Similarity-based retrieval with MPEG-7 3D descriptors: performance evaluation on the Princeton Shape Benchmark

Costantino Grana, Matteo Davolio, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Modena e Reggio Emilia
E-mail: {grana.costantino,davolio.matteo.77,cucchiara.rita}@unimore.it

Abstract

In this work, we describe in detail the new MPEG-7 Perceptual 3D Shape Descriptor and provide a set of tests with different 3D objects databases, mainly with the Princeton Shape Benchmark. With this purpose we created a function library called Retrieval-3D and fixed some bugs of the MPEG-7 eXperimentation Model (XM). We explain how to match the Attributed Relational Graph (ARG) of every 3D model with the modified nested Earth Mover's Distance (mnEMD). Finally we compare our results with the best found in literature, including the first MPEG-7 3D descriptor, i.e. the Shape Spectrum Descriptor.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

3D object description, 3D shape retrieval

1 Introduction

In last years a large amount of audiovisual information is becoming available in digital form on the World Wide Web, and this number intends to grow in the future. The value of information often depends on how easy it can be retrieved; a clarifying metaphor may be a library with a lot of books but without a coherent order: obviously it is not so useful and forces people to lose time in a search task. The question of identifying and managing multimedia content is not just restricted to database retrieval applications such as digital libraries, but extends to areas like broadcast channel selection, multimedia editing, and multimedia directory services. MPEG-7 (ISO/IEC Std. 15938), formally named "Multimedia Content Description Interface", intends to be the answer to this need. MPEG-7 is an ISO/IEC standard developed by the Moving Picture Experts Group, which provides a rich set of standardized tools to describe multimedia content. In this paper, we will examine only the part inherent 3D models descriptions.

In this work, we describe in detail the recently proposed MPEG-7 Perceptual 3D Shape Descriptor, how to decompose a mesh to create an Attributed Relational Graph and the use of the modified nested Earth Mover's Distance to compare two of them. To support our tests, we created a 3D object analysis application called Retrieval-3D and in the process we fixed some bugs of the MPEG-7 eXperimentation Model (XM). Finally we compare its results on the Princeton Shape Benchmark with related work, including the first MPEG-7 3D descriptor, i.e. the Shape Spectrum Descriptor.

2 Perceptual 3D Shape Descriptor

The Perceptual 3D Shape descriptor P3DS (Kim *et alii* 2004) has been recently proposed and adopted as an MPEG-7 standard. P3DS is based on the part-based representation of a given object. The part-based representation is expressed by means of an attributed relational graph (ARG) which can be easily converted into the P3DS descriptor. More specifically, the P3DS descriptor is designed to represent and identify 3-D objects based on the part-based simplified representation using ellipsoidal blobs. Unfortunately the authors recommended using manifold mesh objects with no holes for faster processing and better result. As expected, if the encoder does not produce the part-based representation properly, the retrieval performance would not be good.

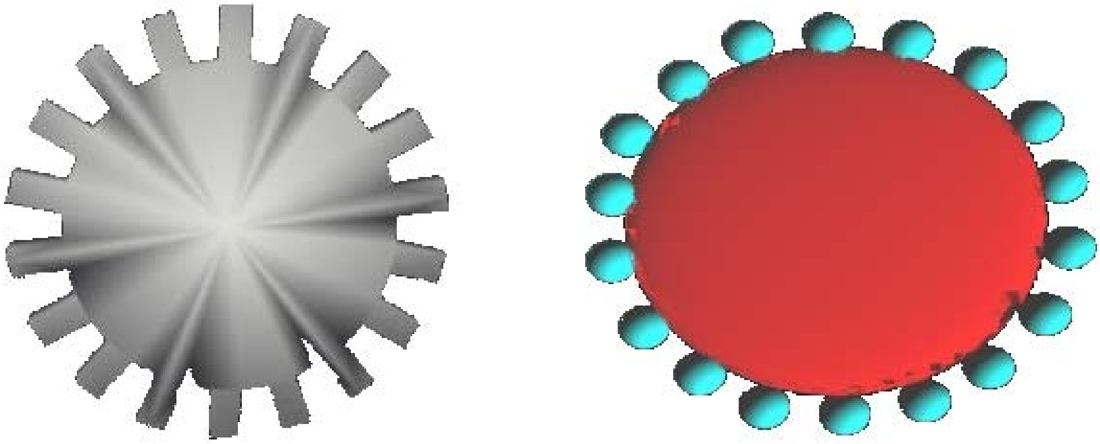


Fig. 1: Gear wheel and it's perceptual descriptor

Unlike the conventional descriptors, the P3DS descriptor supports the functionalities like Query by Sketch and Query by Editing which are very useful in real retrieval system. We now introduced P3DS with a gear wheel 3D model sample: we show it's image, the relative ellipsoidal blobs and the formal syntax of description according to MPEG-7 standard.

2.1 Part subdivision

In order to extract P3DS a decomposition of a given shape is performed, and this matches with that based on human insight. To resolve this problem Kim, Yun and Lee proposed a new shape decomposition method (Kim, Jun and Lee 2005). It is based on a scheme that stands on two psychological rationales: human being recognizes skills and definition of nearly convex shapes. That scheme recursively performs constrained morphological decomposition (CMD) based on the opening operation with ball-shaped structuring elements (SE's) and weighted convexity. Note that the parts to be decomposed are rendered convex or nearly convex by using the ball-shaped SE, since it is convex itself and rotation-invariant. Then, a merging criterion employing the weighted convexity difference, which determines whether adjacent parts are merged or not, is adopted for the sake of providing a compact representation. The proposed scheme consists of three stages, the initial decomposition stage (IDS), the recursive decomposition stage (RDS), and the iterative merging stage (IMS). The aim of all this operation is to provide a representation of the 3D model, that can be easily inserted in an Attributed Relational Graph composed of a few nodes and edges.

2.2 Descriptor extraction

There two equivalent ways to represent a model with the P3DS: with the standard syntax of MPEG-7, like the previous code, or with an Attributed Relational Graph (ARG). An ARG is composed of a few nodes and edges. A node represents a meaningful part of the object with unary attributes, while an edge implies binary relations between nodes. In the descriptor, there are 4 *unary* attributes and 3 *binary* relations which are derived from the geometric relation between the principal axes of the connected nodes. In detail, a *node* is represented by an ellipsoid parameterized by *volume v* , *convexity c*, and two *eccentricity* values e_1 and e_2. More specifically, the convexity is defined as the ratio of the volume in a node to that in its convex hull, and the eccentricity is composed of two coefficients:

$$c = \frac{v_n}{v_{ch}} \quad e_1 = \sqrt{1 - c^2 / a^2} \quad e2 = \sqrt{1 - c^2 / b^2}$$

Where a, b, and c (a ≥ b ≥ c) are the maximum ranges along 1st, 2nd, and 3rd principal axes, respectively. *Edge* features are extracted from the geometric relation between two ellipsoids. They are: *Binary relations* between two nodes, the *distance between centers* of connected ellipsoids and two *angles*. The first angle is between first principal axes of connected ellipsoids and the second one is between second principal axes of them.

```
<?xml version='1.0' encoding='ISO-8859-1' ?>
<Mpeg7 xmlns = "http://www.mpeg7.org/2001/MPEG-7_Schema" xmlns:xsi = "http://www.w3.org/2000/10/XMLSchema-instance">
  <DescriptionUnit xsi:type = "DescriptorCollectionType">
    <Descriptor xsi:type = "Perceptual3DShape">
      <numberOfNodes>1</numberOfNodes>
      <BitsPerAttribute>6</BitsPerAttribute>
      <IsAdjacent>
          1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
          0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0
          0 0 0 0 0 0
          0 0 0 0 0
          0 0 0 0
          0 0 0
          0 0
          0
      </IsAdjacent>
      <Volume>46 0 1 0 0 0 1 0 1 0 0 0 1 1 1 0 0 1 0 </Volume>
      <Center_X>31 2 3 5 6 12 13 20 22 30 32 40 42 48 50 55 57 59 59 </Center_X>
      <Center_Y>31 27 37 17 46 9 54 4 58 2 59 3 58 8 52 15 44 25 35 </Center_Y>
      <Center_Z>31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 </Center_Z>
      <Transform_1>32 58 63 28 55 54 15 42 11 32 32 27 47 17 54 9 58 62 59 </Transform_1>
      <Transform_2>63 50 26 31 9 54 59 57 56 32 63 63 60 60 54 54 49 22 48 </Transform_2>
      <Transform_3>32 33 32 63 32 32 31 48 31 63 32 32 30 30 32 32 35 32 32 </Transform_3>
      <Transform_4>63 30 31 62 31 32 32 25 32 32 32 32 31 33 32 31 29 32 32 </Transform_4>
      <Transform_5>31 31 32 41 31 31 32 16 32 63 32 32 33 33 31 31 28 32 32 </Transform_5>
      <Transform_6>32 63 63 35 63 63 63 59 63 32 63 63 63 63 63 63 63 63 63 </Transform_6>
      <Variance_X>25 4 4 3 4 3 4 5 5 3 3 4 4 4 4 4 4 4 4 </Variance_X>
      <Variance_Y>25 3 3 4 3 3 3 4 3 2 3 3 3 3 3 3 3 3 3 </Variance_Y>
      <Variance_Z>3 3 3 3 3 3 3 3 4 2 2 3 3 3 3 3 3 3 3 </Variance_Z>
      <Convexity>62 61 55 53 55 59 54 59 55 63 63 54 53 57 61 53 56 57 63 </Convexity>
    </Descriptor>
  </DescriptionUnit>
</Mpeg7>
```


Instead the P3DS descriptor contains three *node* attributes, such as *Volume*, *Variance*, and *Convexity*. Next, it contains two *edge* attributes, such as *Center* and *Transform*. Note that Volume, Center, Variance and Convexity are normalized in the interval [0, 1], while Transform is normalized in the interval [-1,1]. Then all are quantizated to 2 rise to the BitsPerAttribute power.

2.3 Modified Nested Earth Mover's Distance

It's not an easy work define a valid distance between two group of ellipsoids that often don't have the same number of elements. Kim, Yun and Lee at this aim present the modified nested Earth Mover's Distance (mnEMD) (Kim, Jun and Lee 2005). However, the ARG matching can be considered as a 2-step procedure: forming a distance matrix with combinatorial differences between each pair of nodes, then the correspondence between nodes in the two ARG's is established based on that matrix by using an appropriate algorithm. So they propose a modified version of nested Earth Mover's Distance algorithm for the ARG matching. This implementation can deal with partially-connected ARG's in an efficient manner and reduce the computational complexity by adopting the notion of imaginary point. The EMD can be introduced intuitively with this example: given two distributions, one can be seen as piles of earth in feature space, the other as a collection of holes in that same space. EMD measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of ground distance. Actually, the mnEMD algorithm has two kinds of EMD's: inner and outer, which correspond to the first step and the second step in the 2-step procedure, respectively. Then, in order to compute both inner and outer EMD's, the weight is identically provided for all nodes as reciprocal of the max number of cluster in the two ARG.

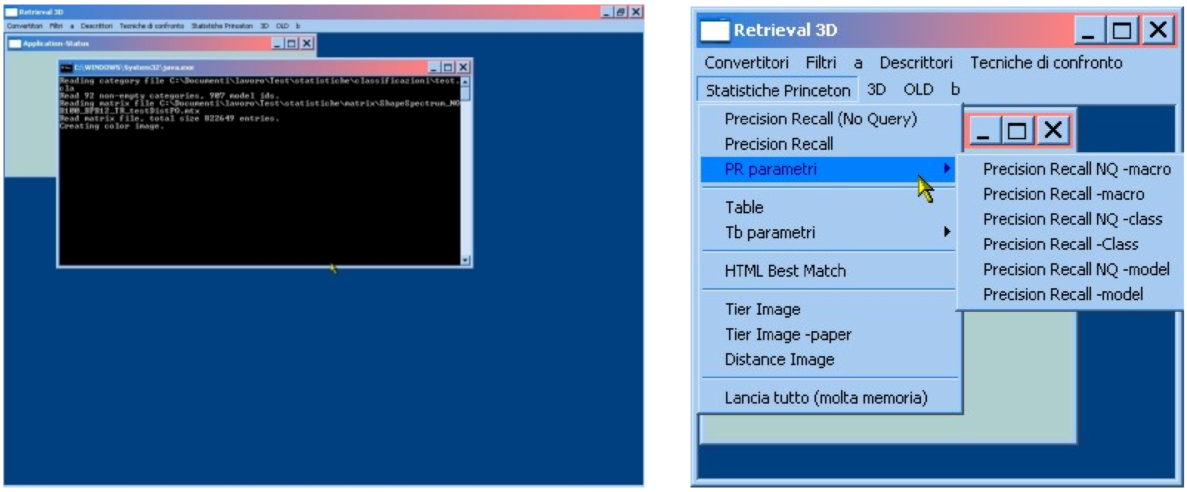


Fig. 2: Retrieval 3D run console application, selection of parameter

2.4 Other descriptors

In previous work, we studied two different implementations of the MPEG-7 3D Shape Spectrum Descriptor (SSD), based on the Shape Index introduced by J.J. Koenderink (Koenderink and van Doorn 1992). One was developed by us (Grana and Cucchiara 2006) and calculates the Principal Curvatures, necessary to know Shape Index, employing the normal at the vertex of the mesh, looking upon the area of the neighboring faces. The other one instead takes the normal at the face, and was suggest by T. Zaharia and F. Prêteux (Zaharia and Prêteux 2001) and implemented in MPEG-7 XM.
In literature D. V. Vranić tested a large number of 3D descriptors with the Princeton Shape Benchmark and other databases. He found that no single descriptor is perfect for every task, but obtained a good mix of performance and precision with an Hybrid Descriptor (Vranić D.V. 2004), obtained by crossbreeding three complementary descriptors: Depth Buffer, Silhouette and Ray-Based Descriptor.
MPEG-7 description tools are implemented in the part 6 of the standard, reference software, and the latest version of this implementation can be obtained as described in (ISO/IEC Std. 15938-6 2001). The reference software, known as the eXperimentation Model (XM), in theory gives the tools to extract the two MPEG-7 3D descriptor from WRL 3D model and to compute the distance between the related features. A few bugs, like memory deallocation and fixed size structures have to be fixed, but the library is overall quite good. In particular most problems may be solved by filtering the mesh to remove duplicated triangles or flat surfaces.

3 A new 3D function library: Retrieval-3D

The developed library allows for mesh filtering removing coincident faces or faces with coincident vertexes; translation of OFF models and non standard WRL models in a WRL file formatted as required by XM. Our WIN32 application provides an integrated environment to execute console applications with the appropriate parameters, obviously making their use simpler, intuitive and fast. Retrieval-3D can execute, other than XM applications, the Princeton Shape Benchmark statistical tools over a set of distance matrices and a classification file, with obviously reduction of the time needed to set all the parameters.

4 The process scheme

In the next page we show the functions and how they work on a generic classified database. With the color orange the software suggest by MPEG-7 eXperimentation Model (XM) is shown, while in violet we have the Princeton console application for statistic purposes. With green color we show the functions supported directly by Retrieval3D, finally with yellow we indicate the file that stored the time necessary for the different process.

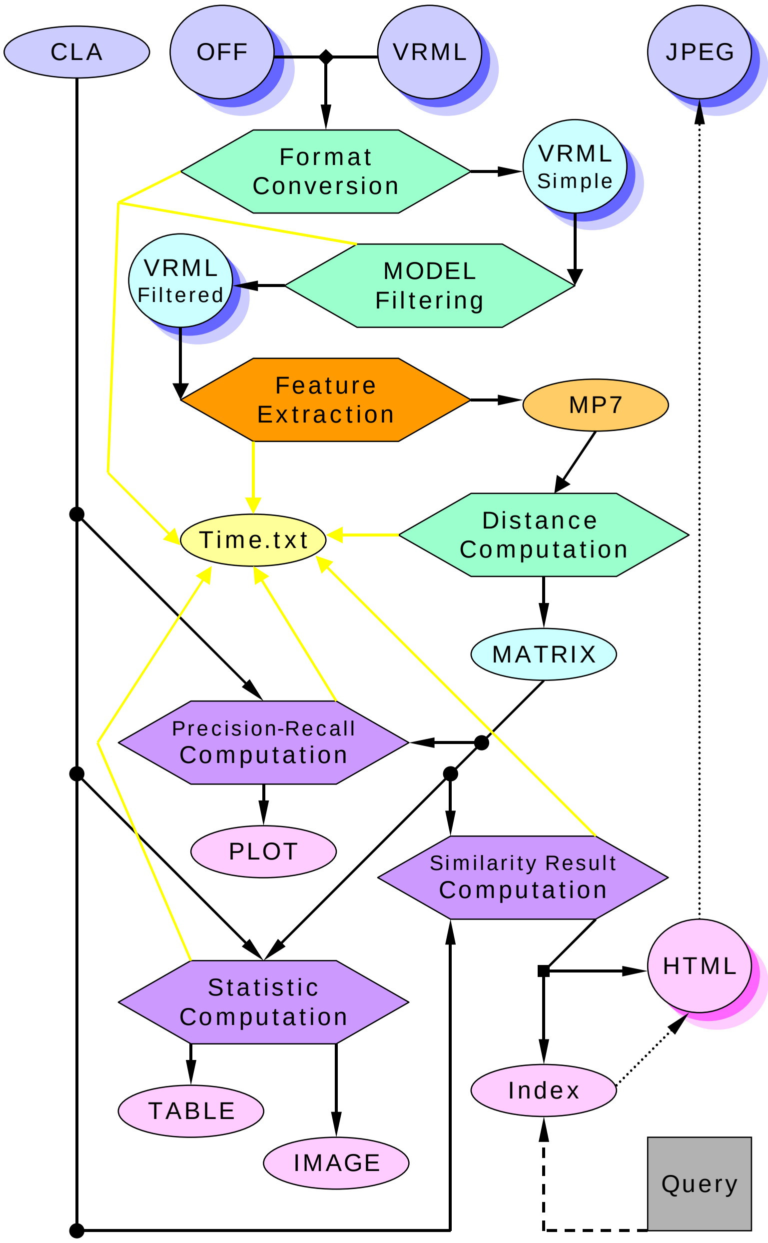


Fig. 3: Retrivial-3D scheme

5 Experimental results

Two series of tests have been conducted. In the first one the input files were extracted from an example VRML database comprising different objects and statues. For each object, different variations were present, constructed by means of geometric transformations and deformations. The database was provided by the Visual Information Processing Lab of the University of Florence. The second test was performed on the Princeton Shape Benchmark (Shilane 2004), a free resource made by Princeton University. This database was created to benchmark different algorithms with the same large number of objects, all classified, by consistent statistic applications. The database is separated in two parts of 907 models each, collected from the web, to allow training of the software on one half, and then verify the result of the best settings on the second part. The polygonal surface geometry of each 3D model is described in an Object File Format (.off) file.
An important note on the P3DS is that the size of the voxel grid is strongly related with the computational complexity and consequently the time required for each object analysis. The evaluation is done by means of the following descriptors: best matches (a web page for each model displaying images of its best matches in rank order. The associated rank and distance value appears below each image, and images of models in the query model's class (hits) are highlighted with a thickened frame), precision-recall plot (a plot describing the

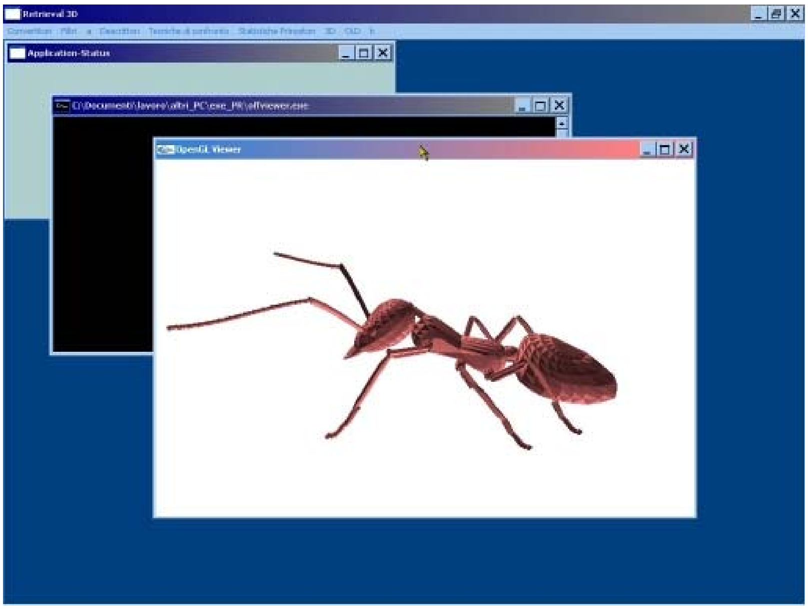


Fig. 4: Retrieval3D run meshview

relationship between precision and recall in a ranked list of matches. For each query model in class C and any number K of top matches, "recall" (the horizontal axis) represents the ratio of models in class C returned within the top K matches, while "precision" (the vertical axis) indicates the ratio of the top K matches that are members of class C), distance image and tier image (the first is an image of the distance matrix where the lightness of each pixel (*i*,*j*) is proportional to the magnitude of the distance between models *i* and *j*. Models are grouped by class along each axis, and lines are added to separate classes, which makes it easy to evaluate patterns in the match results qualitatively, i.e. the optimal result is a set of dark class-sized blocks of pixels along the diagonal indicating that every model matches the models within its class better than ones in other classes. Tier is an image visualizing nearest neighbour, first tier, and second tier matches. This image is often more useful than the distance image because the best matches are clearly shown for every model. The optimal result is a set of black/red, class-sized blocks of pixels along the diagonal indicating that every model matches the models within its class better than ones in other classes).

A table that includes different statistic evaluations is also provided with the following information:

Nearest neighbour: the percentage of the closest matches that belong to the same class as the query.

First-tier and Second-tier: the percentage of models in the query's class that appear within the top K matches, where K depends on the size of the query's class.

E-Measure: a composite measure of the precision and recall for a fixed number of retrieved results. The intuition is that a user of a search engine is more interested in the first page of query results than in later pages.

Discounted Cumulative Gain (DCG): a statistic that 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. Tests have been conducted on a Pentium 4 1.5GHz CPU with 512MB of RAM.

PSB train	Extraction of	descriptor	Calculate distance matrix
	days	Seconds	Seconds
SSD	0,170	14703,984	7,682
P3DS 512	0,064	5520,201	233,145
P3DS 1784	0,205	17719,633	577,320
P3DS 4096	1,479	127792,468	920,354
P3DS 9261	6,233	538530,109	8262,000

Fig. 5: Computational time

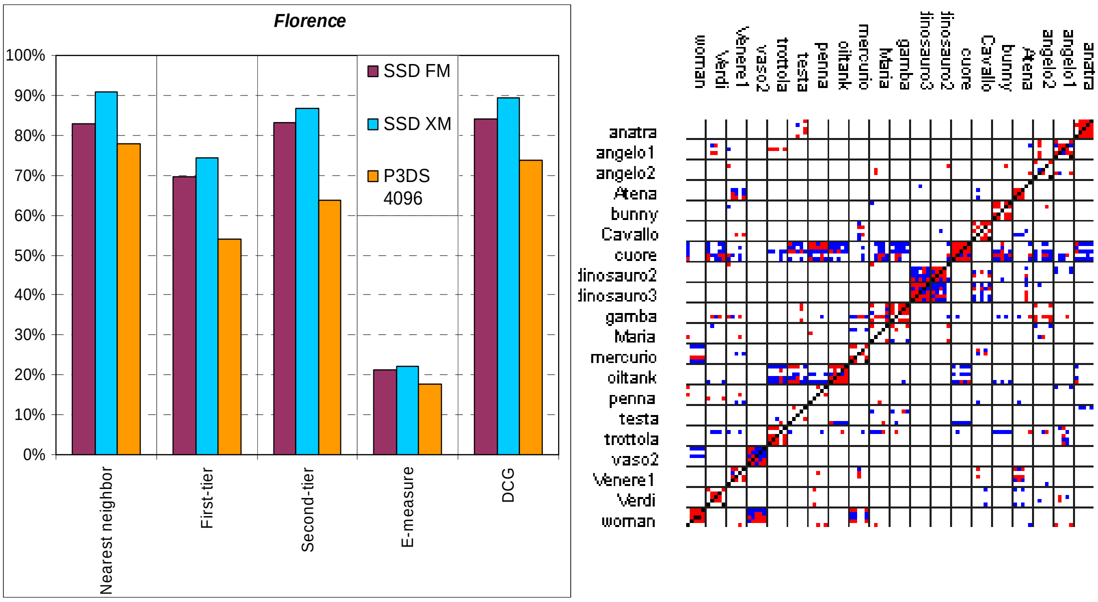


Fig. 6: Table from Florence DB and Tier image for P3DS

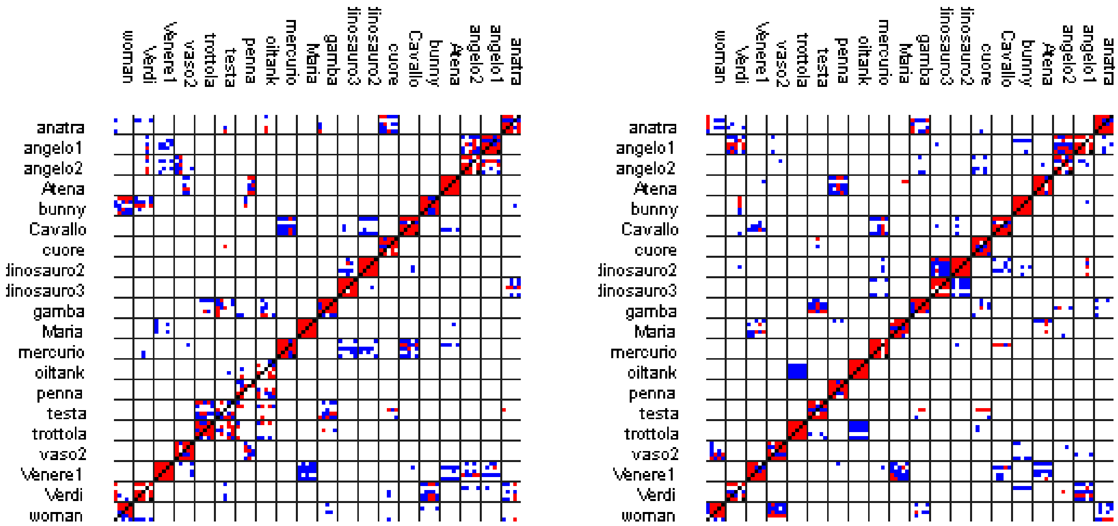


Fig. 7: Tier Image for SSD extract by FindMesh and by XM

All the descriptor have good performance on the Florence small database, which has only 2-manifold models. But the old SSD descriptor is more effective than new P3DS. The XM implementation is better than ours FindMesh.

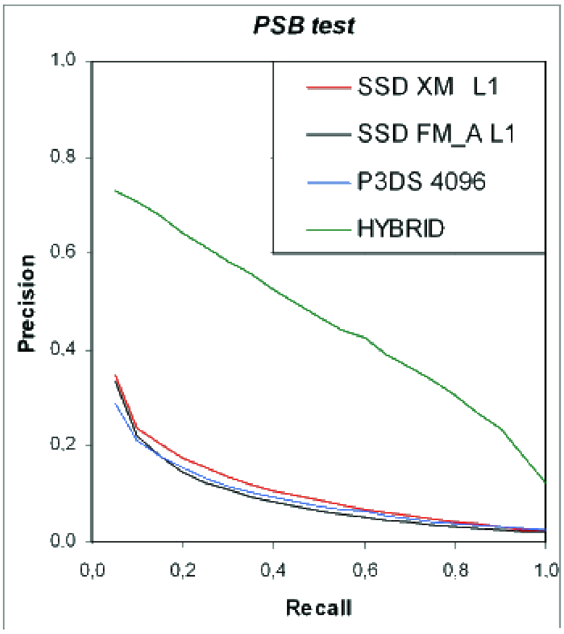


Fig. 8: Vranic's Hybrid [10] VS SSD and P3DS

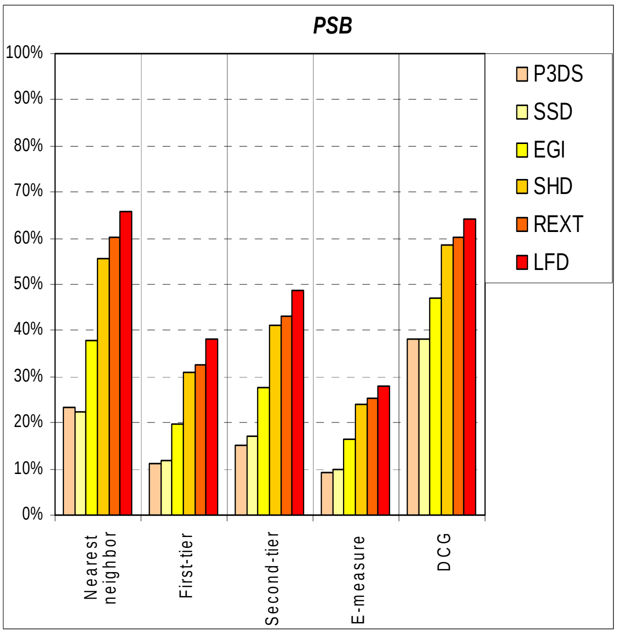


Fig. 9 Table SSD P3DS and other descriptor [7]

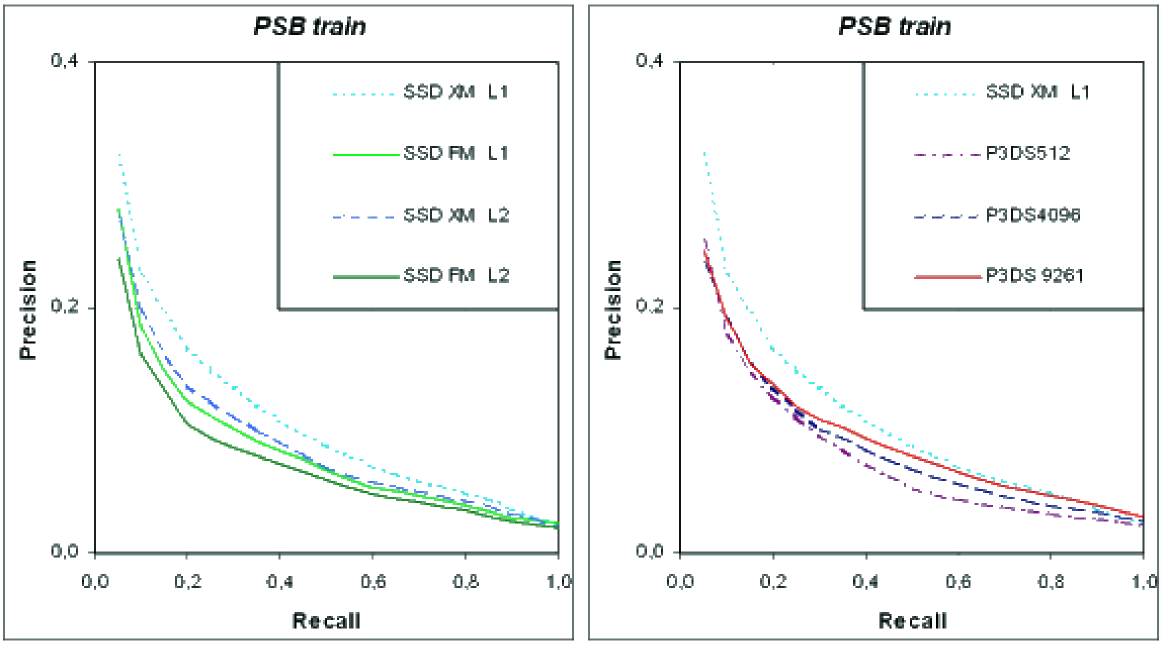
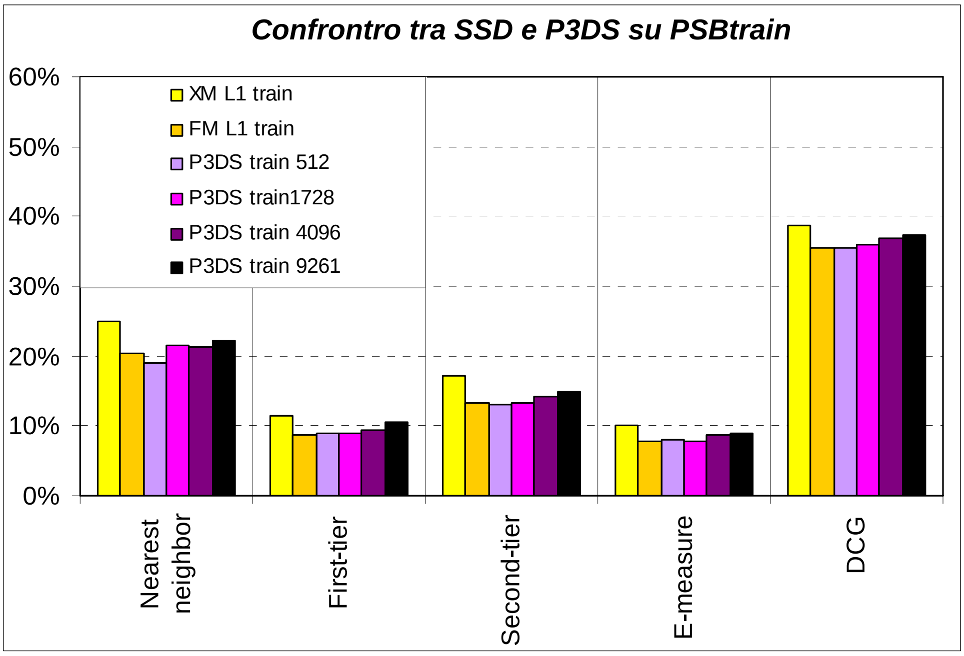


Fig. 10 Precision recall with different distance and different grid dimension

From this Precision Recall Graph and from the table in the next page it's possible to infer that performances of P3DS rise with number of point in the grid used to decompose the model. But the best result still inferior to that of SSD. The best performance for the SSD is obtained using the simple L1 distance and XM implementation. But all this descriptor are clearly less performing than Vranic proposal. Many of the descriptor in literature tested by the authors of Princeton Shape Benchmark are evidently performing better than the choice of MPEG-7. This descriptors are called Extended Gaussian Image EGI, Spherical Harmonic Descriptor SHD, Radialized Spherical Extent Function REXT and Light Field Descriptor LFD.

Descriptors	Nearest neighbour	First-tier	Second-tier	E-measure	DCG
XM L1 train	24,9%	11,5%	17,1%	10,0%	38,6%
FM L1 train	20,3%	8,8%	13,3%	7,8%	35,4%
P3DS train 512	19,0%	8,9%	13,1%	8,0%	35,5%
P3DS train1728	21,6%	9,0%	13,3%	7,9%	35,9%
P3DS train 4096	21,4%	9,5%	14,3%	8,6%	36,9%
P3DS train 9261	22,2%	10,5%	14,9%	8,9%	37,4%
Descriptors	**Nearest neighubour**	**First-tier**	**Second-tier**	**E-measure**	**DCG**
P3DS test 4096	23,3%	11,2%	15,2%	9,2%	38,0%
FM L1 (area) 64 12 test	19,8%	9,0%	13,3%	8,6%	35,8%
XM L1 test	22,5%	11,8%	17,2%	10,0%	38,2%



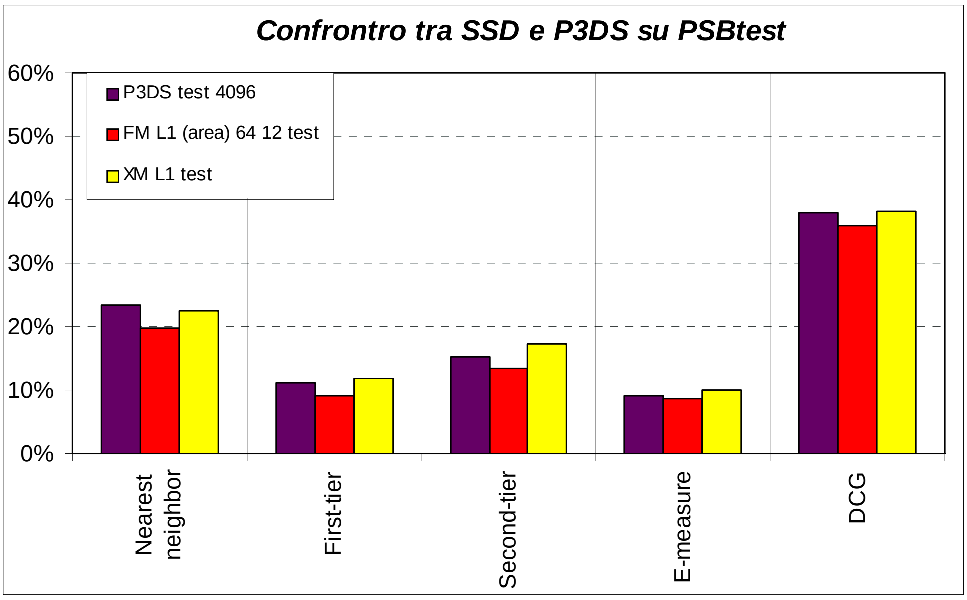


Fig. 11: table with performances of SSD and P3DS descriptor

Best class P3DS XM on PSB train	NN	1st-tier	2nd-tier	E-measure	DCG
body_part___face	82,4%	58,8%	76,1%	50,7%	85,3%
animal___quadruped___apatosaurus	50,0%	50,0%	75,0%	17,1%	69,3%
body_part___brain	42,9%	47,6%	69,0%	25,6%	66,8%
aircraft___balloon_vehicle___dirigible	71,4%	26,2%	31,0%	24,1%	63,2%
animal___arthropod___insect___bee	50,0%	50,0%	50,0%	8,6%	59,9%
chess_piece	52,9%	28,3%	39,3%	26,2%	57,3%
microchip	42,9%	35,7%	38,1%	18,0%	56,5%
aircraft___airplane___fighter_jet	28,0%	15,0%	23,2%	14,7%	55,8%
animal___biped___human	40,0%	13,9%	20,4%	13,3%	55,3%
aircraft___airplane___multi_fuselage	71,4%	26,2%	35,7%	18,8%	55,1%

Best class SSD XM train on PSB train	NN	1st-tier	2nd-tier	E-measure	DCG
ice_cream	66,7%	50,8%	68,2%	34,9%	68,8%
body_part___face	88,2%	30,9%	36,8%	24,5%	66,6%
body_part___skeleton	80,0%	45,0%	70,0%	22,2%	65,9%
animal___biped___human	48,0%	27,3%	42,6%	23,8%	65,6%
aircraft___airplane___fighter_jet	58,0%	23,8%	40,6%	22,1%	65,4%
body_part___brain	57,1%	28,6%	31,0%	14,3%	55,7%
animal___biped___human___human_arms_out	23,8%	20,7%	40,5%	26,7%	53,8%
hat___helmet	70,0%	23,3%	28,9%	17,1%	52,7%
furniture___seat___chair___dining_chair	36,4%	22,7%	30,9%	17,7%	49,5%
animal___arthropod___spider	27,3%	22,7%	31,8%	17,7%	48,5%

Fig. 12: Table of best Class retrieval for P3DS and SSD descriptor

In Fig. 14 it is possible to see that that most of the best result classes are common to both MPEG-7 descriptors, which suggests that the classes in the Princeton benchmark are not equally difficult. The distance list of Fig. 15 and 16 show that generally a bigger grid give best perform, because it captures finer details as the chip connections.

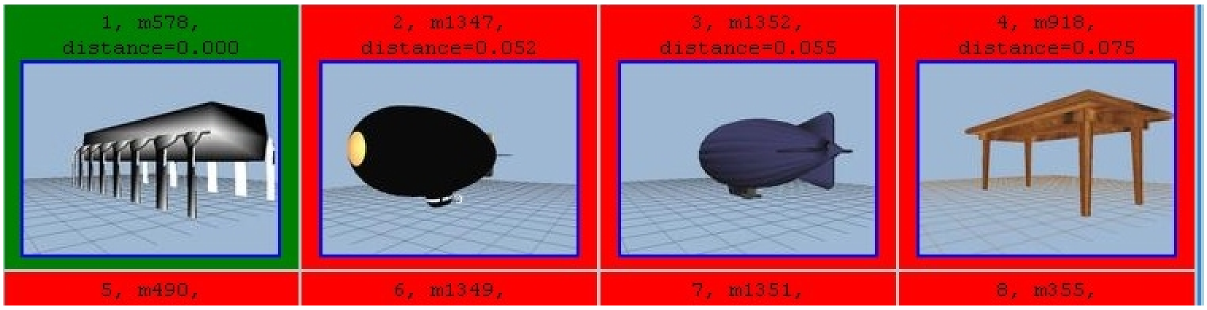


Fig. 13: P3DS descriptor with a grid of 4096 point

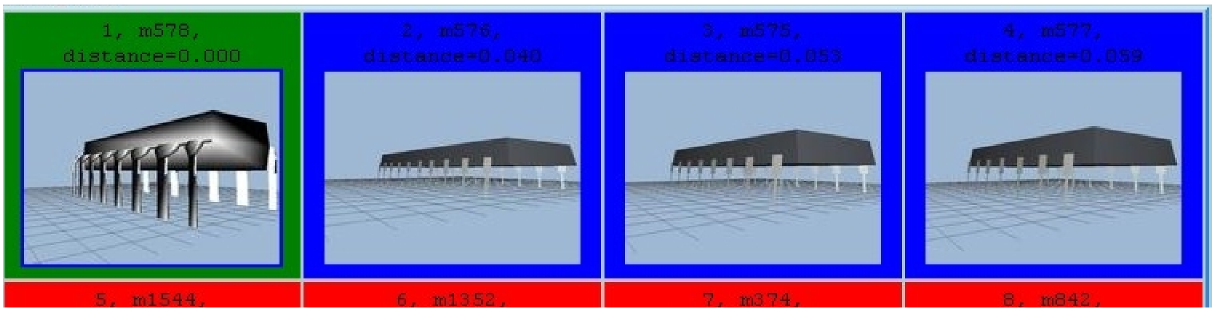


Fig. 14: P3DS descriptor with a grid of 9261 point

6 Conclusion

In this paper, we performed a deep analysis of the MPEG-7 Perceptual 3D Shape Descriptor. As a first conclusion, it is possible to say that XM (ISO/IEC Std. 15938-6) implementation of the P3DS gives results which are even worse then SSD, which uses the implementation provided also in the MPEG-7 eXperimentation Model (XM). Moreover these results do not stand the comparison to other state of the art descriptors.

A second conclusion is that even a bigger grid for the decomposition of the model in part cannot discriminate effectively different P3DS. The use of the new MPEG-7 descriptor seems interesting only for the advantage consisting in the possibility to execute query by sketch due the particular nature of the descriptor. As the authors underline the performance are better for 2-manifold mesh without boundary, but the nature of the 3D model generally found in the web is not like this.

Reference

Duck Hoon Kim, In Kyu Park, Il DongYun and Sang Uk Lee. 2004. A new MPEG-7 standard: Perceptual 3-D shape descriptor. *Lecture Notes In Computer Science (Advances in Multimedia Information Processing - PCM 2004: 5th Pacific Rim Conference on Multimedia),* 3332:238-245

ISO/IEC Std. 15938, MPEG Group, "Information Technology - Multimedia Content Description Interface". *ISO/IEC Final Draft International*, Sydney, July 2001: ISO/MPEG N4285, Standard 15938-1: IT - MCDI - Part 1 "Systems", ISO/MPEG N4288, S. 15938-2: IT - MCDI - Part 2 "Description Definition Language", ISO/MPEG N4358, S. 15938-3 IT: - MCDI - Part 3 "Visual", ISO/MPEG N4224, S. 15938-4 IT: - MCDI - Part 4 "Audio", ISO/MPEG N4242, S. 15938-5 IT: - MCDI - Part 5 "Multimedia Description Schemes".

ISO/IEC Std. 15938-6. MPEG-7 Implementation Studies Group, "Information Technology – Multimedia Content Description Interface - part 6: Reference Software," ISO/IEC FCD 15938-6 / N4006, MPEG-7, Singapore, March 2001.

Grana C., Cucchiara R. 2006. Performance of the MPEG-7 Shape Spectrum Descriptor for 3D objects retrieval. In *Proceedings of the 2nd Italian Research Conference on Digital Library Management Systems*, Padova, Italy.

Kim D. H., Yun I. D. and Lee S. U. 2005. A new shape decomposition scheme for graph-based representation. In *Pattern Recognition*, 38(5):673-689.

Kim D. H., Yun I. D. and Lee S. U. 2004. Interactive 3-D shape retrieval system using the attributed relational graph. In *Proceedings of International Workshop on Multimedia Data and Document Engineering*, 20-31

Koenderink J.J. and van Doorn A.J.. 1992. Surface shape and curvature scales. In *Image and Vision Computing* 10:557-565.

Shilane P., Min P., Kazhdan M. and Funkhouser T. 2004. The Princeton Shape Benchmark. In *Proceedings of the Shape Modeling International*, Genova, Italy, http://shape.cs.princeton.edu/benchmark/

Vranić D.V. 2004. 3D Model Retrieval. *Ph. D. Thesis*, University of Leipzig.

Zaharia T. and Prêteux F. 2001. 3D Shape-based retrieval within the MPEG-7 framework. In *Proceedings of the SPIE Conference on Nonlinear Image Processing and Pattern Analysis XII*, San Jose, CA, USA, 4304:133-145.