2009 Seventh International Workshop on Content-Based Multimedia Indexing

A Compact Multi-View Descriptor for 3D Object Retrieval

Petros Daras, Apostolos Axenopoulos
Informatics and Telematics Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
e-mail: {daras,axenop}@iti.gr

***Abstract*—In this paper, a novel view-based approach for 3D object retrieval is introduced. A set of 2D images (multi-views) are automatically generated from a 3D object, by taking views from uniformly distributed viewpoints. For each image, a set of 2D rotation-invariant shape descriptors is extracted. The global shape similarity between two 3D models is achieved by applying a novel matching scheme, which effectively combines the information extracted from the multiview representation. The proposed approach can well serve as a unified framework, supporting multimodal queries (such as sketches, 2D images, 3D objects). The experimental results illustrate the superiority of the method over similar view-based approaches.**

I. INTRODUCTION

3D models have nowadays become ubiquitous for applications such as games [1], Computer-Aided Design (CAD) [2], molecular biology [3], cultural heritage, etc. The technology innovation in 3D scanners and computer-aided modeling software make it possible to easily construct complete 3D geometry models with relatively low cost and time, which in turn has triggered the rapid enlargement of 3D shape repositories. The latter, along with the explosion of the World Wide Web (WWW), has lead to research in the area of 3D content-based search and retrieval [5] using as query text, sketch and/or 3D object(s).

The existing 3D object retrieval methods can be classified into four main categories: histogram-based, transform-based, graph-based, view-based and, finally, combinations of the above. Histogram-based methods are, in general, easy to implement but usually they are not discriminative enough to make subtle distinctions between classes of shapes. On the other hand, transform-based methods have higher retrieval accuracy. Several transform-based methods [14, 15] usually require a rotation normalization step before the descriptor extraction procedure, while others [10, 6, 7] can achieve rotation invariance. Graph-based methods [11, 12, 13] are more elaborated and complex but they have the potential of encoding geometrical and topological shape properties in a more faithful and intuitive manner.

2D view-based methods [9, 8], consider the 3D shape as a collection of 2D projections taken from canonical viewpoints. Each projection is then described by standard 2D image descriptors like Fourier descriptors [9] or Zernike moments [8]. Ohbuchi et al.[16] recently proposed a view-based 3D model retrieval method based on multi-scale local visual features. The features are extracted from 2D range images of the model viewed from uniformly sampled locations on a view sphere. For each range image, a set of 2D multi-scale local visual features is computed by using the Scale Invariant Feature Transform (SIFT) [17] algorithm. The aforementioned methods have the advantages of being high discriminative, can work for articulated objects, can be very effective for partial matching and can also be beneficial for 2D sketch-based and 2D image-based queries. Their only drawback is that they discard valuable 3D information (due to the self-occlusion).

In this paper, we propose a novel 3D shape retrieval framework supporting multimodal queries (either sketches drawn by a user, or 2D images captured by a user, or 3D objects) by introducing a novel view-based approach able to handle the different types of multimedia data. The proposed Compact Multi-View Descriptor (CMVD) belongs to the category of the 2D view-based approaches and, thus, has the advantages of being high discriminative, can work for articulated objects, can be very effective for partial matching and can support a variety of queries, such as 2D images, hand-drawn sketches and 3D objects. Despite the numerous common advantages, the proposed approach outperforms the existing view-based methods in several aspects:

Compactness. As opposed to the methods presented in [8] and [16], the proposed framework uses significantly less number of different views.

Use of both Binary and Depth Images. The proposed framework generates a set of binary images along with a set of depth images from a 3D object. The depth image can capture more details in a 3D object and increase the shape matching efficiency. This is a significant improvement comparing with the well-known Light Field Descriptor [8], which uses binary images.

Consequently, the method proposed in this paper demonstrates higher retrieval accuracy than the view-based methods presented in [8] and [16].

The rest of the paper is organized as follows: Section 2 analyzes the descriptor extraction procedure. In Section 3, the shape matching framework for both 2D/3D and 3D/3D matching is described. Experimental results evaluating the proposed method and comparing it with other methods are presented in Section 4. Finally, conclusions are drawn in Section 5.

978-0-7695-3662-0/09 $25.00 © 2009 IEEE
DOI 10.1109/CBMI.2009.15

IEEE computer society

II. DESCRIPTOR EXTRACTION

The descriptor extraction procedure can be summarized in the block diagram presented in Figure 1. The input 3D object is a triangulated mesh, in one of the common 3D file formats (VRML, OFF, 3DS, etc.). As a first step, a pose estimation takes place, which includes translation, scaling and rotation of the object. After the pre-processing step, a set of 18 2- dimensional views, taken from the vertices of a bounding 32-hedron is extracted. Both binary (black/white) and depth images are generated. In each of the extracted 2D images, a set of 2D functionals is applied, resulting in a descriptor vector for each view.

A. Pose Estimation

The pose estimation procedure initially involves the translation and scaling of the 3D object. The model is translated so that the center of mass coincides with the center of the coordinate system and scaled in order to lie within a bounding sphere of radius 1.

After translation and scaling, a rotation estimation step is required, since the 3D object may have an arbitrary orientation. In order to achieve the best possible result, a combination of the two dominant rotation estimation methods, PCA [14] and VCA [4], which have been proposed so far in the literature, is utilized.

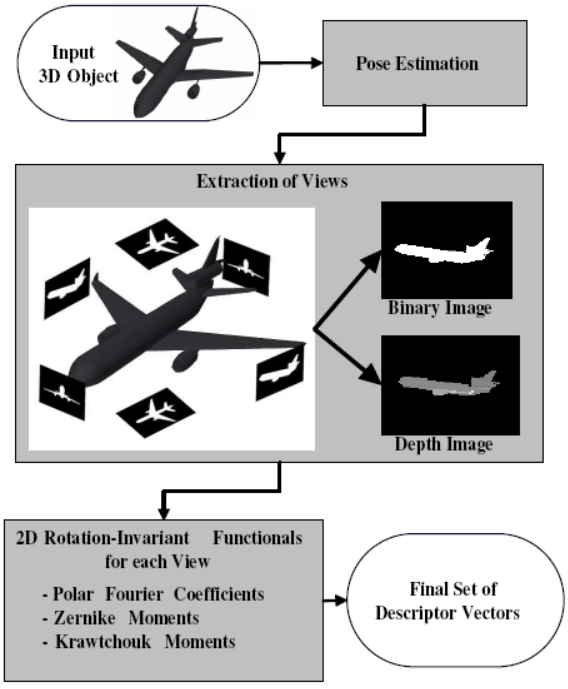

Figure 1. Block Diagram of proposed Descriptor Extraction Method.

The proposed rotation estimation framework leads to the automatic detection of the model's three principal axes with a quite satisfying level of success. However, it does not provide information about the orientation of the principal axes. Taking also into account the fact that the first principal axis may not always be successfully selected among the three principal axes, this leads to a set of $3 \times 8 = 24$ different alignments. The problem of having 24 possible alignments is overcome by appropriately selecting the set of 2D views as well as by introducing an efficient matching method, which will be elaborated in the following sections.

B. A Set of Uniformly Distributed Views

The proposed method is based on the matching of multiple 2D views, which can be extracted from a 3D object by selecting a set of different viewpoints. In order to be uniformly distributed, the viewpoints are chosen to lie at the vertices of a regular polyhedron. The type of the polyhedron and the level of tessellation need to be carefully considered in order to provide the optimal solution. As mentioned in [8], 15 to 20 views can roughly represent the shape of a 3D model. Based on this notion, the 18 vertices of the 32-hedron, which is produced by tessellation of octahedron at the first level, can provide an appropriate set of viewpoints.

In order to render the multi-view images, the camera viewpoints are placed at the 18 vertices of the 32-hedron. Two 2D image types are available: *Binary Images*: the rendered images are only silhouettes, where the pixel values are 1 if the pixel lies inside the model's 2D view and 0 otherwise. *Depth Images*: the pixel intensities are proportional to the distance of the 3D object from each sample point of the corresponding tangential plane.

Although binary images provide an efficient and robust representation of a 2D view, depth images contain more in formation and produce better retrieval results, if appropriately exploited.

C. Computing 2D Functionals on each View

The set of uniformly distributed views, described above, consists of 2D binary images and depth images of size 100×100 pixels. In each image, three rotation-invariant functionals are applied in order to produce the final set of descriptors per view.

Let $f_t(i, j)$ be the 2D image, where $i, j = 0, \ldots, N-1$, $N \times N$ the size of the image, $t = 1, \ldots, N_V$ and N_V the total number of views. The values of $f_t(i, j)$ are either 0 or 1, for the binary images, while in the case of depth images, the values can be any real number between 0 and 1.

2D Polar-Fourier Transform. The Discrete Fourier Transform (DFT) is computed for each $f_t(i, j)$, producing the vectors $FT(k, m)$, where $k, m = 0, \ldots, N-1$. In the DFT, shifts in the spatial domain cause corresponding linear shifts in the phase component. Thus, the DFT magnitude is invariant to circular translation. Therefore, using discrete polar coordinates, rotation is converted to circular translation, which leads to rotation-invariant descriptors. For each $f_t(i, j)$, the first $K \times M$ harmonic amplitudes are considered.

2D Zernike Moments. Zernike moments are defined over a set of complex polynomials which forms a complete orthogonal set over the unit disk and are rotation invariant. The Zernike moments Z_{km} [18], where $k \in N^+$, $|m| \leq k$, are calculated for each $f_t(i, j)$ with spatial dimension $N \times N$, producing a vector of rotation-invariant Zernike descriptors.

2D Krawtchouk Moments. Krawtchouk moments are a set of moments formed by using Krawtchouk polynomials as the basis function set. Following the analysis in [19] and some specifications mentioned in [20], they were computed for each f_t (*i, j*), producing a vector of rotation-invariant Krawtchouk descriptors.

A compact representation of the multi-view descriptor implies a small number of descriptors per view, otherwise the shape matching time would be prohibitive. In an attempt to determine the optimal number of descriptors, we gradually increased the order *k* of Fourier Coefficients, Zernike Moments and Krawtchouk Moments until the performance of each separate 2D functional showed no further improvement. It was found that the optimal order values are k_{FT} = 12, k_{Kraw} = 12 and k_{Zern} = 13, which results in the following descriptor vector dimensions, respectively: N_{FT} = 78, N_{Zern} = 56 and N_{Kraw} = 78. The dimension N_D of the final descriptor vector, per view, is given below:

$$N_D = N_{FT} + N_{Zern} + N_{Kraw} \quad (1)$$

Finally, two types of descriptors are formed: CMVD-Binary that uses binary images and CMVD-Depth that uses depth images.

III. MATCHING METHOD

Similar to existing view-based approaches, the proposed framework measures the similarity between two 3D objects by summing up the similarity from all the corresponding images.

Let $\mathbf{D}_t$ be the descriptor vector of the t^{th} view, which is extracted according to the procedure described in Section II. The dissimilarity metric between a corresponding pair of views of two models *A* and *B* is given by the L1-distance:

$$d_t = \sum_{k=1}^{N_D} \left| D_t^A(k) - D_t^B(k) \right| \quad (2)$$

where N_D is the number of descriptors per view.

A. 3D/3D Matching

Let now *A* and *B* be two 3D models, with descriptor vectors $\mathbf{D}_t^A$ and $\mathbf{D}_t^B$, respectively, where $t = 0, \ldots, N_V$ and N_V the total number of views. The total dissimilarity d between the models *A* and *B* is given by the following equation:

$$d = \sum_{t=1}^{N_V} d_t \quad (3)$$

where d_t is the dissimilarity of the t^{th} view described in (2). Note that the dissimilarity metric does not include matching of all views of model *A* with all views of model *B* ("all-to-all" matching), it includes matching of only the corresponding views (i.e. matching of $View_1^A$ with $View_1^B$, $View_2^A$ with $View_2^B$ and so on). The 3D/3D matching procedure is depicted in Figure 2. The numbering of views has been arbitrarily chosen but it is consistent for every 3D model. This results in a significantly fast matching procedure, however, it requires that rotation normalization provide 100% success, not only in terms of identification of the three principal axes but also in terms of orientation of each axis. Although the combination of PCA and VCA followed in this paper is able to detect the three principal axes, it may confuse the first with the second principal axis, the second with the third, etc. Moreover, it cannot properly identify the orientation of each axis.

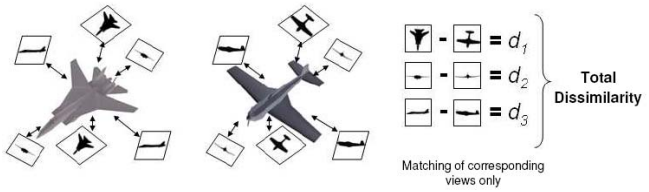

Figure 2. The proposed Similarity Matching Framework. The total dissimilarity between two 3D objects is the sum of the dissimilarities of the corresponding views.

The above inherent limitations of PCA and VCA can be overcome, if, instead of a single set of views, 24 different sets of views are used for the second model. In order to produce these sets of views, the 3D model should be rotated 24 times at intervals of 90 degrees. If the views are taken from the 18 vertices of a 32-hedron, as described in Section II, in all 24 rotations, the viewpoints will always lie at these 18 vertices. Thus, the views (and consequently the 2D descriptors) need to be extracted only once.

The total dissimilarity d between A and B is now modified as:

$$d = \min\left\{d^r\right\} = \min\left\{\sum_{t=1}^{N_V} d_t^r\right\} \quad (4)$$

where r = 1, . . . , 24 is the total number of rotations of the second model, d^r is the dissimilarity of the r^{th} rotation and N_V = 18 is the number of views of the 32-hedron.

B. 2D/3D Matching

Retrieval of 3D models can also be achieved if, instead of a 3D model, a single 2D image is used as query. In order to measure the dissimilarity, the query 2D image is compared to the N_V views of the 3D model and the most similar (to the image) view is selected.

It is obvious that 2D/3D matching cannot be as efficient as 3D-3D matching, since a 2D image is unable to capture the global visual information of an object. However, it is much easier to provide a 2D image as query than a 3D model (e.g. take a photo of an object or draw a sketch).

In Table 1, the average computation times for descriptor extraction and matching procedures are summarized. The times were obtained using a PC with a 2.4 GHz processor and 3GB RAM, running operating system Windows XP.

TABLE I. AVERAGE COMPUTATION TIMES FOR DESCRIPTOR EXTRACTION AND MATCHING PROCEDURES.

Action	Time (msec)
Views Generation	2587
Polar-Fourier Descriptors Extraction	63
Krawtchouk Descriptors Extraction	398
Zernike Descriptors Extraction	811
Matching between 2 models	10

IV. EXPERIMENTS

A. Evaluation of the 3D/3D Matching Method

The proposed method was experimentally evaluated using three different databases. The first one was compiled from the Internet by us and it is called "the ITI database" [21]. It consists of 544 3D models classified in 13 different categories. The second dataset, the "Princeton Shape Benchmark (PSB)", was formed in Princeton University and it consists of 907 3D models classified into 92 categories. Finally, the third dataset, the "Engineering Shape Benchmark (ESB)", contains a total of 867 3D CAD models from the mechanical engineering domain, classified into 44 categories.

To evaluate the proposed method, each 3D model was used as a query object. The retrieval performance was evaluated in terms of "precision" and "recall", where precision is the proportion of the retrieved models that are relevant to the query and recall is the proportion of relevant models in the entire database that are retrieved in the query [14].

The results were compared to those of the following three methods:

- *The Light field descriptor (LFD)*, presented in [8].
- *The Bag-of-Features SIFT algorithm (BF-SIFT)*, which was introduced in [16].
- *DSR*, which is a combination of two view-based methods and a transform-based method [9].

The performance of the first and the third method was computed by using the executables taken from the web pages of the authors, while the results of the second method were directly extracted from the ones presented in [16] and are available only for the PSB database.

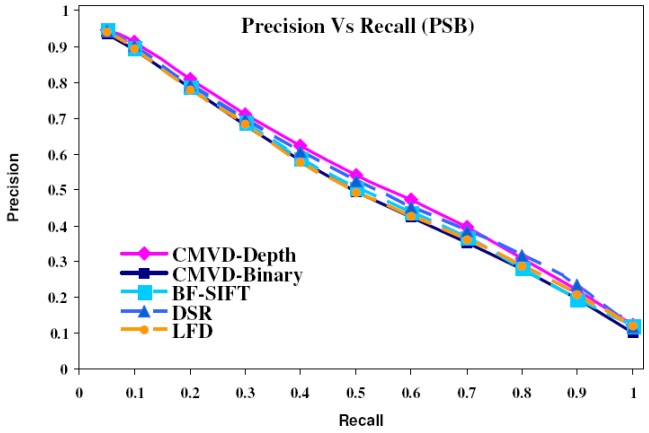

Figure 3. Comparison of the proposed method with LFD, BF-SIFT and DSR in terms of precision-recall, using the PSB database.

Figure 3 contains a numerical precision versus recall comparison of CMVD-Binary and CMVD-Depth with the aforementioned methods using the PSB database. It is clear that the CMVD-Depth descriptor performs better than the CMVD-Binary descriptor. Moreover, CMVD-Depth outperforms all other methods. Similar results are obtained using the ITI and the ESB databases. It is worth to mention that our method is slightly better than DSR, which combines view-based and transform-based information.

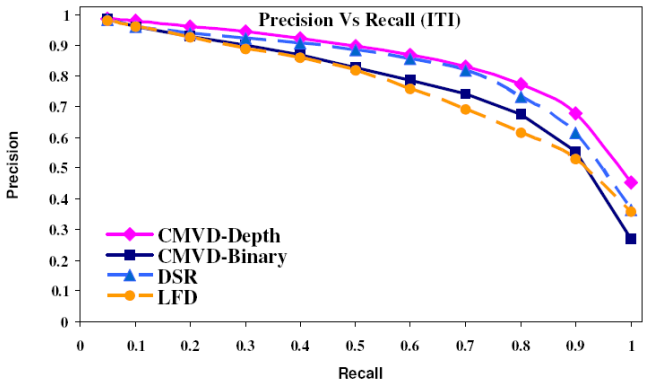

Figure 4. Comparison of the proposed method with LFD and DSR in terms of precision-recall, using the ITI database.

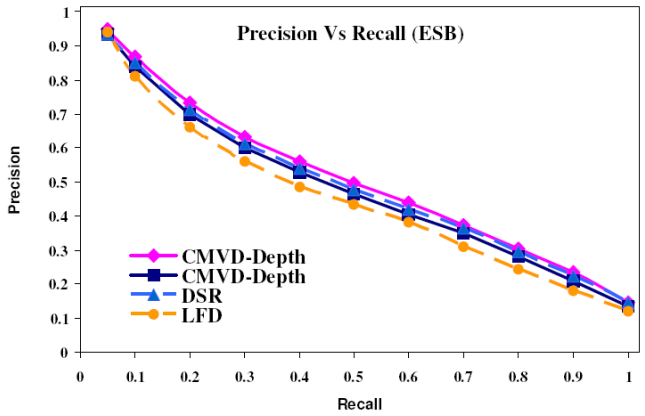

Figure 5. Comparison of the proposed method with LFD and DSR in terms of precision-recall, using the ESB database.

B. Evaluation of the 2D/3D Matching Method

As explained in the previous sections, the proposed framework provides efficient search and retrieval capabilities using only a 2D image or a sketch as a query, when an input 3D model is not available. In this case, the CMVD-Binary descriptor is used, since depth images cannot be easily sketched or retrieved.

In order to support these types of queries, a user-friendly interface has been appropriately designed within the VICTORY project [21]. In Figure 6, screenshots of the VICTORY search and retrieval tool are given. The interface provides a typical drawing tool, allowing the user to easily draw a sketch of the query. Alternatively, the tool provides an extra functionality to load a 2D image and draw a contour of the desired object (Figure 7). This manual segmentation, which separates the query image from the background, is very useful and produces more accurate results. The retrieved 3D objects in the first positions of the rank lists are all similar to the queries, which demonstrates the efficiency of the proposed method to support multiple types of queries.

Figure 6. Retrieved 3D models using as query a hand-drawn sketch

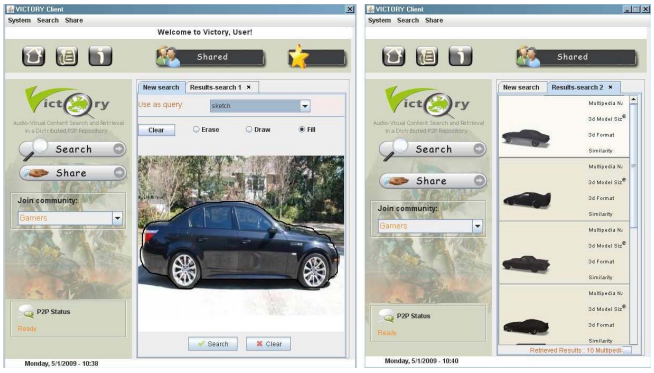

Figure 7. Retrieved 3D models using as query a 2D image

V. CONCLUSIONS

In this paper, a unified framework for 3D object retrieval was presented. The method provides search and retrieval capabilities by supporting multimodal queries (3D objects, 2D images or sketches). The proposed view-based approach creates a compact representation of a 3D object as a set of multiple 2D views (both binary and depth images) taken from uniformly distributed viewpoints. For each view, a set of 2D rotation-invariant shape descriptors is produced. The paper also introduced a novel matching scheme, which calculates the global shape similarity between two 3D models by effectively combining the information extracted from the multi-view representation.

The proposed Compact Multi-View Descriptor (CMVD) was evaluated in terms of retrieval performance using three different databases. The results were compared to those of the best-known retrieval methods in the literature and clearly demonstrate that the proposed method outperforms all others in terms of precision-recall.

ACKNOWLEDGMENT

This work was supported by the EC project VICTORY (http://www.victory-eu.org).

REFERENCES

[1] B. Bustos, D. A. Keim, D. Saupe, T. Schreck, and D. V. Vranic, "Feature-based similarity search in 3D object databases", ACM Comput. Surv., vol. 37, no. 4, pp. 345387, 2005.

[2] S. Jayanti, K. Kalyanaraman, N. Iyer, and K. Ramani, "Developing an engineering shape benchmark for CAD models", Computer-Aided Design, vol. 38, no. 9, pp. 939953, 2006.

[3] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification", IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 3, no. 3, pp. 193207, 2006.

[4] J. Pu, and K. Ramani, "An Approach to Drawing-Like View Generation From 3D Models", In Proc. Of IDETC/CIE 2005, ASME 2005.

[5] B. Bustos, D. Keim, D. Saupe, T. Schreck, "Contentbased 3d object retrieval", IEEE Computer Graphics and Applications 27 (4) pp. 22 27, 2007.

[6] M. Novotni, R. Klein, "3d zernike descriptors for content based shape retrieval", In Proc. of the eighth ACM symposium on Solid modeling and applications, ACM, New York, NY, USA, pp. 216225, 2003.

[7] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors", In Proc. of Symposium on Geometry Processing, Jun. 2003.

[8] D.-Y. Chen, M. Ouhyoung, X.-P. Tian, Y.-T. Shen, , and M. Ouhyoung, "On visual similarity based 3d model retrieval", In Proc. of Eurographics, Granada, Spain, pp. 223232, 2003.

[9] D. Vranic, "3d model retrieval", Ph.D. dissertation, University of Leipzig, 2004.

[10] D.Zarpalas, P.Daras, A.Axenopoulos, D.Tzovaras, and M.G.Strintzis, "3D Model Search and Retrieval Using the Spherical Trace Transform", EURASIP Journal on Advances in Signal Processing Volume 2007, Article ID 23912, 14 pages doi:10.1155/2007/23912, 2007.

[11] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, "Topology matching for fully automatic similarity estimation of 3D shapes", In Proc. of ACM SIGGRAPH, Los Angeles, CA, pp. 203212, USA, 2001.

[12] T. Tung and F. Schmitt, "The augmented multiresolution Reeb graph approach for content-based retrieval of 3D shapes", International Journal of Shape Modeling (IJSM), vol. 11, no. 1, 2005.

[13] A.Mademlis, P.Daras, A.Axenopoulos, D.Tzovaras, and M.G.Strintzis, "Combining Topological and Geometrical Features for Global and Partial 3D Shape Retrieval", IEEE Transactions on Multimedia, Volume 10, Issue 5, pp. 819-831, 2008.

[14] P. Daras, D. Zarpalas, D. Tzovaras, M. G. Strintzis, "Efficient 3-d model search and retrieval using generalized 3-d radon transforms", IEEE Transactions on Multimedia 8 (1), pp. 101114, 2006.

[15] P. Papadakis, I. Pratikakis, S. Perantonis, T. Theoharis, "Efficient 3D shape matching and retrieval using a concrete radialized spherical projection representation", Pattern Recognition 40 (9), pp. 24372452, 2007.

[16] R. Ohbuchi, K. Osada, T. Furuya and T. Banno, "Salient local visual features for shape-based 3D model retrieval", In Proc. og the IEEE International Conference on Shape Modeling and Applications, (SMI 2008), pp. 93-102, 2008.

[17] D.G. Lowe, "Distinctive Image Features from Scale- Invariant Keypoints", International Journal of Computer Vision, 60(2), 2004.

[18] A.P.Vinanco, A.M.Ramirez and F.G.Agustin, "Digital image reconstruction by using Zernike moments", In Proc. of SPIE pp. 281-289, Barcelona, Spain, Sept. 2003.

[19] P.T.Yap, R.Paramesran and S.H.Ong, "Image Analysis by Krawtchouk Moments", IEEE Transactions on Image Processing, Vol. 12, No. 11, pp. 1367-1377, Nov. 2003.

[20] M. R. Teague, "Image analysis via the general theory of moments", Journal of Optical Society of America, vol. 70, pp. 920930, 1979.

[21] The VICTORY 3D Search Engine, http://www.victory-eu.org:8080/victory/results/search.html