EXPLORING THE BAG-OF-WORDS METHOD FOR 3D SHAPE RETRIEVAL

Xiaolan Li[1] *Afzal Godil*[2]
[1] Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, China
[2] National Institute of Standards and Technology, Gaithersburg, U.S.A.
xiaolanli@mail.zjgsu.edu.cn, godil@nist.gov

ABSTRACT

This paper investigates the capabilities of the Bag-of-Words (BW) method in the 3D shape retrieval field. The contributions of this paper are: 1) the 3D shape retrieval task is categorized from different points of view: specific vs. generic, partial-to-global (PG) vs. global-to-global (GG) retrieval, and articulated vs. non-articulated; 2) The spatial information, which is represented as concentric spheres, is integrated into the framework to improve the discriminative capability; 3) the analysis of the experimental results on Purdue Engineering Benchmark (PEB) [4] reveals that some properties of the BW approach make it perform better on the PG task than the GG task. 4) The BW approach is evaluated on non-articulated database PEB [4] and articulated database McGill Shape Benchmark (MSB) [17] and compared to other methods.

Index Terms --- Bag-of-words, spin images, 3D shape retrieval

1. INTRODUCTION

Because of its simplicity, flexibility, and effectiveness, the Bag-of-Words (BW) method, which originated from the document retrieval field, has recently attracted large amount of interest in the computer vision fields. It has been applied in the applications such as image/video classification [7], 3D shape analysis and retrieval [8, 9, 13, 15] and so on. We will explore its performance especially for the 3D shape retrieval task in this paper.

A typical 3D shape retrieval task can be defined as: giving a 3D shape query, to obtain a list of 3D shapes ordered by the similarity between the query object and the one on the list. Several methods are proposed to solve the problem, such as Light Field descriptors [2], spherical harmonics descriptor [6], D2 shape distribution [12], and local feature based methods [9, 13] among others. The performance of the methods varies mainly according to the specific tasks. In fact, from different point of views, the 3D shape retrieval task can be further refined as follow:

1) Differentiated from the object category extent, the task can be discussed in terms of "specific and generic" domains. The representative benchmark of the "generic" one is Princeton Shape Benchmark [16], while CAD [4] and Biometrics [1] analysis are important "specific" domains, which have their own properties. For example, CAD models have more complicated structures with holes and other local features. Although only using the global information can the subtle details be neglected and leads to less than ideal retrieval results.

2) Based on the completeness of the query shape, the task can be divided into two subtasks as "Partial-to-Global Retrieval (PGR)" and "Global-to-Global Retrieval (GGR)". For the former one, every query shape is regarded as an incomplete object, which is used to obtain similar complete objects from the database. This happens in many cases. For example, when using the 3D range scanners to capture 3D data in real time, because of the limitation of the scanner, only parts of the object can be captured during scanning. This incomplete point clouds may be used as the query shape to retrieve the corresponding complete model from an existing database. Most of the global based shape retrieval methods, which require the complete geometry of a 3D object, cannot be applied directly to PGR. To our knowledge there are only a few contributions [8, 10] that solve the PGR problem.

3) Based on the deformability of the shape, there exist "Articulated Shape Retrieval (ASR)" and "Non-Articulated Shape Retrieval (NASR)". Lots of the natural and manmade objects are deformable. For instance, in CAESAR [1], each person is scanned in three postures: standing, sitting with arms open and with arms down. When performing shape retrieval using a standing model of person A as the query model, the preferred result is to obtain the other two different poses of model A, than to retrieve the sitting models of other people. According to the research results in [13], Light Field method [2], which performs great when dealing with NASR problem, produces poor results for ASR task.

To some extent, in [8, 10, 13], the above three different tasks are discussed within the BW framework, but there still lacks through investigation. In this paper, we investigate deeply into all these cases within the framework of BW method with spin images [3] as the low level features, and provide profound experimental results to support the discussion.

2. RELATED WORK

BW method first shows its effectiveness for the PGR task [8, 15]. A visual feature dictionary is constituted by clustering spin images [3]. Then KL divergence is proposed as a similarity measurement in [8], while a probabilistic framework is introduced in [15].

For the ASR task, Ohbuchi et. al. [13] apply the SIFT algorithm to collect visual words. After vector quantization, KL divergence measures the similarities of the models. It also demonstrates that a) given enough samples, the BW method can reach a comparable retrieval result as vision based method like Light Field [2], when dealing with NASR task; b) the BW method performs better than Light Field when dealing with ASR task. In this paper, spin images are used as local features, which can be directly extracted as many as you want in a 3D domain. On the other hand, according to [7], dense features, such as spin images, perform better than sparse features, such as SIFT.

Although the BW method has many advantages, it suffers from its lack of spatial information. Some methods focus on integrating the spatial layout information into the BW method. Lazebnik et. al. [11] first partition the images 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. In [14], 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. Li [9] proposes to treat the model in two different domains, named the feature domain and the spatial domain. The visual word dictionary is built in the feature domain as in the ordinary BW method. On the other side, the whole model is partitioned into several pieces in the spatial domain. Thereafter, each piece of the model is represented as a word histogram. The whole model is recorded as several word histograms along with a geometry matrix which stores the relative distances between every pairs of the pieces. The weighted sums of dissimilarity measurements from these two domains are used to measure the differences between models.

3. CONCENTRIC BAG-OF-WORDS METHOD

We first describe the original formulation of BW representation [7, 8], and then introduce the whole procedure of Concentric Bags-of-Words (CBW) method. Their differences are demonstrated in Figure 1. The left two circles are two shapes represented with one global *BW* model. The first and the second shapes are both composed with 5 different words: *a*, *b*, *c*, *d*, *e*, whose vectors are both (3, 5, 7, 4, 5). That means using BW representations; the first and the second shapes are regarded as the same. The right two circles are the same two shapes, which are represented with *Concentric Bags-of-Words* model. Then along the arrow's direction, counting from the outer sphere to the inner one, their feature vectors are (2 3 1 3 3; 1 1 5 1 1; 0 1 1 0 1) and (0 3 3 2 3; 3 1 2 2 2; 0 1 2 0 0) respectively. Obviously, they are regarded as different shapes under the representation of CBW method.

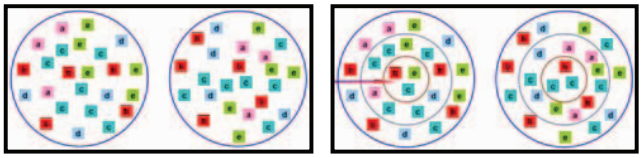

Fig. 1 Comparing BW and CBW representation.

3.1. Bag-of-Words descriptor

Let us use the ordinary 3D shape retrieval as an example to give an explanation of the BW framework. Denote N to be the total number of labels ("*visual words*") in the learned visual dictionary. The 3D shape 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 descriptors, such as spin image [3], are applied to the 3D model 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.

3) The shape of the 3D model is 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. Concentric Bag-of-Words method

Rather than using one global histogram, this paper advocates using more than one histogram along with its related spatial information to reveal the 3D shape in more detail. Specifically, the model is partitioned with several concentric spheres, and all the parts between two neighboring spheres are recorded with original BW descriptor, which leads to the name Concentric Bag-of-Words. A schematic description of the approach is given in Figure 2.

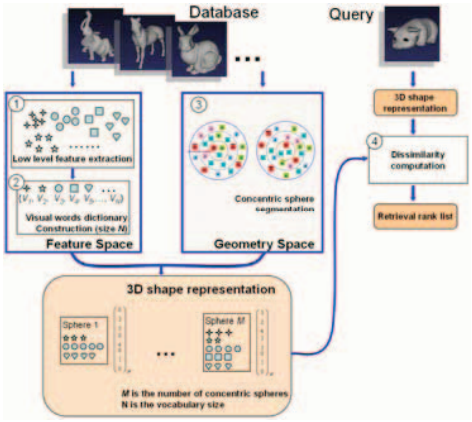

Fig. 2 A schematic description of CBW method

The first block in Figure 2 represents low level feature extraction. We use spin image as the one here. As shown in Figure 3, it characterizes the local 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. The final size of the spin images is defined by the width w and the height h of the spin plane. We uniformly sample N_b oriented basis points and N_s support points on the surface of the model, which satisfies insensitivity to the tessellation and resolution of the mesh.

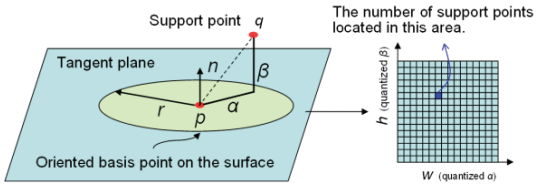

Fig. 3 The demonstration of spin image.

After extracting a set of spin images for each model, we constructed a shape dictionary as shown in the second block, whose size is predetermined as N, by clustering all spin images acquired from the whole training dataset with k-means method.

Instead of representing one model with a histogram of the words from the dictionary, it is partitioned into M regions by grouping the oriented basis points with M concentric spheres as demonstrated in the third block. Thereafter, the model is recorded as a set of histograms.

Because all the models are scaled into a unified scale and the partitioning is also unified, the correspondence between the regions of two models is obvious. It can be constructed from an outer sphere to an inner sphere, as shown in fig 1-right, or reverse. Thus, the CBW feature vector is recorded as

$$\begin{aligned} cfv &= (fv^1, fv^2, \ldots, fv^M) \\ &= (x^1_1, x^1_2, \ldots, x^1_N, \ldots, x^M_1, x^M_2, \ldots, x^M_N). \end{aligned} \quad (1)$$

When performing 3D shape retrieval, the CBW representation of the query shape is constructed on line, and compared with those stored in the database. An ordered retrieval list is obtained according to the dissimilarity metric, which is:

$$Dis(O^A, O^B) = \sum_{i=1}^{M} dis(cfv_i^A, cfv_i^B). \quad (2)$$

where O^A, O^B are two objects A and B respectively; $dis(cfv_i^A, cfv_i^B)$ measures the dissimilarity between two feature vectors, which can be KL divergence [8], cosine distance, L1 and L2 distance, etc.. Thus, for every query object, the objects in the database are all assigned a metric value based on equation (2), which results a sorted retrieval list.

3.3. Experimental results

According to the discussions in [8, 10], several parameters related to the CBW approach are defined as follows:
1) The support range r: r=0.4*R, where R is the radius of the model;
2) The width w and the height h of the spin plane: w=h=12;
3) The number of oriented basis points for one model N_b: N_b =500;
4) The number of oriented basis points for one model N_s: N_s =5000;
5) The size of the dictionary N: N=1500;
6) The number of the concentric spheres M: M<10.

The CBW approach can be applied both in "Specific" and "Generic" domains. Here we demonstrate it on one "Specific" domain, using the Purdue Engineering Benchmark [4]. In figure 4 we compare the Precision Recall curves obtained with CBW and BW using L1 as dissimilarity measurement to those methods defined in [4], which listed from top to bottom are Light Field Descriptor, 2.5D Spherical harmonics, 2D Shape Histogram, 3D Spherical Harmonics, Solid Angle Histogram, 3D Shape Distribution, and Surface Area and Volume. For CBW, we set M=9. Obviously, the concentric sphere partition improves the PR rate, and makes the local feature based method comparable to the global feature based method, such as 2.5D Spherical Harmonics listed as the second best method in [4].

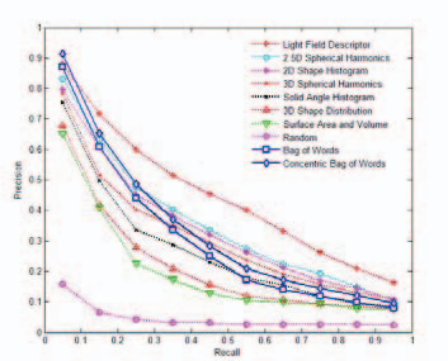

Fig. 4. Precision-recall (PR) plots of CBW, BW and other methods listed in [4].

4. PARTIAL-TO-GLOBAL RETRIEVAL

In CAD domain, the PGR task is specifically important. Suppose that the query partial model is a screw, the target complete model we want to obtain is the same screw with the screw cap on. We design an experiment using PEB [4] to simulate the case described above.
1) Represent the models with the BW method and save the descriptors for the following usage as block 1 and 2 in figure 2.
2) When performing the PG Retrieval, the sampled oriented basis points are grouped into M_p regions according to their geometric positions first. Then one of the groups is chosen as the partial query shape. The BW representation is constructed on line and used to compare with the saved BW descriptors of the complete models. The requirements for dissimilarity measure of the PG retrieval task are quite different than the GG retrieval problem. As described in [8], the dissimilarity between the query data and the target model does not equal to that between the target model and the query data. It means that the dissimilarity metric should be asymmetric. An ordinary symmetric distance measurement, such as L1, L2, is not a suitable choice. KL divergence is chosen here to satisfy the asymmetric property. When using one sixth of the model to be the query shape, the two PR curves in figure 5 demonstrate the improvement introduced by KL comparing to L1.

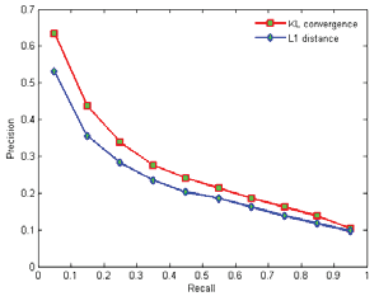

Fig. 5. Precision-recall (PR) plots obtained with KL and L1 dissimilarity measurement when doing PGR.

Figure 6 provides an example comparing the retrieval results of GG Retrieval and PG Retrieval, in which one sixth of a gear is used as the query shape. It shows that the PGR is better than the GGR, since PGR lists more gears on the top of the list than GGR dose. Why does PGR perform better? Recalling the definition of the feature vector will provide some clues to the answer. The feature vector describes the frequency of the visual words appearing in the shape. When using the entire gear model to be the query data, the plane-kind of visual word overwhelms the other features. However, using partial of the object to be the query data, the gear teeth shape dominates the whole shape. So more gears are picked out and listed on the top of the retrieval list.

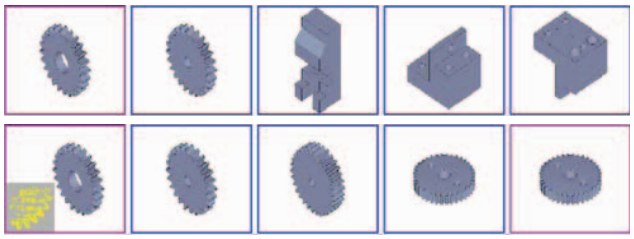

Fig. 6. The top group shows the GGR result using a complete model (the top-left image) as the query. The second group shows the PGR result using 1/6 part of the complete model (the small image shown in the bottom-left corner of the first image) as the query. The top 5 models are listed according to the dissimilarity measurement.

5. ARTICULATED SHAPE RETRIEVAL

The Articulated Shape Retrieval requires that the shape descriptor should be deformation invariant, which is not satisfied by several previous methods [2, 6, 12]. They perform well when dealing with rigid objects, but manifest poor performance when dealing with deformable ones [13]. BW method can still be used effectively for an ASR task. The descriptors for the models are constructed by following the procedure shown as blocks 1 and 2 in figure 2.

We applied the BW method for the ASR task on the McGill Shape Benchmark (MSB) [17]. The configuration of the

parameters is almost the same as those listed in Section 3.3, except that a) the width w and the height h of the spin plane: w=h=*16*; b) the number of oriented basis points for one model N_b: N_b =1000. Since all of the models in MSB are regarded as complete, L1 distance is chosen to measure the dissimilarity.

In figure 7, the BW based retrieval result is compared with several methods described in [5]. The BW method is comparable to the best method EVD. However, except BW, all the other methods are based on geodesic distance computation, which is computational expensive. On the contrary, our method is computational efficient, which is constrained on local area, so it can be applied for on-line retrieval.

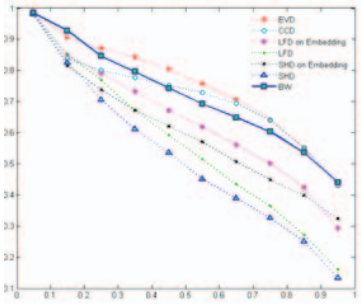

Fig. 7. Precision-recall (PR) plots for various descriptors when applied to the McGill database of articulated shapes [17]. Except BW, all of the other results can be found in [5].

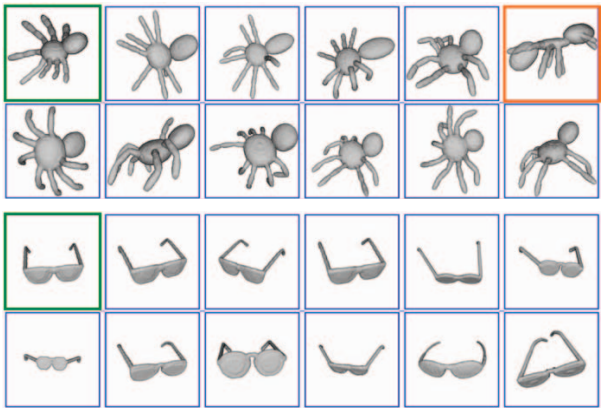

Fig. 8. Two retrieval results (top: spider; bottom: spectacles) from the McGill database [17]: green bold frame defines the query shape, and orange bold frame defines the false pick up.

Figure 8 shows three visual results of the articulated shape retrieval. Only the top 18 results are listed here, in which the green bold framed shape is the query shape, and the orange bold framed shapes are the false recalls. The top group in Figure 8 shows the results of retrieving a spider shape from the database. Among these 18 retrieval shapes, only two shapes do not belong to the spider class but the ant class. The second group in Figure 8 is the result using a spectacle shape as query model. Even though there is quite a large amount of bending in the shapes, the performance is quite good.

6. CONCLUSION AND DISCUSSION

In this paper, we adopted the BW framework to investigate several different tasks in the 3D shape retrieval field, which are classified as specific vs. generic, PG vs. GG retrieval, and articulated vs. non-articulated. For each type, the effectiveness of the BW method is discussed in detail. First, Concentric BW method is introduced to improve the discrimination ability of the original BW representation. Second, BW is applied on PEB to perform PG retrieval task. And several results revealed, for some shape (gear like shape), that PG performs better than GG. Finally, we compared the results of BW to several other methods on the McGill articulated shape database. Our results are comparable to the best results in [5]. More experiments need to be done to verify the influence of the parameters listed in Section 3.3.

REFERENCES

[1] http://store.sae.org/ CAESAR /.

[2] D.Y. Chen, M. Ouhyoung, X.P. Tian, Y.T. Shen. On visual similarity based 3D model retrieval. Eurographics '03, pp. 223-232.

[3] A. E. Johnson, M. Hebert: Using Spin Image for Efficient Object Recognition in Cluttered 3D Scenes. PAMI. Vol.21, No.5, 1999, pp. 433–449.

[4] S. Jayanti, Y. Kalyanaraman, N. Iyer, K. Ramani. Developing an Engineering Shape Benchmark for CAD Models. Computer-Aided Design, Vol. 38, No. 9, 2006, pp. 939-953.

[5] V. Jain, H. Zhang. A Spectral Approach to Shape-Based Retrieval of Articulated 3D Models, Computer-Aided Design, Vol. 39, Issue 5, pp. 398-407, 2007.

[6] M. Kazhdan, T. Funkhouser, S. Rusinkiewicz. Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. Symposium on Geometry Processing. June 2003.

[7] L. Fei-Fei, P. Perona. A Bayesian Hierarchical model for learning natural scene categories. CVPR 2005, pp. 524-531.

[8] Y. Liu, H. Zha, H. Qin. Shape Topics: A Compact Representation and New Algorithms for 3D Partial Shape Retrieval. CVPR'06, 2006, pp. 2025-2032.

[9] X. Li, A. Godil, A. Wagan. Spatially Enhanced Bags of Words for 3D Shape Retrieval. ISVC 2008.

[10] X. Li, A. Godil, A. Wagan. 3D Part Identification Based on Local Shape Descriptors. Permis 2008.

[11] S. Lazebnik, C. Schmid, J. Ponce: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. CVPR'06, 2006, pp. 2169-2178.

[12] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin. Shape Distributions. ACM Transactions on Graphics, 21(4):807-832, 2002.

[13] R. Ohbuchi, K. Osada, T. Furuya, T. Banno. Salient Local visual features for shape-based 3D model retrieval, Proc. IEEE Shape Modeling International 2008, pp. 93-102.

[14] S. Savarese, Li Fei-Fei: 3D Generic Object Categorization, Localization and Pose Estimation. ICCV'07, 2007, pp. 1-8.

[15] Y. Shan, H. S. Sawhney, B. Matei, R. Kumar: Shapeme Histrogram Projection and Matching for Partial Object Recognition. PAMI, Vol. 28, No. 4, 2006, pp. 568-577.

[16] P. Shilane, P. Min, M. Kazhdan, T. Funkhouser. The Princeton shape benchmark. SMI'04, 04(00): 167-178, 2004.

[17] J. Winn, A. Criminisi, T. Minka, Object categorization by learned universal visual dictionary, ICCV 2005, Vol. II

ACKNOWLEDGEMENT

We would like to thank the SIMA and the IDUS program for supporting this work. This work has also been partially supported by NSF Grants of China (60873218) and NSF Grants of Zhejiang (Z1080232).