Spatially Enhanced Bags of Words for 3D Shape Retrieval

Xiaolan Li[1,2], Afzal Godil[1], and Asim Wagan[1]

[1] National Institute of Standards and Technology, USA
[2] Zhejiang Gongshang University, P.R. China
{lixlan,godil,wagan}@nist.gov

Abstract. This paper presents a new method for 3D shape retrieval based on the bags-of-words model along with a weak spatial constraint. First, a two-pass sampling procedure is performed to extract the local shape descriptors, based on spin images, which are used to construct a shape dictionary. Second, the model is partitioned into different regions based on the positions of the words. Then each region is denoted as a histogram of words (also known as *bag-of-words*) as found in it along with its position. After that, the 3D model is represented as the collection of histograms, denoted as bags-of-words, along with their relative positions, which is an extension of an orderless bag-of-words 3D shape representation. We call it as *Spatial Enhanced Bags-of-Words* (SEBW). The spatial constraint shows improved performance on 3D shape retrieval tasks.

1 Introduction

With recent advances in computer aided modeling and 3D scanning technology, the number of 3D models created and stored in 3D shape repositories is increasing very rapidly. 3D models are widely used in several fields, such as computer graphics, computer vision, computer aided manufacturing, molecular biology, and culture heritage, etc. Therefore it is crucial to design effective and efficient methods for retrieving shapes from these 3D shape repositories.

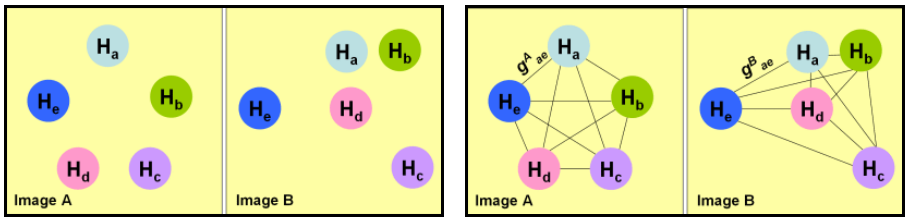
Appearance and geometry are two main aspects of a 3D model. However, most existing search engines focus either on appearance [Chen03] [Osada02] or on geometry [Siddiqi08] alone. Even though for these methods [Papadakis07] [Vranic03], which explore both features, they ask for specifically geometric ordering for the appearance features. They use spherical harmonic transform to combine these two properties together, which requires the order of the sampling positions to be sorted according to two spherical angle coordinates. This requirement reduces the flexibility of the method.

The bag-of-words methods, which represent an image or a 3D shape as an orderless collection of local features, have recently demonstrated impressive level of performance [Li05]. However, it totally disregards the information about the spatial layouts, which limits its descriptive ability. In this paper, we would like to incorporate a weak geometry constraint into the bag-of-words framework, which will compensate for the shortcoming of the framework and keep its flexibility as much as possible.

G. Bebis et al. (Eds.): ISVC 2008, Part I, LNCS 5358, pp. 349–358, 2008.

Our approach is most similar to that of [Liu06] [Shan06]. The similarities include: 1) the spin image is chosen as the local descriptor. 2) the bag-of-words model supports the whole framework of 3D shape retrieval. The main differences between our work and their work are two-fold. First, we incorporate a weak spatial constraint which improves the descriptive capability compared to the original bag-of-words model. Second, we introduce a new similarity metric which accounts for appearance similarity and geometry similarity.

We start by uniformly sampling basis points and support points on the surface of the model, which satisfies insensitivity to the tessellation and resolution of the mesh. After extracting a set of spin images for each model, we construct a large shape dictionary by clustering all spin images acquired from the whole training dataset. Instead of representing one model with a histogram of the words from the dictionary, it is partitioned into several regions by clustering the basis points according to their spatial positions, and then represented as a set of histograms (bags-of-words) with pairwise spatial relations of the regions. After a correspondence based on the appearance is built between two models, the spatial difference, referred to as geometry dissimilarity, between the layouts of the regions is calculated and used as the second part of the dissimilarity metric. That is, as a 2D example shown in figure 1, even though two images containing the same components, the *spatially enhanced bags-of-words* (SEBW) method will differentiate one from the other (1-b). However, the original bags-of-words method will regard them as the same (1-a).



(a) Representing images with *bags-of-words* model. The left and the right image are both partitioned into 5 parts, and each part is represented with a histogram of the words (*bag-of-words*) with no spatial information of the parts taken into account. Under this kind of representation, the left and the right images are regarded as the same

(b) Representing images with *spatially enhanced bags-of-words model*. Besides the bags-of-words representation, the representation of the image also includes the geometric links of pairwise parts, recorded as weighted edges. Under this kind of representation, the left and the right images are different

Fig. 1. Comparing bags-of-words model and the spatially enhanced one

The organization of the paper is as follows. Several related works are summarized in Section 2. The bag-of-words model and the spatially augmented bags-of-words model are presented and defined in Section 3. Then, the procedures of feature extraction and similarity computation are described in Section 4 and 5 respectively. In Section 6, we provide the 3D shape retrieval results on Princeton Shape Benchmark

(PSB) [Shilane04] and analyze the influence of the parameters. Finally, we conclude this paper in Section 7.

2 Previous Work

Designing discriminating 3D shape signatures is an active research area [Iyer05]. Among them, statistics based methods hold a very important position, which can be adopted as part of the machine learning framework.

Statistical properties, frequently represented as histograms [Osada02], have been used to describe the appearance of object/scene for a long time. In [Osada02], several appearance properties are recorded with histograms, including the distances between two randomly selected surface points, areas of the triangles composed of three randomly selected surface points and so on. Instead of calculating the Euclidean distance between every two surface points, Ruggeri [Ruggeri08] compute the histograms based on the geodesic distances, which are effective to retrieve articulated models. Except the histogram, probability density functions [Akgul07] are also used to express the 3D model, which is more robust to the asymmetrical distribution of the triangles.

Because of its simplicity and generality, the bag-of-words method, which is insensitive to deformation, articulation and partially missing data, has attracted lots of interest in 2D [Li05] and 3D [Shan06] [Liu06] [Ohbuchi08] fields. In [Li05], the method is applied to images by using a visual analogue of a word, formed by vector quantizing two regional descriptors: normalized 11*11 pixel gray values and SIFT descriptors. In [Shan06] and [Liu06], a visual feature dictionary is constituted by clustering spin images in small regions. In order to procure partial-to-whole retrieval, Kullback-Leibler divergence is proposed as a similarity measurement in [Liu06], while a probabilistic framework is introduced in [Shan06]. For the sake of collecting visual words, Ohbuchi et al. [Ohbuchi08] apply the SIFT algorithm to depth buffer images of the model captured from uniformly sampled locations on a view sphere. After vector quantization, Kullbak-Leibler divergence measures the similarities of the models.

Besides the many advantages of these methods, they suffer from their extreme simplicity. Because all of the spatial layout information of the features is discarded, their descriptive capability is severely constrained. Lazebnik et al. [Lazebnik06] propose a spatially enriched bags-of-words approach. It works by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. Implicitly geometric correspondences of the sub-regions are built in the pyramid matching scheme [Grauman05]. In [Savarese07], the object is an ensemble of canonical parts linked together by an explicit homographic relationship. Through an optimization procedure, the model, corresponding to the lowest residual error, gives the class label to the query object along with the localization and pose estimation.

Inspired by the work described by [Lazebnik06], [Savarese07], and [Shan06], we propose a *Spatially Enhanced Bags-of-Words* (SEBW) approach. We will elaborate it in the following sections.

3 Spatially Enhanced Bag of Words Model

We first describe the original formulation of bag-of-words [Li05], and then introduce our approach to create a *Spatially Enhanced Bags-of-Words* (SEBW) 3D model representation.

3.1 Bag of Words Model

Let us use the image categorization as an example to give an explanation of the *bag-of-words* model. Denote N be the total number of labels ("*visual words*") in the learned visual dictionary. The image can be represented as a vector with length N, in which the elements count the occurrences of the corresponding label. The procedure can be completed in three steps.

1. Local feature detectors are applied to the images to acquire low level features.
2. *Visual words*, denoted as the discrete set $\{V_1, V_2, \ldots, V_N\}$, are formed by clustering the features into N clusters, so that each local feature is assigned to a discrete label.

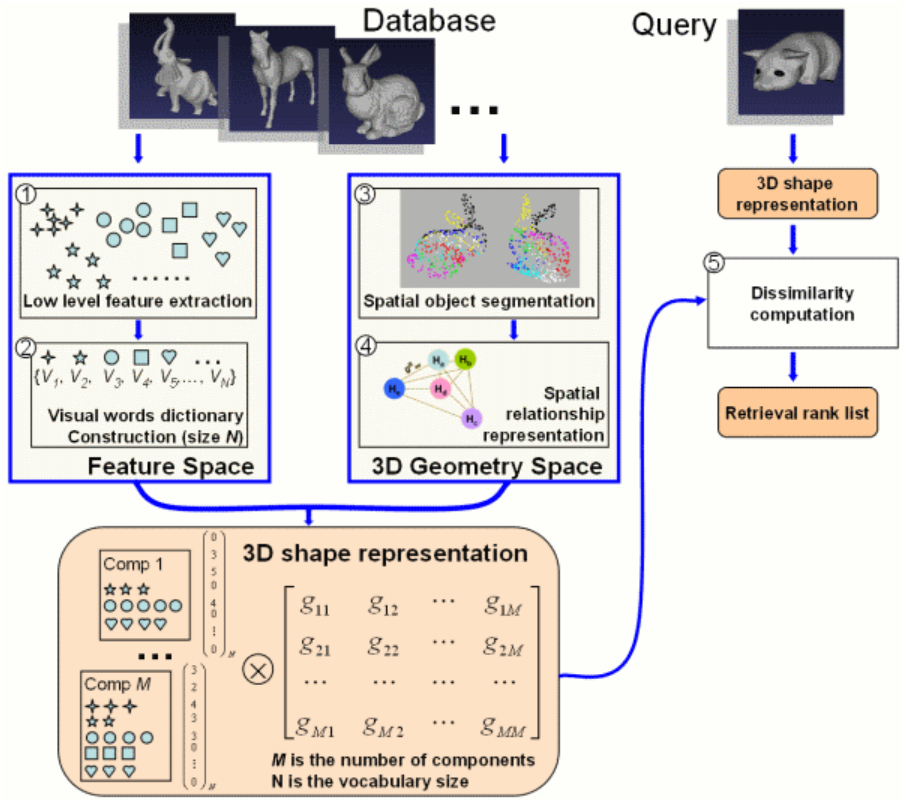


Fig. 2. *Spatially Enhanced Bags-of-Words* (SEBW) method for 3D shape retrieval (best viewed with color)

3. The image contents are summarized with a global histogram ("*bag-of-words*"), denoted as a vector $fv=(x_1, x_2, \ldots, x_N)$, by counting the occurrences of each *visual word.*

3.2 Spatially Enhanced Bags of Words Representation

Rather than using only a global histogram, this paper advocates using more than one histogram along with related spatial information to reveal the 3D shape in more detail. A schematic description of the approach is given in Figure 2. Specifically, after extracting low level features with spin images at oriented basis points, a visual dictionary is formed by clustering them in feature space. Then each 3D shape is partitioned into a predetermined number of regions by clustering the oriented basis points in 3D geometry space. A matrix is used to record the spatial relationship of the regions, while each region is represented with a histogram of visual words. Therefore, the 3D shape is recorded with the *Spatially Enhanced Bags-of-Words* (SEBW) model.

4 Shape Representation

This section elaborates on the ideas introduced in the previous section. Block 1, 2, 3 and 4 are discussed in this section and block 5 is covered in the next section.

4.1 Low Level Feature Extraction

As shown in Figure 3, the Spin image, which is invariant to the rotation and translation transform, characterizes the local appearance properties around its basis point p within the support range r. It is a two-dimensional histogram accumulating the number of points located at the coordinate (α, β), where α and β are the lengths of the two orthogonal edges of the triangle formed by the oriented basis point p, whose orientation is defined by the normal n, and support point q. We choose it as the low level feature descriptor in this paper.

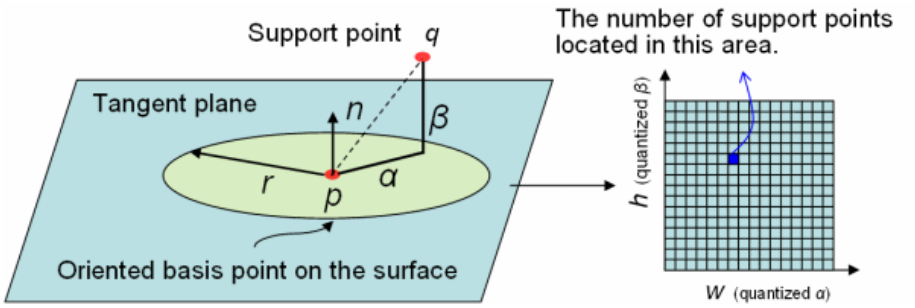


Fig. 3. Extracting low level features with spin images

Because the 3D meshes may have large and small triangles, instead of calculating spin images based on the mesh vertices [Jonson99], a two-pass sampling procedure is performed here. Using a Monte-Carlo strategy [Osada02], for each 3D mesh, N_b oriented basis points p with normal n and N_s support points q are sampled uniformly on

the surface in two passes respectively, where N_b=500, N_s=50000 [Liu06]. Other parameters of the spin image are defined as: 1) r =0.4R, where R is the radius of the mesh. 2) the width and height of spin images, set as w=h=16.

A large number of spin images are collected from the 3D shape database. Each mesh is represented with N_b spin images.

4.2 Visual Words Dictionary Construction

With N_b*N_m spin images, where N_b is as defined previously and N_m is the number of 3D meshes we used for building the visual words dictionary, the k-means algorithm is applied to agglomerate N clusters. This is similar to the procedure we described in section 3.1. Therefore, each spin image is assigned the index of its nearest cluster. Actually, other clustering algorithms [Moosmann08] can be adopted to do the work. Further research needs to be done to analyze the effects of the various clustering methods.

4.3 Spatial Object Segmentation

Although many sophisticated segmentation approaches [Podolak06] [Berretti06] can be exploited to do the work, for the simplicity, the k-means algorithm is performed here to segment the 3D meshes. Evaluating the effects of using different segmentation schemes in our 3D retrieval framework will be a subject for future research.

After step 1, the shape of the 3D mesh is sketched by a set of spin images located at the positions of basis points in 3D geometry space. For each 3D mesh, the spatial object segmentation is achieved by clustering the basis points into M clusters based on their spatial positions, where M is a predetermined number.

4.4 Spatially Enhanced Object Representation

Once the object has been partitioned into M components, the associated spin images of the shape are also split into M groups. Referred to the visual words dictionary, whose vocabulary size is N, each spin image corresponds to a visual word. Therefore, each component is represented by a vector fv=(x_1, x_2, …, x_N) counting visual word frequencies. A M*N Feature Matrix FM depicts the appearance of the object, where

$$FM = [\, fv_1, fv_2, \ldots, fv_M \,]^T, \tag{1}$$

We outline the geometric properties of the model with a matrix GM:

$$GM = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1M} \\ g_{21} & g_{22} & \cdots & g_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ g_{M1} & g_{M2} & \cdots & g_{MM} \end{bmatrix}_{M*M} \tag{2}$$

where g_{ij} is the Euclidean distance between pair of the centers of the components, i, j = 1, 2, …, M.

Therefore, each object is recorded with M visual words histograms associated with one Geometry Matrix, as shown in Figure 2.

5 Dissimilarity Measure

When performing 3D shape retrieval, the spatially enhanced representation of the query shape is constructed on line, and compared with those in the database. A retrieval list is ordered according to the dissimilarity metric. In our paper, it is made up of two parts, which is formulated as:

$$Dis(O^A, O^B) = \alpha \cdot Dis_a(FM^A, FM^B) + (1-\alpha) \cdot Dis_g(GM^A, GM^B), \tag{3}$$

$$Dis_a(FM^A, FM^B) = \sum_{i=1}^{M} \min_{\pi(i)}(dist(fv_i^A, fv_{\pi(i)}^B)), \tag{4}$$

$$Dis_g(GM^A, GM^B) = \sum_{i=1}^{M}\sum_{j=1}^{M} \left| g_{i,j}^A - g_{\pi(i),\pi(j)}^B \right|, \tag{5}$$

where O^A, O^B are two objects A and B respectively; α is the weight to balance the effects of appearance features and geometry features, satisfying $0 \leq \alpha \leq 1$; $dist(fv_i^A, fv_{\pi(i)}^B)$ is the distance between two feature vectors. The distance can be measured with Kullback-Leibler divergence [Liu06], cosine distance, L1 and L2 distance, etc.

The pseudo-code for measuring the dissimilarity between a 3D query object A and another 3D shape B from the database is listed as follows.

```
for each component i of A
  get corresponding component's index (i) for ob-
  ject B resulting mini feature distance dist_f^i;
end
summarize dist_f^i of all the components i(eq. (4));
calculate the geometry distance (eq. (5));
measure the dissimilarity with eq. (3);
```

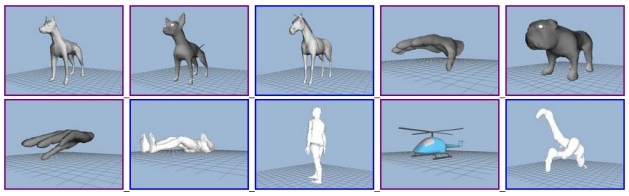
Then corresponding to the query object, every object in the database is assigned a metric value. Accordingly, a retrieval rank list is obtained based on it.

6 Experiments

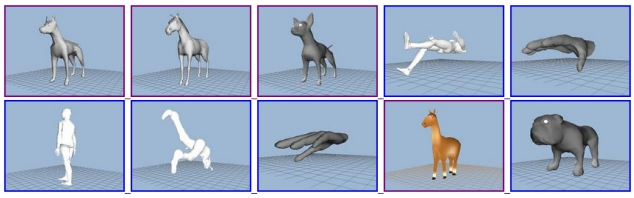
Princeton Shape Benchmark (PSB) [Shilane04] is chosen as the 3D shape database. It is divided into two sets: training set and testing set. We conduct the experiments on the test set, which contains 907 3D models belonging to 92 classes at the finest classification granularity.

We compare our approach (SEBW) with the bag-of-words method (BW). For SEBW, the number of components is set to be 10, and dissimilarity weight α = 0.7. Figure 4 shows the retrieval lists using m92 in the PSB. Comparing a with b, it is obvious that with our SEBW method, the wrong shape, helicopter, is eliminated from the retrieval list.

Table 1 lists the five statistics: Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), E-Measure (E-M), and Discounted Cumulative Gain (DCG) as described in



(a) Retrieval results using BW method



(b) Retrieval results using SEBW method

Fig. 4. Compare BW and SEBW with the first 9 retrieval results using the same query shape placed at the top left position

[Shilane04]. It shows that the Nearest Neighbor improved with the spatial information. The Precision-Recall curves in Figure 5 also affirm that. However the improvement is not large. The reason behind could be: 1) The number of the basis points is not sufficient to adequately cover the shape; 2) The segmentation method, k-means, is sensitive to initialization, and not particularly robust across variations in the shape. A more sophisticated segmentation method might improve the result.

Table 1. The retrieval statistics about two retrieval methods

	NN	FT	ST	E-M	DCG
BW	0.335	0.173	0.247	0.155	0.446
SEBW	0.338	0.160	0.226	0.141	0.433

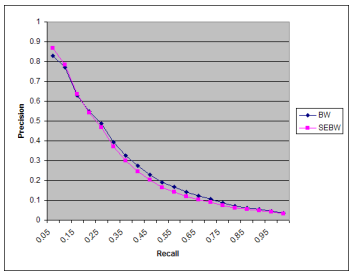
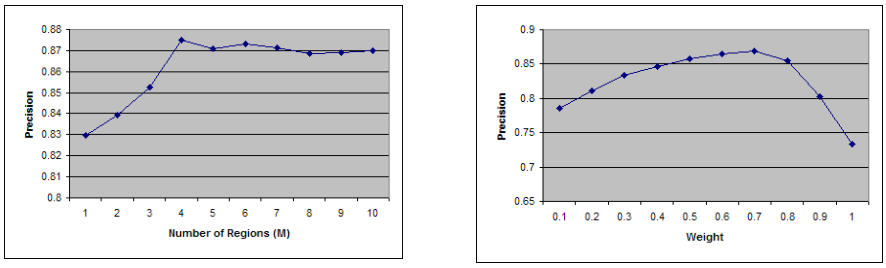


Fig. 5. The precision-recall curve of two methods: Bag-of-Words (BW), Spatially Enhanced Bags-of-Words (SEBW)



(a) Effects of M (b) Effects of α

Fig. 6. How does the parameter affect the retrieval precision

Two important parameters, which will affect the retrieval precision, are investigated. The first parameter is the number of components M. To balance between efficiency and effectiveness, M is limited to less than 11. Actually, beyond a certain threshold, even if the number of the components is increased, the precision is relatively stable. As shown in Figure 6-a, the threshold is 4. Note when M=1, SEBW is the same as BW. Therefore, BW can be regarded as a special case of SEBW.

Fixing the number of regions (M) as 10, the second parameter is discussed. It is the dissimilarity weight α defined in eq. (3). From the equation, we see that when α is too large, the appearance features dominate the dissimilarity measure; while when α is too small, the geometry features dominate. Both will decrease the retrieval precision. The experiment results verify the conclusion. As shown in Figure 6-b, when α is set to be larger than 0.9 or less than 0.2, the retrieval precision decreases rapidly. The highest retrieval precision is obtained with α set at 0.7.

7 Discussion

In this paper, we propose a means by which to incorporate spatial information into the bag-of-words model for 3D shape retrieval. The enhanced method is compared with the original bag-of-words method. Two key parameters for the approach are discussed in detail. However, this is only the initial investigation into combining the spatial information with BW method. Besides using the simple distance matrix to infer the geometric properties, other sophisticated geometric properties, even the topologic structure can be explored in the future.

References

[Akgul07] Akgul, C.B., Sankur, B., Yemez, Y., Schmitt, F.: Density-Based 3D shape descriptors. EURASIP Journal on Advances in Signal Processing ID 32503, 16 pages (2007)

[Berretti06] Berretti, S., Bimbo, A.D., Pala, P.: Partitioning of 3D meshes using reeb graphs. In: ICPR 2006 (2006)

[Chen03] Chen, D.Y., Ouhyoung, M., Tian, X.P., Shen, Y.T.: On visual similarity based 3D model retrieval. In: Computer Graphics Forum (Eurographics 2003), vol. 03, pp. 223–232 (2003)

[Grauman05] Grauman, K., Darrell, T.: Pyramid match kernels: Discriminative classification with sets of image features. In: ICCV 2005 (2005)

[Iyer05] Iyer, N., Jayanti, S., Lou, K., Kalyanaraman, Y., Ramani, K.: Three-dimensional shape searching: state-of-the-art review and future trends. Computer-Aided Design 37(2005), 509–530 (2005)

[Johnson99] Johnson, A.E., Hebert, M.: Using Spin Image for Efficient Object Recognition in Cluttered 3D Scenes. PAMI 21(5), 433–449 (1999)

[Lazebnik06] Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: CVPR 2006, pp. 2169–2178 (2006)

[Li05] Fei-Fei, L., Perona, P.: A Bayesian Hierarchical model for learning natural scene categories. In: CVPR 2005, pp. 524–531 (2005)

[Liu06] Liu, Y., Zha, H., Qin, H.: Shape Topics: A Compact Representation and New Algorithms for 3D Partial Shape Retrieval. In: CVPR 2006, pp. 2025–2032 (2006)

[Moosmann08] Moosmann, F., Nowak, E., Jurie, F.: Randomized clustering forests for image classification. PAMI 30(9), 1632–1646 (2008)

[Ohbuchi08] Ohbuchi, R., Osada, K., Furuya, T., Banno, T.: Salient local visual features for shape-based 3D model retrieval. In: Proc. IEEE Shape Modeling International, pp. 93–102 (2008)

[Osada02] Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Transaction on Graphics 21(4), 807–832 (2002)

[Papadakis07] Papadakis, P., Pratikakis, I., Perantonis, S., Theoharis, T.: Efficient 3D shape matching and retrieval using a concrete radicalized spherical projection representation. Pattern Recognition 40(2007), 2437–2452 (2007)

[Ruggeri08] Ruggeri, M.R., Saupe, D.: Isometry-invariant matching of point set surfaces. In: Eurographics workshop on 3D object retrieval, 8 pages (2008)

[Savarese07] Savarese, S., Fei-Fei, L.: 3D Generic Object Categorization, Localization and Pose Estimation. In: ICCV 2007, pp. 1–8 (2007)

[Shan06] Shan, Y., Sawhney, H.S., Matei, B., Kumar, R.: Shapeme Histrogram Projection and Matching for Partial Object Recognition. PAMI 28(4), 568–577 (2006)

[Shilane04] Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The Princeton shape benchmark. In: Proc. of the Shape Modeling International 2004 (SMI 2004), vol. 04(00), pp. 167–178 (2004)

[Siddiqi08] Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bioux, S., Dickinson, S.: Retrieving articulated 3D models using medial surfaces. In: Machine Vision and Applications (MVA), vol. 19(4), pp. 261–275 (July 2008)

[Sivic05] Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collections. In: ICCV 2005, pp. 370–377 (2005)

[Vranic03] Vranic, D.V.: 3D model retrieval. Ph. D. Dissertation (2003)