Noname manuscript No.
(will be inserted by the editor)

3D Object Retrieval via Range Image Queries in a Bag-of-Visual-Words Context

Konstantinos Sfikas · Theoharis Theoharis · Ioannis Pratikakis

Received: date / Accepted: date

Abstract 3D object retrieval based on range image queries that represent partial views of real 3D objects is presented. The complete 3D models of the database are described by a set of panoramic views and a Bag-of-Visual-Words model is built using SIFT features extracted from them. To address the problem of partial matching, we suggest a histogram computation scheme, on the panoramic views, that represents local information by taking into account spatial context. Furthermore, a number of optimization techniques are applied throughout the process, for enhancing the retrieval performance. Its superior performance is shown by evaluating it against state-of-the-art methods on standard datasets.

Keywords 3D Object Retrieval · Range Scans · Panoramic Views · SIFT · Bag-of-Visual-Words Model

1 Introduction

In the past few years, the increasing availability of low-cost 3D scanners has resulted in the creation of large 3D model repositories, thus making content-based retrieval a key operation. 3D model retrieval has considerably matured and a number of very accurate and robust descriptors have been proposed by our team [32, 2, 35] and others [9, 25, 43, 28]. These methodologies use a 3D model query to search a database of 3D models in a content-based manner. However, in practical situations, it is often difficult to come up with a suitable 3D model query in the first place: this has either to be found or built, a random and time-consuming action, respectively.

Nowadays, 3D scanners that typically produce *range images* (also called *range scans* and/or *depth buffers*) from real world 3D objects are becoming common and cheap, e.g. Microsoft Kinect [38]. It would thus, be beneficial, to use the range scans of real objects as queries on the 3D model repositories.

However, a number of challenges exist. First, a range image represents only a partial object. Thus, it is not straightforward to effectively match such data against a complete 3D model representation, since an important part of it may be missing. Second, range images can be rough and noisy. Third, it is not straightforward how to bridge the gap between the 3D model representation and the range image, i.e. how to produce descriptors that can be relatively invariant to these two representations. The representation gap makes it difficult to extract a signature that will be (at least partially) similar when presented with a complete, clean 3D model and when presented with a partial and noisy range image of a similar query object.

In the proposed approach, we have extended our previous work [34] and addressed the aforementioned challenges in the following way. For the complete database 3D models a set of panoramic views is extracted and consecutively the SIFT algorithm [27] is applied on hierarchically divided spatial areas of the views. These SIFT descriptors are used to feed a Bag-of-*Visual*-Words (BoVW) model [12], similar to the ones used for the categorization of textual infor-

Konstantinos Sfikas
Computer Graphics Laboratory, Department of Informatics
and Telecommunications, University of Athens, Athens, Greece
E-mail: ksfikas@di.uoa.gr

Theoharis Theoharis
Computer Graphics Laboratory, Department of Informatics
and Telecommunications, University of Athens, Athens, Greece
E-mail: theotheo@di.uoa.gr
and
IDI, NTNU, Norway
E-mail: theotheo@idi.ntnu.no

Ioannis Pratikakis
Department of Electrical and Computer Engineering,
Democritus University of Thrace,
GR-67100 Xanthi, Greece
E-mail: ipratika@ee.duth.gr

mation [20,40]. Using the trained BoVW model, for each 3D model of the database, as well as for every range image of the query objects, a signature in the form of spatial histogram is defined. This signature is generated from the same type of information representation (i.e. range images) and contains local information taking into account spatial context, thus bridging the representation gap. The matching between a query model and each of the database models is based on these signatures.

The remainder of the paper is structured as follows. In Section 2, recent work in 3D model retrieval based on range image queries is presented. Section 3 details the proposed method and Section 4 presents experimental results achieved in the course of the method's evaluation. Finally, conclusions are drawn in Section 5.

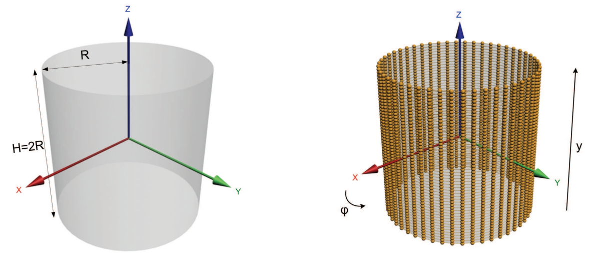
2 Related Work

Over the past few years, the number of works addressing the problems of multimodal 3D object retrieval and recognition (and particularly those based on range image queries), have increased significantly. Although this task still remains non-trivial, the quality of existing works shows that very important steps have been made in the field. Common retrieval scenarios deal with two different query image types: (i) directly captured range images from real 3D objects (i.e. using a 3D range scanner) and (ii) artificially produced range images via depth buffer capturing of complete 3D models. The first query image type is closer to real-world applications and is being increasingly adopted as 3D scanning becomes common place and corresponding datasets are created.

Hetzel et al. [22] explore a view based approach for the recognition of free-form objects in range images. They combine a set of local features (pixel depth, surface normal and curvature metrics) in a multidimensional histogram in order to achieve classification. Johnson and Hebert [24], use a spin image representation scheme in order to achieve simultaneous recognition of multiple 3D objects in cluttered scenes. The spin image representation is used for matching surface points. Chen and Bhanu [10] introduce a local surface descriptor for 3D model recognition. This descriptor is computed on feature points of a 3D surface, where large shape variations occur. The local surface descriptor is characterized by its centroid, its local surface type and a 2D histogram. The latter shows the frequency of occurrence of shape index values (calculated from principal curvatures) vs the angles between the normal of the reference feature point and those of its neighbors. Ruiz-Correa et al. [33] propose a method for recognizing 3D objects in real range image scenes. Initially, shape class components are learnt and extracted from range images and then the spatial relationships among the extracted components are used to form a model that consists of a three-level hiererchy of classifiers. Adan et al. [1] explore the use of Depth Gradient Image (DGI) models for the recognition of 3D models. The DGI representation synthesizes both surface and contour information, for a specific viewpoint, by mapping the distance between each contour point and the edge of the viewpoint image in terms of internal and external object pixels. This measure is computed for the entire model, taken from the nodes of a tessellated sphere. Frome et al. [18] introduced two regional shape descriptors, the 3D generalization of the 2D shape context descriptor and the harmonic shape descriptor. The authors evaluate the performance of the proposed descriptors in recognizing similar objects in scenes with noise or clutter.

Ohbuchi et al. [30] proposed the Multiple Orientation Depth Fourier Transform (MODFT) descriptor where the model is projected from 42 viewpoints to cover all possible view aspects. Each depth buffer is then transformed to the $r-\theta$ domain and the Fourier transform is applied. To compare two models, all possible pairs of coefficients are compared which inevitably increases comparison time. Stavropoulos et al. [41] present a retrieval method based on the matching of salient features between the 3D models and query range images. Salient points are extracted from vertices that exhibit local maxima in terms of protrusion mapping for a specific window on the surface of the model. A hierarchical matching based scheme is used for matching. The authors experimented on range images acquired from the SHape REtrieval Contest 2007 (SHREC'07) *Watertight models* [21] and the Princeton Shape Benchmark (PSB) standard [37] datasets. Chaouch and Verroust-Blondet [7] present a 2D/3D shape descriptor which is based on either silhouette or depth-buffer images. For each 3D model a set of six projections in calculated for both silhouette and depth-buffers. The 2D Fourier transform is then computed on the projection. Furthermore, they compute a relevance index measure which indicates the density of information contained in each 2D view. The same authors in [8] propose a method where a 3D model is projected to the faces of its bounding box, resulting in 6 depth buffers. Each depth buffer is then decomposed into a set of horizontal and vertical depth lines that are converted to state sequences which describe the change in depth at neighboring pixels. Experimentations were conducted on range images artificially acquired from the PSB dataset. Shih et al. [36] proposed the elevation descriptor where six depth buffers (elevations) are computed from the faces of the 3D model's bounding box and each buffer is described by a set of concentric circular areas that give the sum of pixel values within the corresponding areas. The models were selected from the standard PSB dataset.

Experimenting on the SHREC'09 *Querying with Partial Models* [15] dataset, Daras and Axenopoulos in [13] present a view-based approach for 3D model retrieval. The 3D model is initially pose normalized and a set of binary



(a) (b)

Fig. 1: (a) A projection cylinder for the acquisition of a 3D model's panoramic view and (b) the corresponding discretization of its lateral surface to the set of points $s(\phi_u, y_v)$

(silhouette) and range images are extracted from predefined views on a 32-hedron. The set of features computed on the views are the Polar-Fourier transform, Zernike moments and Krawtchouk moments. Each query image is compared to all the extracted views of each model of the dataset. Ohbuchi et al. [31] extract features from 2D range images of the model viewed from uniformly sampled locations on a view sphere. For every range image a set of multi-scale 2D visual features are computed using the Scale Invariant Feature Transform (SIFT) [27]. Finally, the features are integrated into a histogram using the Bag-of-Features approach [20]. The same authors enhanced their approach by pre-processing the range images, in order to minimize interfere caused by any existing occlusions, and also and by refining the positioning of SIFT interest points, so that higher resolution images are favored [19,29]. Their works have experimented on and have participated on both corresponding SHREC'09 *Querying with Partial Models* and SHREC'10 *Range Scan Retrieval* [16] contests. Wahl et al. [44] propose a four-dimensional feature that parameterizes the intrinsic geometrical relation of an oriented surface point pair (surflets). For a 3D model a set of surflet pairs is computed over a number of uniformly sampled viewing directions on the surrounding sphere. This work was one of the two contestants of the SHREC'10 *Range Scan Retrieval* track.

3 The Proposed Method

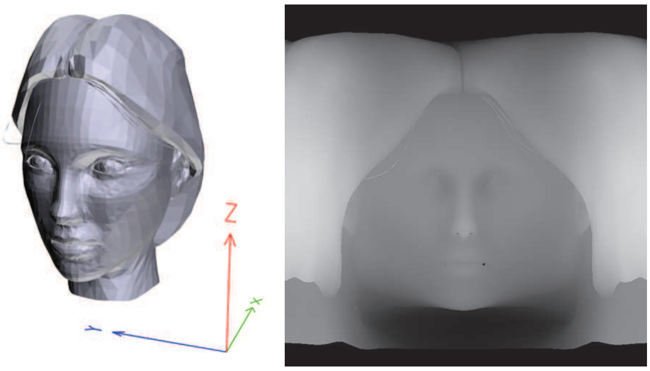
In the sequel, the proposed methodology on 3D object retrieval via range image queries will be detailed. Initially, let us consider that the complete 3D model dataset and the query range image set define two different entities of the retrieval pipeline, due to their difference in representation and processing strategies. At each step, we will indicate which of the two sets is regarded. Key concepts that will be discussed in what follows: Panoramic Views computation, SIFT descriptors extraction, Bag-of-Visual-Words modelling and Spatial Histograms.

3.1 Panoramic Views Computation

For each 3D model of the database a number of panoramic views (or cylindrical projections) are extracted. These projections are computed on cylindrical axes that are perpendicular to and uniformly distributed over the surface of the 3D model's circumscribed sphere, in accordance with the PANORAMA [32] projection methodology.

To obtain a panoramic view, we project the 3D model to the lateral surface of a cylinder of radius R and height $H = 2R$, centered at the origin with its axis parallel to one of the selected axes (in this example the principal axis z, see Fig. 1a). We set the value of R to $2 * d_{max}$ where d_{max} is the maximum distance of the model's surface from its centroid. In the following, we parameterize the lateral surface of the cylinder using a set of points $s(\phi, y)$ where $\phi \in [0, 2\pi]$ is the angle in the xy plane, $y \in [0, H]$ and we sample the ϕ and y coordinates at rates $6B$ and B, respectively (we set $B = 360$). Thus we obtain the set of points $s(\phi_u, y_v)$ where $\phi_u = u * 2\pi/(6B)$, $y_v = v * H/B$, $u \in [0, 6B-1]$ and $v \in [0, B-1]$. These points are shown in Fig. 1b.

The next step is to determine the value at each point $s(\phi_u, y_v)$. The computation is carried out iteratively for $v = 0, 1, \ldots, B-1$, each time considering the set of coplanar $s(\phi_u, y_v)$ points, i.e. a cross section of the cylinder at height y_v and for each cross section we cast rays from its center c_v in the ϕ_u directions. To capture the position of the model's surface, for each cross section at height y_v we compute the distances from c_v to the intersections of the model's surface with the rays at each direction ϕ_u.



(a) (b)

Fig. 2: (a) An example 3D model and (b) its corresponding cylindrical projection on the Z-axis.

Let $pos(\phi_u, y_v)$ denote the distance of the furthest from c_v point of intersection between the ray emanating from c_v in the ϕ_u direction and the model's surface; then $s(\phi_u, y_v) =$

$pos(\phi_u, y_v)$. The value of a point $s(\phi_u, y_v)$ lies in the interval $[0, R]$, where R denotes the radius of the cylinder.

A cylindrical projection can be viewed as a 2D gray-scale image where pixels correspond to the $s(\phi_u, y_v)$ intersection points in a manner reminiscent of cylindrical texture mapping [42] and their values are mapped to the $[0, 1]$ interval. The number of extracted cylindrical projections for each complete 3D model is 60, which maintains acceptable processing speed and coverage of the surface of the 3D model's circumscribed sphere. In Fig. 2a, we show an example 3D model and in Fig. 2b the unfolded visual representation of its corresponding cylindrical projection $s(\phi_u, y_v)$.

3.2 SIFT Descriptors Extraction

After the panoramic view extraction, the SIFT (Scale Invariant Feature Transform) [27] descriptor is calculated on the produced cylindrical depth images. The first step to the SIFT descriptor computation is the definition of an interest point set on the image, upon which the descriptors are calculated. The original implementation by Lowe, defines these interest points through the Difference of Gaussians (DoG) algorithm, which is geared towards enhancing the edges and other details present in the image. It has been experimentally found that the calculation of the SIFT descriptors over the complete image for a large number of randomly selected points [19,5,6] (frequently defined as Dense SIFT/ DSIFT, in the literature), instead of selecting a limited number of interest points, yields better results in terms of retrieval accuracy.

At each interest point, an image descriptor is computed. The SIFT descriptor is defined as a position-dependent histogram of local gradient directions around the interest point. To achieve scale invariance of the descriptor, the size of this local neighbourhood is normalized in a scale-invariant manner. To achieve rotational invariance of the descriptor, the dominant orientation of the neighbourhood is determined from the orientations of the gradient vectors in this neighbourhood and is used for orienting the grid over which the position-dependent histogram is computed.

One recently proposed improvement technique for SIFT, by Arandjelovic and Zisserman [3], aims at enhancing the similarity measure used when comparing the descriptors. The authors show that using a square root (Hellinger) kernel (also known as the Bhattacharyya's coefficient [4]) instead of the standard Euclidean distance measure, for the comparison of the SIFT descriptors (or SIFT histograms), increases performance. The intuition behind this proposal is based on the observation that Euclidean distance can be dominated by large bin values, whereas Hellinger distance is more sensitive to smaller bin values.

The Hellinger kernel for two L^1 normalized histograms, x and y is defined as:

$$H(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i} \tag{1}$$

To compare the SIFT vectors with a Hellinger kernel is a simple two-step algebraic manipulation (thus easy to implement in any existing SIFT implementation). First L^1 normalization of the SIFT vector, which originally has L^2 norm [27, 3], and then square-rooting each of its elements. Computing Euclidean distance in the feature map space is equivalent to Hellinger distance in the original space:

$$\sqrt{x}^T \sqrt{y} = H(x,y) \tag{2}$$

3.3 Bag-of-Visual-Words Modelling

Once the panoramic views extraction and the SIFT descriptor calculation steps are complete, the Bag-of-Visual-Words (BoVW) visual model for the database is built. In visual information retrieval, the BoVW model defines that each image contains a number of local *visual features*. Since every visual feature, or collection of similar visual features, can appear with different frequencies on each image, matching the visual feature frequencies of two images, achieves correspondence. In our case, the SIFT descriptors are defined as the BoVW model's visual features.

The basic step in the process of building the BoVW model for the 3D model database is the generation of a *codebook* (or a vocabulary), a collection of visual features that appear on each image. The codebook is generated by considering the visual features of a representative number of training database models. To achieve greater flexibility, rather than generating the codebook by selecting individual visual features of the training models, the corresponding panoramic views are clustered into several similar patches, the *codewords*. One simple method is performing *k-means* clustering [26] over all the visual features. Codewords are then defined as the centers of the learned clusters. The number of the clusters is the codebook size. Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.

After the codebook generation procedure, the next step is the description of the database 3D models using the corresponding codewords. In a similar manner, for each panoramic view of the database 3D model set, the visual features are computed and matched to their closest codewords, by comparing them to the corresponding clusters, generated in the previous step. Again, the *k-means* algorithm is used for matching. Note that the *k-means* clustering method makes use of the Euclidean distance for the comparison between the clusters and the visual features. According to the RootSIFT transformation, applied on the visual features, this results in using the Hellinger distance for the operation. The set of histograms describing the frequency of occurrence of the generated codewords, for each 3D model's panoramic views, is stored as the corresponding 3D model's signature.

3.4 Spatial Histograms

In an extension of the standard Bag-of-Visual-Words model, described in the previous subsection, and targeting the matching between a complete 3D model and a range image representing a partial query, we have modified the histogram generation in the following manner. Since a panoramic view of a complete 3D model contains a 360°of information, an attempt to match it to a query range image, which contains only a portion of that information, will produce poor results in the majority of cases. To alleviate this problem we suggest a progressive partitioning scheme for the panoramic views of the database 3D models. Each panoramic view image is iteratively split along the horizontal dimension (width), which is perpendicular to the axis of the corresponding projection cylinder (see Fig. 3).

The spatial histograms are then computed for each resulting subpart of the image. For example, on a first level of progressive partitioning, the spatial histogram is computed on the complete panoramic view image of the 3D model. On the second level, the panoramic view image is partitioned once along the horizontal dimension and two spatial histograms are computed for the resulting left and right subimage, respectively. On the third level, the complete panoramic view image is partitioned twice along the horizontal dimension and three spatial histograms are computed for each of the three resulting subimages. The process continues until a certain level of progressive partitioning is reached, which in our case has been selected to be 6.

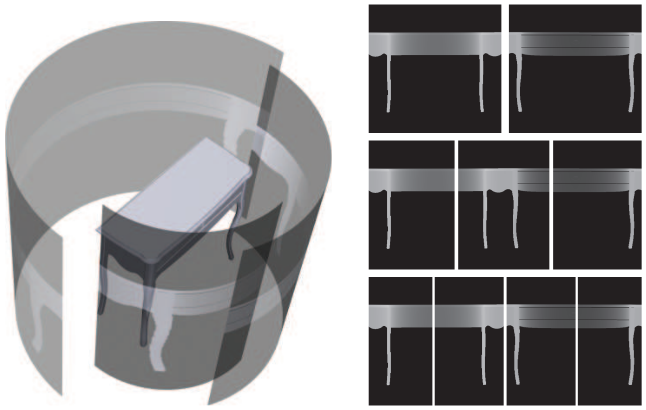

(a)
(b)

Fig. 3: (a) simulated rendering of three progressive partitioning levels for a complete 3D model. At each level only one subimage is displayed. (b) unwrapped cylindrical projections of the aforementioned progressive partitioning levels with all subimages illustrated.

As the complete panoramic view images contain 360°of information, each level of progressive partitioning produces spatial histograms that reflect a fraction of that information (i.e. 180°, 120°, 90°, ..., etc). Therefore, the matching between each spatial histogram of a 3D model and the corresponding histograms of a query range image needs to be weighted based on a ratio that measures the possibility of achieving a match between the contained information of the two (spatial) histograms, at each level of progressive partitioning:

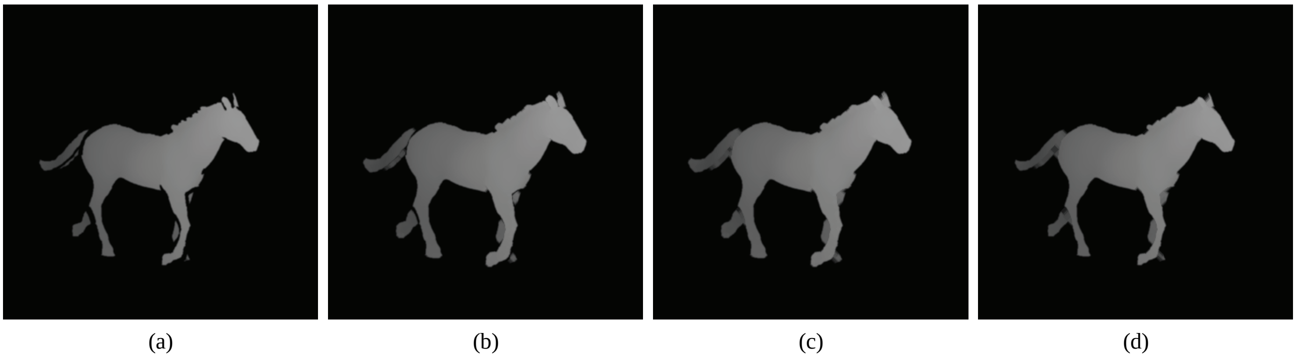
$$PQ_{ratio}(i) = \left| \frac{Q_{angle}}{P_{angle}(i)} \right| \tag{3}$$

where $P_{angle}(i)$ is the Field-of-View [42] of the panoramic view (sub)image at the i^{th} level of progressive partitioning and Q_{angle} is the Field-of-View of the query range image, measured in degrees. The Q_{angle} is based on the properties of the camera used for capturing the range images and in our experiments, we have estimated this angle at 60°, which simulates the projection of the query range image to one of the faces of a hexagon.

3.5 Range Image Matching

Based on the constructed BoVW model, for the database 3D models, the matching of the query range images is performed. The query range images are usually very noisy, due to the capturing process and an extra preprocessing step is often necessary before the actual matching. Here, we have followed a simple strategy that attempts to fill any holes, resulting from the object scanning and/or eliminate any outliers that do not belong to the actual objects (i.e. parts of the background).

Initially, morphological dilation [14] is performed on the original query range image. This step grows the range image regions, so that any moderate holes, that could have occurred due to errors in the capturing process, become small enough, in order to be considered as candidates for closing (see Fig. 4b). On the second step, morphological closing [14] is performed on the dilated image, in order to achieve closing of any small open areas that have been produced due to the digitization process (see Fig. 4c). The next step of the preprocessing strategy is the morphological erosion [14] of the image, so that it returns to its original form, with any small to moderate holes closed (see Fig. 4d). All of



(a) (b) (c) (d)

Fig. 4: An example query range image (a) before preprocessing, (b) after dilation, (c) closing and (d) erosion steps, in consecutive order.

the aforementioned morphological operators are applied using disk-shaped structuring elements of size 3. The final step of the preprocessing strategy involves smoothing the range image by convolving it with an isotropic Gaussian kernel, which ensures that any rough edges are leveled and any small outlying regions, remote from the main object, are discarded.

Following the preprocessing, the range images are used as queries for the 3D model database. In a strategy similar to the histogram computation for the panoramic views of the database 3D models, the queries are compared to the models of the database. The SIFT descriptor is extracted from the range image, RootSIFT transformation is applied on the descriptor and finally, based on the generated Bag-of-Visual-Words, a histogram describing the codebook frequencies of occurrence is computed as the query's descriptor.

The similarity between the spatial histograms H of a 3D model and the histogram h of a query range image is calculated as follows: for each level l of progressive partitioning, the spatial histograms H_l are compared to the query range image histogram h using as a metric the Normalized Histogram Intersection distance ($D_{NHI}(H,h)$), defined by [11] as:

$$D_{NHI}(H,h) = \sum_{i=1}^{n} \frac{\min(H(i),h(i))}{H(i)+h(i)} \tag{4}$$

The best match is recorded and weighted by PQ_{ratio} for the corresponding level. Then, for all progressive partitioning levels, the best matches are summed to create the final distance between the query and the corresponding database model. We define max_l to be the maximum level of progressive partitioning set to 6.

$$Dist(H,h) = \sum_{l=1}^{max\_l} PQ_{ratio}(l)[\arg\min_{H_l} (D_{NHI}(H_l,h))] \tag{5}$$

The complete descriptor extraction and matching algorithm is outlined in **Algorithm 1** and **Algorithm 2**, respectively.

Algorithm 1 Bag_of_Visual_Words Model Building Algorithm

```
Input: $3D\_models$ /*Database 3D model set*/
$n\_models = card(3D\_models)$ /*The number of 3D models of the database*/
$n\_axes = 60$ /*Define 60 random axis points*/
$max\_l = 6$
for $n = 1 \rightarrow n\_models$ do
    $axes = rand(n\_axes)$
    for $m = 1 \rightarrow n\_axes$ do
        $pan(n,m) =$ EXTRACT_PVIEW($3D\_model(n), axes(m)$)
        $sift(n,m) =$ DSIFT($pan(n,m)$)
        $rsift(n,m) =$ ROOTSIFT($sift(n,m)$)
    end for
end for
$train\_set =$ rand($rsift, n\_model/10$)
$codebook =$ TRAIN_BOVW($train\_set$)
for $n = 1 \rightarrow n\_models$ do
    for $m = 1 \rightarrow n\_axes$ do
        for $l = 1 \rightarrow max\_l$ do
            $hist_l(n,m)$ = EXTRACT_HIST($codebook, rsift(n,m)$)
        end for
    end for
end for
return $codebook, hist$
```

Note that the EXTRACT_PVIEW, DSIFT, ROOTSIFT, TRAIN_BOVW and EXTRACT_HIST functions refer to the panoramic view extraction, Dense SIFT descriptor calculation, RootSIFT transformation, Bag-of-Visual-Words codebook generation and (spatial) Histogram extraction operations of the pipeline, respectively.

4 Evaluation

The datasets that we used for the experimental evaluation of our method are the following: (i) SHREC'09 *Querying*

Algorithm 2 Query Range Image Matching Algorithm

```
Input: $codebook, hist, query\_image$
Input: $n\_models, n\_axes$ /*The number of 3D models of the database
    and the random axis points, as defined in Algorithm 1*/
$siftq = \text{DSIFT}(query\_image)$
$rsiftq = \text{ROOTSIFT}(siftq)$
$histq = \text{EXTRACT\_HIST}(codebook, rsiftq(n))$
for $n = 1 \to n\_models$ do
    for $m = 1 \to n\_axes$ do
        $query\_dist(n,m) = Dist(hist(n), histq)$
    end for
end for
$final\_dist = min(query\_dist)$
return $final\_dist$
```

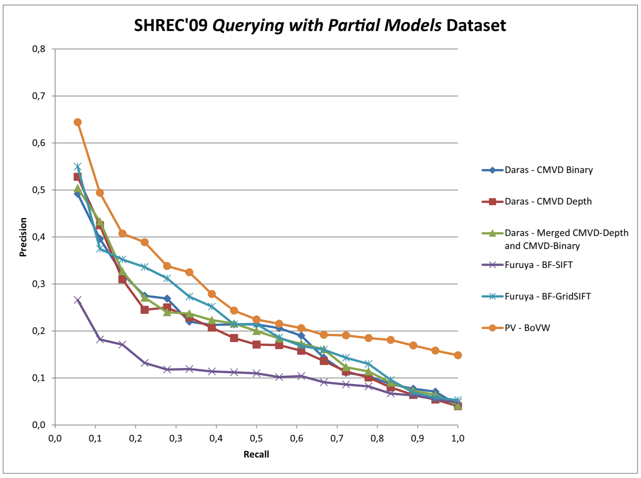


Fig. 5: Comparative results based on the average P-R scores for the SHREC'09 *Querying with Partial Models* dataset.

with Partial Models [15], (ii) SHREC'10 *Range Scan Retrieval* [16] and (iii) SHREC'11 *Shape Retrieval Contest of Range Scans* [39]. The target subset of the datasets used is based on the generic shape benchmark constructed at NIST [17].

On the first two datasets, SHREC'09 *Querying with Partial Models* and SHREC'10 *Range Scan Retrieval*, we compared against existing results of the participating contestants. More specifically, on the SHREC'09 *Querying with Partial Models* we compared against the variations of CMVD (Compact MultiView Descriptor) by Daras and Axenopoulos [13] and the BF-SIFT and BF-GridSIFT methods by Furuya and Ohbuchi. On the SHREC'10 *Range Scan Retrieval* dataset we compared against the variations of the BF-DSIFT-E method proposed by Ohbuchi and Furuya [29] and the variations of the SURFLET method proposed by Hillebrand et al. [44]. Furthermore, on this dataset we compared against the initial version of our method (without Bag-of-Visual-Words) presented in [34]. In the SHREC'11 *Shape Retrieval Contest of Range Scans* competition, due to lack of participants no results were published. We publish our results on this dataset for future reference.

Our experimental evaluation is based on Precision-Recall plots and five quantitative measures: Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), E-measure (E) and Discounted Cumulative Gain (DCG) [37] for the classes of the corresponding datasets. For every query model that belongs to a class C, recall denotes the percentage of models of class C that are retrieved and precision denotes the proportion of retrieved models that belong to class C over the total number of retrieved models. The best score is 100% for both quantities. Nearest Neighbor (NN) indicates the percentage of queries where the closest match belongs to the query class. First Tier (FT) and Second Tier (ST) statistics, measure the recall value for the $(D-1)$ and $2(D-1)$ closest matches respectively, where D is the cardinality of the query's class. E-measure combines precision and recall metrics into a single number and the DCG statistic weights correct results near the front of the list more than correct results later in the ranked list under the assumption that a user is less likely to consider elements near the end of the list [23,37].

According to the SHREC'09 classification scheme, the target subset is composed of 720 complete 3D models, classified into 40 classes, each of which contains 18 models. The query set is composed of 20 range images taken from 20 objects from arbitrary view directions.

Figure 5, using the experimental results given in [15], illustrates the P-R scores for the complete SHREC'09 *Querying with Partial Models* dataset for the proposed 3D model retrieval method (PV - BoVW) and the methods by Daras and Furuya. Table 1 shows the corresponding five quantitative measures for the same methods.

Table 1: Comparison between the proposed method and the methods presented on the SHREC'09 *Querying with Partial Models* track using five quantitative measures. All measures are in the interval $[0,1]$.

Method	NN	FT	ST	E	DCG
CMVD-BINARY	0.350	0.217	0.283	0.200	0.521
CMVD-DEPTH	0.450	0.197	0.267	0.174	0.511
Merged CMVD-DEPTH and CMVD-BINARY	0.350	0.211	0.281	0.192	0.526
BF-SIFT	0.150	0.114	0.186	0.116	0.423
BF-GridSIFT	0.450	0.225	**0.297**	0.204	0.532
PV - BoVW	**0.600**	**0.251**	0.292	**0.206**	**0.553**

Both the P-R scores of Fig. 5, as well as the quantitative measures of Table 1 illustrate that the proposed method achieves superior performance compared to the variations of the CMVD, as well as both the BF-SIFT and the BF-GridSIFT retrieval methods.

Table 2: Comparison between the proposed method and the methods presented on the SHREC'10 *Range Scan Retrieval* track using five quantitative measures. All measures are in the interval [0,1].

Method	NN	FT	ST	E	DCG
BF-DSIFT-E (LFE)	0.573	0.380	0.524	0.367	0.683
Closing_3x3_BF-DSIFT-E (LFE)	0.598	0.393	0.535	0.382	0.696
Closing_6x6_BF-DSIFT-E (LFE)	0.650	**0.424**	0.569	**0.398**	0.713
Dilation_3x3_BF-DSIFT-E (LFE)	0.675	0.405	0.557	0.392	0.713
Dilation_6x6_BF-DSIFT-E (LFE)	0.547	0.395	0.550	0.386	0.696
SURFLET - mean	0.325	0.244	0.363	0.252	0.556
SURFLET - meanraw	0.171	0.153	0.242	0.163	0.462
SURFLET - meansqrd	0.231	0.197	0.322	0.213	0.513
SURFLET - median	0.282	0.226	0.325	0.224	0.528
SURFLET - mediansqrd	0.282	0.226	0.325	0.224	0.528
PanoramicViews - SIFT	0.512	0.374	0.466	0.256	0.598
PV - BoVW	**0.691**	0.413	**0.570**	0.386	**0.720**

The next dataset, SHREC'10 *Range Scan Retrieval* is composed of the following two subsets: the target subset which contains 800 complete 3D models, classified into 40 classes, each of which has 20 models and the query subset which contains 120 range images that have been acquired by capturing 3 range scans of 40 objects from arbitrary view directions.

In Figure 6, using the experimental results given in [16], we show the P-R scores for the complete SHREC'10 *Range Scan Retrieval* dataset for the proposed 3D object retrieval method and the methods by Ohbuchi and Hillebrand. Table 2 shows the corresponding five quantitative measures for the same methods.

Table 2 shows that the proposed method has the highest scores on three out of five measures (and is also close on the remaining two measures, by a small margin). The P-R scores of Fig. 6, illustrate that the proposed method outperforms the track contestants, as well as our previously proposed method (PanoramicViews - SIFT), presented in [34].

Finally, in the SHREC'11 classification scheme, the target subset is composed of 1000 complete 3D models, categorized into 50 classes, each of which contains 20 models. The query set is composed of 150 range images acquired by capturing 3 range scans, of each of 50 objects that correspond to the above classes, from arbitrary view directions.

In Table 3 we show the five quantitative measures for the complete SHREC'11 *Shape Retrieval Contest of Range Scans* dataset for the proposed 3D model retrieval method.

Table 3: Five quantitative measures for the proposed 3D object retrieval method on the SHREC'11 *Shape Retrieval Contest of Range Scans* dataset. All measures are normalized.

Method	NN	FT	ST	E	DCG
PV - BoVW	0.512	0.374	0.466	0.256	0.598

The proposed method was tested on a Core2Quad 2.5 GHz system, with 6 GB of RAM, running Matlab R2012a. The system was developed in a hybrid Matlab/C++/OpenGL architecture, which resulted in low computational times. The average computation time for *the Bag_of_Visual_Words Model Building* algorithm (see **Algorithm 1**) on a 1,000 3D model database is about 4,5 hours (an offline process). The average computation time required for querying a range image object on the aforementioned database (see **Algorithm 2**) is about 5 seconds.

5 Conclusions

We proposed a spatial histograms strategy in a Bag-of-Visual-Words context that fits the information present in panoramic views of 3D objects to the task of partial matching. Special attention has been given to the query range image preprocessing stage, where a number of consecutive filters are applied on the images in order to alleviate problems introduced by the digitization process. This improved 3D object retrieval methodology, was evaluated not only against our previous approach [34] and the corresponding SHREC'10 *Range Scan Retrieval* track dataset but also on standard datasets from the SHREC'09 *Querying with Partial Models* track and the corresponding state-of-the-art 3D object retrieval methodologies and the SHREC'11 *Shape Retrieval Contest of Range Scans* track. In every case, the proposed 3D object retrieval method outperforms competing descriptors.

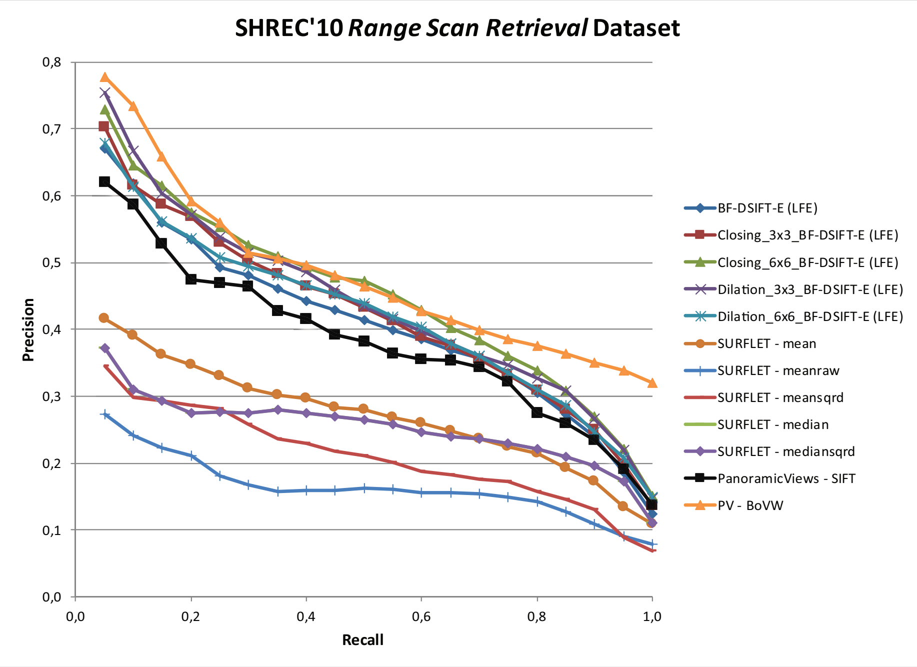


Fig. 6: Comparative results based on the average P-R scores for the SHREC'10 *Range Scan Retrieval* dataset.

Acknowledgements This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES-3DOR (MIS 379516). Investing in knowledge society through the European Social Fund.

European Union
European Social Fund
OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY
NSRF
2007-2013
programme for development
EUROPEAN SOCIAL FUND
Co-financed by Greece and the European Union

References

1. Adán, A., Merchán, P., Salamanca, S.: 3D scene retrieval and recognition with depth gradient images. Pattern Recognition Letters **32**(9), 1337 – 1353 (2011). DOI 10.1016/j.patrec.2011.03.016
2. Agathos, A., Pratikakis, I., Papadakis, P., Perantonis, S., Azariadis, P., Sapidis, N.S.: 3D articulated object retrieval using a graph-based representation. The Visual Computer **26**(10), 1301–1319 (2010). DOI 10.1007/s00371-010-0523-1
3. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR, pp. 2911–2918. IEEE (2012)
4. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society **35**, 99–109 (1943)
5. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via plsa. In: A. Leonardis, H. Bischof, A. Pinz (eds.) ECCV (4)'06, *Lecture Notes in Computer Science*, vol. 3954, pp. 517–530. Springer (2006)
6. Bosch, A., Zisserman, A., Muñoz, X.: Image classification using random forests and ferns. In: ICCV'07, pp. 1–8. IEEE (2007)
7. Chaouch, M., Verroust-Blondet, A.: Enhanced 2D/3D approaches based on relevance index for 3D-shape retrieval. In: SMI, p. 36. IEEE Computer Society (2006). DOI 10.1109/SMI.2006.11
8. Chaouch, M., Verroust-Blondet, A.: A new descriptor for 2D depth image indexing and 3D model retrieval. In: Image Processing, 2007. ICIP 2007. IEEE International Conference on, vol. 6, pp. 373–376 (2007). DOI 10.1109/ICIP.2007.4379599
9. Chen, D.Y., Tian, X.P., Shen, Y.T., Ouhyoung, M.: On visual similarity based 3D model retrieval. Comput. Graph. Forum pp. 223–232 (2003)
10. Chen, H., Bhanu, B.: 3D free-form object recognition in range images using local surface patches. Pattern Recognition Letters **28**(10), 1252 – 1262 (2007). DOI 10.1016/j.patrec.2007.02.009
11. Cheng, E., Xie, N., Lin, H., Bakic, P.R., Maidment, A.D.A., Megalooikonomou, V.: Mammographic image classification using histogram intersection. In: ISBI, pp. 197–200. IEEE (2010)
12. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
13. Daras, P., Axenopoulos, A.: A compact multi-view descriptor for 3D object retrieval. In: Proceedings of the 2009 Seventh International Workshop on Content-Based Multimedia Indexing, CBMI '09, pp. 115 –119. IEEE Computer Society, Washington, DC, USA (2009). DOI 10.1109/CBMI.2009.15
14. Dougherty, E.R., of Photo-optical Instrumentation Engineers, S.: An introduction to morphological image processing. SPIE Optical Engineering Press, Bellingham, Wash., USA (1992)
15. Dutagaci, H., Godil, A., Axenopoulos, A., Daras, P., Furuya, T., Ohbuchi, R.: SHREC'09 track: Querying with partial models. In: M. Spagnuolo, I. Pratikakis, R. Veltkamp, T. Theoharis (eds.) Eurographics Workshop on 3D Object Retrieval, pp. 69–76. Eurographics Association, Munich, Germany (2009). DOI 10.2312/3DOR/3DOR09/069-076
16. Dutagaci, H., Godil, A., Cheung, C.P., Furuya, T., Hillenbrand, U., Ohbuchi, R.: SHREC'10 track: Range scan retrieval.

In: M. Daoudi, T. Schreck (eds.) Eurographics Workshop on 3D Object Retrieval, pp. 109–115. Eurographics Association, Norrköping, Sweden (2010). DOI 10.2312/3DOR/3DOR10/109-115

17. Fang, R., Godil, A., Li, X., Wagan, A.: A new shape benchmark for 3D object retrieval. In: G. Bebis, R.D. Boyle, B. Parvin, D. Koracin, P. Remagnino, F.M. Porikli, J. Peters, J.T. Klosowski, L.L. Arns, Y.K. Chun, T.M. Rhyne, L. Monroe (eds.) ISVC (1), *Lecture Notes in Computer Science*, vol. 5358, pp. 381–392. Springer (2008). DOI 10.1007/978-3-540-89639-5_37
18. Frome, A., Huber, D., Kolluri, R., Blow, T., Malik, J.: Recognizing objects in range data using regional point descriptors. In: T. Pajdla, J. Matas (eds.) Proceedings of the European Conference on Computer Vision (ECCV), *Lecture Notes in Computer Science*, vol. 3023, pp. 224–237. Springer Berlin Heidelberg (2004). DOI 10.1007/978-3-540-24672-5_18
19. 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, CIVR ’09, pp. 26:1–26:8. ACM, New York, NY, USA (2009)
20. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **36**(1), 3–42 (2006)
21. Giorgi, D., Biasotti, S., Paraboschi, L.: SHape REtrieval contest 2007: Watertight models track (2007)
22. Hetzel, G., Leibe, B., Levi, P., Schiele, B.: 3D object recognition from range images using local feature histograms. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 2, pp. 394–399. IEEE Computer Society (2001). DOI 10.1109/CVPR.2001.990988
23. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002). DOI 10.1145/582415.582418
24. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Anal. Mach. Intell. **21**(5), 433–449 (1999). DOI 10.1109/34.765655
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, SGP ’03, pp. 156–164. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland (2003)
26. Lloyd, S.P.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2), 129–137 (1982)
27. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV ’99, pp. 1150–. IEEE Computer Society, Washington, DC, USA (1999)
28. Mohamed, W., Ben Hamza, A.: Reeb graph path dissimilarity for 3D object matching and retrieval. The Visual Computer **28**(3), 305–318 (2012). DOI 10.1007/s00371-011-0640-5
29. Ohbuchi, R., Furuya, T.: Scale-weighted dense bag of visual features for 3D model retrieval from a partial view 3D model. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pp. 63 –70 (2009). DOI 10.1109/ICCVW.2009.5457716
30. Ohbuchi, R., Nakazawa, M., Takei, T.: Retrieving 3D shapes based on their appearance. In: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, MIR ’03, pp. 39–45. ACM, New York, NY, USA (2003)
31. Ohbuchi, R., Osada, K., Furuya, T., Banno, T.: Salient local visual features for shape-based 3D model retrieval. In: Shape Modeling International, pp. 93 –102. IEEE (2008). DOI 10.1109/SMI.2008.4547955
32. Papadakis, P., Pratikakis, I., Theoharis, T., Perantonis, S.J.: PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval. International Journal of Computer Vision **89**(2-3), 177–192 (2010)
33. Ruiz-Correa, S., Shapiro, L.G., Meila, M.: A new paradigm for recognizing 3-D object shapes from range data. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, *ICCV ’03*, vol. 2, pp. 1126–. IEEE Computer Society, Washington, DC, USA (2003)
34. Sfikas, K., Pratikakis, I., Theoharis, T.: 3D Object Retrieval via Range Image Queries based on SIFT descriptors on Panoramic Views. In: M. Spagnuolo, M. Bronstein, A. Bronstein, A. Ferreira (eds.) Eurographics Workshop on 3D Object Retrieval, pp. 9–15. Eurographics Association, Cagliari, Italy (2012)
35. Sfikas, K., Theoharis, T., Pratikakis, I.: Non-rigid 3D object retrieval using topological information guided by conformal factors. The Visual Computer **28**, 943–955 (2012). DOI 10.1007/s00371-012-0714-z
36. Shih, J.L., Lee, C.H., Wang, J.T.: A new 3D model retrieval approach based on the elevation descriptor. Pattern Recognition **40**(1), 283 – 295 (2007). DOI 10.1016/j.patcog.2006.04.034
37. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Proceedings of the Shape Modeling International 2004, SMI ’04, pp. 167–178. IEEE Computer Society, Washington, DC, USA (2004). DOI 10.1109/SMI.2004.63
38. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’11, pp. 1297–1304. IEEE Computer Society, Washington, DC, USA (2011). DOI 10.1109/CVPR.2011.5995316
39. SHREC2011: Shape retrieval contest of range scans. http://www.itl.nist.gov/iad/vug/sharp/contest/2011/RangeScans/. Accessed on 08/2012
40. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**, 591–606 (2009)
41. Stavropoulos, T.G., Moschonas, P., Moustakas, K., Tzovaras, D., Strintzis, M.G.: 3D model search and retrieval from range images using salient features. IEEE Transactions on Multimedia **12**(7), 692–704 (2010). DOI 10.1109/TMM.2010.2053023
42. Theoharis, T., Papaioannou, G., Platis, N., Patrikalakis, N.M.: Graphics and Visualization: Principles & Algorithms. A. K. Peters, Ltd., Natick, MA, USA (2007)
43. Vranic, D.V.: DESIRE: a composite 3D-shape descriptor. In: ICME, pp. 962–965. IEEE (2005)
44. Wahl, E., Hillenbrand, U., Hirzinger, G.: Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification. In: 3DIM, pp. 474–481. IEEE Computer Society (2003). DOI 10.1109/IM.2003.1240284