Eurographics Workshop on 3D Object Retrieval (2009)
I. Pratikakis, M. Spagnuolo, T. Theoharis, and R. Veltkamp (Editors)

Visual vocabulary signature for 3D object retrieval and partial matching

R. Toldo[1] and U. Castellani[1] and A. Fusiello[1]

[1] Dipartimento di Informatica, Università di Verona,
Strada Le Grazie 15, 37134 Verona, Italy

Abstract

In this paper a novel object signature is proposed for 3D object retrieval and partial matching. A part-based representation is obtained by partitioning the objects into subparts and by characterizing each segment with different geometric descriptors. Therefore, a Bag of Words *framework is introduced by clustering properly such descriptors in order to define the so called 3D visual vocabulary. In this fashion, the object signature is defined as a histogram of 3D* visual word *occurrences. Several examples on the Aim@Shape watertight dataset demonstrate the versatility of the proposed method in matching either 3D objects with articulated shape changes or partially occluded or compound objects. In particular, a comparison with the methods that participated to the Shape Retrieval contest 2007 (SHREC) reports satisfactory results for both object retrieval and partial matching.*

1. Introduction

In the last years, the proliferation of large databases of 3D models caused a surge of interest in methods for content-based object retrieval [IJL*05, FKMS05, TV04]. One of major challenges in the context of data retrieval is to elaborate a suitable canonical characterization of the entities to be indexed. In the literature, this characterization is referred to as a *descriptor* or *signature*. Since the descriptor serves as a key for the search process, it decisively influences the performance of the search engine in terms of computational efficiency and relevance of the results. Roughly speaking, there are two categories of descriptors: (i) *global* and (ii) *local*. Global descriptors consists in a set of features that effectively and concisely describe the entire 3D model [FMK*03]. Local descriptors are instead collections of local features of relevant object subparts [SF06].

In this paper we present a local methods inspired to the *Bag-of-Words* (BoW) framework for textual document classification and retrieval. In this approach, a text is represented as an unordered collection of words, disregarding grammar and even word order. The extension of such approach to visual data requires the building of a *visual vocabulary*, i.e., the set of the visual analog of words. For example, in [CDF*04] 2D images are encoded by collecting interest points which represent local salient regions. This approach has been extended in [GD07] by introducing the concept of *pyramid* kernel matching. Instead of building a fixed vocabulary, the visual words are organized in a hierarchical fashion in order to reduce the conditioning of the free parameter definition (i.e., the number of bins of the histogram). Finally, in [LMSR08] the BoW paradigm has been introduced for human actions categorization from real movies. In this case, the visual words are the quantized vectors of spatiotemporal local features. The extension of the BoW paradigm to 3D objects is non-trivial and has been proposed only in few recent works [OkOFB08, LZQ06, LGW08]. In [OkOFB08] range images are synthetically generated from the full 3D model, then salient points are extracted as for the 2D (intensity) images. In [LZQ06, LGW08] Spin Images are chosen as local shape descriptors after sampling the mesh vertices.

In our approach a 3D visual vocabulary is defined by extracting and grouping the geometric features of the object sub-parts from the dataset, after 3D object segmentation. Note that usually local techniques are defined by point-based features rather than by segmentation. Only recently [SSSCO08] proposed a part-based retrieval method by partitioning an object to meaningful segments and finding analogous parts in other objects. Thank to this *part-based* representation of the object we achieve pose invariance, i.e., insensitivity to transformation which change the articulations of the 3D object [GSCO07]. Moreover, our method is able to discriminate objects with similar skeletons, a feature that is shared by very few other works like [TL07]. Its main steps are:

Object sub-parts extraction (Sec. 2). Spectral clustering is used for the selection of seed-regions. Being inspired by the *minima-rule* [HR87], the adjacency matrix is tailored in order to allow convex regions to belong to the same segment. Furthermore, a multiple-region growing approach is introduced to expand the selected seed-regions, based on a weighted fast marching. The main idea consist on reducing the speed of the front for concave areas which are more likely to belong to the region boundaries. Then, the segmentation is recovered by combining the seeds selection and the region-growing steps.

Object sub-parts description (Sec. 3). Local region descriptors are introduced to define a compact representation of each sub-part. Working at the part level, as opposed to the whole object, enables a more flexible class representation and allows scenarios in which the query model is significantly deformed. We focus on region descriptors easy to compute and partially available from the previous step (see [SF06] for an exhaustive overview of shape descriptors).

3D visual vocabularies construction (Sec. 4). The set of region descriptors are properly clustered in order to obtain a fixed number of 3D visual *words* (i.e., the set of clusters centroids). In practice, the clustering defines a vector quantization of the whole region descriptor space. Note that the vocabulary should be large enough to distinguish relevant changes in object parts, but not so large as to distinguish irrelevant variations such as noise.

Object representation and matching (Sec. 5). Each 3D object is encoded by assigning to each object sub-part the corresponding visual word. Indeed, a BoW representation is defined by counting the number of object sub-parts assigned to each word. In practice, a histogram of visual words occurrences is built for each 3D object which represent its *global* signature [CDF*04]. Matching is accomplished by comparing the signatures.

2. Objects segmentation

Due to its wide ranging applications, 3D object segmentation has received a great attention lately. The recent survey by [Sha08] and the comparative study by [AKM*06] have thoroughly covered the several different approaches developed in literature.

In the following we present a novel mesh segmentation technique that provides a consistent segmentation of similar meshes complying with the cognitive *minima rule* [HR87]. In addition, the overall approach depends on very few parameters and is very fast.

The minima rule states that human perception usually divides a surface into parts along the concave discontinuity of the tangent plane [HR87]. Therefore this suggests to cluster in the same set convex regions and to detect boundary parts as concave ones. A concise way to characterize the shape in terms of principal curvatures is given by the *Shape Index* [Pet02].

$$s = -\frac{2}{\pi}\arctan\left(\frac{k_1+k_2}{k_1-k_2}\right) \qquad k_1 > k_2 \tag{1}$$

where k_1, k_2 are the principal curvatures of a generic vertex $x \in V$. The Shape Index varies in $[-1,1]$: a negative value corresponds to concavities, whereas a positive value represents a convex surface.

The key idea behind our algorithm is the synergy between two main phases: (i) the detection of similar connected convex regions, and (ii) the expansion of these seed-regions using a multiple region growing approach. According to the minima-rule the Shape Index is employed in both phases.

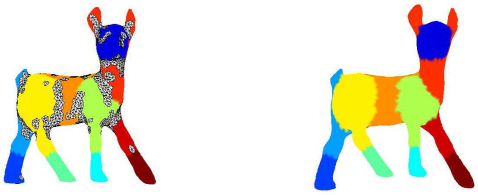
2.1. Seed-regions detection by Spectral Clustering

The extraction of the seed-regions is accomplished with Normalized Graph Cuts [SM00]. This approach has been firstly applied to image segmentation although it is stated as a general clustering method on weighted graphs. In our case, the weight matrix is built using the Shape Index at each vertex:

$$w(x_i, x_j) = e^{-|s(x_i)-s(x_j)|} \tag{2}$$

where the vertices with negative Shape Index – i.e., those corresponding to concave regions – have been previously discarded. In this way we cluster together vertices representing the same convex shape.

The number of clusters, needed by the Spectral clustering approach, is linked, but not equal, to the number of final segments. Indeed, clusters are not guaranteed to be connected in the mesh. This happens because we do not take into account any geodesic distance information at this stage: we cluster



(a) Seed regions found with spectral clustering.

(b) Final Segmentation.

Figure 1: *An example of segmentation.*

only according to the curvature value at each vertex. Hence, we impose connection as a post-processing step: the final seed regions are found as connected components in the mesh graph, with vertices belonging to the same cluster. An example of seed regions found by the algorithm is shown in figure 1(a).

2.2. Multiple region growing by weighted fast marching

Once the overall seed regions are found, we must establish a criteria to cluster the vertices that don't belong to any initial seed region. The key idea is to expand the initial seeds region using a *weighted* geodesic distance. Again, the weight at each vertex is chosen according to the minima-rule. In formulae, given two vertices $x_0, x_1 \in V$, we define the *weighted geodesic distance* $d(x_0,x_1)$ as

$$d(x_0,x_1) = min_\gamma \left\{ \int_0^1 \|\gamma'\| w(\gamma(t)) dt \right\} \tag{3}$$

where $w(\cdot)$ is a weight function (if $w(\cdot) = 1$ this is the classic geodesic distance) and γ is a piecewise regular curve with $\gamma(0) = x_0$ and $\gamma(1) = x_1$. Our weight function is based on the Shape Index s:

$$w(x) = e^{\alpha s(x)} \tag{4}$$

where α is an arbitrary constant. An high α value heavily slow down the front propagation where the concavity are more prominent. In our experiments we used a fixed $\alpha = 5$ to obtain consistent segmentations.

An example segmentation along with starting seed regions is shown in figure 1(b). Several other examples of segmentation on different objects are shown in figure 2. Similar parts seem to be segmented in a similar manner (provided that the parameters of the segmentations are equal).

3. Segment descriptors

We chose four type of descriptors to represent each extracted region. The first three are local and a value is computed for every point of the region, namely:

- **Shape Index** *si*. As explained before, the Shape Index provides a local categorization of the shape into primitive forms such as spherical cap and cup, rut, ridge, trough, or saddle.
- **Radial Geodesic Distance** *rg*. Radial geodesic distance measures the geodesic distance of a surface point to the geodesic centroid of the region. In our case, for computation efficiency, we approximate the geodesic centroid as the closest point on the mesh to the Euclidean centroid.
- **Normal Direction** *n*. This is the unit normal vector at a surface point. We represent it as a pair (θ,α) where θ is the angle between the normal vector and the XZ-plane and α is the angle between the positive X-Vector and the projection of the normal vector on the XZ-plane. The normal n is scale invariant but not pose invariant.

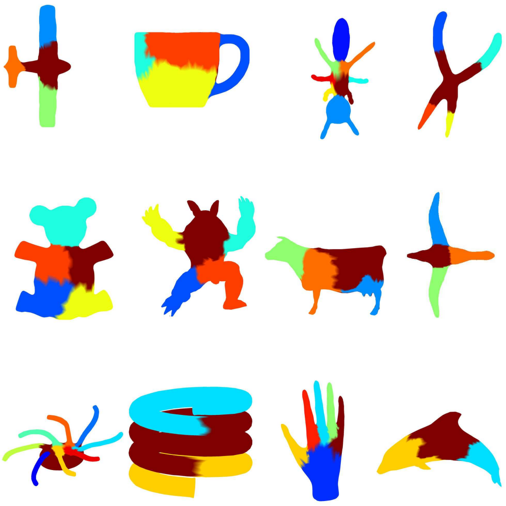


Figure 2: *Examples of segmentation of some objects from the Aim@Shape Dataset.*

The three descriptors *SI*, *RG*, *N* are defined as the normalized histograms of the observed values in the region vertices, respectively. The fourth descriptor depends on the relative positions of the regions and thus it's a context descriptor. Precisely, the histogram of the **Geodesic Context** *GC* descriptor for a region is built computing the geodesic distance between its centroid and the centroids of the other regions. The *GC* descriptor, defined for regions, resembles the shape context descriptor [BM00], defined for points.

Please note that the number of bins chosen for each histogram of the four descriptors is a critical choice. A small number reduce the capability of the region descriptor in discriminating among different segments. On the other hand, a high number increases the noise conditioning. Hence we introduce, for each descriptor, histograms with different number of bins in order to obtain a *coarse-to-fine* regions representation.

4. 3D visual vocabularies construction

The different sets of region descriptors must be clustered in order to obtain several visual words. Since we start with different segmentations and different types of descriptors, we adopted a multi-clustering approach rather than merging descriptors in a bigger set. Before the clusterization, the sets of descriptors are thus split in different subsets as illustrated in figure 3. The final clusters are obtained with a k-means algorithm. Again, instead of setting a fixed free parameter k, namely the number of cluster, we carry out different clusterizations while varying this value.

Once the different clusters are found we retain only their centroids, which are our *visual words*. In figure 4 an exam-

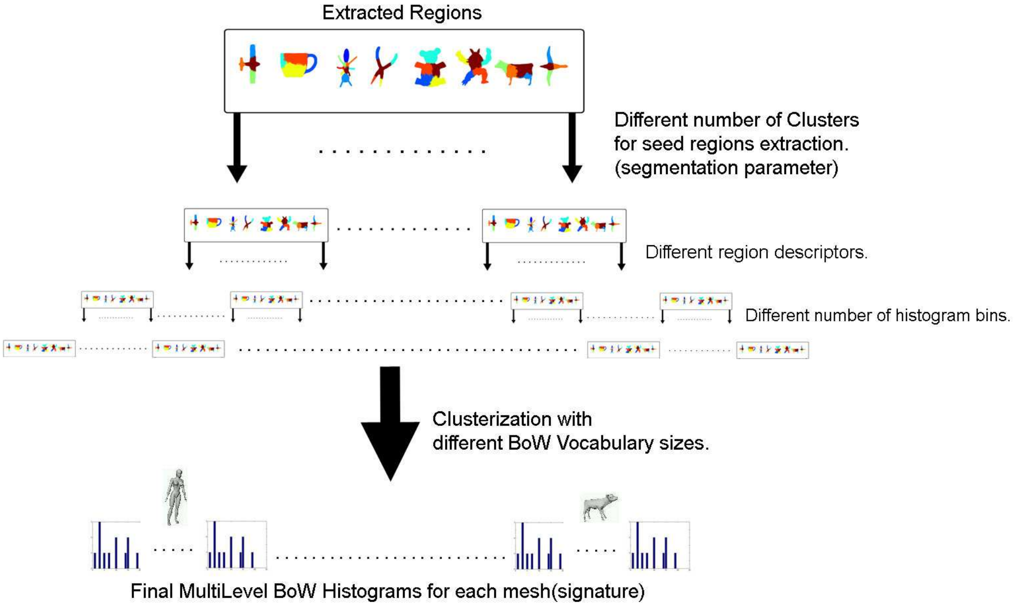

Extracted Regions
Different number of Clusters for seed regions extraction. (segmentation parameter)
Different region descriptors.
Different number of histogram bins.
Clusterization with different BoW Vocabulary sizes.
Final MultiLevel BoW Histograms for each mesh(signature)

Figure 3: *The construction of the vocabularies is performed in a multilevel way. At the beginning we have all region extracted for different numbers of seed regions (variable segmentation parameter). For every region, different descriptors are attached. The different region descriptors are divided by the type of descriptor and its number of bins. The final clusterizations are obtained with varying number of clusters. At the end of the process we obtain different Bag-of-Words histograms for each mesh.*

ple of descriptors subset clusterization with relative distance from centroid is shown. Note that object sub-parts from different categories may fall in the same cluster since they share similar shape.

More in details, at the end of this phase we obtain the set of visual vocabularies $V_s^{d,b,c}$, where:

- s identifies the index of the multiple 3D segmentation (variable segmentation parameter $s \in \{6,8,10,12,14\}$),
- d identifies the region descriptor types ($d \in \{SI, RG, N, GC\}$),
- b identifies the refined level of the region descriptor (number of histogram bins $b \in \{20,30,40,50\}$),
- c identifies the refined level of the vocabulary construction (number of clusters).

5. 3D representation and matching

In order to construct a Bag-of-Words histogram of a new 3D object, we compare its regions descriptors with the visual words of the corresponding visual vocabulary. In practice, each segment is assigned to the most similar visual words. Indeed, by counting the number of segment assigned to each word the Bag-of-Words representation is obtained. The resulting signature is a very sparse vector of occurences. Finally, the objects matching is obtained by comparing their respective signature by using standard metric for histograms. Note that, as observed in [GD07] the proposed method implicitly encodes the sub-parts matching since corresponding segments are likely to belong to the same histogram bin. If a new category of objects is added to the dataset, the visual vocabularies need to be updated.

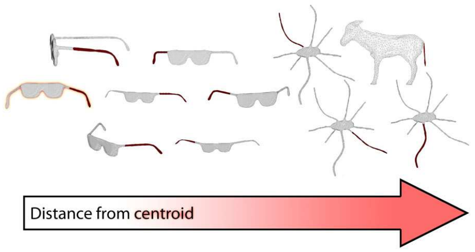


Figure 4: *Example of a Bag-of-Words cluster for SI descriptors. The centroid is highlighted with red and others region in the same cluster are sorted by distance from centroid. Note that sub-parts of meshes from different categories may fall in the same cluster since they share similar shape.*

6. Results

In order to prove the effectiveness and the generalization capability of the proposed paradigm we tested it with two different tasks. The first one is a classical retrieval task in which the dataset consists of 400 meshes of 20 different classes. In the second task, using the previous dataset as ground truth, it is required to classify 30 queries composed with different parts from the ground truth meshes.

6.1. Retrieval Task

The Aim@Shape Watertight dataset has been used for various retrieval contests [VtH07]. It contains 20 categories each composed of 20 meshes. The entire dataset is shown in figure 5. We compared our method with the participant of the Aim@Shape Watertight 2007 contest [VtH07]. We used precision and recall to evaluate our results, that are two fundamental measures often used in evaluating search strategies. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database, while precision is the ratio of the number of relevant records retrieved to the size of the return vector [SM83]. In table 1 the precision and recall of our approach along with the results of the other methods are reported, while in figure 6 the precision vs recall plot of our method is shown. The results divided by category are shown in figure 5. The algorithm fails with some meshes, but the overall rate of success is still fairly good. The dataset is tough since there are many categories and objects inside the same category can be very different.

Precision after	20	40	60	80
Ideal	1	0.5	0.333	0.25
Tung et al.	0.714	0.414	0.290	0.225
Our Approach	**0.648**	**0.379**	**0.270**	**0.210**
Akgul et al.	0.626	0.366	0.262	0.205
Napoleon et al.	0.604	0.366	0.262	0.205
Daras et al.	0.564	0.346	0.252	0.199
Chaouch et al.	0.546	0.329	0.241	0.190
Recall after	20	40	60	80
Ideal	1	1	1	1
Tung et al.	0.714	0.828	0.872	0.902
Our Approach	**0.648**	**0.758**	**0.808**	**0.841**
Akgul et al.	0.626	0.732	0.786	0.821
Napoleon et al.	0.604	0.732	0.788	0.822
Daras et al.	0.564	0.692	0.756	0.798
Chaouch et al.	0.546	0.658	0.724	0.763

Table 1: *Precision and Recall after 20, 40, 60 and 80 retrieved items*

6.2. Partial Matching Task

The ground-truth dataset is again the Aim@Shape Watertight. The query test models are 30 and each query model shares common subparts with (possibly) more than one model belonging to the ground-truth dataset. The query set is shown in figure 7. Again, we compared our method with the participant of the Aim@Shape Partial Matching 2007 contest [VtH07]. In this case we didn't employ the Geodesic Context descriptor, since it's global and the Normal Direction descriptor, since it's not pose invariant. In order to evaluate the performance, a set of highly relevant, marginally relevant and non-relevant models belonging to the dataset has been associated to each query model (table 2). The performance indicator used is the Normalized Discounted Cumulated Gain vector (NDCG) [JK02], which is recursively defined as

$$DCG[i] = \begin{cases} G[i] & \text{if i = 1} \\ DCG[i-1] + G[i]\log_2(i)) & \text{otherwise} \end{cases} \quad (5)$$

where G[i] represents the value of the gain vector at the position i. In our case, for a specific query, G(i) equals 2 for highly relevant models, 1 for marginally relevant models and 0 for non-relevant models. The normalized discounted cumulated gain vector NDCG is obtained by dividing DCG by the ideal cumulated gain vector. In figure 8 the NDCG of our approach along with the results of the other methods are reported. We can notice how our method performs better than the other methods considered.

6.3. Timing

The entire pipeline is computationally efficient in each stage. We used an entry level laptop at 1.66*Ghz* to perform tests. The code is written in Matlab with some parts in C. An entire mesh segmentation of 3500 vertices is computed in less than 5 seconds, of which $\sim 2.8s$ are necessary to extract all the seed regions, and $\sim 2.1s$ are needed to compute the entire

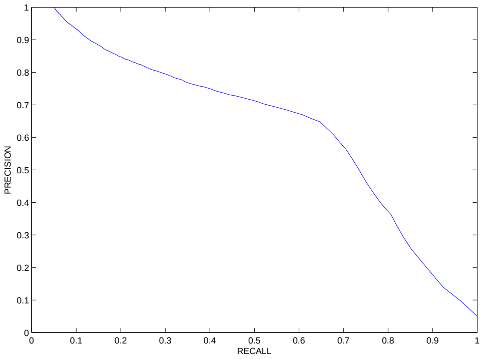
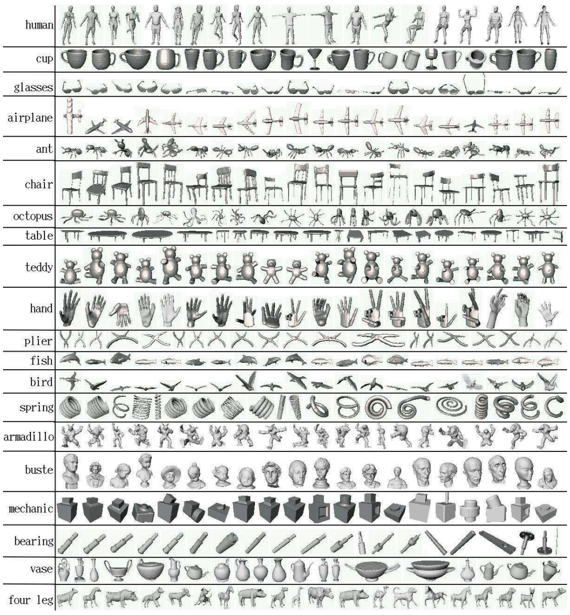
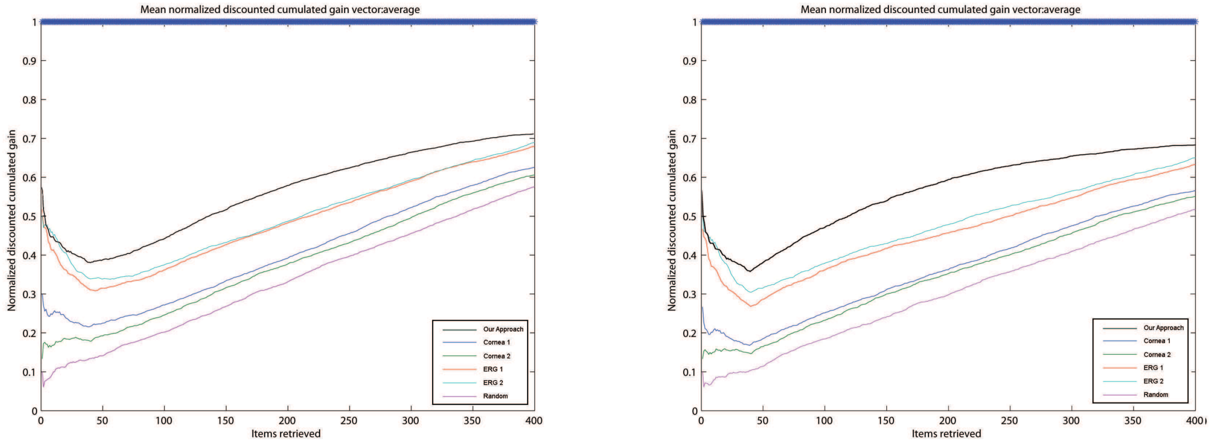


Figure 6: *Precision-recall of our method.*



Category	Precision after 20
Human	0,53
Cup	0,46
Glasses	0,90
Airplane	0.73
Ant	0.92
Chair	0.57
Octopus	0.61
Table	0.52
Teddy	0.94
Hand	0.32
Plier	0.99
Fish	0.8
Bird	0.4
Spring	0.96
Armadillo	0.94
Buste	0.57
Mechanic	0.80
Bearing	0.44
Vase	0,8
Four Legs	0.32

Figure 5: *Aim@Shape Watertight Dataset objects divided by category and retrieval precision for each category after 20 retrieved items*



(a)

(b)

Figure 8: *Overall Normalized Discount Cumulated Gain considering only highly relevant models 8(a) and both highly relevant and marginally relevant models 8(b).*

Query Number	Relevant Classes	Marginally Relevant Classes
1	cup, teddy	vase, four legs
2	human, table	armadillo, chair
3	buste, mechanic	
4	plier, spring	airplane, bird
5	ant, glasses	octopus
6	four legs, airplane	bird, plier, teddy
7	armadillo, vase, bearing	human, cup
8	fish, bird mechanic	airplane, plier
9	chairs, bearings	tables
10	human, table	armadillo, chair
11	fish, hand	
12	human, octopus	armadillo, ant
13	hand, spring	
14	human, fish	armadillo
15	four legs, vase	cup, teddy
16	bird, buste	airplane, plier
17	chair, plier	airplane, bird, table
18	ant, octopus	
19	airplane, armadillo	human, bird, plier
20	teddy, spectacle	four legs
21	cup, springs	vase
22	four legs, cup	vase, teddy
23	armadillo, bearing, bird	human, airplane, plier
24	airplane, bird	plier
25	head, vase	cup
26	chair, table	
27	teddy, hand	four legs
28	octopus, plier	bird, airplane, ant
29	airplane, mechanical	bird, plier
30	four legs, human	armadillo, teddy

Table 2: *The category ground-truth for each query model.*

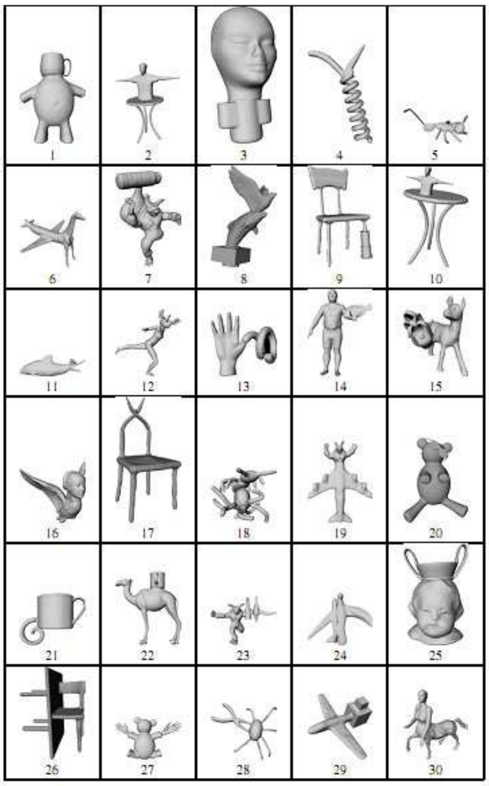


Figure 7: *Aim@Shape Partial Matching query objects.*

.

hierarchical segmentation. Region descriptors are computed efficiently: on the average it takes $\sim 0.5s$ to extract all the four descriptors of a single region. As for the k-means clusterization, 10 clusters for 300 points each composed of 200 feature are extracted in less than one second.

7. Conclusions

In this paper a new approach for 3D object retrieval and partial matching is introduced basing on the Bag-of-Words paradigm. The main steps of the involved pipeline have been carefully designed by focusing on both the effectiveness and efficiency.

The Bag-of-Words approach fits naturally with sub-parts encoding by combining segment descriptors into several visual vocabularies. This allows the retrieval of objects which heavily deform their shape and change significantly their pose. Moreover, our methods is able to satisfy query models of composed objects.

The experimental results are encouraging. Our framework is versatile in reporting satisfying performances for both object retrieval and partial matching as shown in the comparison with other methods.

Acknowledgments

This paper was partially supported by PRIN 2006 project 3-SHIRT.

References

[AKM*06] ATTENE M., KATZ S., MORTARA M., PATANE G., SPAGNUOLO M., TAL A.: Mesh segmentation - a comparative study. In *Proceedings of the IEEE International Conference on Shape Modeling and Applications* (2006), IEEE Computer Society, p. 7.

[BM00] BELONGIE S., MALIK J.: Matching with shape contexts. *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on* (2000), 20–26.

[CDF*04] CRUSKA G., DANCE C. R., FAN L., WILLAMOWSKI J., BRAY C.: Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision* (2004), pp. 1–22.

[FKMS05] FUNKHOUSER T., KAZHDAN M., M. P., SHILANE P.: Shape-based retrieval and analysis of 3D models. *Communications of the ACM 48*, 6 (2005).

[FMK*03] FUNKHOUSER T., MIN P., KAZHDAN M., CHEN J., HALDERMAN A., DOBKIN D.: A search engine for 3D models. *ACM Transactions on Graphics 22* (2003).

[GD07] GRAUMAN K., DARRELL T.: The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research 8*, 2 (2007), 725–760.

[GSCO07] GAL R., SHAMIR A., COHEN-OR D.: Pose-oblivious shape signature. *IEEE Transaction on Visualization and Computer Graphics 13*, 2 (2007), 261–271.

[HR87] HOFFMAN D. D., RICHARDS W. A.: Parts of recognition. *Cognition* (1987), 65–96.

[IJL*05] IYER N., JAYANTI S., LOU K., KALYNARAMAN Y., RAMANI K.: Three dimensional shape searching: State-of-the-art review and future trend. *Computer Aided Design 5*, 37 (2005), 509–530.

[JK02] JÄRVELIN K., KEKÄLÄINEN J.: Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst. 20*, 4 (2002), 422–446.

[LGW08] LIN X., GODIL A., WAGAN A.: Spatially enhanced bags of words for 3d shape retrieval. In *ISVC '08: Proceedings of the 4th International Symposium on Advances in Visual Computing* (2008), vol. 5358, Springer-Verlag, pp. 349–358.

[LMSR08] LAPTEV I., MARSZA M., SCHMID C., ROZENFELD B.: Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[LZQ06] LI Y., ZHA H., QIN H.: Sapetopics: A compact representation and new algorithm for 3d partial shape retrieval. In *International Conference on Computer Vision and Pattern Recognition* (2006).

[OkOFB08] OHBUCHI R., K. OSADA, FURUYA T., BANNO T.: Salient local visual features for shape-based 3d model retrieval. In *International Conference on Shape Modelling and Applications* (2008).

[Pet02] PETITJEAN S.: A survey of methods for recovering quadrics in triangle meshes. *ACM Computing Surveys 34*, 2 (2002).

[SF06] SHILANE P., FUNKHOUSER T.: Selecting distinctive 3D shape descriptors for similarity retrieval. In *International Conference on Shape Modelling and Applications* (2006), IEEE Computer Society.

[Sha08] SHAMIR A.: A survey on mesh segmentation techniques. *Computer Graphics Forum* (2008).

[SM83] SALTON G., M.MCGILL.: *Introduction to modern information retrieval*. McGraw Hill, 1983.

[SM00] SHI J., MALIK J.: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8 (2000), 888–905.

[SSSCO08] SHALOM S., SHAPIRA L., SHAMIR A., COHEN-OR D.: Part analogies in sets of objects. In *Eurographics Workshop on 3D Object Retrieval* (2008).

[TL07] TAM G. K. L., LAU W. H. R.: Deformable model retrieval based on topological and geometric signatures. *IEEE Transaction on Visualization and Computer Graphics 13*, 3 (2007), 470–482.

[TV04] TANGELDER J. W., VELTKAMP R. C.: A survey of content based 3d shape retrieval methods. In *International Conference on Shape Modelling and Applications* (2004), pp. 145–156.

[VtH07] VELTKAMP R. C., TER HAAR F. B.: *SHREC 2007 3D Retrieval Contest*. Technical Report UU-CS-2007-015, Department of Information and Computing Sciences, 2007.