}_1}^{\sigma}$ and $\mathbf{G}_{\mathbf{u}_2}^{\sigma}$, respectively. Here, the similarity measure is defined in terms of the scale-dependent descriptors, but the definition for the scale-invariant descriptors is the same with $\widehat{\mathbf{G}}$ substituted for $\mathbf{G}$.

5.5 Advantages of the Scale-Dependent/Invariant Local Shape Descriptors

The scale-dependent/invariant descriptors have a number of advantages over other previously proposed 3D shape descriptors such as *splash* (Stein and Medioni 1992), *point signatures* (Chua and Jarvis 1997) and *spin images* (Johnson 1997). The dense nature of the scale-dependent/invariant descriptor means that it does not suffer from the sensitivity to the sampling rate as the other descriptors do. An extension to *spin images* (Johnson et al. 1998) shows robustness to the sampling rate. Also, unlike the other descriptors, the size of the support region for the proposed descriptor can be canonically determined and is set proportional to the intrinsic scale of the feature point it is describing. This further implies that these other descriptors are not as robust against occlusion and clutter.

In the spin images descriptor, for example, setting the support region too large for a range image containing multiple objects occluding each other can result in the descriptor using points corresponding to multiple objects. Thus such a descriptor corrupted by clutter cannot be relied upon for an accurate match and nor can a descriptor with a support region too small as this results in a lot of similar descriptors. We present a comparison on the performance of our approach with that of the spin images approach for the purpose of 3D object recognition in cluttered scenes in Sect. 7.

6 Range Image Registration

The novel scale-dependent and scale-invariant local 3D shape descriptors contain rich discriminative information regarding the local geometric structures. As such, these descriptors provide strong cues for matching and aligning 3D geometric data. We will demonstrate this by focusing on registering range images, an integral step in various 3D geometry processing applications such as 3D model construction.

6.1 Pairwise Matching and Alignment

The hierarchical structure of the set of scale-dependent local 3D shape descriptors can be leveraged when aligning a pair of range images $\{\mathbf{R}_1, \mathbf{R}_2\}$ with the same global scale. Note that if we know that the range images are captured with the same range scanner, or if we know the metrics of the 3D coordinates, e.g. centimeters or meters, we can safely assume that they have, or we can covert them to, the same global scale.

Once we have a set of scale-dependent local 3D shape descriptors for each range image, we construct a set of possible correspondences by matching each descriptor to the n most similar.[8] The consistency of the global scale allows us to consider only those correspondences at the same scale in the geometric scale-space, which greatly decreases the number of correspondences that must be later sampled. We find the best pairwise rigid transformation between the two range images by randomly sampling this set of potential correspondences and determining the one that maximizes the area of overlap between the two range images, similar to RANSAC (Fischler and Bolles 1981). However, rather then sampling the correspondences at all scales simultaneously, we instead sample in a coarse-to-fine fashion, beginning with the descriptors with the coarsest scale and ending with descriptors with the finest scale. This enables us to quickly determine a rough alignment between two range images, as there are, in general, fewer features at coarser scales.

For each scale σ_i, starting from the coarsest scale σ_{max}, we randomly construct $N(\sigma_{max} - \sigma_i + 1)$ sets of 3 correspondences, where each correspondence has a scale between σ_{max} and σ_i. For each correspondence set $\mathcal{C}$ we estimate a rigid transformation $\mathcal{T}$, using the method proposed by Umeyama (1991), and then add to $\mathcal{C}$ all those correspondences $(\mathbf{u}_j, \mathbf{v}_j, \sigma_j)$ where $\|\mathcal{T} \cdot \mathbf{R}_1(\mathbf{u}_j) - \mathbf{R}_2(\mathbf{v}_j)\| \le \alpha$ and $\sigma_j \le \sigma_i$. Throughout the sampling process we keep track of the transformation and correspondence set that yield the maximum area of overlap. Once we begin sampling the next finer scale σ_{i-1} we initially test whether the correspondences at that scale increase the area of overlap induced by the current rigid transformation. This allows us to quickly add a large number of correspondences at finer scales efficiently without drawing an excessive number of samples.

Figure 9(a) shows the results of applying our pairwise registration algorithm to two range images of the Happy Buddha captured from different views.[9] The number of correspondences is quite large and the correspondences are distributed across all scales. Although the result is an approximate alignment, since for instance slight perturbations in the scale-dependent feature locations may amount to slight shifts in the resulting registration, the large correspondence set established with the rich shape descriptors leads to very accurate estimation of the actual transformation.

We may align a pair of range images $\{\mathbf{R}_1, \mathbf{R}_2\}$ with different global scales using the scale-invariant local 3D shape descriptors, which amounts to estimating the 3D similarity transformation between the range images. Since we no longer know the relative global scales of the range images, we must consider the possibility that a feature in one range image may correspond to a feature detected at a different scale in the second range image. Our algorithm proceeds by first constructing a potential correspondence set that contains, for each scale-invariant local 3D shape descriptor in the first range image $\mathbf{R}_1$, the n most similar in the second range image $\mathbf{R}_2$. We find the best pairwise similarity transformation by applying RANSAC to this potential correspondence set. For each iteration the algorithm estimates the 3D similarity transformation (Umeyama 1991) and com-

[8] In our experiments, n is set in the range of 5–10.

[9] For this pairwise registration of the Happy Buddha, our Python/C++ implementation on a commodity 2.66 GHz Intel Core2Duo machine required on average 1 minute to complete the registration process including computation of the geometric scale space, corners, descriptors and pairwise alignment with 5000 RANSAC iterations.

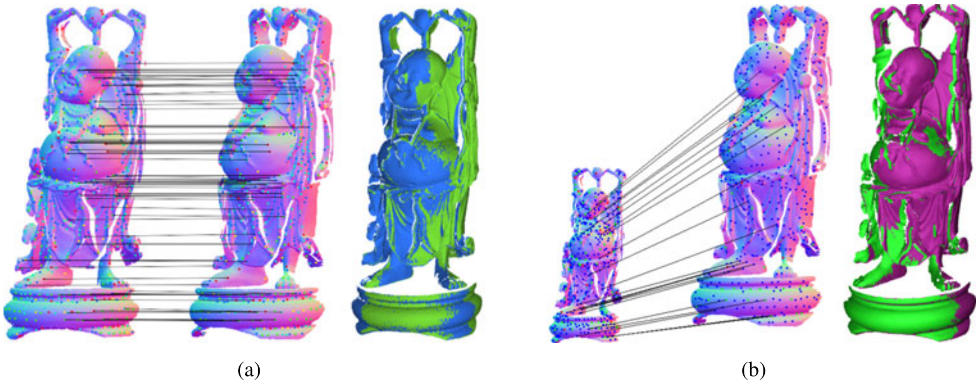


Fig. 9 (**a**) Aligning two range images with the same global scale using a set of scale-dependent local 3D shape descriptors. On the *left* we show the 67 point correspondences found with our matching algorithm and on the *right* the result of applying the rigid transformation estimated from the correspondences. (**b**) Aligning two range images with inconsistent global scales and resolutions using a set of scale-invariant local 3D shape descriptors. On the *left* we show the point correspondences found with our matching algorithm and on the *right* the results of applying the estimated 3D similarity transformation. Both the scale-dependent and -invariant descriptors realize very accurate and efficient automatic pairwise registration of range images

putes the area of overlap. The transformation which results in the maximum area of overlap is considered the best.

Figure 9(b) shows the result of applying our algorithm to two range images of the Buddha model of different views with a relative global scale and resolution difference of approximately 1.8. The two range images were obtained by resampling and scaling the vertex coordinates of one of the range images by a factor of 0.55. Despite the considerable difference in the relative global scale and resolution, we can recover the similarity transformation accurately without any initial alignments or assumptions about the models and their global scales.

6.2 Multiview Matching and Alignment

Armed with the pairwise registration using scale-dependent/-invariant descriptors, we may derive a fully automatic range image registration method that exploits the geometric scale-variability.

Given a set of range images $\{\mathbf{R}_1, \ldots, \mathbf{R}_n\}$, our fully automatic range image registration algorithm first constructs the geometric scale-space of each range image. Scale-dependent features are detected at discrete scales and then combined into a single comprehensive scale-dependent feature set, where the support size of each feature follows naturally from the scale in which it was detected. Each feature is encoded in either a scale-dependent or scale-invariant local shape descriptor, depending on whether the input range images have a consistent global scale or not. We then apply the appropriate pairwise registration algorithm, presented in the previous sections, to all pairs of range images in the input set to recover the pairwise transformations. We augment each transformation with the area of overlap resulting from the transformation. Next we construct a graph similar to the model graph (Huber and Hebert 2003), where each range image is represented with a vertex and each pairwise transformation and area of overlap is encoded in a weighted edge. We prune edges with an area of overlap less than a predetermined threshold. In order to construct the final set of meshes $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ we compute the maximum spanning tree of the model graph and register range images in each connected component using their estimated corresponding transformations. The alignment obtained by our algorithm is approximate yet accurate enough to be directly refined by any ICP-based registration algorithm without any human intervention, resulting in a fully automatic range image registration algorithm. We show the effectiveness of exploiting geometric scale-variability in range image registration with a number of examples in the following section.

6.3 Range Image Registration Results

The novel scale-dependent and scale-invariant local 3D shape descriptors contain rich discriminative information regarding the local geometric structures. As a practical example, we show the effectiveness of these descriptors in range image registration, one of the fundamental steps in geometry processing. We show that the scale-dependent and scale-invariant descriptors can be used to register a set of range images both with and without global scale variations, without any human intervention. Most importantly, we show that we can register a mixed set of range images corresponding to multiple 3D models (with each range image containing a single view of a single model) simultaneously and fully automatically.[10]

[10] In all our experiments, we randomized the order of the range images to ensure that no a priori information is given to the algorithm.

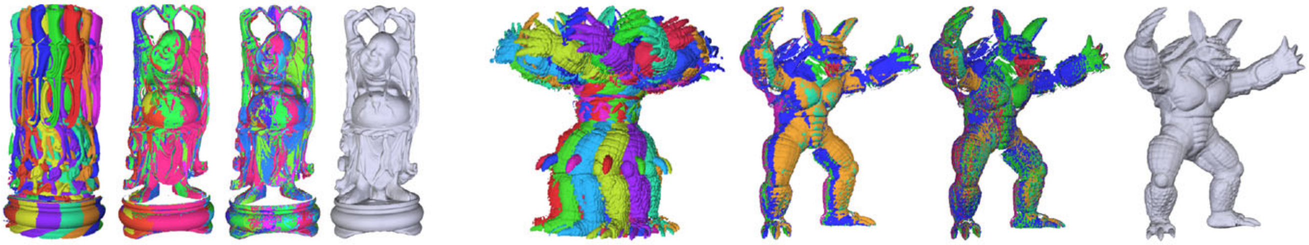


Fig. 10 Fully automatic registration of 15 views of the Happy Buddha model and 12 views of the Armadillo model using scale-dependent local descriptors. For each object, the *first image* shows the initial set of range images. Note that no initial alignment is given and they are situated as is. The *second image* shows the approximate registration obtained with our framework, which is further refined with multi-view ICP (Nishino and Ikeuchi 2002) in the *third image*. Finally a water tight model is built using a surface reconstruction algorithm (Kazhdan et al. 2006) in the *fourth image*. The approximate registration obtained with our framework is very accurate and enables direct refinement with ICP-based methods which otherwise require cumbersome manual initial alignment

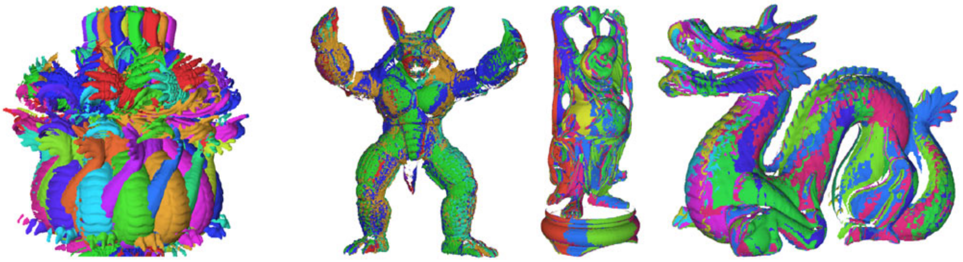


Fig. 11 Automatic registration of a set of range images of multiple objects: total 42 range images shown on the *left most*, consisting of 15 views of the Happy Buddha model, 12 views of the Armadillo, and 15 views of the Dragon model. The scale-dependent local 3D shape descriptors contain rich discriminative information that enables automatic discovery of the three disjoint models from the mixed range image set. Note that, to show the accuracy of our registration, the results shown here have not been post-processed with a global registration algorithm

6.3.1 Range Images with Consistent Global Scale

Figure 10 illustrates the results of applying our framework independently to 15 views of the Happy Buddha model and 12 views of the Armadillo model, with consistent global scales. Scale-dependent local shape descriptors were detected at 5 discrete scales, $\sigma = \{0.5, 1, 1.5, 2, 2.5\}$, in the geometric scale-space. The approximate registration results after applying our matching method using scale-dependent local 3D shape descriptors are refined using multi-view ICP (Nishino and Ikeuchi 2002) and a watertight model is computed using a surface reconstruction method for oriented points (Kazhdan et al. 2006). We may quantitatively evaluate the accuracy of our approximate registration using the local 3D shape descriptors by measuring the displacement of each vertex in each range image from the final watertight model. The average distances for all the vertices in all range images for the Armadillo and Happy Buddha models, relative to the diameter of the models, were 0.17 % and 0.29 % percent, respectively. The results show that the scale-dependent local 3D shape descriptors provide rich information leading to accurate approximate registration that enables fully automatic registration without any need of initial estimates.

Next, we demonstrate the ability of our framework to simultaneously register range images corresponding to multiple 3D models. In order to automatically discover and register the individual models from a mixed set of range images, we prune the edges on the model graph that correspond to transformations with an area of overlap less then some threshold. In practice, we found this threshold easy to set as our framework results in approximate alignments that are very accurate. Figure 11 summarizes the results. Note that no refinement using global registration algorithms has been applied to these results to display the accuracy of our method, but can easily be applied without any human intervention.

6.3.2 Range Images with Inconsistent Global Scale

Our framework is also capable of fully automatically registering a number of range images with unknown global scales. Figure 12 illustrates the results of applying our framework to 15 views of the Happy Buddha and Dragon models. Each range image was globally scaled by a random factor between 0.75 and 1 and subsampled at the same rate using bilinear interpolation to simulate the reduction in range image resolution when scanning objects at increasing

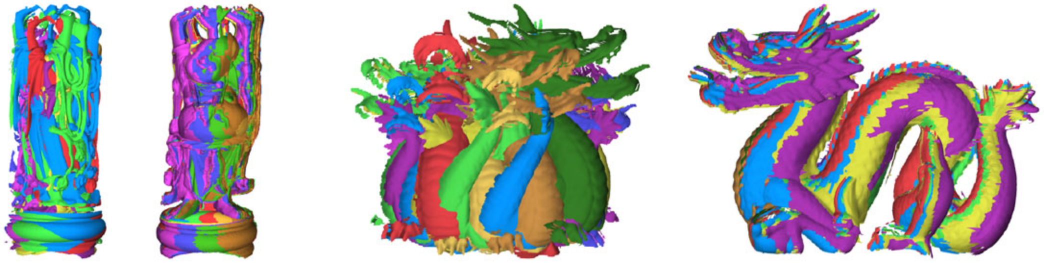


Fig. 12 Automatic registration of 15 views of the Happy Buddha and Dragon models each with a random global scaling from 0.75 to 1. Each view was also subsampled by the same factor to reduce the range image resolution. For each model we visualize the initial set of range images on the *left* and the approximate alignment obtained by our framework on the *right*, prior to any post-processing with ICP. Even with the substantial variations in the global scale and resolution, the scale-invariant local 3D shape descriptors enables us to obtain accurate (approximate) registrations without any assumptions about the initial poses

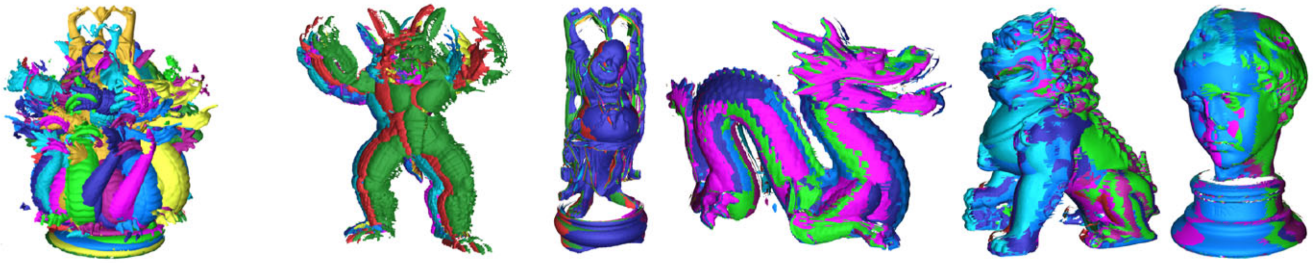


Fig. 13 Automatic approximate registration of 66 randomly scaled and subsampled range images consisting of 15 views each of the Happy Buddha and Dragon models, and 12 views each of the Armadillo, Chinese Lion, and Eros models (shown on the *left* together). Each range image was scaled and subsampled at a randomly chosen rate between 0.75 and 1.0. The scale-invariant local 3D shape descriptors enable automatic approximate registration of the 5 models from this mixed set of range images without any prior information even with significant scale and resolution variations

distances. For each pair of adjacent range images the average errors in the estimated scales after our approximate registration using scale-invariant local 3D shape descriptors were 0.48 % for the Dragon and 1.02 % for the Happy Buddha model. These results show that even with substantial variations in the global scale, our method successfully aligns the range images with high accuracy, which is good enough for subsequent refinement with ICP-based methods as in the examples shown in Fig. 10 without any manual intervention.

Figure 13 shows the results of applying our framework to 66 range images corresponding to views of five different models that have been randomly scaled by a factor between 0.75 and 1. Each view was also subsampled at the same rate to reduce the range image resolution. Despite the scale and resolution variations, our scale-invariant representation of the underlying local geometric structures enables us to automatically register all five models without any human intervention.

7 Three-Dimensional Object Recognition

Next, we show how the hierarchy induced by the scale variation of local geometric structures may be employed in aiding accurate 3D object recognition. The goal of 3D object recognition is to correctly identify objects that are present in a 3D scene, usually in a depth/range image, and to estimate the location and orientation of each object. 3D object recognition has been of interest for industrial automatic assembly but with the availability of portable laser range scanners and especially consumer range scanners such as the Microsoft Kinect, scene understanding of cluttered range image scenes is an important problem to address. We show that by fully leveraging the additional information provided by the scale variability in the matching phase in addition to employing our scale-dependent/invariant local 3D shape descriptors that encode the natural support region for each feature, we can achieve accurate recognition results even in highly cluttered range image scenes.

7.1 Scale-Dependent Model Library and Scenes

We first construct a model library of the 3D models of the objects we are interested in recognizing in the target scenes. In order to compute a scale-dependent representation of each object, we first synthesize range images from a number of uniformly distributed views of the 3D model of the object. The number of views are chosen so that there is overlap between each adjacent pair of views such that all areas of the 3D model are captured in at least one of the synthesized range images. For each synthesized range image, we compute scale-dependent corners at a number of discrete scales. To determine the set of scales to use for the geometric scale space analysis, we choose five proportionately spaced discrete scales such that only 5 % to 10 % of the detected scale-dependent corners are from the coarsest scale. As a consequence, only the most salient geometric features are detected at the coarsest scale. We then compute a scale-invariant local 3D shape descriptor for each scale-dependent corner.

We then represent each object in the model library with its 3D model and a single consolidated set of scale-dependent corners that captures all views of the object and their corresponding descriptors. To do this, each subset of scale-dependent corners computed from each view of the object are brought to a single coordinate frame by using the known transformations between the synthesized views. We remove any duplicate features resulting from the overlaps between any two views of the object. Any two corners within a small distance threshold of each other, detected at the same intrinsic scale and with a degree of similarity above a certain threshold value are considered to be a single feature and one of them is removed.

The scenes to be recognized are range images and thus do not require any preprocessing beside the computation of scale-dependent corners and their corresponding scale-invariant local 3D shape descriptors. The set of scales used to construct the geometric scale-space are determined in the same way as the model scales.

7.2 Constrained Interpretation Tree Matching

Given the scale-dependent representations of the models and scene, we perform matching using an interpretation tree structure (Grimson et al. 1990; Flynn and Jain 1991) that embodies all possible correspondences between model and scene features. For each model M_i to be searched for in a scene S, we create an interpretation tree IT_i. At the root of the tree, there are no correspondences. We build each successive level of the tree by picking a scale-dependent corner from the model and representing its correspondences with highly similar corners from the scene as nodes in the tree. Each node in the tree embodies a hypothesis regarding the presence of the given model in the scene, formed by the set of correspondences at that node and all its parent nodes. And descent in the tree implies an increasing level of commitment to a particular hypothesis (Flynn and Jain 1991).

The search space of all correspondences represented by the entire interpretation tree may be exponentially large for complex scenes (Grimson 1988). For example, for a model with m primitives and a scene with n primitives, there may be n nodes at the first level of an unconstrained tree, n^2 nodes at the second level and so on. Hence constraining and pruning the tree becomes crucial to keep the search space tractable. Here, we exploit the rich discriminative information encoded in the scale-dependent features to impose the following constraints on the tree to keep the search space tractable. From here on, we refer to a scale-dependent corner computed at location $\mathbf{u}$ and with scale σ for a model M_i and scene S as $\mathbf{M}^{\sigma}_{i,\mathbf{u}}$ and $\mathbf{S}^{\sigma}_{\mathbf{u}}$ and their corresponding scale-invariant local 3D shape descriptor as $\hat{\mathbf{M}}^{\sigma}_{i,\mathbf{u}}$ and $\hat{\mathbf{S}}^{\sigma}_{\mathbf{u}}$, respectively.

7.2.1 *Scale Hierarchy*

The scale-dependent corners induce a hierarchy among the set of computed corners based on the intrinsic scale of each corner. The scale-dependent corners detected at coarser scales represent variations in the underlying geometry that are of prominent size and those at finer scales represent smaller variations. Correspondingly, the scale-invariant local 3D shape descriptors corresponding to the scale-dependent corners detected at the coarser scales also encode a larger neighborhood around the detected corner and convey relatively greater discriminative information. We prioritize such feature by matching the scale-dependent corners detected at the coarsest scale first, followed by those detected at increasingly finer scales. As shown in Fig. 14, any pair of model corners $\mathbf{M}^{\sigma_1}_{i,\mathbf{u}_1}$ and $\mathbf{M}^{\sigma_2}_{i,\mathbf{u}_2}$ used to build the successive levels of the tree are chosen so that $\sigma_1 \geq \sigma_2$. This lends a hierarchical structure to the interpretation tree and does away with ambiguities regarding which model primitive to choose to build the next level of the tree.

7.2.2 *Valid Correspondences*

Any two scale-dependent corners that represent the same underlying geometric structure must have the same intrinsic scale. Therefore, we only allow correspondences between two corners if both have the same intrinsic scale. Thus, a correspondence between $\mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o}$ and $\mathbf{S}^{\sigma_b}_{\mathbf{u}_p}$ may be valid only when they have the same intrinsic scale, $\sigma_a = \sigma_b$. We forgo this constraint for scenes that do not necessarily contain the models at the same scale.

We also take advantage of the high discriminability of the scale-invariant local 3D shape descriptors and allow correspondences to be established between features only when their corresponding descriptors are highly similar. A correspondence between $\mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o}$ and $\mathbf{S}^{\sigma_b}_{\mathbf{u}_p}$ is considered valid only when the similarity measure between their corresponding scale-invariant local 3D shape descriptors $\hat{\mathbf{M}}^{\sigma_a}_{i,\mathbf{u}_o}$ and $\hat{\mathbf{S}}^{\sigma_b}_{\mathbf{u}_p}$ is above a similarity threshold s_t.

To account for the possibility that a model corner $\mathbf{M}^{\sigma}_{i,\mathbf{u}}$ might not be present in a scene, we establish a correspondence between each $\mathbf{M}^{\sigma}_{i,\mathbf{u}}$ and a NULL entity as in Flynn and Jain (1991), Grimson (1988) and add this correspondence to

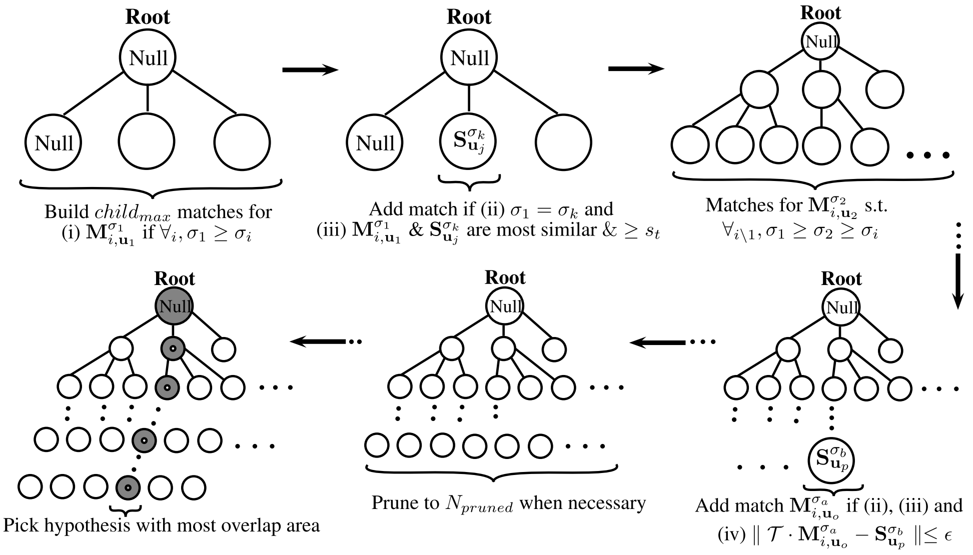

Root
Null
Null
Build $child_{max}$ matches for
(i) $\mathbf{M}^{\sigma_1}_{i,\mathbf{u}_1}$ if $\forall_i, \sigma_1 \geq \sigma_i$
Root
Null
Null
$\mathbf{S}^{\sigma_k}_{\mathbf{u}_j}$
Add match if (ii) $\sigma_1 = \sigma_k$ and
(iii) $\mathbf{M}^{\sigma_1}_{i,\mathbf{u}_1}$ & $\mathbf{S}^{\sigma_k}_{\mathbf{u}_j}$ are most similar & $\geq s_t$
Root
Null
Matches for $\mathbf{M}^{\sigma_2}_{i,\mathbf{u}_2}$ s.t.
$\forall_{i\setminus 1}, \sigma_1 \geq \sigma_2 \geq \sigma_i$
Root
Null
$\mathbf{S}^{\sigma_b}_{\mathbf{u}_p}$
Add match $\mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o}$ if (ii), (iii) and
(iv) $\parallel \mathcal{T} \cdot \mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o} - \mathbf{S}^{\sigma_b}_{\mathbf{u}_p} \parallel \leq \epsilon$
Root
Null
Prune to N_{pruned} when necessary
Root
Null
Pick hypothesis with most overlap area

Fig. 14 Schematic of our scale-hierarchical interpretation tree. For each level, a new model corner with the highest intrinsic scale is chosen and at most $child_{max}$ matches with the most similar scene corners that satisfy the scale, similarity and geometric constraints are added for each branch in the previous level. The hypothesis with the most overlap area is chosen for verification

the tree as a child node for every node in the previous level. Furthermore, we also set a limit $child_{max}$, on the number of valid correspondences that can be added as child node to any particular node in the previous level.

7.2.3 *Geometric Constraint*

Since each node in the tree represents a set of correspondences at that node and all its parent nodes, we can compute a transformation $\mathcal{T}$ for any such node so that the pairs of model and scene corner points that form the set of correspondences are aligned with each other. As a correct set of correspondences should yield an accurate transformation, any correspondence c_{new} between model corner $\mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o}$ and scene corner $\mathbf{S}^{\sigma_b}_{\mathbf{u}_p}$ being considered to be added to the tree as a node must be consistent with the transformation $\mathcal{T}$ for its potential parent node. We enforce this constraint by only allowing correspondences to be added to the tree that satisfy $\|\mathcal{T} \cdot \mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o} - \mathbf{S}^{\sigma_b}_{\mathbf{u}_p}\| \leq \epsilon$, where ϵ is a threshold value.

7.2.4 *Pruning*

We prune the tree when the number of nodes in any level of the tree goes above a threshold value N_{max}. Only N_{pruned} nodes which represent the strongest hypotheses are then kept in the tree. We define the strength of a hypothesis by the cardinality of its correspondence set $|C|$ and the average transformation error induced by its corresponding transformation $\mathcal{T}$, in aligning model and scene corner points in the correspondence set C. To facilitate this, we sort all nodes in the level of the tree to be pruned based on the cardinality of the correspondence set represented by each node in a descending order. Within this sorted list of nodes, the nodes with correspondence sets of the same size are then further sorted in an ascending order based on the average transformation error induced by the hypothesis. The first N_{pruned} nodes in this sorted list is then kept with the rest pruned off.

7.3 Hypothesis Verification and Segmentation

Among the hypotheses represented by the leaf nodes of the tree IT_i, we choose only h_{max} of the strongest hypothesis for verification which entails using the geometric transformation $\mathcal{T}$ defined by it to transform the 3D model of our library object M_i into the scene and evaluating its accuracy given by the area of overlap $A(H_n)$ between the transformed model and the scene. We then choose the hypothesis that produces the maximum area of overlap as the best hypothesis H_{best}, which we refine using ICP. We compute the accuracy of H_{best} as,

$$\alpha(H_{best}) = \frac{A(H_{best})}{M_a(H_{best})}, \tag{15}$$

where, $A(H_{best})$ is the area of overlap between model M_i transformed by H_{best} and the scene S, and $M_a(H_{best})$ is the total visible surface area of the model M_i, within the bounding box of the scene S, after being transformed by H_{best}.

We then accept H_{best} as being correct if $\alpha(H_{best})$ is above a threshold α_t, otherwise we reject it. This essentially means that at least $100\alpha_t$ % of the transformed model within the scene boundaries needs to be visible in the scene. If H_{best} is rejected, then we conclude that model M_i is not present in the scene S. If H_{best} is accepted, we segment the scene S by removing vertices that fall in the overlapping region referenced by $A(H_{best})$. We remove all scale-dependent corners from the scene that fall in $A(H_{best})$ from consideration for the recognition of the next model M_{i+1} in our model library. As a result, the space of all possible correspondences for subsequent recognition of the remaining models in our library, is vastly reduced.

We then proceed with the recognition process by building a new constrained interpretation tree IT_{i+1} for the next model M_{i+1} in our library. We continue this process either until we have built an interpretation tree for all the models in the model library or until there are two or fewer scale-dependent corners available in the scene as a result of the segmentation of the scene, in which case a unique hypothesis cannot be computed.

7.4 Scale-Dependent Recognition Results

We demonstrate the effectiveness of our approach by performing recognition experiments on cluttered real range image scenes from two datasets, the University of Western Australia (UWA) (Mian et al. 2006) and Queen's University (Qu-een's) (Taati et al. 2007; Lab QURCV 2009). These two datasets are the most comprehensive publicly available datasets to date, containing scenes, scanned from a single viewpoint, with multiple 3D objects occluding each other and thus creating clutter.

In our implementation, we relax the constraint for a correspondence between $\mathbf{M}^{\sigma_a}_{i,\mathbf{u}_o}$ and $\mathbf{S}^{\sigma_b}_{\mathbf{u}_p}$ to be considered as valid based on its intrinsic scales and instead regard their correspondence as valid if σ_a and σ_b are within a single relative intrinsic scale of each other. For all the experiments, we use the following set of parameter values: $s_t = 75$ % of self-similarity measure, $child_{max} = 5$ (including the NULL node), $\epsilon =$ between 3 or 4 times the resolution of the synthesized range images used in building the model library, $N_{max} = 2000000$, $N_{pruned} = 2000$, $h_{max} = 20$ and $\alpha_t = 0.3$. Also, to compute the scale-dependent representation of each model, we synthesize range images from eight uniformly distributed views around the vertical axis of the model.

7.4.1 UWA Dataset

The UWA dataset contains five 3D models and 50 real scenes with four or five of the models causing clutter and occlusion. Figure 15(a) and (b) shows the recognition rate on the 50 real scenes, as a function of occlusion and clutter respectively. We define occlusion and clutter for each model in a scene as Mian et al. (2006):

$$\text{occlusion} = 1 - \frac{\text{model surface patch area in scene}}{\text{total model surface area}}, \tag{16}$$

$$\text{clutter} = 1 - \frac{\text{model surface patch area in scene}}{\text{total surface area of scene}}. \tag{17}$$

We manually segmented each of the scenes to compute the ground truth occlusion and clutter values for each object in each scene. We were able to recognize objects with significant occlusion and clutter as shown in Fig. 16. The average recognition rate of our approach was 93.58 % which is comparable to the 95 % overall recognition rate achieved by Mian et al. (2006) on real and synthetic data. Their overall recognition rate, however, was based not only on the 50 cluttered real scenes used here but also on a much larger number of synthetic scenes of simple clutterless views of single objects which we did not have access to.

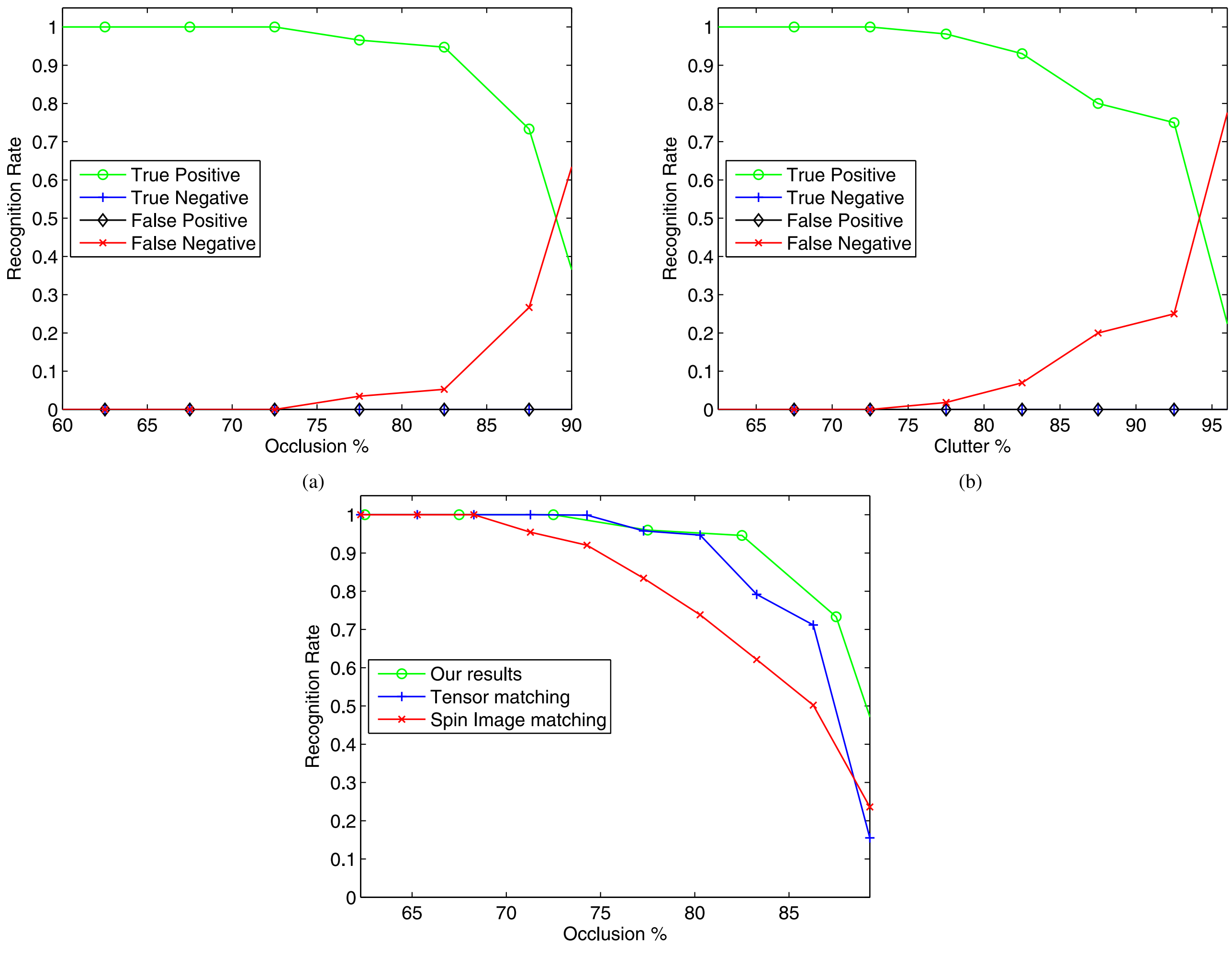
To achieve rigorous and fair evaluation in comparison, we compare our results with the recognition results on the exact same dataset of cluttered scenes presented in Mian et al. (2006)[11] for their tensor matching approach and the spin images recognition algorithm (Johnson and Hebert 1999). As in Mian et al. (2006), we exclude the Rhino model from our recognition results as the spin images algorithm completely failed to recognize the Rhino in any of the scenes as it contained large holes as a result of being scanned from insufficient views. The recognition rate of our approach in this case was 97.5 % and we outperform tensor matching and spin images, which have recognition rates of 96.6 % and 87.8 % respectively, with up to 84 % occlusion. Figure 15(c) shows the recognition rate of our approach as a function of occlusion on the 50 real scenes, excluding results from the Rhino model, for a direct comparison with tensor matching and spin images reported in Mian et al. (2006).

7.4.2 Queen's Dataset

The Queen's dataset contains five models and 80 scenes in point cloud format. Each model was present in four scenes with a single model, 18 scenes with three models, 16 scenes with four models and 10 scenes with all five models. For each model, the normals were available for each of the vertices. The scenes, however, were in an unstructured point cloud form and the vertices did not have associated normals. The scenes in this dataset, hence, point out a limitation of the presented approach, that of the requirement of a range image or a structured point cloud. We, therefore, needed to preprocess these models and scenes such that we may analyze their geometric scale-space and perform 3D object recognition on the scenes.

To that end, we compute a triangulated surface mesh for each model with the Marching Cubes algorithm using MeshLab (2010). Similarly, we obtain a triangulated surface mesh for each of the scenes after estimating the normal for each vertex based on its local neighborhood. Note that, unlike for the models, surface normals were unavailable in the dataset for the scenes. We then use the triangulated surface mesh to synthesize a range image for each scene. For a number of scenes, however, some regions of the scene were lost during

[11] This was confirmed by correspondence with Mian et al. (2006).



(a)

(b)

(c)

Fig. 15 Recognition rates of our scale-dependent approach on 50 real scenes from the UWA dataset with respect to (**a**) occlusion and (**b**) clutter. There are no false positives and the false negatives occur close to 100 % occlusion. Our method achieves consistently high recognition rate across different amounts of occlusion and clutter. Results excluding the rhino are presented in (**c**) for direct comparison with Mian et al. (2006), which we outperform. Please refer to Fig. 19(b) in Mian et al. (2006) for accurate plots for tensor matching and spin images as the plots here are approximate from Mian et al. (2006)

the surface generation process and thus these regions were absent from the generated surface mesh.

Figure 18 shows our recognition results for four scenes from the Queen's dataset. As illustrated in Fig. 18(c) and (d), significant regions of the scene corresponding to some of the models were lost during the surface generation process, the Kid and Zoe models in this instance. Not surprisingly, the Kid and Zoe models could not be recognized in these scenes. Despite this and the significant amount of clutter and occlusion present in the scenes, our average recognition rate for all the models in the Queen's dataset on the 80 scenes was 82.43 %. The recognition rates for the individual models were: 77.08 % for the Angel, 87.5 % for the BigBird, 87.5 % for the Gnome, 83.33 % for the Kid and 76.6 % for the Zoe model. Figure 17 shows the recognition rate of our approach on the 80 scenes as a function of occlusion and clutter.

For a direct comparison with the results reported in Taati (2009), we report our results for the same subset of 55 scenes used for evaluation in Taati (2009).[12] Our average recognition rate in this case was 81.96 %. In comparison, the best reported average recognition rate in Taati (2009) using the spin images descriptor was 53.8 % and using another proposed shape descriptor, the vector quantized variable dimensional local shape descriptor (VD-LSD(VQ)), was 83.8 %. We believe the slightly lower recognition rate for our approach can largely be attributed to the loss of surface regions during the surface generation step we employed. It can also partially be attributed to the fact that we

[12] This was confirmed by correspondence with Taati (2009).

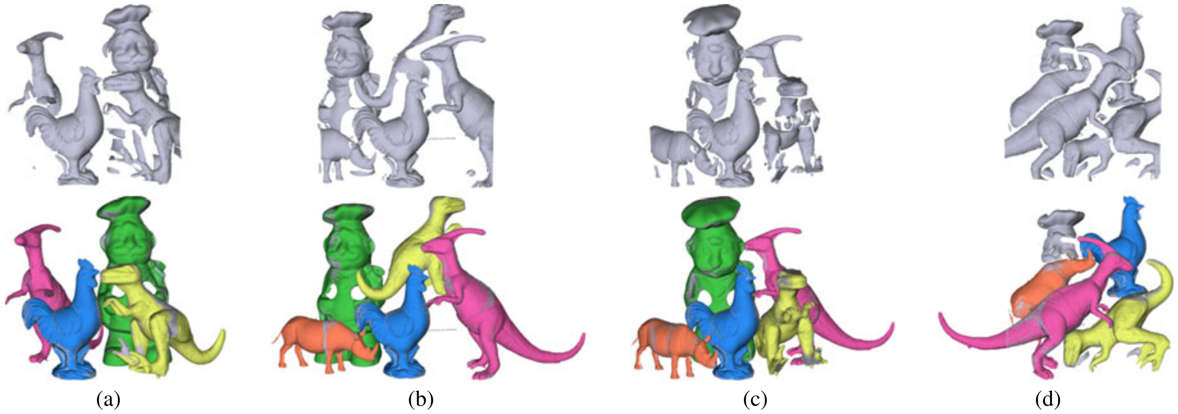
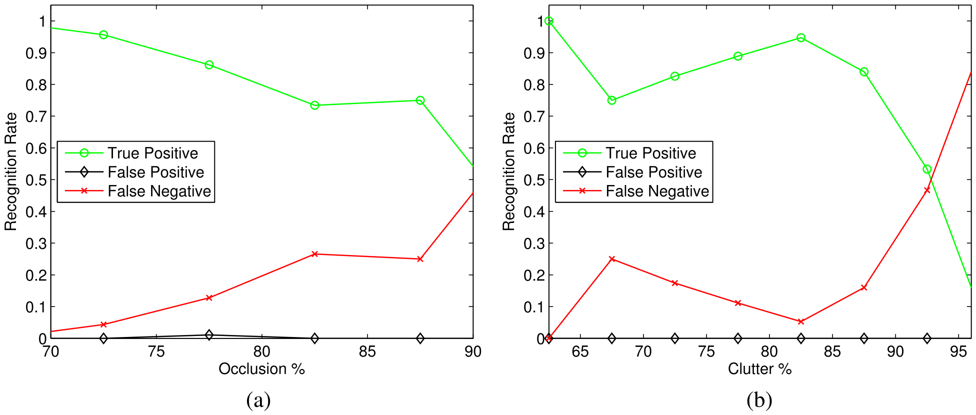


Fig. 16 Scale-dependent recognition results on four scenes from the UWA dataset. All objects that have been recognized are replaced with their 3D models in different colors. Only the Chef in (**d**), which was over 92 % occluded, was not recognized. Our method successfully recognizes the remaining objects despite the significant clutter and occlusion, and localizes each object very accurately

Fig. 17 Recognition rates of our scale-dependent approach on 80 real scenes from the Queen's dataset with respect to (**a**) occlusion and (**b**) clutter



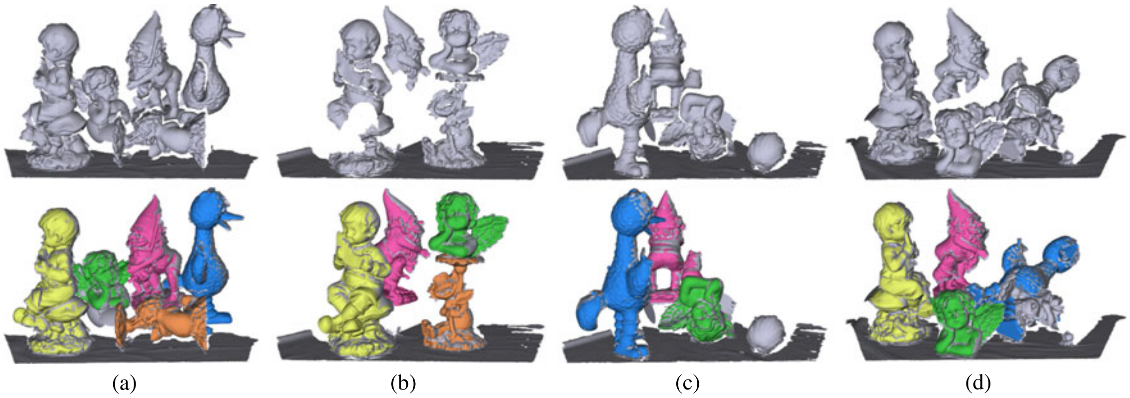


Fig. 18 Scale-dependent recognition results on four scenes from the Queen's dataset. All objects that have been recognized are replaced with their 3D models in different colors. Only the Kid in (**c**) and Zoe in (**d**), the *yellow* and *brown* colored models in (**a**), (**b**) respectively, were not recognized. In both these cases, significant regions of the scene corresponding to these models were lost during the surface generation process. All other models were recognized and localized accurately despite the significant clutter and occlusion (Color figure online)

only synthesized range image views around the vertical axis of each model while building our model library and thus did not explicitly capture the top and bottom views of the models, which were present in some of the scenes. And although our analysis assumes that connectivity between 3D points in the data is given, which was not the case for the original data in this case, we were still able to achieve recognition rates much higher than that for the spin images.

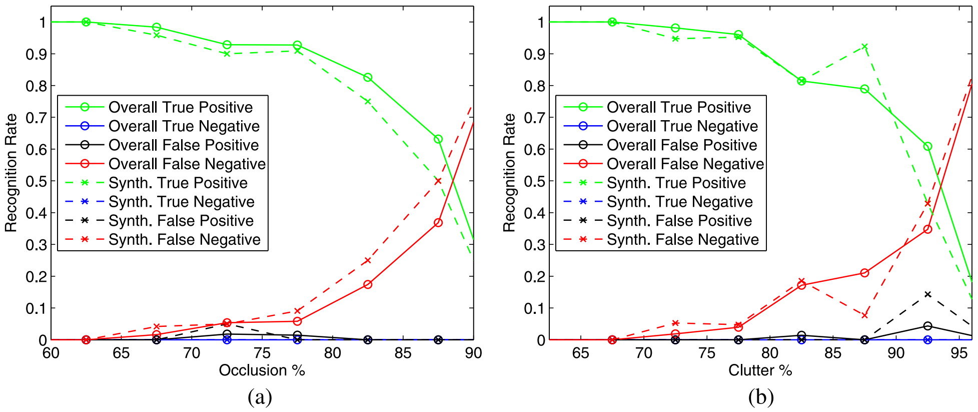


Fig. 19 Recognition rates of our scale-invariant approach with respect to (**a**) occlusion and (**b**) clutter, on real scenes and synthetic scenes containing globally scaled library objects. To our knowledge, we are the first to show systematic results on scale-invariant 3D object recognition

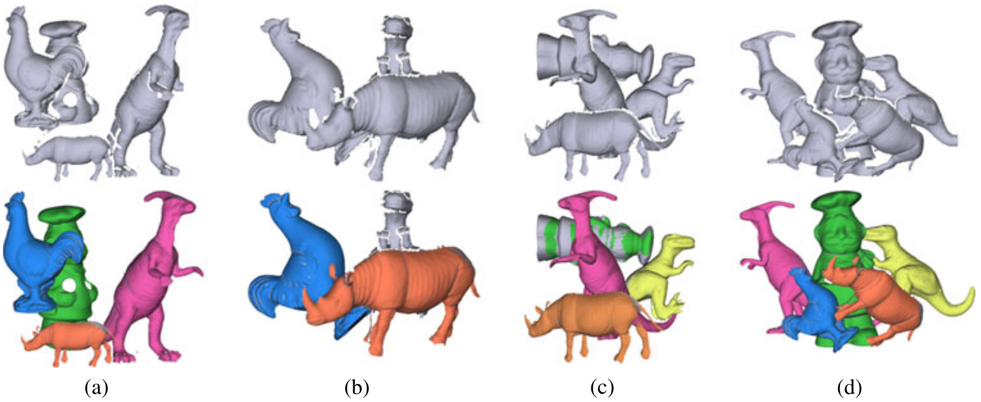


Fig. 20 Four synthetic scenes with objects randomly scaled between 60 % to 150 % of their original sizes. Despite the global scale variation, occlusion and clutter, our method successfully recognizes most objects in the scene and localizes them very accurately

7.5 Scale-Invariant Recognition Results

We perform recognition experiments on all 50 real scenes from the UWA dataset as well as 30 synthesized range image scenes, which we created, where models from the UWA dataset are scaled from 60 to 150 percent of their size. For this set of experiments, we allow for correspondences between corners to be established across all scales and use a similarity transformation $\hat{\mathcal{T}}$ to allow for scale-invariant recognition. All parameter values are set the same as the previous experiments.

As illustrated in Fig. 20, we are able to recognize scaled library objects in range image scenes with significant occlusion and clutter. Figure 19(a) and (b) shows the recognition rate of our scale-invariant approach as a function of occlusion and clutter respectively. We achieve a recognition rate of 89.08 % on the synthetic scenes and an overall recognition rate of 89.29 %. The reduced recognition rate in comparison to the case of same global scale between models and scene (in Sect. 7.4.1) can be interpreted as the direct consequence of the increased search space of correspondences by allowing scaling as part of the transformation.

We have demonstrated that our framework is capable of performing scale-invariant recognition tasks in complex scenes as well. To our knowledge, we are the first to show systematic results on scale-invariant 3D object recognition. We believe our scale-invariant recognition approach has broad practical implications as the model library may be built with a suitably scaled object model and scaled objects can be accurately recognized in a scene. Such an ability is crucial given the wide spread use of range sensors with various modalities ranging from laser range sensing to stereo and to consumer depth cameras, which will lead to abundant depth data of various global scales.

8 Conclusion

In this paper, we proposed the use of the scale variation of local geometric structures as an additional dimension that can be exploited for various computer vision applications. To that end, we presented a novel framework for analyzing and leveraging the scale variability of geometric structures that are captured in a range image. In particular, we introduced the geometric scale-space to unveil and analyze the geometric scale variability, derived methods for detecting geometric features of varying scales and for identifying their scales. We also introduced a novel local shape descriptor that encodes the discriminative local geometric structures according to their natural scales.

Furthermore, we demonstrated how this added dimension provided by the scale of local geometric structures may be exploited as a discriminative property in establishing correspondences between local geometric structures in range images. We showed the power of geometric scale analysis by using the scale-dependent/invariant local 3D shape descriptors in range image registration, which shows that even fully automatic registration of multiple 3D objects from an unordered set of mixed range images can be achieved. We further demonstrated the effectiveness of geometric scale-space analysis by performing accurate 3D object recognition in highly cluttered range image scenes containing multiple objects occluding each other in varying degrees. The scale of local geometric structures provide an added layer of discriminability when used together with local shape descriptors. We believe the geometric scale-space analysis as well as the resulting features and descriptors provide a solid foundation for fully leveraging scale variability as another dimension of geometric data in various computer vision applications. Software that implements the 3D geometric scale-space analysis introduced in this paper, including feature and descriptor computation, and fully automatically range image registration, can be downloaded from https://www.cs.drexel.edu/~kon/3DGSS.

Acknowledgements This work was supported in part by National Science Foundation CAREER Award IIS-0746717 and IIS-0803670. The Armadillo, Happy Buddha, and Dragon models were provided courtesy of Stanford University Computer Graphics Laboratory. The Chinese Lion and Eros models were provided courtesy of INRIA by the AIM@SHAPE Shape Repository. The Sitting Buddha model was provided courtesy of VCG-ISTI by the AIM@SHAPE Shape Repository. The Chef, Chicken, Parasaurolophus, T-Rex and Rhino models and 50 scenes containing them were provided courtesy of the University of Western Australia. The Angel, BigBird, Gnome, Kid and Zoe models and 80 point cloud scenes containing them were provided courtesy of Queen's University, Canada.

References

Akagunduz, E., & Ulusoy, I. (2007). Extraction of 3D transform and scale invariant patches from range scans. In *IEEE int'l conf. on computer vision and pattern recognition* (pp. 1–8).

Bariya, P., & Nishino, K. (2010). Scale-hierarchical 3D object recognition in cluttered scenes. In *IEEE conf. on computer vision and pattern recognition*.

Brady, M., Ponce, J., Yuille, A., & Asada, H. (1985). *Describing surfaces* (Technical Report AIM-822). MIT Aritificial Intelligence Laboratory Memo.

Bro, R., Acar, E., & Kolda, T. G. (2008). Resolving the sign ambiguity in the singular value decomposition. *Journal of Chemometrics 22*, 135–140.

Carmo, M. P. D. (1976). *Differential geometry of curves and surfaces*. Prentice Hall: New York.

Chang, K. C., Ding, W., & Ye, R. (1992). Finite-time blow-up of the heat flow of harmonic maps from surfaces. *Journal of Differential Geometry*, *36*(2), 507–515.

Chen, C., Hung, Y., & Cheng, J. (1999). RANSAC-based DARCES: a new approach to fast automatic registration of partially overlapping range images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(11), 1229–1234.

Chua, C., & Jarvis, R. (1997). Point signatures: a new representation for 3D object recognition. *International Journal of Computer Vision*, *25*(1), 63–85.

Cohen, R., Hardt, R., Kinderlehrer, D., Lin, S., & Luskin, M. (1987). *Minimum energy configurations for liquid crystals: computational results* (pp. 99–121). New York: Springer.

Coron, J. M. (1990). Nonuniqueness for the heat flow of harmonic maps. *Annales de l'Institut Heri Poincaré*, *7*(4), 335–344.

Dinh, H. Q., & Kropac, S. (2006). Multi-resolution spin-images. In *IEEE int'l conf. on computer vision and pattern recognition*, Washington, DC, USA (pp. 863–870). Los Alamitos: IEEE Comput. Soc.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Flynn, P., & Jain, A. (1991). Bonsai: 3D object recognition using constrained search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(10), 1066–1075.

Freire, A. (1995). Uniqueness for the harmonic map flow in two dimensions. *Calculus of Variations and Partial Differential Equations*, *3*(1), 95–105.

Frome, A., Huber, D., Kolluri, R., Bulow, T., & Malik, J. (2004). Recognizing objects in range data using regional point descriptors. In *European conf. on computer vision*.

Gelfand, N., Mitra, N., Guibas, L., & Pottmann, H. (2005). Robust global registration. In *Symposium on geometry processing*.

Grimson, W. (1988). The combinatorics of object recognition in cluttered environments using constrained search. In *IEEE int'l conf on computer vision* (pp. 218–227).

Grimson, W., Lozano-Perez, T., & Huttenlocher, D. (1990). *Object recognition by computer: the role of geometric constraints*. Cambridge: MIT Press.

ter Haar, F., & Veltkamp, R. (2007). Automatic multiview quadruple alignment of unordered range scans. In *IEEE shape modeling and applications*, Washington, DC, USA (pp. 137–146). Los Alamitos: IEEE Comput. Soc.

Hardt, R. (1991). Singularities of harmonic maps. *Bulletin of the American Mathematical Society*, *34*, 15–34.

Huber, D., & Hebert, M. (2003). Fully automatic registration of multiple 3d data sets. *Image and Vision Computing*, *21*(7), 637–650.

Johnson, A. (1997). Spin-images: a representation for 3-D surface matching. Ph.D. thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

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

Johnson, A., Carmichael, O., Huber, D., & Hebert, M. (1998). Toward a general 3-D matching engine: multiple models, complex scenes, and efficient data filtering. In *Proceedings of the 1998 image understanding workshop (IUW)* (pp. 1097–1107).

Kazhdan, M., Bolitho, M., & Hoppe, H. (2006). Poisson surface reconstruction. In *Eurographics symp. on geometry processing*, Eurographics Association, Aire-la-Ville, Switzerland (pp. 61–70).

Kimmel, R. (1997). Intrinsic scale space for images on surfaces: the geodesic curvature flow. In *Scale-space theory in computer vision* (pp. 212–223).

Kimmel, R., & Sethian, JA (1998). Computing geodesic paths on manifolds. In *Proceedings of national academy of sciences USA* (pp. 8431–8435).

Koenderink, J. (1984). The structure of images. *Biological Cybernetics*, *50*, 363–370.

Lab QURCV (2009). Queen's range image and 3-D model database. http://rcvlab.ece.queensu.ca/~qridb/

Lalonde, J., Unnikrishnan, R., Vandapel, N., & Hebert, M. (2005). Scale selection for classification of point-sampled 3-D surfaces. In *Int'l conf. on 3-D digital imaging and modeling*.

Li, X., & Guskov, I. (2005). Multi-scale features for approximate alignment of point-based surfaces. In *Symposium on geometry processing*.

Li, X., & Guskov, I. (2007). 3D object recognition from range images using pyramid matching. In *IEEE int'l conf. on computer vision workshop on 3D representation for recognition* (pp. 1–6).

Lindeberg, T. (1994). *Scale-space theory in computer vision*. Dordrecht: Kluwer Academics.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, *30*, 77–116.

Makadia, A., Patterson, A., & Daniilidis, K. (2006). Fully automatic registration of 3D point clouds. In *IEEE int'l conf. on computer vision and pattern recognition*, Washington, DC, USA (pp. 1297–1304). Los Alamitos: IEEE Comput. Soc.

MeshLab (2010). http://meshlab.sourceforge.net/.

Mian, A., Bennamoun, M., & Owens, R. (2004). From unordered range images to 3D models: a fully automatic multiview correspondence algorithm. In *Theory and practice of computer graphics*, Washington, DC, USA (pp. 162–166). Los Alamitos: IEEE Comput. Soc.

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

Mokhtarian, F., Khalili, N., & Yuen, P. (2001). Curvature computation on free-form 3-D meshes at multiple scales. *Computer Vision and Image Understanding*, *83*, 118–139.

Morita, S., Kawashima, T., & Aoki, Y. (1992). Hierarchical shape recognition based on 3D multiresolution analysis. In *European conf. on computer vision*.

Nishino, K., & Ikeuchi, K. (2002). Robust simultaneous registration of multiple range images. In *Proc. of fifth Asian conference on computer vision, ACCV 02* (pp. 454–461).

Novatnack, J., & Nishino, K. (2007). Scale-dependent 3D geometric features. In *IEEE int'l conf. on computer vision*.

Novatnack, J., & Nishino, K. (2008). Scale-dependent/invariant local 3D shape descriptors for fully automatic registration of multiple sets of range images. In *European conference on computer vision* (pp. 440–453).

Pauly, M., Kobbelt, L. P., & Gross, M. (2006). Point-based multi-scale surface representation. *ACM Transactions on Graphics*, *25*(2), 177–193.

Ponce, J., & Brady, M. (1985). Toward a surface primal sketch. *The International Journal of Robotics Research*, *2*, 420–425.

Schlattmann, M. (2006). Intrinsic features on surfaces. In *Central European seminar on computer graphics* (pp. 169–176).

Skelly, L. J., & Sclaroff, S. (2007). Improved feature descriptors for 3-d surface matching. In: *Proc. SPIE conf. on two- and three-dimensional methods for inspection and metrology* (Vol. 6762, pp. 63–85).

Stein, F., & Medioni, G. (1992). Structural indexing: efficient 3-D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 125–145.

Struwe, M. (1985). On the evolution of harmonic mappings of riemannian surfaces. *Commentarii Mathematici Helvetici*, *60*(1), 558–581.

Sun, Y., & Abidi, M. (2001). Surface matching by 3D point's fingerprint. In *IEEE int.'l conf. on computer vision* (Vol. 2, pp. 263–269).

Taati, B. (2009). Generation and optimization of local shape descriptors for point matching in 3-D surfaces. Ph.D. thesis, Queen's University, Ontario, Canada.

Taati, B., Bondy, M., Jasiobedzki, P., & Greenspan, M. (2007). Variable dimensional local shape descriptors for object recognition in range data. In *IEEE int'l conf. on computer vision workshop on 3D representation for recognition*. Los Alamitos: IEEE Comput. Soc.

Tang, B., Sapiro, G., & Caselles, V. (2000). Diffusion of general data on non-flat manifolds via harmonic maps theory: the direction diffusion case. *International Journal of Computer Vision*, *36*(2), 149–161.

Taubin, G. (1995). A signal processing approach to fair surface design. In *ACM SIGGRAPH* (pp. 351–358).

Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(4), 376–380.

Unnikrishnan, R., & Hebert, M. (2008). Multi-scale interest regions from unorganized point clouds. In *IEEE int'l conf. on computer vision and pattern recognition: workshop on search in 3D (S3D)*.

Weickert, J., Ishikawa, S., & Imiya, A. (1999). Linear scale-space has first been proposed in Japan. *Journal of Mathematical Imaging and Vision*, *10*(3), 237–252.

Witkin, A. (1984). Scale-space filtering: a new approach to multi-scale description. In *IEEE int'l conf. on acoustics, speech, and signal processing* (pp. 150–153).

Zhang, D., & Hebert, M. (1999). Harmonic maps and their applications in surface matching. In *IEEE int'l conf. on computer vision and pattern recognition* (Vol. 2).

Zou, G., Hua, J., Lai, Z., Gu, X., & Dong, M. (2009). Intrinsic geometric scale space by shape diffusion. *IEEE Transactions on Visualization and Computer Graphics*, *15*, 1193–1200.