Scale-Hierarchical 3D Object Recognition in Cluttered Scenes

Prabin Bariya Ko Nishino
Department of Computer Science
Drexel University
{pb339,kon}@drexel.edu

Abstract

3D object recognition in scenes with occlusion and clutter is a difficult task. In this paper, we introduce a method that exploits the geometric scale-variability to aid in this task. Our key insight is to leverage the rich discriminative information provided by the scale variation of local geometric structures to constrain the massive search space of potential correspondences between model and scene points. In particular, we exploit the geometric scale variability in the form of the intrinsic geometric scale of each computed feature, the hierarchy induced within the set of these intrinsic geometric scales, and the discriminative power of the local scale-dependent/invariant 3D shape descriptors. The method exploits the added information in a hierarchical coarse-to-fine manner that lets it cull the space of all potential correspondences effectively. We experimentally evaluate the accuracy of our method on an extensive set of real scenes with varying amounts of partial occlusion and achieve recognition rates higher than the state-of-the-art. Furthermore, for the first time we systematically demonstrate the method's ability to accurately localize objects despite changes in their global scales.

1. Introduction

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. This is a challenging task especially since the scene may be cluttered and the objects in the scene may be occluding each other.

Traditional approaches to 3D object recognition generally comprise of two phases: feature extraction and matching. In the feature extraction phase, representative features are chosen or computed from the data. Local features are preferred in order to handle occlusion. In the matching phase, correspondences between the features from the models that are to be recognized and those from the scene are established. The characteristics of the features play a significant role in how the matching can be performed. The faithfulness of the computed features for representing the underlying 3D surface data and the discriminative power of the features are key components in the accuracy of any 3D object recognition system.

In the past, various primitives ranging from raw point data [5] to high-level geometric properties such as curvature and torsion [13] have been used for the purpose of 3D object recognition. However, the fact that geometric structures that characterize the surface geometry have natural support regions of varying sizes and carry significant discriminative information in themselves has been overlooked in the past. The scale variation of the geometric structures in the 3D data provide additional information which can be leveraged for 3D object recognition. Recently, Novatnack and Nishino [12] have analyzed the geometric scale-space of range images and demonstrated its usefulness in range image registration.

In this paper, we present an integrated framework that exploits the rich discriminative information provided by the scale-variability of local geometric structures to recognize and localize objects in cluttered 3D scenes. We build a model library of all objects that are to be recognized, and represent each object and scene by a set of scale-dependent corners and their scale-invariant local 3D shape descriptors. We perform recognition by using an interpretation tree based method with a single tree constructed for each model in the model library. The nodes in the tree represent correspondences between a model feature and scene feature, with each branch representing a hypothesis about the presence/absence and pose of that model in the scene.

Our key idea is to capitalize on the rich discriminative information offered by these scale-dependent features to aid in the matching phase. We show how the exponentially large space of correspondences [6] between model and scene features can be culled effectively with novel constraints based on the added geometric scale information. We use the intrinsic scale of each scale-dependent corner to restrict its possible correspondences to only those corners that are also detected at the same intrinsic scale. The

978-1-4244-6983-3/10/$26.00 ©2010 IEEE

robust nature and discriminative capability of the scale-dependent/invariant local 3D shape descriptor allow us to further limit the correspondences to only between corners with a high degree of similarity. Furthermore, we show how the inherent scale hierarchy of local geometric structures can be used to impose a hierarchical coarse-to-fine structure to the tree-based matching.

We demonstrate the effectiveness and accuracy of the proposed method by performing recognition experiments on 50 real scenes with varying levels of occlusion and clutter. We achieve a recognition rate of 97.5% with up to 84% occlusion which outperforms the state of the art reported on the same extensive data set [10]. Our overall recognition rate is 93.58%, for all levels of occlusion. Furthermore, we show that the proposed framework enables 3D object recognition in scenes where objects from the library are present but in different global scales. We perform recognition experiments in 50 real scenes plus 30 synthesized range images containing scaled versions of the models in our library in the presence of occlusion and clutter in addition to the real scenes, and achieve an overall recognition rate of 89.29%. This paper is the first to report a systematic study of 3D object recognition for scaled objects, which we believe is an important capability in practical scenarios.

2. Related Work

Past approaches have varied widely in the type of features and their representations used for 3D object recognition. Stein and Medeoni [13] use the distribution of normals, called 'splash', around a point of interest, usually in a high curvature area. Chua and Jarvis [2] use the point signature which encodes the minimum distances of points on a 3D contour to a reference plane. This approach is, however, sensitive to the sampling rate as well as noise. Dorai and Jain [3] use measures such as gaussian curvature, mean curvature, shape index and curvedness along with the spectral extension of the shape measure in their view dependent recognition system (COSMOS). Their approach, however, cannot be used for recognition of occluded objects. Johnson and Herbert [9] use point features and the spin image representation which encodes the 2D histograms of the 3D points around the feature. Spin images, however, suffer from low discriminating capability and sensitivity to resolution and sampling rate, which were later improved by Carmichael *et al*. [1]. Many other approaches also suffer from a number of limitations including robustness to occlusion and clutter, discriminative power of the feature used, sensitivity to noise and sampling, etc. Moreover, none of the past approaches have explicitly explored the use of geometric scale-variability of local surface structures present in the data for 3D object recognition.

As for the matching phase, tree-based methods have been used extensively in object recognition [4, 5, 7]. By representing correspondences between a pair of model and scene primitives as nodes in a tree, the space of all possible correspondences between model and scene primitives can be organized and searched in a structured manner. Greenspan [5] uses a test and verify approach with a binary decision tree classifier and feature extraction is avoided by using low-level point data. Grimson and Lozano-Perez [7] use an interpretation tree structure to represent all possible pairings of model and scene segments. They prune off most of these combinations through the use of distance and angular constraints. Grimson [6] shows that the expected complexity of recognizing objects in a cluttered scene is exponential in the size of the correct interpretation. Flynn and Jain [4] prune this space by using various unary and binary predicates for 3D recognition of objects with planar, cylindrical, and spherical surface types.

Mian *et al*. [10] use multidimensional table representations (tensors) for recognition in scenes in the presence of clutter and occlusion and achieve remarkable recognition rate which, to our knowledge, is the state-of-the-art demonstrated on an extensive data set. Later in this paper, we compare our results with Mian *et al*. [10] and also with the spin images approach [9]. There also has been some work done in recognition in scenes with scaled free-form library objects. Mokhtarian *et al*. [11] used a geometric-hashing based approach to recognize some partially occluded and scaled library objects. However extensive results and recognition rates for their approach are not available.

3. Scale-Dependent Model Library and Scenes

We first construct a model library of objects we wish to recognize and represent each object with a suitable set of features. For this, we exploit the scale-variability of local geometric structures in the 3D data and use features that accurately portray this scale-variability. We then compute a scale-dependent representation for each model that is to be recognized. Similarly, we represent scenes with their scale-dependent representation.

3.1. Geometric Scale Variability

The geometric scale-space analysis of range images were proposed by Novatnack and Nishino in [12], in which they compute corners on the 3D surface that capture the natural scales of the underlying geometric structure. These features along with their local 3D shape descriptor were then used to automatically align a mixed set of range images to reconstruct multiple objects at once.

The geometric scale-space of a range image can be constructed by filtering its surface normal field with Gaussian kernels of increasing standard deviation using the geodesic distance, which correspond to the set of discrete scales used for the scale-space analysis. 3D geometric corners are then

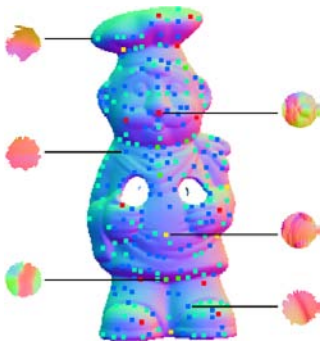


Figure 1. Scale-dependent corners and scale-invariant local 3D shape descriptors computed on range images synthesized to represent model objects, based on geometric scale-space analysis. Red, yellow, green, turquoise and blue colors indicate the corners detected from the coarsest to finest scales.

detected by using a corner detector at each discrete scale and by searching for spatial local maxima of the corner detector responses. The intrinsic scale of each 3D geometric corner is identified by searching for the local maxima of the corner detector responses across the set of discrete scales. 3D shape descriptors can then be computed at each detected corner by encoding the surface normals within a local surface region proportional to the scale of the corners using the exponential map.

We choose these scale-dependent corners and their scale-invariant local 3D shape descriptors to represent the models and scenes in our framework, as these have been shown to accurately represent the scale-variability of the local geometric structures in the 3D data. The scale-dependent corners detected at the finer scales represent subtle characteristics of the underlying geometry whereas those detected at increasingly coarser scales represent salient features of larger scales. Figure 1 shows the scale-dependent corners computed on range images of a model object in our library. Scale-invariant local 3D shape descriptors for corners computed at different scales are also shown.

For correspondences to be established between scale-dependent features from a model and a scene, we must be able to compute the distance between the respective scale-invariant local 3D shape descriptors. For this purpose, we use the similarity measure defined by Novatnack and Nishino [12]. We refer to the scale-invariant local 3D shape descriptor as $\hat{\mathbf{G}}^{\sigma}_{\mathbf{u}}$, for a scale-dependent corner computed at location $\mathbf{u}$ and with scale σ. The similarity measure is then defined as the angular normalized cross-correlation between the two sets of surface normals in their overlapping area,

$$\mathcal{S}(\hat{\mathbf{G}}^{\sigma_a}_{\mathbf{u}_k}, \hat{\mathbf{G}}^{\sigma_b}_{\mathbf{u}_l}) = \frac{\pi}{2} - \frac{1}{|A \cap B|} \sum_{\mathbf{v} \in A \cap B} \arccos(\hat{\mathbf{G}}^{\sigma_a}_{\mathbf{u}_k}(\mathbf{v}) \cdot \hat{\mathbf{G}}^{\sigma_b}_{\mathbf{u}_l}(\mathbf{v})), \tag{1}$$

where A and B are the set of points in the descriptors $\hat{\mathbf{G}}^{\sigma_a}_{\mathbf{u}_k}$ and $\hat{\mathbf{G}}^{\sigma_b}_{\mathbf{u}_l}$, respectively.

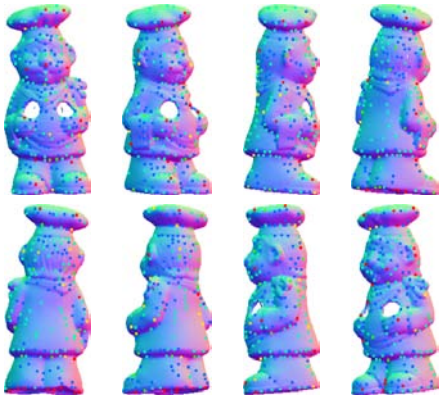


Figure 2. Synthesized range images of eight uniformly distributed views of the Chef model. The scale-dependent corners computed from these are consolidated into a single set, one for each model in the library.

3.2. Model Library

The model library comprises 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 represent each object with a set of range images. We synthesize range images from a number of uniformly distributed views of the 3D model of the object. As illustrated in Figure 2, 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 discrete scales to use in the geometric scale-space analysis, we compute the percentage of total scale-dependent corners that are detected at the coarsest scale from among the set of discrete scales. 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 compute a scale-invariant local 3D shape descriptor for each scale-dependent corner.

We then represent each object in the model library with a single set of scale-dependent corners that captures all views of the object. 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. Due to overlaps between any two views of the object, duplicate features may be present. To avoid such redundancy, 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. At the end, each object in the model library is represented with its 3D model and a

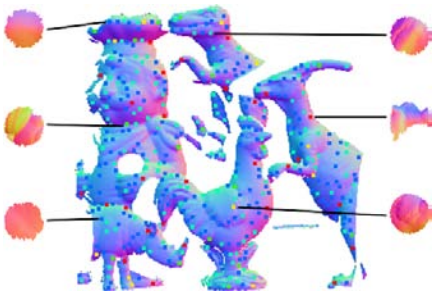


Figure 3. Scale-dependent corners and scale-invariant descriptors computed on a real range image, based on its scale-space analysis. The descriptors for a corner detected at a coarser level encodes a relatively larger neighborhood around the corner.

single consolidated set of scale-dependent corners and their corresponding scale-invariant local 3D shape descriptors.

3.3. Scenes

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. Figure 3 shows scale-dependent corners and some of their corresponding scale-invariant descriptors computed on a scene with clutter and occlusion.

4. Scale-Dependent Interpretation Tree

Given the scale-dependent representations of the models and scene, we perform matching using a tree structure that embodies all possible correspondences between model and scene features. We search for each object in the scene one at a time, with a constrained interpretation tree that exploits the rich discriminative information made available by the scale-dependent corners. Any successful search result can then be used to prune off scene features from areas of the scene that have been recognized and segmented, so that these are no longer used in any subsequent search for any other object.

4.1. Interpretation Tree

An interpretation tree approach [8] matches model primitives with scene primitives by representing a correspondence between them as a node in a tree structure. At the root of the tree, there are no correspondences. With each increasing level of the tree, a new model primitive is chosen and its correspondence with all available scene primitives form nodes at that level. 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. Descent in the tree implies an increasing level of commitment to a particular hypothesis [4].

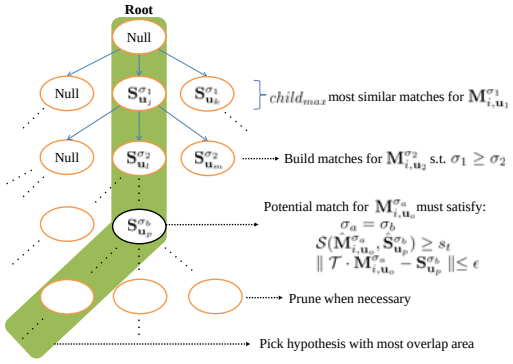


Figure 4. Schematic of our scale-hierarchical interpretation tree. For each level, a new model corner with the highest intrinsic scale is chosen and at most $child_{max}$ matches with the most similar scene corners that satisfy the scale, similarity and geometric constraints, are added for each branch in the previous level. The hypothesis with the most overlap area is chosen as most probable.

The search space of all correspondences represented by the entire interpretation tree may be exponentially large for complex scenes [6]. 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. Our key idea is to impose constraints on the nodes to be added to the tree by exploiting the rich discriminative information encoded in the scale-dependent corners.

4.2. Constrained Interpretation Tree Formation

For each model M_i to be searched for in a scene S, we create an interpretation tree IT_i. We build successive levels of the tree by picking a scale-dependent corner from the model and representing its correspondences with similar corners from the scene as nodes in the tree. The scale-dependent nature of the computed corners then allows us to impose constraints on which nodes can be added to the tree during the tree formation. We also make distinctions between the constraints placed on the tree for scale-dependent object recognition in scenes with objects from the model library of the same global scale versus scale-invariant object recognition in scenes which may contain globally scaled library objects.

In keeping with the notation for scale-invariant local 3D shape descriptor defined earlier in Equation 1, 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.

4.2.1 Scale Hierarchy

One of our insights is that 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 the finer scales represent small variations in the underlying geometry whereas those that are detected at increasingly coarser scales represent variations that are more prominent in size. 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 give priority to such corners by matching the scale-dependent corners detected at the coarsest scale first, followed by those detected at increasingly finer scales. As shown in Figure 4, any pair of model corners $\mathbf{M}_{i,\mathbf{u}_1}^{\sigma_1}$ and $\mathbf{M}_{i,\mathbf{u}_2}^{\sigma_2}$ 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.

4.2.2 Valid Correspondences

Our second key insight is to utilize the intrinsic scale of each scale-dependent corner to limit the space of correspondences. The intrinsic scale of a scale-dependent corner is given by the scale at which it was computed from the set of discrete scales used for scale-space analysis. Any two scale-dependent corners that represent the same underlying geometric structure must have the same intrinsic scale. Therefore, a correspondence between $\mathbf{M}_{i,\mathbf{u}_o}^{\sigma_a}$ and $\mathbf{S}_{\mathbf{u}_p}^{\sigma_b}$ may be valid only when they have the same intrinsic scale, $\sigma_a = \sigma_b$.

In contrast, for scenes containing scaled versions of objects in the library, the scale-dependent corners from the model and the scene that represent that same underlying geometric structure may not have the same intrinsic scale. Therefore to perform scale-invariant recognition in such scenes, we forgo the above constraint and allow for correspondences to be established across all intrinsic scales.

We also take advantage of the high disriminability of the scale-invariant local 3D shape descriptors and the fact that any two such descriptors that represent the same underlying surface geometric structures must be highly similar. A correspondence between $\mathbf{M}_{i,\mathbf{u}_o}^{\sigma_a}$ and $\mathbf{S}_{\mathbf{u}_p}^{\sigma_b}$ is considered valid only when their similarity measure between their corresponding scale-invariant local 3D shape descriptors $\hat{\mathbf{M}}_{i,\mathbf{u}_o}^{\sigma_a}$ and $\hat{\mathbf{S}}_{\mathbf{u}_p}^{\sigma_b}$, as defined by Equation 1 is above a similarity threshold s_t,

$$\mathcal{S}(\hat{\mathbf{M}}_{i,\mathbf{u}_o}^{\sigma_a}, \hat{\mathbf{S}}_{\mathbf{u}_p}^{\sigma_b}) \geq s_t. \tag{2}$$

$\mathcal{S}(\hat{\mathbf{M}}_{i,\mathbf{u}_o}^{\sigma_a}, \hat{\mathbf{S}}_{\mathbf{u}_p}^{\sigma_b})$ is essentially the average angular difference of the surface normals encoded in $\hat{\mathbf{M}}_{i,\mathbf{u}_o}^{\sigma_a}$ and $\hat{\mathbf{S}}_{\mathbf{u}_p}^{\sigma_b}$, so this thresholding translates into an angular cutoff of the average normal differences in the local neighborhood around the corners $\mathbf{M}_{i,\mathbf{u}_o}^{\sigma_a}$ and $\mathbf{S}_{\mathbf{u}_p}^{\sigma_b}$. In our experiments, we set s_t to 75% of the self-similarity measure.

To account for the possibility that a model corner $\mathbf{M}_{i,\mathbf{u}}^{\sigma}$ might not be present in a scene, we establish a correspondence between each $\mathbf{M}_{i,\mathbf{u}}^{\sigma}$ and a NULL entity as in [4, 6] and add this correspondence to 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. In our experiments, we set $child_{max}$ to five, including the NULL node.

4.2.3 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 $\mathbf{M}_{i,\mathbf{u}_o}^{\sigma_a}$ and scene corner $\mathbf{S}_{\mathbf{u}_p}^{\sigma_b}$ 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

$$\| \mathcal{T} \cdot \mathbf{M}_{i,\mathbf{u}_o}^{\sigma_a} - \mathbf{S}_{\mathbf{u}_p}^{\sigma_b} \| \leq \epsilon, \tag{3}$$

where ϵ is a threshold value. We set ϵ to three times the resolution of the synthesized range images used in building the model library.

For scenes that preserve the scale of the objects in the model library, the transformation $\mathcal{T}$ is a rigid transformation and entails the computation of a rotation matrix $\mathcal{R}$ and a translation t. However for scale-invariant object recognition in scenes that may contain scaled objects from the model library, a 3D similarity transformation $\hat{\mathcal{T}}$ must be estimated which entails the computation of a scale factor s in addition to $\mathcal{R}$ and t. These transformations can be computed using the method proposed by Umeyama in [14].

4.2.4 Pruning

As mentioned earlier, the space of correspondences represented by the tree is exponential and hence the tree must be pruned to keep the search space tractable. 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 sorted list is then kept with the rest pruned off. In our experiments, we set N_{max} and N_{pruned} to 2000000 and 2000 respectively.

4.3. Hypothesis Verification and Segmentation

Once the tree IT_i is fully formed, the many hypotheses in it must be verified to check for their accuracy. Verification of a hypothesis 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.

As it is infeasible to verify all the hypotheses given by the last level of the tree, we prune the last level of the tree so that only h_{max} of the strongest hypothesis remain. In our experiments, we set h_{max} to 20. We then verify each remaining hypothesis H_n by computing 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,

$$\alpha(H_{best}) = \frac{A(H_{best})}{M_a(H_{best})}, \tag{4}$$

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, otherwise we reject it. In our experiments, we set this threshold to 0.3, which essentially means that at least 30% 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.

5. Experimental Results

We built a model library comprising of five models of real objects used by Mian *et al*. in [10] and compute their scale-dependent representation using a set of five discrete scales. We perform recognition experiments on a number of real and synthetic 3D scenes containing multiple objects from our model library. To compare our recognition results as a function of occlusion and clutter, we define occlusion and clutter for each model object in a scene according to Mian *et al*. [10]:

$$\text{occlusion} = 1 - \frac{\text{model surface patch area in scene}}{\text{total model surface area}}, \tag{5}$$

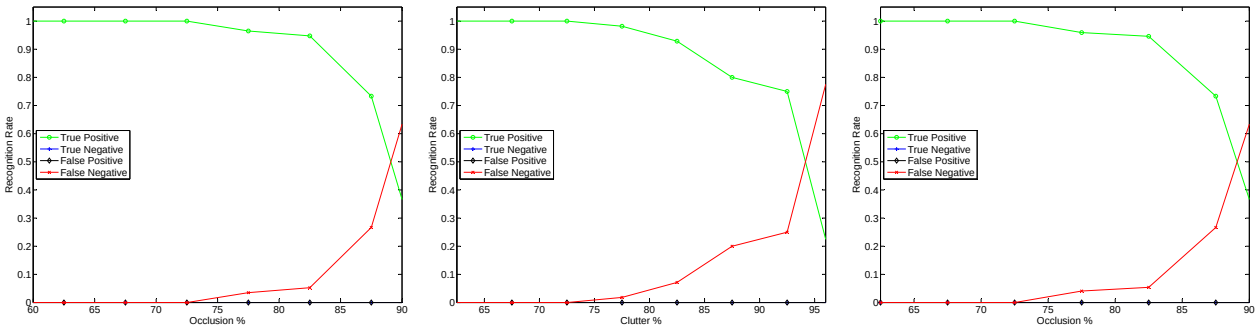
$$\text{clutter} = 1 - \frac{\text{model surface patch area in scene}}{\text{total surface area of scene}}. \tag{6}$$

5.1. Scale-Dependent Recognition

We perform scale-hierarchical 3D object recognition on the same set of 50 real scenes as used in [10], which contain multiple objects causing clutter and occlusion. 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. We manually segmented each of the scenes to compute the ground truth occlusion and clutter values for each object in each scene.

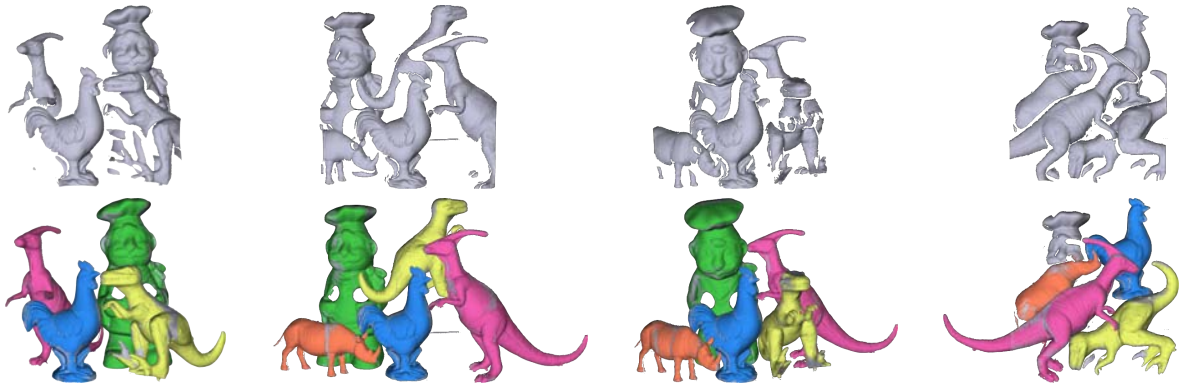
Figure 5 (a) and (b) show the recognition rate on the 50 real scenes, as a function of occlusion and clutter respectively. We were able to recognize objects with significant occlusion and clutter as shown in Figure 6. The average recognition rate of our approach was 93.58% which is comparable to the 95% recognition rate achieved by Mian *et al*. [10], but their dataset was augmented with a large number of synthetic scenes of simple clutterless views of single objects.

To achieve rigorous and fair evaluation in comparison to past methods, we compare our results with the recognition results on the exact same dataset presented by Mian *et al*. [10] for their tensor matching approach and the spin images recognition algorithm [9]. Similar to [10], we exclude the rhino from our recognition results as the spin images algorithm completely failed to recognize the rhino in any of the scenes as the rhino model 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 5(c) shows the recognition rate of our approach as a function of occlusion on the 50 real scenes, excluding results from the rhino model. We encourage the reader to compare Figure 5(c) to Figure 19(b) in



(a) (b) (c)

Figure 5. Recognition rates of our scale-dependent approach on 50 real scenes 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 comparision with [10], which we outperfom.



(a) (b) (c) (d)

Figure 6. Scale-dependent recognition results on four real scenes. 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.

[10] for a direct comparison with tensor matching and spin images recognition algorithms.

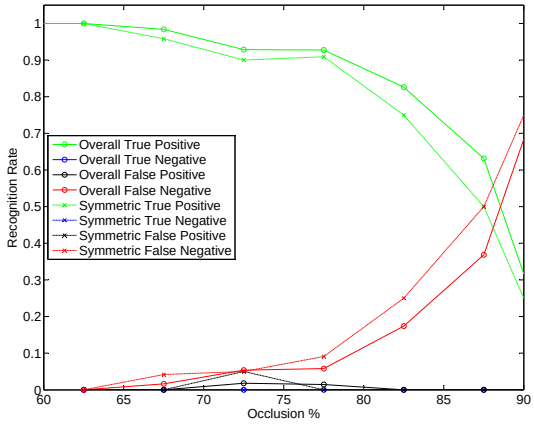
5.2. Scale-Invariant Recognition

We perform recognition experiments on all 50 real scenes used for the previous experiment as well as 30 synthesized 3D scenes in which objects from the model library were scaled from 60 to 150 percent of its size. We use the same set of parameter values as in the previous set of experiments. We allow for correspondences between corners to be established across all scales and use a similarity transformation $\hat{T}$ to allow for scale-invariant recognition.

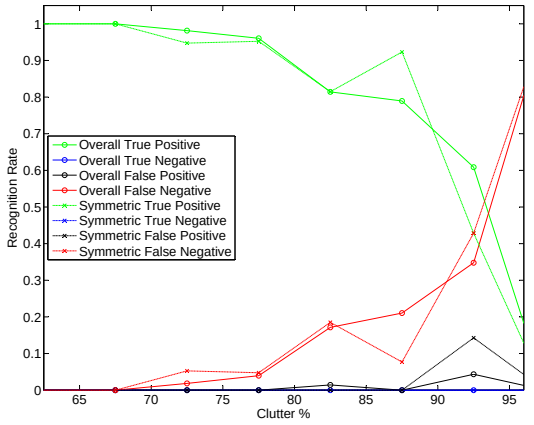
As illustrated in Figure 8, we are able to recognize scaled library objects in 3D scenes with significant occlusion and clutter. Figure 7 (a) and (b) shows the recognition rate of our scale-invariant approach as a function of occlusion and clutter respectively. We achieve a recognition rate of 89.08% 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 can be interpreted as the direct consequence of the increased search space of correspondences by allowing scaling as part of the transformation.

6. Conclusion

In this paper we presented an automatic 3D object recognition method that is able to accurately recognize highly occluded objects in scenes with significant clutter. Our key contribution is to exploit the scale-variability of local geometric structures in the data to effectively constrain the space of all possible correspondences between model and scene primitives. We performed experiments on 50 real scenes and achieved a recognition rate of 97.5% with upto 84% occlusion, which outperforms the state of the art. Furthermore, for the first time, we systematically demonstrate that our framework is capable of performing scale-invariant recognition tasks in complex scenes as well. Experiments on real and synthetic scenes with scaled library objects were performed and a recognition rate of 89.29% was achieved. We believe our scale-invariant recognition approach has broad practical implications as the model library may be

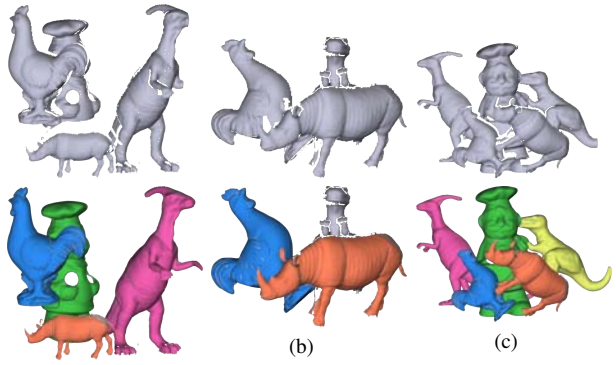


(a)



(b)

Figure 7. 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.



(a) (b) (c)

Figure 8. Three 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.

built with a suitably scaled object model and scaled objects can be accurately recognized in a scene. We hope that our work will invigorate more interest in scale-invariant 3D object recognition as we see it vital in real-world 3D recognition scenarios.

Acknowledgement

This work was supported in part by National Science Foundation CAREER Award IIS-0746717 and IIS-0803670.

References

[1] O. Carmichael, D. Huber, and M. Hebert. Large Data Sets and Confusing Scenes in 3-D Surface Matching and Recognition. In *Proc. of 3DIM*, pages 358–367, October 1999.

[2] C. Chua and R. Jarvis. Point Signatures: A New Representation for 3D Object Recognition. *Int'l Journal of Computer Vision*, 25(1):63–85, 1997.

[3] C. Dorai and A. Jain. COSMOS - A Representation Scheme for 3D Free-Form Objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(10):1115–1130, 1997.

[4] P. Flynn and A. Jain. Bonsai: 3D Object Recognition using Constrained Search. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(10):1066–1075, 1991.

[5] M. Greenspan. The Sample Tree: A Sequential Hypothesis Testing Approach to 3D Object Recognition. *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 772–779, 1998.

[6] W. Grimson. The Combinatorics of Object Recognition in Cluttered Environments using Constrained Search. *Proc. Int'l Conf. Computer Vision*, pages 218–227, 1988.

[7] W. Grimson and T. Lozano-Perez. Model-Based Recognition and Localization from Sparse Range or tactile data. *International Journal of Robotics Research*, 3(3):3–35, 1984.

[8] W. Grimson, T. Lozano-Perez, and D. Huttenlocher. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press, 1990.

[9] A. Johnson and M. Hebert. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[10] A. Mian, M. Bennamoun, and R. Owens. Three-Dimensional Model-Based Object Recognition and Segmentation in Cluttered Scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2006.

[11] F. Mokhtarian, N. Khalili, and P. Yuen. Multi-Scale Free-Form 3D Object Recognition using 3D Models. *Image and Vision Computing*, 19(5):271–281, April 2001.

[12] J. Novatnack and K. Nishino. Scale-Dependent/Invariant Local 3D Shape Descriptors for Fully Automatic Registration of Multiple Sets of Range Images. In *European Conf. on Computer Vision*. IEEE Computer Society, 2008.

[13] F. Stein and G. Medioni. Structural Indexing: Efficient 3-D Object Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(2):125–145, 1992.

[14] S. Umeyama. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.