Evaluating 3D Spatial Pyramids for Classifying 3D Shapes

R. J. López-Sastre, A. García-Fuertes, C. Redondo-Cabrera, F. J. Acevedo-Rodríguez, S. Maldonado-Bascón

GRAM, Department of Signal Theory and Communications. University of Alcalá. Spain

Abstract

This paper focuses on the problem of 3D shape categorization. For a given set of training 3D shapes, a 3D shape recognition system must be able to predict the class label for a test 3D shape. We introduce a novel discriminative approach for recognizing 3D shape categories which is based on a 3D Spatial Pyramid (3DSP) decomposition. 3D local descriptors computed on the 3D shapes have to be extracted, to be then quantized in order to build a 3D visual vocabulary for characterizing the shapes. Our approach repeatedly subdivides a cube inscribed in the 3D shape, and computes a weighted sum of histogram of visual word occurrences at increasingly fine sub-volumes. Additionally, we integrate this pyramidal representation with different types of kernels, such as the Histogram Intersection Kernel and the extended Gaussian Kernel with χ^2 distance. Finally, we perform a thorough evaluation on different publicly available datasets, defining an elaborate experimental setup to be used for establishing further comparisons among different 3D shape categorization methods.

Keywords:
3D shape recognition, 3D Spatial Pyramids, 3D SURF descriptors

1. Introduction

3D shape classification is a fundamental task to access existing 3D models on the level of object categories. This is specially important, if we take into account that the number of 3D models is growing rapidly, due to the fast evolution in both graphics hardware and software for 3D model acquisition and manipulation (*e.g.* [1, 2, 3, 4]).

Recently, a novel approach, the 3D Spatial Pyramid Matching Kernel (3DSPMK) [5], has been introduced for object recognition in point clouds. Inspired by this work, we extend the original approach to be used in the context of category-level 3D shape recognition. First, we generalize the formulation of the 3DSPMK to arbitrary kernels, note that in [5] only the Histogram Intersection Kernel (HIK) is considered. This way, we propose a holistic representation for 3D shapes defining a general 3D Spatial Pyramid (3DSP) decomposition which can be used with multiple kernels, such as the extended Gaussian Kernel with the χ^2 distance. Note these kernels have shown promising results in image categorization [6].

We formulate a discriminative approach for recognizing 3D shape categories which is depicted in Figure 1. We start extracting 3D local descriptors(*e.g.* 3D SURF [7] descriptors) from 3D shapes. These descriptors are then quantized, *e.g.* using K-means, so as to obtain a 3D visual vocabulary. Essentially, we build a Bag-of-Words (BoW) representation [8, 9], which is a popular strategy for representing images, within the context of image categorization. Therefore, this visual vocabulary is used to represent the shapes following a BoW approach. The 3DSP repeatedly subdivides a cube inscribed in the 3D shape, and computes a weighted sum of histogram of visual word occurrences at increasingly fine sub-volumes. Selective volume decomposition strategies are used, as in [5], which drastically reduce the volume to consider, while the performance does not decrease.

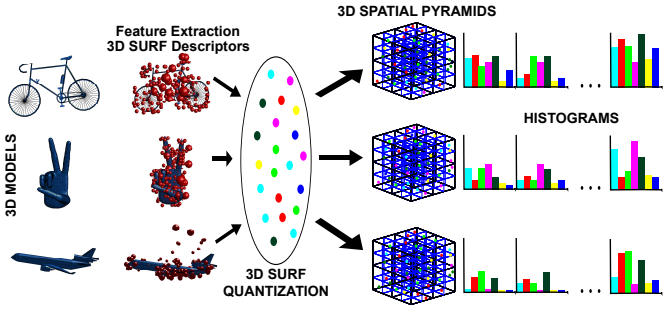


Figure 1: Proposed approach using a 3D Spatial Pyramid (3DSP) for 3D shape class recognition. We quantize 3D local descriptors, extracted from 3D shapes, into 3D visual words. This codebook is used to represent the shapes in a BoW approach. The 3DSP repeatedly subdivides a cube inscribed in the 3D shape, and computes a weighted sum of histogram at increasingly fine sub-volumes.

In order to offer to the research community a clear benchmark for establishing further comparisons among different 3D shape categorization methods, we also propose an elaborate experimental setup using different publicly available datasets (SHREC'12 [10], Princeton Shape Benchmark [11], TOSCA [12] and Sumner[13]). We perform a thorough evaluation of our novel approach on this experimental setup.

The rest of the paper is organized as follows. Section 2 describes related work. The 3DSP is detailed in Section 3. The

experimental setup and results are presented in Sections 4 and 5, respectively. We conclude in Section 6.

2. Related Work

The problem of 3D shape class recognition has been extensively explored in the literature, and both local and global features have been proposed. A considerable variety of global descriptors have been detailed, such as the shape moments [14] or the shape histograms [15]. However, neither partial shapes, nor intra-class variations are successfully handled by global descriptions.

In the 2D case, it is well-known that the use of local features is beneficial for the object recognition problem. In the literature, there are also 3D shape categorization methods using local features. Local 3D features can be extracted directly from the 3D volume (voxels) (*e.g.* [7, 16, 17, 18]) or from 2D surfaces embedded in the 3D space (3D mesh) (*e.g.* [19, 20, 21, 22]). Within the first group, scale-dependent and scale-invariant local 3D shape descriptors are proposed in [16], variants of SIFT [23] and SURF[24] are introduced in [17] and [7] respectively, and a localized version of the volumetric feature SHD [25] is proposed in [18]. Mian *et al.* introduce the use of local tensors [26]. Additionally, we also find works where the descriptors are extracted from range data, *e.g.* [27] where 3D shape context descriptors are extracted in 3D from the point cloud which emerges from the depth image.

Knopp *et al.* [7] introduce the 3D SURF descriptors in combination with a probabilistic Hough voting framework for the purpose of 3D shape class recognition. Our approach significantly differs from [7]. First, their model does not introduce any 3D pyramid representation for the shape. Second, instead of providing a discriminative approach with a SVM framework, a generative approach inspired by the Implicit Shape Model [28] is presented.

A BoW for 3D shape categorization can be found in [29]. Toldo *et al.* [29] describe 3D shapes by splitting them into segments, which are then described on the basis of their curvature characteristics. These novel descriptors attached to the regions are then vector-quantized into multiple visual vocabularies. For each shape a BoW representation per codebook is build, and multiple SVMs are used for classification. The main differences between our approach an [29] are the following. In [29] a standard BoW characterization approach is used in conjunction with the HIK for a classification with SVMs. That is, the approach in [29] does not build any 3D spatial pyramid representation which is able to enrich the BoW representation with coarse-grained geometric cues. Furthermore, instead of using just a single visual codebook, in [29] up to 108 different visual vocabularies are needed for the categorization of each particular 3D shape.

The method closest to ours is that of Redondo-Cabrera *et al.* [5]. They introduce the 3DSPMK, using only the HIK kernel, and for the particular problem of recognizing objects in depth images. However, we proceed to extend this approach to the problem of 3D shape recognition. Moreover, instead of just using the HIK kernel, we generalize the formulation to define a 3D spatial pyramid decomposition which can be integrated with different type of kernels in a discriminative approach using a SVM framework. This way, we are able to combine the 3D spatial pyramid with an extended Gaussian kernel using χ^2 distances. Our results confirm the convenience of this extension so as to increase the shape class recognition performance.

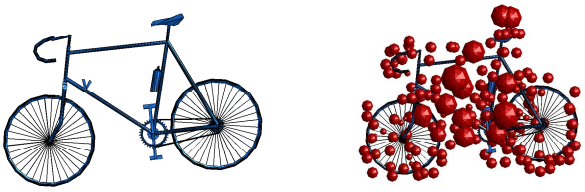


Figure 2: Extraction of 3D SURF descriptors from a 3D shape.

3. Categorizing 3D Shapes

3.1. 3D Shape Class Representation

For the 3DSP, we propose to characterize each 3D shape using local features. As it is shown in Figure 1, the approach starts from a 3D shape of the object of interest. Each shape is characterized by a set of 3D local descriptors, *e.g.* 3D SURF descriptors [7]. Figure 2 shows an example of extraction of 3D SURF descriptors from a 3D shape. In contrast to a random or dense coverage of the shape with spin images [19], the 3D SURF is equipped with a 3D interest point detector, which picks out a repeatable and salient set of interest points in the shapes. The local 3D SURF descriptors are computed in these points via uniformly sampling Haar-wavelet responses. Then, by following a traditional BoW approach, we quantize these 3D descriptors, into 3D visual words. Finally, each 3D shape can be characterized by a histogram of its 3D visual words.

3.2. Categorizing 3D Shapes with the 3D Spatial Pyramid

We proceed to generalize the formulation of the 3DSPMK introduced in [5]. Let us assume we model a 3D shape $\mathcal{S}$ by an orderless set of 3D visual words. That is, if we define a visual codebook C of size K, each 3D feature is associated to a codebook label $\{1, \ldots, K\}$. We could characterize each shape $\mathcal{S}$ with a histogram $H(\mathcal{S})$ quantizing the occurrences of the 3D visual words.

However, the 3DSP representation should be able to capture the spatial distribution of such labels at different scales and locations in a working volume $\Omega^{(0)}$. Therefore, we define a pyramid structure by partitioning $\Omega^{(0)}$ into fine sub-cubes (see Figure 3). For each level l of the pyramid, the volume of the previous level, $\Omega^{(l-1)}$, is decomposed into eight sub-cubes, hence a pyramid $P(L)$ of L levels contains $D = 8^L$ sub-cubes.

Before building the spatial pyramid representation, and in order to achieve a spatial distribution of 3D visual words that occupies the greatest possible proportion of working volume $\Omega^{(0)}$, we perform a centering and scaling process of the initial

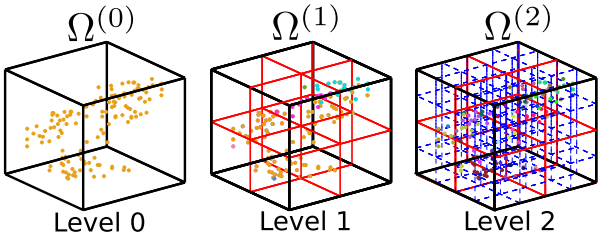


Figure 3: Example of a 3DSP of three levels. The working cube $\Omega^{(0)}$ is recursively decomposed into eight sub-cubes. The dots represent the positions where the local features have been extracted for a particular 3D shape.

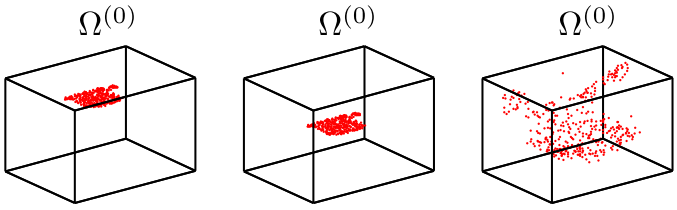


Figure 4: Example of centering and scaling process of a spatial distribution of codewords extracted from a 3D shape. In the first $\Omega^{(0)}$ cube, the initial spatial distribution of 3D visual words is represented. Second $\Omega^{(0)}$ cube shows a centered spatial distribution of codewords, this spatial distribution is then scaled to fit the $\Omega^{(0)}$ cube. The final results can be observed in the third $\Omega^{(0)}$ cube.

spatial distribution of 3D visual words. This process is detailed in Figure 4.

Once a pyramid $P(L)$ is composed, we define the 3DSP representation of a particular 3D shape $\mathcal{S}$ by a weighted ensemble of histograms $H(\mathcal{S})$ as follows,

$$H(\mathcal{S}) = \left[\omega_0 H^0(\mathcal{S}), \omega_1 H^1(\mathcal{S}), ..., \omega_L H^L(\mathcal{S})\right], \quad (1)$$

where $H^l(\mathcal{S})$ is the histogram of the features in the level l of the pyramid. Each $H^l(\mathcal{S})$ is obtained by concatenating 8^l histograms computed in all of the 8^l sub-cubes for level l. In order to penalize the future matches (between histogram bins) found in larger volumes, we define the weight ω_l as

$$\omega_l = \frac{1}{2^{L-l}}. \quad (2)$$

Equation (1) contains the general formulation of the 3DSP representation for any shape. In order to use the 3DSP representation in a discriminative approach, we can incorporate different kernels into the formulation. This way, based on the fundamental concept of defining similarities between objects, these representations allow the integration of the 3DSP in a SVM classifier, for example. In particular, we propose to incorporate two different kernels: the Histogram Intersection Kernel (HIK) and the extended Gaussian kernel with χ^2 distances.

The 3DSP-HIK kernel $K_{3DSP-HIK}$ is formulated as follows. When a pyramid decomposition $P(L)$ is constructed, we are able to perform a pyramid matching in 3D of two 3DSP $H(\mathcal{S}_X)$ and $H(\mathcal{S}_Y)$, computed for shapes $\mathcal{S}_X$ and $\mathcal{S}_Y$. The 3DSP-HIK kernel is defined as

$$K_{3DSP-HIK}(H(\mathcal{S}_X), H(\mathcal{S}_Y)) = \sum_{i=1}^{N} \min(H(\mathcal{S}_X)_i, H(\mathcal{S}_Y)_i), \quad (3)$$

where N is the number of components of histograms $H(\mathcal{S}_X)$ and $H(\mathcal{S}_Y)$, and $H(\mathcal{S}_X)_i$ represents the value of the i-th bin of the histogram.

Additionally, we can formulate the 3DSP-χ^2 kernel $K_{3DSP-\chi^2}$. Starting from a pyramid decomposition $P(L)$ and two 3DSP representations $H(\mathcal{S}_X)$ and $H(\mathcal{S}_X)$ for shapes $\mathcal{S}_X$ and $\mathcal{S}_Y$, we first define the χ^2 distance between them as

$$D_{\chi^2}(H(\mathcal{S}_X), H(\mathcal{S}_Y)) = \frac{1}{2}\sum_{i=1}^{N} \frac{(H(\mathcal{S}_X)_i - H(\mathcal{S}_Y)_i)^2}{H(\mathcal{S}_X)_i + H(\mathcal{S}_Y)_i}, \quad (4)$$

and we formulate the 3DSP-χ^2 kernel as follows,

$$K_{3DSP-\chi^2}(H(\mathcal{S}_X), H(\mathcal{S}_Y)) = \exp\left(-\tfrac{1}{A} D_{\chi^2}(H(\mathcal{S}_X), H(\mathcal{S}_Y))\right), \quad (5)$$

where A is a scalar which normalizes the distances. In the experiments, one can set A to the average χ^2 distance between all elements of the training set.

Note that the HIK and the extended Gaussian with χ^2 distances kernels satisfy the Mercer's conditions, as it has been proved in [30] and [31] respectively.

3.2.1. Selective 3DSP

The 3DSP has one clear disadvantage: its high computational cost. For a pyramid of L levels and a vocabulary of size K, we will obtain a vector of dimensionality $K\sum_{l=0}^{L} 8^l$, that is 2^l times more bins in each level with respect to to the 2D version introduced in [32]. With the aim of jointly increasing the classification accuracy and the computational efficiency of the 3DSP, we can incorporate to our approach the equivalent *selective* volume decomposition schemes based on representative and discriminative (sub-)volume selection processes detailed in [5]. The main objective of these approaches is to reduce the large number of uninformative sub-cubes that yield unnecessary long histograms, while the performance does not decrease.

We define the 3DSP-K-Repre as the 3DSP with kernel K using the Representativeness-based selection method in [5]. This selective pyramid decomposition will incorporate into the pyramid only those (sub-)cubes that are likely to represent shape classes in our dataset. Let $\Omega^{(0)}$ be the working cube for level zero. We first perform the pyramid decomposition until level L, so we obtain $\Omega_i^{(L)}$ sub-volumes, where $i = 1, \ldots, 8^L$. We now define the working volume of level zero as $\hat{\Omega}^{(0)}$, where the decomposition only includes those sub-cubes $\hat{\Omega}_i^{(L)}$ in which a percentage p of the 3D shape models are represented. We consider that a 3D shape is represented in a sub-cube if there is at least one feature for this shape falling in the sub-cube. Note that this pyramid volume selection process is performed at the beginning of the training, once all the 3D features have been

extracted. This way, the new working volume $\hat{\Omega}^{(0)}$ can be used to build all the features to represent the different shapes.

We also define the 3DSP-K-Disc as the 3DSP with kernel K using the Discriminative Feature-based Selection approach in [5]. Although the representativeness-based selective method drastically reduces the working volume, it does not exploit the fact that the sub-volume selected may contain features that are not discriminative for the classes of interest. The objective of the Discriminative Feature-based selection scheme is to select those cubes that are likely to contain discriminative features. This time, we consider all the training shapes of all the classes to compute. Given a pyramid $P(L)$, we inspect all the sub-volumes in level L, *i.e.* $\Omega_i^{(l)}$ for $i = 1, \ldots, 8^L$. For each sub-volume and each 3D shape class, we measure the proportion of shape models that contain at least one *discriminative* feature in each sub-volume. If this ratio is greater than an empirically fixed threshold, then the sub-volume $\Omega_i^{(l)}$ is considered as discriminative for the analyzed object class. The final discriminative decomposition is obtained by merging all the discriminative sub-volumes for each category. When do we consider a feature discriminative? We follow the feature score formulated in Equation (3) of [5]: the ratio between the percentage of descriptors that belong to a particular feature for a shape class, and the proportion of descriptors that belong to the same feature when all the 3D shape categories are considered. That is, we are able to measure how informative for a particular 3D shape class a feature is. Subsequently, we select only those sub-volumes that contain this type of discriminative features.

Note that the proposed approaches are feature selection methodologies that do not affect the kernel formulations proposed.

4. Experimental Setup

For the 3D shape categorization problem, we propose an elaborate experimental setup with different publicly available datasets. Our aim is to provide to the research community a clear benchmark so as to establish further comparisons among different methods. We start describing the databases, and then how the evaluation of the results is going to be performed.

4.1. The Databases

The following state-of-the-art publicly available databases are going to be used: SHREC'12 [10], Princeton [11], and TOSCA + Sumner [12, 13]. All these datasets consists of clean and segmented 3D shapes (see Figure 5).

The SHREC'12 - Generic 3D Shape Retrieval contest dataset [10] offers 1200 different 3D models distributed across 60 classes. Specifically, for each class, 10 models are used for training and 10 models for testing. We have done a random distribution of the 20 models per class in order to obtain the training and testing subsets. This distribution of data results interesting to analyze the performance of different approaches when only limited training data is available.

The challenging Princeton Shape Benchmark database [11] offers 1800 shapes of 7 classes. For the 3D shape classification experiment, we propose to use the coarse level two, with the subsets for training and testing proposed in [7]: for each class, half of the 3D shapes are used for training and half for testing.

Finally, the TOSCA [12] and the Sumner [13] databases are jointly used. This combination offers 474 shapes for a total of 12 classes. The 3D shapes appear in a variety of poses and with deformations. 66 randomly selected models are used for testing, and the rest for training, as in [7].

The proposed datasets define an experimental setup where more than 3400 3D shapes can be used for the performance evaluation of the different methods. We publicly distribute [1] this experimental setup, including: the annotations and the training and testing subsets described; and a set of tools for accessing and managing the database annotations. Our aim is to establish a new benchmark for evaluating 3D shape categorization algorithms. By making this experimental setup available, we make it effortless for future researchers to perform similar performance analysis of their methods. Furthermore, a reference implementation of the code for reproducing all the results reported in this paper is also released.

4.2. Evaluation of Results

For each database, we have clearly defined two main subsets: training and testing. Although the ground truth is offered for both subsets, the testing data must be used strictly for reporting of results alone, *i.e.* it must not be used in any way to train or tune the proposed approaches. Only the training data can be used for parameter tuning or feature selection, *e.g.* using n-fold cross-validation.

The proposed experimental setup offers a multi-class problem, and we propose two evaluations measures in order to compare different methods. First, the performance p of the classifier defined as

$$p = \frac{TP}{TP + FP}, \tag{6}$$

where TP and FP are the number of true positives and false positives, respectively. Second, we propose to compute the confusion matrix for each method, and to calculate the mean of the elements on the main diagonal, a measure we refer to as Mean Correct Classification (MCC).

5. Results

We evaluate our 3D shape categorization approach on all the dataset proposed in Section 4. In the experiments, we use a visual vocabulary of different sizes ($K = 200$, $K = 400$ and $K = 1000$). The visual vocabulary is obtained performing a K-means clustering on a subset of the 3D SURF [7] descriptors extracted from the training 3D shapes. We represent each 3D shape by a 3D spatial pyramid. Typical pyramid level values for our experiments are $L = \{0, 1, 2\}$. Note that when $L = 0$, we simply have a standard BoW, but in our case in 3D. We report the performance of the 3DSP using the full volume of pyramid

[1] The experimental setup described can be downloaded from http://agamenon.tsc.uah.es/Personales/rlopez/data/3dsr

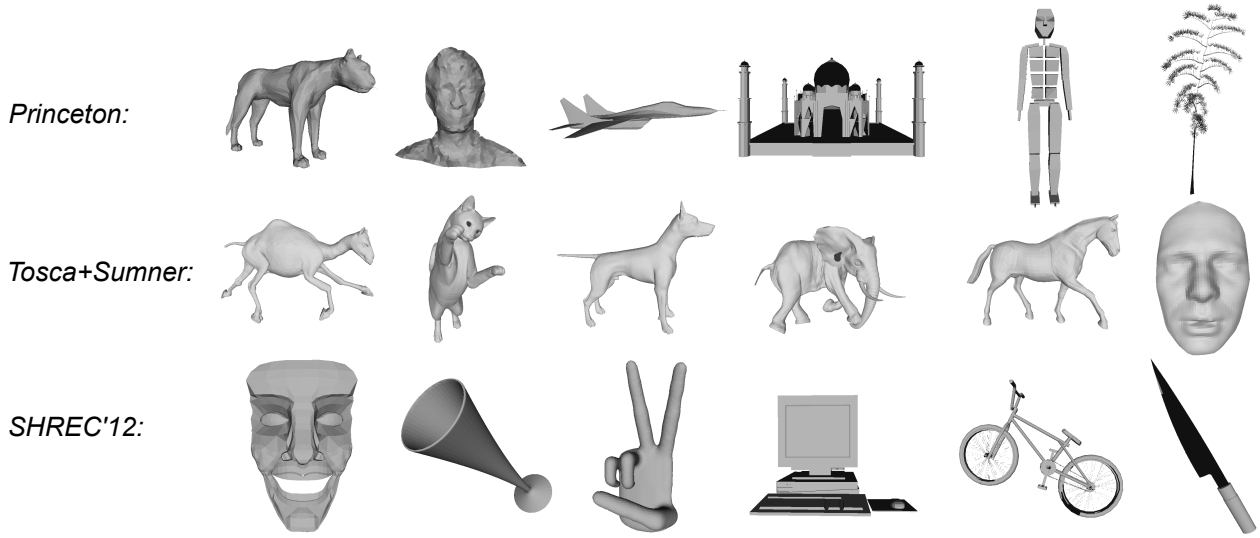


Figure 5: Samples of 3D shapes from the datasets Princeton [11], TOSCA + Sumner [12, 13] and SHREC'12 [10].

and also following the selective algorithms described in Section 3.2.1[2].

For the extraction of 3D SURF descriptors we use the original implementation provided in [7][3]. Specifically, we start scaling each 3D shape to fit a cube with a side of length 256. Then, each shape is voxelized into the cube grid using the intersection of faces with the grid-bins. With the aim of covering the full 3D shape with local descriptors, we have experimentally chosen the following parameters for the 3D SURF descriptors: the distance between triangle mesh and the border of the cube is fixed to 30, and the threshold is fixed to 10^{-8}.

For classification we use Support Vector Machines (SVMs). We explore how different kernel functions perform categorizing shapes. Specifically, we combine the 3DSP pyramid decomposition with HIK (3DSP-HIK) and χ^2 (3DSP-χ^2) kernels, which have shown promising results in image categorization [6]. The multi-class classification problem is solved training the SVM using the one-against-one strategy. We follow the approach in [33], and train $N(N-1)/2$ classifiers (being N the number of classes) where each one is trained on data from only two classes. For testing, we follow the *Max Wins* voting strategy [33]: if one of the classifiers votes for the class i, then the vote for the i-th class is added by one. The class with the highest number of votes is selected for each image. In case that two classes have identical votes, we select the one with smaller index. Specifically, we use libSVM [34] for training and testing the classifiers. A 5-fold cross-validation on the training set to tune SVM parameters is conducted.

[2]For the representativeness method, we fix the parameter p to 0.1. For the Discriminative Feature-based selection method, we fix τ and β to 0.7 and 0.5 respectively.

[3]The binaries for computing 3D SURF descriptors can be downloaded from `http://homes.esat.kuleuven.be/~jknopp/codes/index_codes.html`

5.1. SHREC'12

The results obtained by our method for the SHREC'12 data are show in Table 1. The best result is obtained for the 3DSP-χ^2, with a vocabulary size of 1000 and $L = 0$. These results reveal that, for vocabularies of size 200 or 400, the higher the level of the 3DSP, the better the results. Both the -Disc and -Repre approaches significantly reduce the computation time, while, generally, the performance does not decrease. In this experiment we can observe that the SHREC'12 is a challenging dataset due to the high number of classes and the low number of training 3D shape examples (only 10 per class). In the winner configuration, only for two classes, `Plier` and `Non Flying Insect`, we obtain a classification accuracy of 100%, and for the classes, `Door` and `Truck Non Container` the classification rate is 0%. Interestingly, when the vocabulary size is fixed to 1000, an increment in the pyramid level does not improve the classification results. Actually, the best results have been obtained by a 3DSP with $L = 0$, *i.e.* a *standard* BoW approach. As we shall see in the experimental validation with the rest of datasets, this behavior is only observed with the SHREC'12 database. We believe this may have been caused by the following reasons: first, this dataset offers a high variability in terms of rotation and changes of viewpoint of the different models, a fact that definitely does not benefit our 3DSP approach when $L > 0$ (we provide more details in Section 5.5.2); and second, the experimental setup designed for the SHREC'12 dataset is very challenging, offering just 10 examples to train each of the 60 classes.

5.2. Princeton

The results obtained by our method for the Princeton Shape Benchmark data are show in Table 2. The best result is obtained for the 3DSP-χ^2, with a vocabulary size of 1000 and $L = 1$. Again, we observe that the χ^2 kernel is obtaining the best results. Systematically, the -Repre approach is also casting better results than the -Disc based version.

Table 1: Comparison of different approaches of the 3DSP using different shape representations and different kernels on the SHREC'12 dataset, measured as MCC (%).

K	L	3DSP-HIK	3DSP-HIK-Repre	3DSP-HIK-Disc	3DSP-χ^2	3DSP-χ^2-Repre	3DSP-χ^2-Disc
200	0	63	n/a	n/a	63.33	n/a	n/a
200	1	63.7	63.7	63.7	64.33	64.33	64.33
200	2	64.7	64.7	64.7	65	65.17	64.83
400	0	62.7	n/a	n/a	64.83	n/a	n/a
400	1	62.5	62.5	62.5	63.33	63.33	63.33
400	2	63.7	64.17	63.83	64.83	61.17	61.33
1000	0	65	n/a	n/a	**65.67**	n/a	n/a
1000	1	63.83	63.83	63.83	63.83	63.83	63.83
1000	2	63.33	63.33	63.33	62.83	62.17	62

Table 2: Comparison of different approaches of the 3DSP using different shape representations and different kernels on the Princeton Shape Benchmark dataset, measured as MCC (%).

K	L	3DSP-HIK	3DSP-HIK-Repre	3DSP-HIK-Disc	3DSP-χ^2	3DSP-χ^2-Repre	3DSP-χ^2-Disc
200	0	60.11	n/a	n/a	61.43	n/a	n/a
200	1	64.11	64.11	64.11	63.30	63.30	63.30
200	2	64.22	64.22	62.13	63.17	63.17	60.62
400	0	61.92	n/a	n/a	63.35	n/a	n/a
400	1	66.01	66.01	66.01	65.74	65.74	65.74
400	2	65.65	65.65	63.19	64.57	64.57	61.50
1000	0	63.91	n/a	n/a	64.67	n/a	n/a
1000	1	65.79	65.79	65.79	**66.31**	**66.31**	**66.31**
1000	2	66.29	66.29	63.50	64.80	64.80	59.76

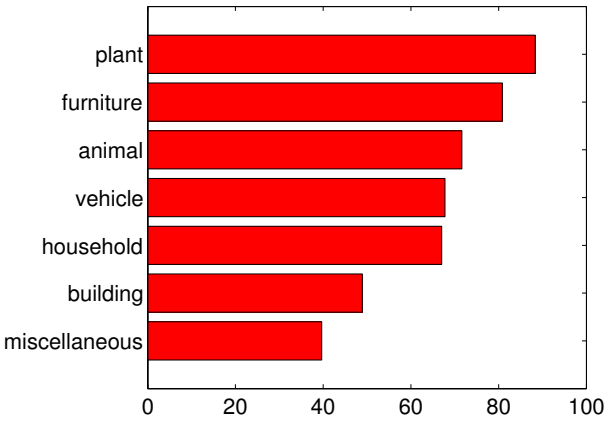


Figure 6: Classification accuracy for each class in the Princeton database. Results for the 3DSP-χ^2 with $K = 1000$ and $L = 1$.

Figures 6 and 7 show the classification accuracy and the confusion matrix, respectively, for the best approach, *i.e.* 3DSP-χ^2 for $L = 1$ and $K = 1000$. For the `Miscellaneous` class is where our approach incurs the maximum confusion, and this is due to its high variability. The best recognition performance is achieved for the classes `Plant` and `Furniture`.

The confusion matrices and graph bars for all the approaches included in Table 2 can be inspected in the Experiment Code Item 1 in the Collage Platform.

5.3. TOSCA and Sumner

The results obtained by our method for the TOSCA and Sumner databases are show in Table 3. Our best result is 95.7%, which is obtained by several parameters configurations of our method. Figures 8 and 9 show the results per class for the 3DSP-HIK with $K = 1000$ and $L = 1$. First, one can observe that for 9 classes, our method obtains a classification rate of 100%. Furthermore, for all the classes, this percentage is above 80%. The confusion matrices and graph bars for all the approaches included in Table 3 can be inspected in the Experiment Code Item 1 in the Collage Platform.

5.4. A comparison with the state-of-the-art

In Table 4 we compare our results with the results reported in [7] for 3D shape classification. The 3DSP based approach improves the state-of-the-art for the Princeton database. It is worth to mention that this dataset is very challenging, not only due to the number of shapes, but because it presents a very high variation amongst the classes (*e.g.* within the class `Animal`, the dataset provides models for ants and fishes).

For the TOSCA+Sumner dataset, our best 3DSP based approach, *i.e.* the 3DSP-HIK with $K = 1000$ and $L = 1$, is able to retrieve 63 shapes (of 66) correctly. Note that in [7], authors

Table 3: Comparison of different approaches of the 3DSP using different shape representations and different kernels on the TOSCA and Sumner dataset, measured as MCC (%).

K	L	3DSP-HIK	3DSP-HIK-Repre	3DSP-HIK-Disc	3DSP-χ^2	3DSP-χ^2-Repre	3DSP-χ^2-Disc
200	0	92.3	n/a	n/a	92.3	n/a	n/a
200	1	94.4	94.4	94.4	94.4	94.4	94.4
200	2	93	93	91.6	93	93	91.6
400	0	**95.7**	n/a	n/a	**95.7**	n/a	n/a
400	1	94.4	94.4	94.4	94.4	94.4	94.4
400	2	90.2	90.2	90.2	90.2	90.2	90.2
1000	0	**95.7**	n/a	n/a	**95.7**	n/a	n/a
1000	1	**95.7**	**95.7**	**95.7**	94.4	94.4	94.4
1000	2	93	92.4	94.4	91.6	91.6	91.6

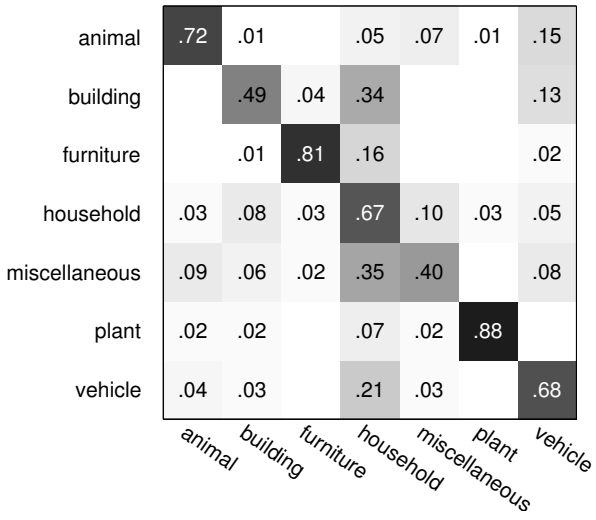


Figure 7: Confusion matrix for the 7 categories in the Princeton database. Average classification rates for individual categories are listed along the main diagonal. Results for the 3DSP-χ^2 with $K = 1000$ and $L = 2$.

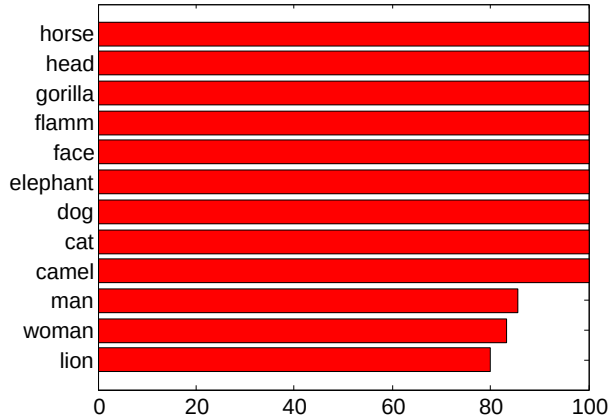


Figure 8: Classification accuracy for each class in the TOSCA+SUMNER database. Results for the 3DSP-HIK with $K = 1000$ and $L = 1$.

claim they use 66 shapes for testing, but they only report results for 57, so the results for this dataset are not comparable.

In the novel SHREC'12 dataset, and to the best of our knowledge, we are the first reporting results for generic 3D shape categorization. We achieve a performance of 63.83% for the following configuration of our approach: 3DSP-χ^2, $K = 1000$ and $L = 1$. This dataset offers a high number of classes, and the experimental setup designed provides very few shapes for training and testing, 10 per class. This makes the problem of training a SVM based approach, such as the 3DSP, really hard.

5.5. Discussion

After this thorough performance evaluation, let us discuss the most relevant aspects of the 3DSP approach within the context of 3D shape categorization.

5.5.1. Influence of the model parameters

This paper introduces a novel and holistic approach for 3D shape categorization. The 3DSP approach has shown promising results on three diverse datasets. Apart from the parameters of the feature extraction stage and the kernels, two are the parameters that completely characterize the 3DSP approach: the pyramid levels (L) and the size of the vocabulary (K).

First, let us examine the behavior of the 3DSP when L increases. For all the kernels used, and when the vocabulary is small (*e.g.* 200), the categorization results improve as we go from $L = 0$ to a multi-level pyramid structure ($L = 1$), in all the datasets. If we continue increasing the pyramid levels to $L = 2$, the results do not generally improve. Actually, for the three datasets, one can observe how the performance of the entire 3DSP remains essentially identical or even decreases. This means that the highest level of the 3DSP if too finely subdivided in subcubes (for $L = 2$ the number of subcubes is 64), which yield too few matches between the features within them. It is worth to mention that a similar behavior was observed in [32] but for the 2D spatial pyramids. To summarize, when using the 3DSP a good choice is to use $L = 1$, because: a) higher values do not always guarantee better results, and b) the computational cost for 3DSP with $L \geq 2$ increases.

In any BoW based approach the size of the visual vocabulary matters, and the 3DSP is no exception. In the experiments,

Table 4: A comparison of the performance of our approach with the state-of-the-art methods, measured as precision p.

	Princeton			TOSCA + Sumner			SHREC'12		
Method	#TP	#FP	p	#TP	#FP	p	#TP	#FP	p
ISM [7]	529	378	58.3%	56	1	98%	n/a	n/a	n/a
BOF-knn [7]	491	416	54.1%	56	1	98%	n/a	n/a	n/a
BOF-SVM [7]	472	435	52.0%	41	16	72%	n/a	n/a	n/a
3DSP	601	306	**66.26%**	63	3	95.8%	383	217	**63.83** %

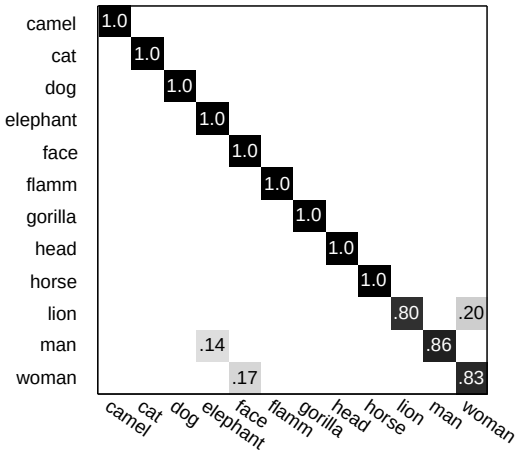


Figure 9: Confusion matrix for the 12 categories in the TOSCA+SUMNER database. Average classification rates for individual categories are listed along the main diagonal. Results for the 3DSP-HIK with $K = 1000$ and $L = 1$.

we have increased the size of the vocabulary from $K = 200$ to $K = 400$ and $K = 1000$. It is interesting to observe that increasing the size of the codebook for $L = 0$ results in a small performance increase, if we compare it with the results obtained by smaller vocabularies used with a 3DSP structure of higher levels. For instance, in the Princeton dataset, we observe that a 3DSP with $L = 2$ and $K = 200$ obtains a higher performance $(64, 22\%)$ than a simple BoW (*i.e.* 3DSP with $L = 0$) with a vocabulary of size 400 $(61, 92\%)$ or 1000 $(63, 91\%)$. In general, the geometric cues provided by the 3DSP have a similar or even greater discriminative power than an enlarged visual vocabulary. For all the datasets the best results have been obtained by the biggest vocabularies. It is worth to recall that the dimensionality of the histogram-based feature of the 3DSP increases with K and L, so the smaller these parameters, the less the computational cost of the approach.

With respect to the feature extraction and kernel parameters, we can conclude that: a) in general, the performance of the χ^2 version of the kernel is better, although the runtime for the computation of the HIK is the lowest; b) the performance of the 3DSP-K-Repre approaches is slightly better than for 3DSP-K-Disc versions. Note that these two selective approaches significantly reduce the dimensionality of the histogram-based representation, while the performance does not worsen.

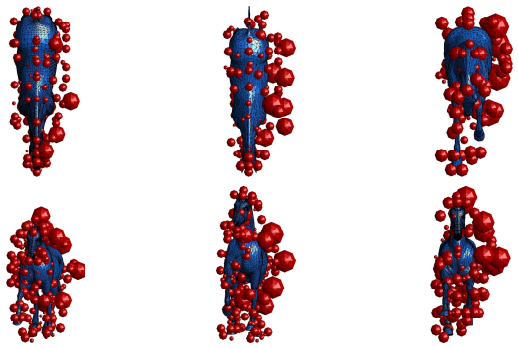


Figure 10: Examples correctly recognized for the class `horse` in the test set for the TOSCA+SUMNER dataset. Observe the deformations and changes of orientation of the different models. The 3DSP is able to correctly classify all of them correctly, even with a pyramid structure with $L = 2$.

5.5.2. Invariance to rotation and deformation

Definitely, one of the limitations of the 3DSP representation is its ability to deal with isometric transformations and deformations of the 3D shapes. It is important to analyze these aspects, because, when dealing with 3D data, the objects are rarely observed in a canonical frame of reference with respect to orientation. This is specially relevant to 3D categorization systems, where the test 3D shapes are generally given in arbitrary scale, position and orientation in 3D-space. Furthermore, these arbitrary orientations do not necessarily correspond to the orientations of the training samples.

In this section, we analyze the influence of the different parameters of the 3DSP approach on the recognition performance under rotations and deformations of the 3D shapes. For this analysis, we have decided to use the TOSCA+SUMNER dataset (this database presents a high variability in terms of both deformations and changes of orientation of the 3D models).

First, note that in the pipeline proposed for the 3DSP, we do not control the orientation of the 3D shapes given, *i.e.* the scaling and centering process shown in Figure 4 does not modify the original orientation of the shape. The 3DSP is able to capture the spatial distribution of the local features extracted from the training 3D shapes at different scales and locations in a predefined working volume. Because the 3DSP only learns the geometric cues from the training data, it has some rotational variability.

If we inspect the 3D shape categorization results in the TOSCA + SUMNER dataset we observe that the 3DSP is able to deal

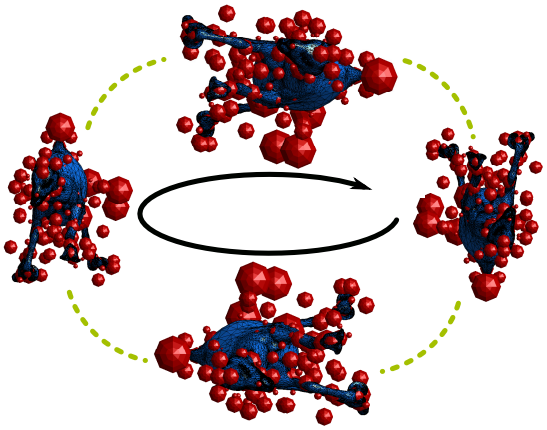


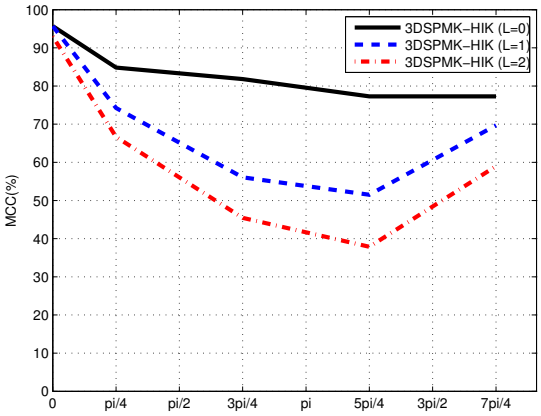
Figure 11: All the testing 3D shape models are rotated incrementally, in steps of $\frac{\pi}{4}$radians. This figure shows an example of these rotations for a `camel` shape.

quite well with the deformations and rotations of the models. For instance, as it can be seen in the confusion matrix provided in Figure 9, for the class `horse` all the testing 3D shapes are correctly recognized. Figure 10 shows all the test 3D shape for the class horse, note the changes of orientation and deformations. We explain this performance as follows.

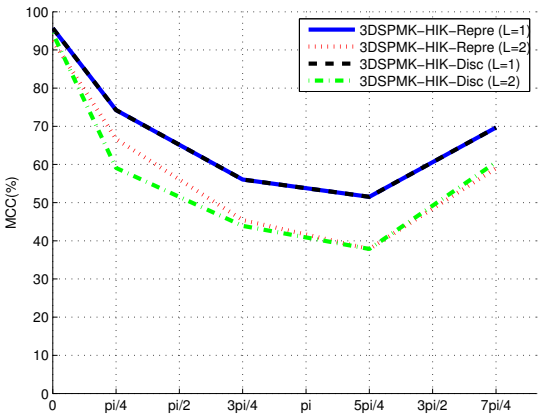
The variance of the 3DSP to rotation will specially depend on the number of levels of the pyramid structure. Essentially, when $L = 0$, our 3DSP is a standard BoW approach. Such an approach is invariant to rotation, if the local features extracted are also invariant under rotation and scale, which is the case for the 3D SURF features used. When $L > 0$ the variance to rotation can augment. First, we have to recall that the 3DSP representation is a weighted ensemble of the histograms at each of the levels of the pyramid, including $L = 0$ (see Equation 1). This means that, even for a 3DSP of $L > 0$, the representation includes the invariant to rotation histogram for level 0. Additionally, it might happen that the rotation (or deformation) is so slight that the features involved do not move to different sub-volumes within the pyramid. Furthermore, the training data might provide similar rotation and deformation configurations to the ones observed during testing. These reasons explain the results of the 3DSP model in the TOSCA+SUMNER dataset.

In order to thoroughly evaluate the rotational variability of our approach, we have performed an additional experiment. It consist of the following steps. First, we take the previously trained models on the TOSCA+SUMNER dataset with the HIK kernel, for $L = 1$ and $L = 2$, and with a vocabulary of size 1000. For all the test 3D shapes, we incrementally rotate them from 0 to $\frac{7\pi}{4}$, in steps of $\frac{\pi}{4}$ radians (see Figure 11). After each rotation, the 3D SURF descriptors are computed and the 3DSP representation is build. In Figure 12 we show the classification performance versus the change in orientation.

First, Figure 12a shows how the classification accuracy varies for 3DSP representations when no feature selection methods are used. It is interesting to observe the performance of the



(a)



(b)

Figure 12: Variation of the classification performance versus rotation variations.

configuration 3DSPMKHIK for $L = 0$, *i.e.* a standard BoW approach where no spatial pyramid is used. This configuration also shows a decrease of the performance under severe rotations of the models, which indicates that 3D SURF descriptors are not totally rotation invariant. We experimentally observe that the higher the level of the pyramid, the higher its rotational variability. Second, Figure 12b shows that the rotational variability slightly increases for the Discriminative Feature-based approach.

We can conclude that the level of the pyramid is the most significant parameter. So, as for the 2D spatial pyramid [32], the 3DSP is not fully invariant to rotations and deformations. Even if the local features used are invariant to rotation, it is important that all further steps along the 3D shape categorization pipeline are as well. As a solution, any technique for automatically aligning the 3D shapes into a canonical coordinate frame (*e.g.* [35, 36]) could be incorporated to our approach as a preprocessing stage.

5.6. Timing

The code has been written in Matlab with some parts in C. To perform the test we used an Intel Core 2 Quad CPU Q6600 @ 2.40GHz, running OS Ubuntu 12.04. The entire approach is computationally efficient. Recall that the 3DSP representation uses histogram vectors which are extremely sparse. Three are the parameters that most affect the runtime of the proposed pipeline: the vocabulary size, the pyramid levels and the type of Kernel (HIK or χ^2). We again used the TOSCA+SUMNER dataset for this evaluation of the timing information. The overall process of building and testing the 3DSP representations for the 66 test models in the TOSCA+SUMNER dataset takes the times detailed in Table 5. In general, the runtime slightly increases with the vocabulary size and the pyramid levels. The results also confirm that the HIK is more efficient than the χ^2 kernel.

5.7. Testing the 3DSP approach with my own 3D shapes

We encourage the readers to try our methods through the Collage Platform. In Experiment Code Item 2, readers are allowed to upload the 3D SURF descriptors extracted from their own 3D shapes. With these descriptors, our algorithms will estimate a shape class. We refer to Experiment Data Item 2 to know more details on how to compute the 3D SURF descriptors, and how to use them with our trained models.

6. Conclusion

In this paper, we introduced the 3DSP representation in combination with two kernel definitions (the 3DSP-HIK and the 3DSP-χ^2) for the problem of 3D shape categorization. A thorough evaluation of these kernels has been carried out, and it demonstrates the power of the classification framework proposed on state-of-the-art databases. Rather than simply releasing a set of classification results, we defined an elaborate experimental setup, which we hope will allow to establish further comparisons with other methods dealing with the challenging problem of 3D shape class recognition. Last but not least, we have released a publicly available version of all the codes and data needed to reproduce the results.

Bringing in some weak form of textured information (if available in the 3D shape) is one interesting avenue of future research that might bring us closer to our goal. One way of doing so is combining the 3DSP approach with appropriate local 3D features, which also capture information from the texture.

Acknowledgments

We want to thank the reviewers for their constructive and helpful suggestions. This work was partially supported by projects TIN2010-20845-C03-03, UAH2011/EXP-030 and IPT-2011-1366-390000.

References

[1] Adán A, Salamanca S, Merchán P. A hybrid humancomputer approach for recovering incomplete cultural heritage pieces. Computers & Graphics 2012;36:1–15.

[2] Catalano C, Mortara M, Spagnuolo M, Falcidieno B. Semantics and 3d media: Current issues and perspectives. Computers & Graphics 2011;35:869 –77.

[3] Pessoa SA, de S. Moura G, Lima JPSM, Teichrieb V, Kelner J. Rpr-sors: Real-time photorealistic rendering of synthetic objects into real scenes. Computers & Graphics 2012;36:50 – 69.

[4] Weise T, Wismer T, Leibe B, Van Gool L. Online loop closure for real-time interactive 3D scanning. Computer Vision and Image Understanding 2011;115:635 –48.

[5] Redondo-Cabrera C, Lopez-Sastre RJ, Acevedo-Rodriguez J, Maldonado-Bascon S. SURFing the point clouds: Selective 3D spatial pyramids for category-level object recognition. In: IEEE CVPR. 2012,.

[6] Zhang J, Marszalek M, Lazebnik S, Schimd C. Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV 2007;73(2):213–38.

[7] Knopp J, Prasad M, Willems G, Timofte R, Van Gool L. Hough transform and 3D SURF for robust three dimensional classification. In: Proceedings of the 11th European conference on Computer vision. 2010,.

[8] Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision. 2004,.

[9] Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos. In: ICCV. 2003,.

[10] Li B, Godil A, Aono M, Bai X, Furuya T, Li L, et al. SHREC'12 track: Generic 3d shape retrieval. In: 3DOR: Eurographics Workshop on 3D Object Retrieval. 2012,.

[11] Shilane P, Min P, Kazhdan M, Funkhouser T. The princeton shape benchmark. In: Shape Modeling International. 2004,.

[12] Bronstein A, Bronstein M, Kimmel R. Numerical Geometry of Non-Rigid Shapes. Springer; 2009.

[13] Sumner RW, Popović J. Deformation transfer for triangle meshes. ACM Transactions on Graphics 2004;23:399–405.

[14] Saupe D, Vranic DV. 3D model retrieval with spherical harmonics and moments. In: DAGM-Symposium on Pattern Recognition. 2001,.

[15] Osada R, Funkhouser T, Chazelle B, Dobki D. Shape distributions. ACM Transactions on Graphics 2002;21(4).

[16] Novatnack J, Nishino K. Scale-dependent/invariant local 3D shape descriptors for fully automatic registration of multiple sets of range images. In: ECCV. 2008,.

[17] Scovanner P, Ali S, Shah M. A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th international conference on Multimedia. 2007,.

[18] Shilane P, Funkhouser T. Distinctive regions of 3d surfaces. ACM Trans Graph 2007;26.

[19] Johnson A, Hebert M. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 1999;21(5):433–49.

[20] Gal R, Cohen-Or D. Salient geometric features for partial shape matching and similarity. ACM Trans Graph 2006;25(1):130–50.

[21] Furuya T, Ohbuchi R. Dense sampling and fast encoding for 3d model retrieval using bag-of-visual features. In: Proceedings of the ACM International Conference on Image and Video Retrieval. 2009,.

[22] Darom T, Keller Y. Scale invariant features for 3d mesh models. IEEE Trans Image Processing 2012;21(5):2758–69.

[23] Lowe DG. Distinctive image features from scale-invariant keypoints. IJCV 2004;60(2):91–110.

[24] Bay H, Ess A, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. CVIU 2008;110:346–59.

[25] Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3d shape descriptors. In: Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing. 2003,.

[26] Mian A, Bennamoun M, Owens R. Three-dimensional model-based object recognition and segmentation in cluttered scenes. PAMI 2006;28.

[27] Frome A, Huber D, Kolluri R, Blow T, Malik J. Recognizing objects in

Table 5: Average time per 3D shape in seconds that takes the overall process of building and testing the 3DSP representation. The 66 test models in the TOSCA+SUMNER dataset have been used.

K	L	3DSP-HIK	3DSP-HIK-Repre	3DSP-HIK-Disc	3DSP-χ^2	3DSP-χ^2-Repre	3DSP-χ^2-Disc
200	0	0.08	n/a	n/a	0.07	n/a	n/a
200	1	0.1	0.1	0.1	0.1	0.1	0.1
200	2	0.1	0.1	0.1	0.2	0.1	0.2
1000	0	0.1	n/a	n/a	0.1	n/a	n/a
1000	1	0.2	0.2	0.2	0.1	0.1	0.1
1000	2	0.4	0.4	0.4	1.2	1.1	1.1

range data using regional point descriptors. In: ECCV (3)'04. 2004, p. 224–37.

[28] Leibe B, Leonardis A, Schiele B. Robust object detection with interleaved categorization and segmentation. IJCV 2008;77(1-3):259–89.

[29] Toldo R, Castellani U, Fusiello A. The bag of words approach for retrieval and categorization of 3D objects. The Visual Computer: International Journal of Computer Graphics 2010;26(10):1257–68.

[30] Grauman K, Darrell T. The pyramid match kernel:discriminative classification with sets of image features. In: ICCV. 2005, p. 1458–65.

[31] Fowlkes C, Belongie S, Chung F, Malik J. Spectral grouping using the nyström method. IEEE TPAMI 2004;26:214–25.

[32] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006,.

[33] Hsu CW, Lin CJ. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 2002;13:415–25.

[34] Chang CC, Lin CJ. LIBSVM: a library for support vector machines; 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[35] Fu H, Cohen-Or D, Dror G, Sheffer A. Upright orientation of man-made objects. ACM Trans Graph 2008;27(3).

[36] Chaouch M, Verroust-Blondet A. Alignment of 3d models. Graphical Models 2009;71(2):63–76.