Eurographics Workshop on 3D Object Retrieval (2013)
S. Biasotti, I. Pratikakis, U. Castellani, T. Schreck, A. Godil, and R. Veltkamp (Editors)

Compact Vectors of Locally Aggregated Tensors for 3D shape retrieval

Hedi Tabia[1] , David Picard[1] , Hamid Laga[2,3] and Philippe-Henri Gosselin[1,4]

[1]ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051, France
[2]Phenomics and Bioinformatics Research Centre, University of South Australia, Australia
[3]Australian Centre of Plant Functional Genonomics (ACPFG), Australia
[4]INRIA Rennes Bretagne Atlantique, France

Abstract

During the last decade, a significant attention has been paid, by the computer vision and the computer graphics communities, to three dimensional (3D) object retrieval. Shape retrieval methods can be divided into three main steps: the shape descriptors extraction, the shape signatures and their associated similarity measures, and the machine learning relevance functions. While the first and the last points have vastly been addressed in recent years, in this paper, we focus on the second point; presenting a new 3D object retrieval method using a new coding/pooling technique and powerful 3D shape descriptors extracted from 2D views. For a given 3D shape, the approach extracts a very large and dense set of local descriptors. From these descriptors, we build a new shape signature by aggregating tensor products of visual descriptors. The similarity between 3D models can then be efficiently computed with a simple dot product. We further improve the compactness and discrimination power of the descriptor using local Principal Component Analysis on each cluster of descriptors. Experiments on the SHREC 2012 and the McGill benchmarks show that our approach outperforms the state-of-the-art techniques, including other BoF methods, both in compactness of the representation and in the retrieval performance.

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Information storage and retrieval]: Information Search and Retrieval—I.5.4 [Pattern recognition]: Applications —

1. Introduction

Databases of 3D models available in the public domain have created the need for shape analysis and retrieval algorithms capable of finding similar shapes in the same way a search engine responds to text and image queries. State-of-the-art 3D model retrieval methods [TV08] often start by extracting local features and descriptors, then aggregating the features into compact signatures, and finally compare the signatures with standard distances. Recently, Bag of Features (BoF) representations, that have been widely adopted by the computer vision community for image retrieval and scene understanding, are gaining popularity in 3D shape analysis and retrieval [BBGO11, LGS10, TDVC10]. One of the advantages of using BoF representations [JPD*12] is that one can benefit from the rich literature of powerful local descriptors, such as SIFT [Low04], Histogram of Oriented Gradients (HoG) [DT05], and many others [MS05]. Second, BoF representations can be compared with standard distance measures. Third, although BoF vectors can have large dimensions when used for retrieval in large databases [JPD*12], they are often sparse and inverted lists can be used to implement efficient search [JPD*12, SZ03]. BoF-based approaches start by building a dictionary of K visual words from a set of training samples. The visual words are usually obtained by k-means clustering of the local features extracted from all the 3D models in the training set. Each 3D model is then represented with the statistics of the distribution of the visual words in the 3D model. These vector representations can be compared with standard distances, and be subsequently used by robust classification methods such as Support Vector Machines. Most of BoF-based 3D retrieval techniques proposed so far in the literature represent a 3D model with a vector of frequencies of occurrences of visual words [LG09, OOFB08, LGS10, TCF10, BBGO11, Lav12]. This corresponds to 0-order statistics of the distribution of

DOI: 10.2312/3DOR/3DOR13/017-024

the visual words. In this paper, we explore the usage of high-order statistics and show that this enables the computation of compact descriptors while outperforming 0-order statistics in terms of retrieval and classification performance.

Our approach starts with the extraction of dense features from each shape. The advantage of using dense descriptors, in contrast to few descriptors computed at sparse locations, is well acknowledged by the computer vision community as it enables efficient 3D model retrieval and classification. The gain in performance comes at the expense of significant increase in computation time and memory requirement. In this paper we study and evaluate four descriptor aggregation procedures, namely: (1) the standard BoF approach based on 0-order statistics, (2) the Vectors of Locally Aggregated Tensors (VLAT) method [NPG12b], which sums tensor products of the local descriptors, (3) Principal Component Analysis (PCA)-based VLAT descriptors, which reduce the size of the dictionary while improving further their discrimination power, and (4) we propose to further increase the compactness of the reduced-size signature by projection in a well chosen sub-dimensional space. Our evaluation of these approaches on 3D generic shapes and on 3D non-rigid shapes show that aggregation with high order statistics significantly outperforms standard BoF approaches both in terms of retrieval performance and in terms of compactness of the representation. The remainder of the paper is organized as follows. In section 2, the related works are presented. In section 3, the method is detailed. Then, in Section 4 the experiments are presented. Conclusions and future developments (Section 5) end the paper.

2. Related work

While Bag of Features approaches are very popular in 2D image analysis, few works have been introduced for 3D object recognition. Most of them are straightforward extension of the 2D approaches to 3D data. Existing approaches differ in the type of local features used in the construction of the visual vocabulary, and in the way these features are aggregated into signatures. Some of the approaches represent a 3D model by a set of 2D views which are indexed using bags of 2D SIFT features [OOFB08, LGS10]. Other techniques use features extracted directly on the surface of 3D shapes. Liu et al. [LZQ06] and Li and Godil [LG09] use Spin Image descriptors computed on a dense set of feature points uniformly sampled on the surface of the 3D model. Toldo et al. [TCF10] segment the shape into regions and describe each region with several descriptors. A 3D shape is then modeled as a histogram of sub-parts occurrences. Segmentation-based BoF have the advantage of capturing some semantics of the shapes. Tabia et al. [TDCV12] represent each 3D object as a vector of weighted occurrences of features. In all these works, spatial relationships between features are lost in the construction of the signatures. Lavoue [Lav12] and Bronstein et al. [BBGO11] add spatial relationships to take into account the spatial information between feature points. Most of BoF-based 3D retrieval approaches focus on the type of features and little attention is given to the way these features are aggregated into signatures. Consequently the BoF vector that results from the representation is of very high dimension, particularly when dealing with large-scale databases. Our main contribution in this paper is a new step for descriptor aggregation into compact signatures, while improving the high discrimination power of the descriptor. The challenge in using dictionary-based approaches is to find a good trade-off between discrimination power, the size of the descriptor, and the scalability to large object databases. First, most of the methods used in 3D retrieval represent a 3D model with a histogram of frequency of occurrences of the visual words in the model. This is a direct adaptation from the basic BoF approach used by the computer vision and pattern recognition communities for image retrieval and categorization [SZ03]. To the best of our knowledge, other variants of BoF approaches have never been applied to the problem of 3D model retrieval. For example, these methods have been generalized using coding/pooling schemes, and achieve good performances in image categorization with the same feature size, using Locality-constrained Linear Coding (LLC) [WYY*10]. Representing a 3D model with a vector of frequencies of occurrences of the visual word corresponds to 0-order statistics of the distribution of the visual words. In this paper, we explore the usage of high order statistics. The most popular method is the Fisher Vectors (FV) [PSM10]. FV describes how the set of descriptors deviates from an average distribution of descriptors, modeled by a parametric generative model, usually a Gaussian Mixture Model. The FV achieves better results than BoF approaches [CLVZ11], but at the cost of a larger index and higher search time. Jerge et al. [JPD*12] proposed a set of approximation and compression techniques so that FV can be used on large image datasets. Negrel et al. [NPG12b] build efficient signatures by linearizing the kernel function on bags. A Kernel function on bags aims at computing a similarity analog to vote-based systems, but with respect to Mercer's condition. Their huge computational complexity can be highly reduced using linearization techniques based on tensors, called Vectors of Locally Aggregated Tensors (VLAT). VLAT performances are comparable to Fisher Vectors [NPG12b]. The size of the dictionary produced by VLAT can be significantly reduced using Compact VLAT [NPG12a].

In this paper, we study for the first time the performance of these different variants of BoF techniques on 3D model retrieval tasks. We will show that Vectors of Locally Aggregated Tensors (VLAT) and Principal Component Analysis-based VLAT outperform significantly the standard BoF techniques in terms of retrieval performance. We propose also Compact VLAT, which reduces significantly the size of the shape descriptor, and show that it achieves relatively good performance which makes it suitable for broad classification

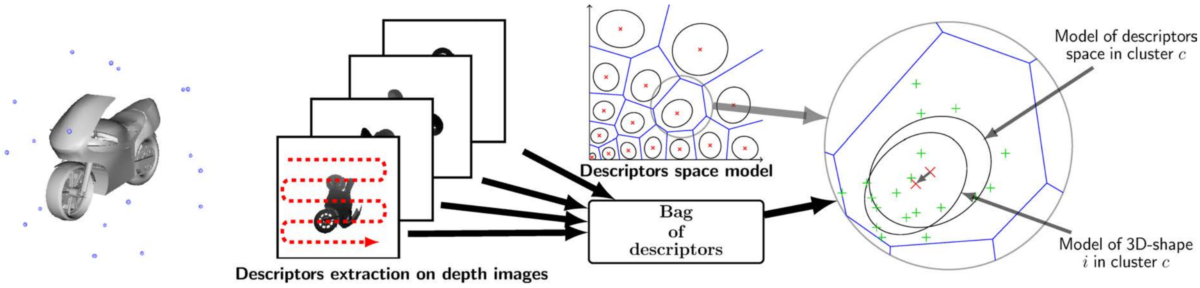
Descriptors extraction on depth images
Descriptors space model
Bag of descriptors
Model of descriptors space in cluster c
Model of 3D-shape i in cluster c

Figure 1: *Method overview: First we compute depth images of the model captured from cameras localized on the unit sphere. Then, local features* (e.g. *HoG) are extracted from the images. Finally, we compute the VLAT signature using deviations between covariance matrices of the codebook and covariance matrices of the descriptors.*

of shapes. We evaluate the performance of these descriptors on two different types of databases: the SHREC 2012 Generic 3D Shape Retrieval benchmark [LGA*12] and the Mcgill dataset [SZM*08] for articulated 3-D models.

3. Method overview

Figure 1 gives an overview of the proposed method. First in a pre-processing step, we normalize the models to ensure that the extracted descriptors are invariant to translation and scale. Then, we render depth maps of the object from n views uniformly sampled on the surface of a bounding unit sphere. We represent each depth map as a collection of dense Histogram of Oriented Gradients (HoG) descriptors. HoG have the advantage of being compact and easy to compute. We then build a shape signature by aggregating the descriptors using the BoF paradigm. In this paper we explore and evaluate various BoF schemes, namely: (1) the standard BoF approaches, which describe a 3D shape as a vector of frequencies of occurrences of the visual words, (2) the Vector of Locally Aggregated Tensors (VLAT), which have been originally proposed for image analysis [NPG12a]. They have the advantage of being compact and efficient for large-scale image retrieval, (3) we further reduce the size of descriptors constructed with VLAT, while improving further their discrimination power, by using Principal Component Analysis (PCA) on the VLAT vectors, and (4) finally, we proposed a new version called compact VLAT, which builds a compact vector and show that it is suitable for rough classification of 3D shapes. The methods proposed in this paper have several advantages. They are invariant to rigid transformations and some articulated deformations and are robust to geometrical and topological noise. They are robust to the level of tessellation of 3D models. They handle any type of 3D shape representations as long as depth images can be rendered. The important deviation from the state-of-the-art is that they rely on compact dictionaries (256 or 512 visual words) but achieve significantly better performance than classical BoF methods.

3.1. Dense feature extraction

Prior to feature extraction, we need to proceed to a robust normalization of pose and scale of the 3D objects in order to remain invariant to two geometrical transformations (translation, scaling). Here, we do not perform the pose normalization for rotation because the locations of local features are completely ignored in our method. For the center and the scale, we use the smallest enclosing sphere [LGS10]. The use of the smallest enclosing sphere has several advantages: it is fast to compute, it allows the maximization of the model size inside the unit sphere. We represent a 3D object with a set of depth maps captured by virtual cameras distributed uniformly around the object, see Figure 2. In order to capture all the important features of the object, we use a large number of views (80 in our implementation) and capture depth images of size 256 × 256. From the depth images, we extract a dense set of HoG descriptors on a dense regular grid. Every two pixels we compute one HoG descriptor at four different scales (16×16, 24×24, 32×32 and 40×40 pixels). We then obtain a large unordered set of local descriptors.

Let $\mathbf{B}_i = \{\mathbf{b}_{ir}, r = 1 \dots n\}$ be the set of descriptors extracted from the depth images of model i. Unlike [LGS10], we do not maintain a separate set of descriptors for each view. Instead, we put all the descriptors in a single bag and use them to build our compact dictionary. This makes the descriptor invariant to rotations of the 3D model. Also, given a sufficiently large number of views and a densely sampled HoG descriptors, we insure that the most important features of an object are captured by the representation. One method to map the set of descriptors into a single vector that can be used for indexation is the Bag of Features (BoF) [SZ03]. It involves the construction of a visual codebook (visual words) and the count of occurrences of these words in the 3D model. Applying in a straightforward manner the traditional BoF paradigm to 3D models that are represented with a large and dense set of descriptors will result in large dictionaries. We propose to use Vectors of Locally Aggregated Tensors (VLAT) recently proposed in [NPG12a] for image

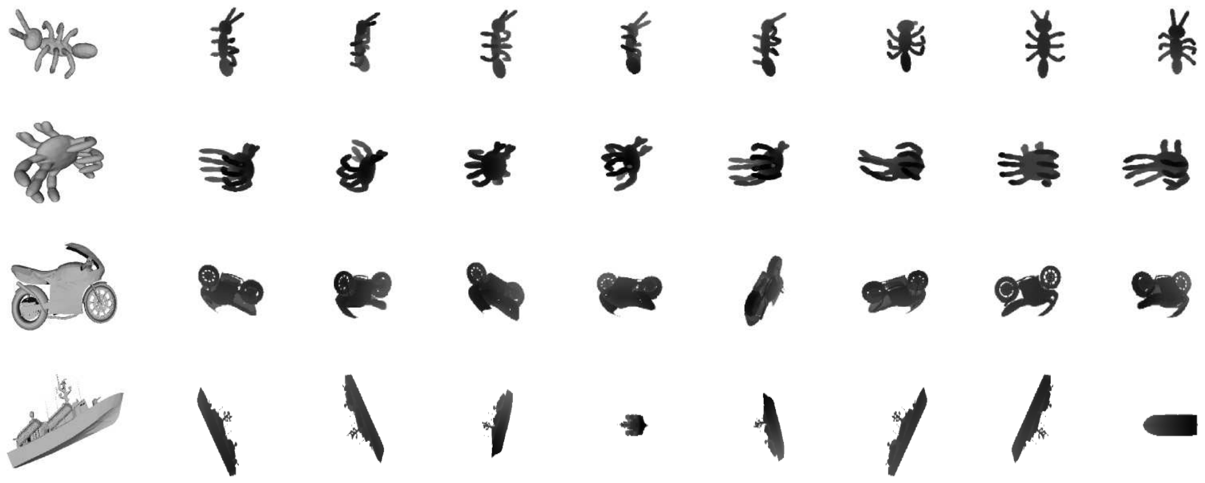

Figure 2: *Example of 3D models from the McGill (models in the two first rows) and the Shrec12 (models in the two other rows) data-sets. The first column shows renderings of the models, while the other columns show depth images extracted randomly from the corresponding models.*

analysis to build compact yet very discriminative shape signatures using a small dictionary size.

3.2. VLAT: Tensor based aggregation

Let $\mathbf{B}_i = \{\mathbf{b}_{ir}\}$ and $\mathbf{B}_j = \{\mathbf{b}_{jr}\}, r = 1 \ldots n$, be two bags of features representing the sets of descriptors in two 3D shapes i and j. An effective method to compute the similarity between two bags is based on kernel functions. Thanks to mathematical properties like Mercer conditions, these kernel functions on bags can be used with many powerful kernel-based learning techniques. The novelty in this paper is that we consider a kernel function on bags for each cluster c:

$$K(\mathbf{B}_i, \mathbf{B}_j) = \sum_c K_B(\mathbf{B}_{ic}, \mathbf{B}_{jc}) \qquad (1)$$

with $\mathbf{B}_{ic} = \{\mathbf{b}_{icr}\}_r$ the descriptors of model i that belong to cluster c. K_B is a kernel on the bags $\mathbf{B}_{ic}$ and $\mathbf{B}_{jc}$, which we define as the sum of Gaussian kernels on each pair of descriptors belonging to $\mathbf{B}_{ic}$ and $\mathbf{B}_{jc}$.

$$K_B(\mathbf{B}_{ic}, \mathbf{B}_{jc}) = \sum_{r,s} e^{-\frac{1}{\sigma^2}||\mathbf{b}_{icr} - \mathbf{b}_{jcs}||^2} \qquad (2)$$

Computing a Gaussian kernel on each pair of descriptors is computationally prohibitive, especially for large bags ($O(n^2)$ for a dictionary of size n). Thus our second idea is to linearize the Kernel functions on bags. To do so, we first normalize all the descriptors to have a unit length. Then we expand it using Taylor series, and finally linearize the kernel function using tensor products:

$$\begin{aligned} K_B(\mathbf{B}_{ic}, \mathbf{B}_{jc}) &= e^{-\frac{2}{\sigma^2}} \sum_{r,s} e^{\frac{2\langle \mathbf{b}_{icr}, \mathbf{b}_{jcs} \rangle}{\sigma^2}} \\ &= e^{-\frac{2}{\sigma^2}} \sum_{r,s} \sum_p \frac{\alpha_p}{\sigma^{2p}} \langle \mathbf{b}_{icr}, \mathbf{b}_{jcs} \rangle^p \\ &= e^{-\frac{2}{\sigma^2}} \sum_p \frac{\alpha_p}{\sigma^{2p}} \langle \sum_r \otimes^p \mathbf{b}_{icr}, \sum_s \otimes^p \mathbf{b}_{jcs} \rangle \end{aligned}$$

with $\otimes^p \mathbf{x}$ the tensor product of order p of the vector $\mathbf{x}$.

When stopped at $p = 2$, the expansion corresponds to the second order statistics of the set $\mathbf{B}_{ic}$. To extends the range of the similarity measure, we propose to center the obtained tensors by the mean tensors of each cluster c. The mean and tensors of the cluster have been learned during the dictionary construction. We proceed as follows;

- First, we compute the mean descriptor μ_c and mean tensor matrix $\mathcal{T}_c$ of each cluster c:

$$\mu_c = \frac{1}{|c|} \sum_i \sum_r \mathbf{b}_{icr} \qquad (3)$$

$$\mathcal{T}_c = \frac{1}{|c|} \sum_i \sum_r (\mathbf{b}_{icr} - \mu_c)(\mathbf{b}_{icr} - \mu_c)^T \qquad (4)$$

  where $|c|$ is the number of descriptors in cluster c and $\mathbf{b}_{icr}$ are the descriptors of model i that belong to c. The quantities μ_c and $\mathcal{T}_c$ provide statistical summaries of the shape cluster and capture the main variabilities.
- Next, for every 3D model i we compute one feature per cluster c. We compute a tensor $\mathcal{T}_{ic}$ as an aggregation of

the centered tensors of centered descriptors:

$$\mathcal{T}_{ic} = \sum_r (\mathbf{b}_{irc} - \mu_c)(\mathbf{b}_{icr} - \mu_c)^\top - \mathcal{T}_c \tag{5}$$

where μ_c is the center of cluster c. We flatten the matrix $\mathcal{T}_{ic}$ into a feature vector $\mathbf{x}_{ic}$, the descriptor of the 3D shape i with respect to cluster c.

- Finally, we concatenate all the vectors $\mathbf{x}_{ic}$ of object i to form a single feature $\mathbf{X}_i = \{\mathbf{x}_{ic}\}, c = 1 \dots n$, where n is the size of the dictionary. $\mathbf{X}_i$ is called the VLAT signature of model i.

For best performance, we perform a normalization step of the $\mathbf{X}_i$ signature.

$$\forall j, \mathbf{X}'_i[j] = sign(\mathbf{X}_i[j]) \, |\mathbf{X}_i[j]|^\alpha, \tag{6}$$

$$\mathbf{V}_i = \frac{\mathbf{X}'_i}{||\mathbf{X}'_i||} \tag{7}$$

With $\alpha = 0.05$ typically. VLAT performs very good results in similarity search and automatic indexing of 3D objects with linear metric, but leads to large feature vectors. The size of VLAT features is $n \times d \times d$, where n is the number of clusters and d is the dimension of the descriptor.

3.3. PCA-based VLAT (PVLAT)

To reduce the dimension of the VLAT descriptor, we perform a tensor decomposition using Takagi's factorization:

$$\mathcal{T}_c = \mathcal{A}_c D_c \mathcal{A}_c^\top \tag{8}$$

Where D_c is a real non-negative diagonal matrix containing the eigenvalues of $\mathcal{T}_c$ and $\mathcal{A}_c$ is unitary composed of eigenvectors of $\mathcal{T}_c$. Then we project the centered descriptors belonging to c on the eigenvectors:

$$\mathbf{b}'_{icr} = \mathcal{A}_c^\top (\mathbf{b}_{icr} - \mu_c). \tag{9}$$

We can then deduce form equation 5 the new signature called PVLAT of the model i in cluster c as the sum of tensors of projected descriptors $\mathbf{b}'_{icr}$ belonging to cluster c, centered by D_c:

$$\mathcal{T}_{ic} = \sum_r \mathbf{b}'_{icr} \mathbf{b}'^\top_{icr} - D_c. \tag{10}$$

Similar to VLAT, we concatenate and normalize each cluster signature. The size of the final PVLAT signature depends on the number of eigenvector selected in each cluster.

3.4. Compact PCA-based VLAT (CPVLAT)

We propose to further reduce the size of the PVLAT signature while retaining its discriminative power. Given a set S of N 3D models, we compute the Gram matrix G of PVLAT signatures:

$$\mathbf{G}_{i,j} = \mathbf{V}_i^\top \mathbf{V}_j, \quad i, j \in S \tag{11}$$

Then, we compute the eigenvalues and eigenvectors of the Gram matrix G. Then, we compute a low rank approximation of G. We denote by $\mathbf{L}_t$ the matrix with the t largest eigenvalues on the diagonal:

$$\mathbf{L}_t = diag(\lambda_1 ... \lambda_t), \tag{12}$$

and we denote by Λ_t the matrix of the first t eigenvectors:

$$\mathbf{U}_t = [\Lambda_1 ... \Lambda_t], \tag{13}$$

We can then define $\mathbf{G}_t$ as an approximation of G:

$$\mathbf{G}_t = \mathbf{U}_t \mathbf{L}_t \mathbf{U}_t^\top \tag{14}$$

Then, we compute the projectors of PVLAT signatures in approximated subspace:

$$\mathbf{P}_t = \mathbf{V} \mathbf{U}_t \mathbf{L}_t^{-1/2} \tag{15}$$

with $\mathbf{V} = [\mathbf{V}_1 ... \mathbf{V}_N]$ is the matrix of PVLAT signatures. This method is analogous to Kernel-PCA using a dot product Kernel. For each 3D model, we compute the projection of PVLAT in the sub-space as:

$$\mathbf{Y}_i = \mathbf{P}_t^\top \mathbf{V}_i \tag{16}$$

$\mathbf{Y}_i$ contains an approximate and low dimensional version of $\mathbf{V}_i$. The subspace defined by the projectors preserves most of the similarity even for very small dimension because the PVLAT optimization has concentrated the information in a small number of dimensions. One can notice that this procedure is analogous to that of a kernel PCA with a linear kernel.

4. Experiments and results

In our method implementation, for each 3D shape in the dataset, we capture a set of depth maps from different view points. To generate this set of depth maps for a model, we create 2D projections from multiple viewpoints. These viewpoints are equally spaced on the unit sphere. In our current implementation, we use 80 depth maps. Actually, each model was normalized for size by rescaling it so that the average Euclidean distance from points on its surface to the center of mass is 0.5. Then, all models were normalized for translation by moving their center of mass to the origin. Then, for each depth map, we extract HOG features on a dense grid, one HOG feature every two pixels, with four different scales 16×16, 24×24, 32×32 and 40×40 pixels. For dictionary construction, we use the K-Means Algorithm. To evaluate our method, we used two different 3D model databases. The first one is the SHREC 2012 Generic 3D Shape Retrieval benchmark [LGA*12] which is a standard shape benchmark widely used in shape retrieval community. The dataset contains 1200 three-dimensional models, classified into 60 object categories based mainly on visual similarity. We also present our results on a non-rigid database, the Mcgill dataset provided by Siddiqi et al. [SZM*08] for articulated 3-D models. It contains 255 objects divided into ten classes (Ant, Crabs, Hands, Humans, Octopuses, Pliers, Snakes, spectacles, Spiders and Teddy). Each class contains similar 3D shapes under a variety of poses. We conducted four different experiments. In the first one (Table 1), we evaluate the performance of the three proposed signatures and analyze the effect of the dictionary size on their performances. In the second experiment (Table 2), we compare the performance of the proposed signatures to standard bag of features ones. In the Third experiment, we compare our methods to state-of-the-art methods benchmarked in SHREC 2012 [LGA*12]. Finally, we investigate the robustness of the proposed PVLAT signature against non-rigid transformations. The first three experiments were conducted on the SHREC 2012 [LGA*12] dataset. In the last experiment we describe results on the McGill dataset. To objectively evaluate our method we use a statistical tool provided by the SHREC 2012 and the McGill datasets to compare 3-D retrieval methods. Given a classification and a distance matrix computed with any shape matching algorithm, a suite of tools produces statistics and visualizations that facilitate the evaluation of the match results. It includes five evaluation metrics, namely: the Nearest Neighbor (NN), First-tier (1-Tier), Second-tier (2-Tier), E-Measures and Discounted Cumulative Gain (DCG). We also report the precision-recall graph.

4.1. Generic shapes

We first show the contribution of the VLAT, PVLAT and CPVLAT signatures to retrieval performance on generic 3D shapes from SHREC2012 [LGA*12]. From Table 1, we see that our signatures have the same behaviour with dictionary size change. Proposed signatures are fairly stable when moving from 256 to 512 visual words. We can also see from Table 1 that the PVLAT based method performs the best, followed by the PCVLAT and the VLAT one. We observe a gain of 5.75% between VLAT and PVLAT which highlights the improvements brought by the signature optimization. Note that the size of the VLAT feature is $n \times d \times d$, with n the number of clusters and d the size of descriptors. That means, when $d = 128$ and $n = 512$, the size of the VLAT signature is $8M$. When using the PVLAT, the size of the signature is reduced according to the number of eigenvector selected in each cluster. For the CPVLAT, the size of the signature depends on the low rank approximation of the Gram matrix. Here, we drastically reduce its dimension by keeping only 128 dimensions. So that, the final CPVLAT size is set to 128. Table 2 shows the performance of our method compared to standard BoF approach. We can clearly see that our method achieves significantly better retrieval performance. With only a dictionary of size 512 our method achieves 0.83 in the Nearest Neighbour measure against 0.68 for a standard BoF with a dictionary of 16384 visual words.

Table 3 shows a comparative evaluation of our method and five other methods presented in [LGA*12]. As one can see our method performs better than LSD-sum, 3DSP and ZFDR. The DG1SIFT performs the best, followed by DVD+DB. Note that DG1SIFT combines three different methods for descriptor sampling: dense regular sampling, random sampling, and one global SIFT descriptor. DG1SIFT also uses the BoF paradigm, with a vocabulary size exceeding 13k, which is much higher than the approach we propose in this paper. These experiments show that our approach achieves good retrieval performances on standard 3D databases, while using a compact signature with few hundreds of dimensions. This makes our approach suitable for retrieval in web-scale databases. Figure 3 presents the Precision vs Recall plots of our method and the state of the art methods from [LGA*12]. It is interesting to notice that our method and the ZFDR one present quite comparable performances, however our method's precision is slightly higher for low recall values. Our method clearly outperforms LSD-sum and 3DSP methods.

4.2. Articulated Shapes

In order to evaluate the performance of our 3D shape retrieval approach on non-rigid 3D models, we use the McGill Articulated Shape Benchmark database. We have compared the performance of the PVLAT method with three recent algorithms on the McGill Database: Hybrid BoW approach from [Lav12], the graph-based approach from Agathos et al. [APP*09] and the hybrid 2D/3D approach from Papadakis et al. [PPT*08]. Table 4 presents the retrieval performance of the three algorithms compared to our PVLAT method. From this table, one can notice that the graph-based algorithm [APP*09] performs the best. This is because of the robustness of the structural representation used in the algorithm against the articulation deformations. However, not only the computation complexity of graph matching algorithms, graph based representation have a limited discriminating power when searching for generic shapes, because only topology is taken into account. Moreover, for graph based methods, minor changes in topology may result in significant differences in similarity. So that, they cannot be applied to arbitrary meshes, because topological problems like holes disturb the computation of the effective structure of the objects. Although the PVLAT method does not con-

Methods	Dictionary size (n)	NN	1-Tier	2-Tier	e-Measure	DCG
VLAT	**256**	**0.7733**	**0.4088**	**0.5021**	**0.3467**	**0.7017**
VLAT	**512**	**0.7767**	**0.4074**	**0.5158**	**0.3654**	**0.7173**
PVLAT	**256**	**0.8125**	**0.4647**	**0.5953**	**0.4207**	**0.7661**
PVLAT	**512**	**0.8342**	**0.4844**	**0.6234**	**0.4421**	**0.7807**
CPVLAT	**256**	**0.8125**	**0.3474**	**0.4253**	**0.3017**	**0.6761**
CPVLAT	**512**	**0.8133**	**0.3628**	**0.4323**	**0.3105**	**0.6821**

Table 1: *Proposed method performances with respect to the dictionary size.*

Methods	Dictionary size (n)	NN	1-Tier	2-Tier	e-Measure	DCG
BOF	256	0.3267	0.1591	0.2297	0.1560	0.4722
BOF	512	0.3392	0.1643	0.2431	0.1647	0.4814
BOF	4096	0.5908	0.2873	0.3889	0.2706	0.6109
BOF	16384	0.6858	0.3436	0.4637	0.3240	0.6688
VLAT	**512**	**0.7767**	**0.4074**	**0.5158**	**0.3654**	**0.7173**
PVLAT	**512**	**0.8342**	**0.4844**	**0.6234**	**0.4421**	**0.7807**
CPVLAT	**512**	**0.8133**	**0.3628**	**0.4323**	**0.3105**	**0.6821**

Table 2: *Comparison with Bag of Features.*

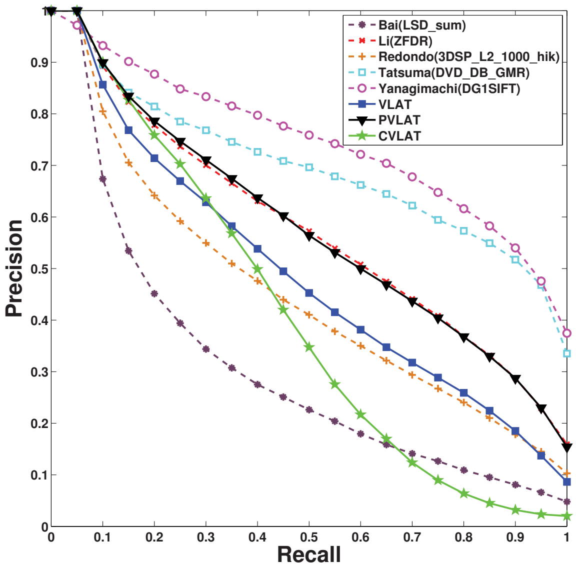

Figure 3: *Precision recall graph for our approach on the SHREC 2012 dataset compared to state-of-the-art graphs.*

sider the structural information of the shape, it obtains very comparable results to the graph-based algorithm and slightly better than the two other algorithms. The robustness of the PVLAT under non-rigid shapes can be probably due to the local HOG feature's robustness against the small isometric transformations.

5. Conclusion

We have presented a novel approach to generic 3D object retrieval using feature vectors constructed from local descriptors of object depth maps. In our approach, the descriptors are aggregated using Vectors of Locally Aggregated Tensors technique which showed to be a good approximation of insightful similarity measures between descriptors. We further reduce their size using Principal Component Analysis (PCA) on the VLAT vectors, while improving further their discrimination power. We also propose to increase the compactness of the reduced-size signature by projection in a well chosen sub-dimensional space. This subspace is obtained by a low rank approximation of the Gram matrix on a training set. The approach has been evaluated on two standard 3D shape retrieval benchmarks, demonstrating the method is suitable for generic shapes and produces very high retrieval accuracy even with presence of non-rigid transformations. A lot of future work remains. As the proposed signatures are very powerful and highly compacted, we plan to test it on web-scale datasets.

Acknowledgements

This work has been partly supported by the Culture3DCloud project. Hamid Laga is supported by the South Australian State Government through its Premiers Science and Research Fund.

References

[APP*09] Agathos A., Pratikakis I., Papadakis P., Perantonis S. J., Azariadis P. N., Sapidis N. S.: Retrieval of 3d articulated objects using a graph-based representation. In *3DOR* (2009), Eurographics Association, pp. 29–36. 6, 8

[BBGO11] Bronstein A. M., Bronstein M. M., Guibas L. J., Ovsjanikov M.: Shape google: Geometric words and expressions for invariant shape retrieval. *ACM Trans. Graph. 30* (2011). 1, 2

[CLVZ11] Chatfield K., Lempitsky V., Vedaldi A., Zisserman A.: The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC* (2011), vol. 76, pp. 1–12. 2

Methods	NN	1-Tier	2-Tier	e-Measure	DCG
DG1SIFT [LGA*12]	0.879	0.661	0.799	0.576	0.871
PVLAT	**0.8342**	**0.4844**	**0.6234**	**0.4421**	**0.7807**
DVD+DB [LGA*12]	0.831	0.496	0.634	0.450	0.785
ZFDR [LGA*12]	0.818	0.491	0.621	0.442	0.776
CPVLAT	**0.8133**	**0.3628**	**0.4323**	**0.3105**	**0.6821**
VLAT	**0.7767**	**0.4074**	**0.5158**	**0.3654**	**0.7173**
3DSP_L3_200_hik [LGA*12]	0.708	0.361	0.481	0.335	0.679
LSD-sum [LGA*12]	0.517	0.232	0.327	0.224	0.565

Table 3: *Comparison with state-of-the-art methods as reported in the shape retrieval context SHREC 2012 [LGA*12] .*

Methods	NN	1-Tier	2-Tier	DCG
Graph-based algorithm [APP*09]	0.976	0.741	0.911	0.933
PVLAT	**0.969**	**0.658**	**0.781**	**0.894**
Hybrid BoW algorithm [Lav12]	0.957	0.635	0.790	0.886
Hybrid 2D/3D algorithm [PPT*08]	0.925	0.557	0.698	0.850

Table 4: *Results on McGill dataset.*

[DT05] Dalal N., Triggs B.: Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01* (2005), CVPR '05, pp. 886–893. 1

[JPD*12] Jegou H., Perronnin F., Douze M., Sanchez J., Perez P., Schmid C.: Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell. 34*, 9 (2012), 1704–1716. 1, 2

[Lav12] Lavoué G.: Combination of bag-of-words descriptors for robust partial shape retrieval. *The Visual Computer 28*, 9 (2012), 931–942. 1, 2, 6, 8

[LG09] Li X., Godil A.: Exploring the bag-of-words method for 3d shape retrieval. In *Proceedings of the 16th IEEE international conference on Image processing* (2009), ICIP'09, IEEE Press, pp. 437–440. 1, 2

[LGA*12] Li B., Godil A., Aono M., Bai X., Furuya T., Li L., López-Sastre R. J., Johan H., Ohbuchi R., Redondo-Cabrera C., Tatsuma A., Yanagimachi T., Zhang S.: Shrec'12 track: Generic 3d shape retrieval. In *3DOR* (2012), Eurographics Association, pp. 119–126. 3, 6, 8

[LGS10] Lian Z., Godil A., Sun X.: Visual similarity based 3d shape retrieval using bag-of-features. In *Proceedings of the 2010 Shape Modeling International Conference* (Washington, DC, USA, 2010). 1, 2, 3

[Low04] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (Nov. 2004), 91–110. 1

[LZQ06] Liu Y., Zha H., Qin H.: Shape topics: A compact representation and new algorithms for 3d partial shape retrieval. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (2006), CVPR '06, IEEE Computer Society, pp. 2025–2032. 2

[MS05] Mikolajczyk K., Schmid C.: A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 10 (Oct. 2005), 1615–1630. 1

[NPG12a] Negrel R., Picard D., Gosselin P.: Compact tensor based image representation for similarity search. In *ICIP* (Orlando, Florida, USA, September 2012). 2, 3

[NPG12b] Negrel R., Picard D., Gosselin P.: Using spatial pyramids with compacted vlat for image categorization. In *ICPR* (Tsukuba Science City, Japan, November 2012). 2

[OOFB08] Ohbuchi R., Osada K., Furuya T., Banno T.: Salient local visual features for shape-based 3d model retrieval. In *Shape Modeling International* (2008), pp. 93–102. 1, 2

[PPT*08] Papadakis P., Pratikakis I., Theoharis T., Passalis G., Perantonis S.: 3d object retrieval using an efficient and compact hybrid shape descriptor. In *3DOR* (2008), Eurographics Association. 6, 8

[PSM10] Perronnin F., Sánchez J., Mensink T.: Improving the fisher kernel for large-scale image classification. In *ECCV* (2010), pp. 143–156. 2

[SZ03] Sivic J., Zisserman A.: Video google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV '03* (Washington, DC, USA, 2003), pp. 1470–. 1, 2, 3

[SZM*08] Siddiqi K., Zhang J., Macrini D., Shokoufandeh A., Bouix S., Dickinson S.: Retrieving articulated 3-d models using medial surfaces. *Mach. Vision Appl. 19*, 4 (2008), 261–275. 3, 6

[TCF10] Toldo R., Castelllani U., Fusiello A.: The bag of words approach for retrieval and categorization of 3D objects. *The Visual Computer 26*, 10 (2010), 1257–1268. 1, 2

[TDCV12] Tabia H., Daoudi M., Colot O., Vandeborre J.-P.: Three-dimensional object retrieval based on vector quantization of invariant descriptors. *SPIE Journal of Electronic Imaging 21*, 2 (April-June 2012), 023011–1–023011–8. 2

[TDVC10] Tabia H., Daoudi M., Vandeborre J.-P., Colot O.: Local visual patch for 3D shape retrieval. In *ACM International Workshop on 3D Object Retrieval (in conjunction with ACM Multimedia 2010)* (Firenze, Italy, October 25 2010). 1

[TV08] Tangelder J. W. H., Veltkamp R. C.: A survey of content based 3d shape retrieval methods. In *Multimedia Tools and Applications* (2008), pp. 441–471. 1

[WYY*10] Wang J., Yang J., Yu K., Lv F., Huang T., Gong Y.: Locality-constrained linear coding for image classification. In *CVPR* (2010), pp. 3360–3367. 2