Performance of the MPEG-7 Shape Spectrum Descriptor for 3D objects retrieval

Costantino Grana, Rita Cucchiara
Dipartimento di Ingegneria dell'Informazione
Universit degli Studi di Modena e Reggio Emilia
Email: {grana.costantino,cucchiara.rita}@unimore.it

***Abstract*— In this work, we describe in detail the MPEG-7 Shape Spectrum Descriptor and provide a set of tests with different 3D objects databases. To verify if the literature reported low performance of this descriptor were due to the comparison employed, we also used the Earth Movers Distance which allows much more detailed histograms comparisons. Finally we compare our outcomes with the best results in related work.**

I. Introduction

As the World Wide Web becomes ever-present and the total information produced by digital means rapidly grows, techniques which enable quick and focused achievement of the interest data are getting more and more fundamental. In parallel, the role of multimedia is also increasingly important in many real-world applications such as e-commerce, communication, teaching, biomedicine, digital library, and journalism. Content-based retrieval of audiovisual data is a topic that attracts more and more researchers. Instead of relying on textual annotations, which are mostly interactively created, certain low-level features of multimedia objects can be automatically analyzed and used for describing the content. First solutions for content-based retrieval were related to images, movies, and audio sequences, while first techniques for retrieval of 3D-mesh models have recently appeared [1]. To this aim standards have been produced, MPEG-7 being the most ambitious one, ranging from semantics to low level features for static images, audio and video sequences, and also covering one 3D descriptor [2].

In this work, we aim at describing, matching, and retrieve 3D objects by content. In particular to provide a fully compliant MPEG-7 description of 3D models, to test the retrieval performance against a common database, to interact with a web based system, to provide remote querying and mesh comparisons.

II. 3D objects description

For retrieving purposes we elected to use the MPEG-7 3D descriptor [2], the Shape Spectrum Descriptor (SSD), which looks for intrinsic shape description for 3D mesh models and exploits some local attributes of the 3D surface, extracted from the principal curvatures. The mathematical definition of the shape index is:

$$I_p = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_p^1 + k_p^2}{k_p^1 - k_p^2} \qquad (1)$$

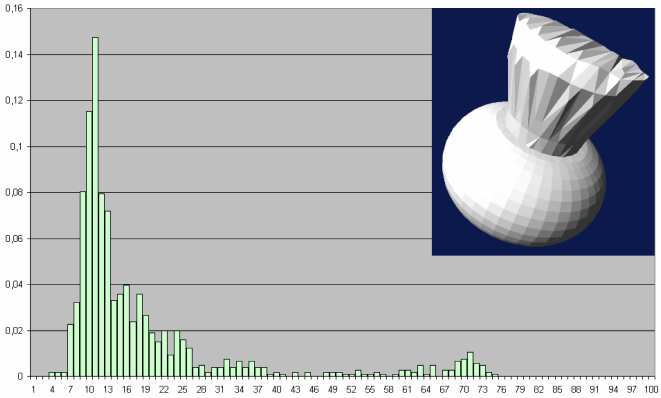

Fig. 1. Example of the Shape Spectrum Descriptor of a vase.

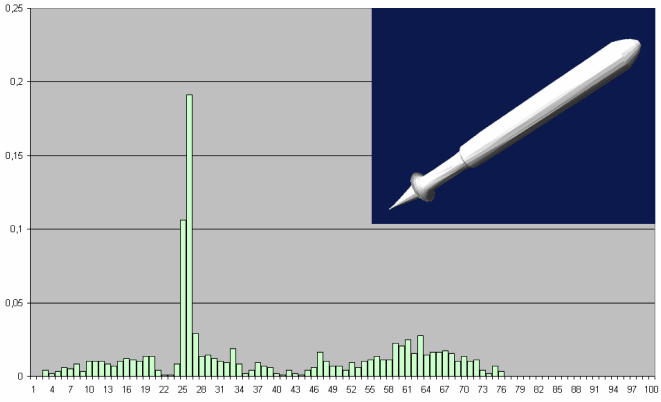

Fig. 2. Example of the Shape Spectrum Descriptor of a pen.

where k_p^1 and k_p^2 are the principal curvatures of the surface, with $k_p^1 \geq k_p^2$. All shapes can be mapped into the interval $I_p \in [0, 1]$ and every distinct shape corresponds to a unique value of I_p, except the planar shape (Fig. 1 and 2). Points on a planar surface have an indeterminate shape index since $k_p^1 = k_p^2 = 0$. Using MPEG-7 to store these description, the number of planar and singular surfaces is stored along the other values (Fig. 3). The shape index of a rigid surface is not only independent of its position and orientation in space, but also independent of its scale and is unitless.

An object's shape can now be characterized quantitatively in

```
<?xml version="1.0" encoding="UTF-8"?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
       xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
    <Description xsi:type="ContentEntityType">
        <MultimediaContent xsi:type="MultimediaType">
            <Multimedia>
               <MediaLocator>
                   <MediaURI>Statua3D.wrl</MediaURI>
               </MediaLocator>
               <MediaSourceDecomposition>
                   <Segment xsi:type="StillRegion3DType">
                        <VisualDescriptor xsi:type="Shape3DType">
                             <Spectrum>12 130 148 22 301</Spectrum>
                             <PlanarSurfaces>0</PlanarSurfaces>
                             <SingularSurfaces>0</SingularSurfaces>
                        </VisualDescriptor>
                   </Segment>
               </MediaSourceDecomposition>
            </Multimedia>
        </MultimediaContent>
    </Description>
</Mpeg7>
```

Fig. 3. MPEG-7 description of a statue with the Shape Spectrum descriptor.

terms of its shape spectrum. It characterizes the shape content of an object by summarizing the area on the surface of an object at each shape index value. The shape spectrum of an object view is obtained from its range data by constructing a histogram of the shape index values -e.g. 0.01 as the bin width, as defined in the default settings for the MPEG-7 standard- and accumulating all the object pixels that fall into each bin. The proposed shape spectrum of a view can be computed from any collection of (x, y, z) points on which the fundamental notions of metric, tangent space, curvature, and natural coordinate frames can be suitably defined.

Since the shape spectrum of a view is constructed directly using the original shape index values computed at each pixel in its image, segmentation of object surfaces is avoided. Nevertheless if surfaces are available, as VRML or MPEG-4 descriptions for instance, the normal direction may be extracted at every vertex in the mesh, by accumulating the normal direction of every surface which touches the vertex, and then normalizing.

Given the normal direction at every vertex, we need a descriptor for the object at that point. At a vertex on a surface, there are infinitely many directions in which the curvature can be measured. The directions are chosen to be in the tangent plane to the surface at the vertex. The minimum and maximum curvatures are called the principal curvatures. The corresponding unit-length tangent vectors are called the principal directions.

For an implicit surface $F(x, y, z) = 0$, a unit-length normal vector field is

$$\mathbf{N}(x, y, z) = \frac{\nabla F(x, y, z)}{|\nabla F(x, y, z)|} \tag{2}$$

The derivative matrix of the normal vector field is

$$D\mathbf{N}(x, y, z) = \left(I - \mathbf{N}(x, y, z)\mathbf{N}(x, y, z)^{\mathrm{T}}\right) \frac{D^2F(x, y, z)}{|\nabla F(x, y, z)|} \tag{3}$$

Intuitively, $\mathbf{N}$ measures the first-order rate of change of F normalized by the length of the gradient. A similar interpretation applies to $D\mathbf{N}$. It measures the second-order rate of change of F normalized by the length of the gradient, but within the tangent space at the point. The matrix $I - \mathbf{N}\mathbf{N}^{\mathrm{T}}$ is the projection matrix onto the tangent plane.

Given a unit-length normal vector $\mathbf{N}$, we may choose unit-length vectors $\mathbf{U}$ and $\mathbf{V}$ such that $\mathbf{U}, \mathbf{V}, \mathbf{N}$ is an orthonormal set. Define the matrix $J = [\mathbf{U}|\mathbf{V}]$, a 3×2 matrix whose columns are the specified vectors. The shape matrix is

$$S = J^{\mathrm{T}}(D\mathbf{N})J \tag{4}$$

and is a 2×2 matrix. The matrix describes locally the shape of the surface in that its eigenvalues are the principal curvatures. If κ is a principal curvature, let $\mathbf{E}$ be the 2×1 eigenvector of S associated with it. The 3×1 principal direction vector is $J\mathbf{E}$.

In the case of a triangle mesh, we can estimate $\mathbf{N}$, and then choose vectors $\mathbf{U}$ and $\mathbf{V}$. We do not know $D\mathbf{N}$, but instead estimate its values from what we do know. The intuition is as follows. All the vertices have estimated surface normals, these normals representing a first-order rate of change of F locally. The rate of change of the vertex normals between a pair of adjacent vertices is a measure of a second-order rate of change from which we can estimate $D\mathbf{N}$.

The rate of change of $\mathbf{N}$ in a specified unit-length tangent direction $\mathbf{W}$ is the directional derivative

$$\mathbf{R} = (D\mathbf{N})\mathbf{W}, \tag{5}$$

a vector-valued quantity. If we can estimate $\mathbf{R}$ at vertices, then we can estimate $D\mathbf{N}$. Let $\mathbf{P}_i$ and $\mathbf{P}_j$ be two adjacent

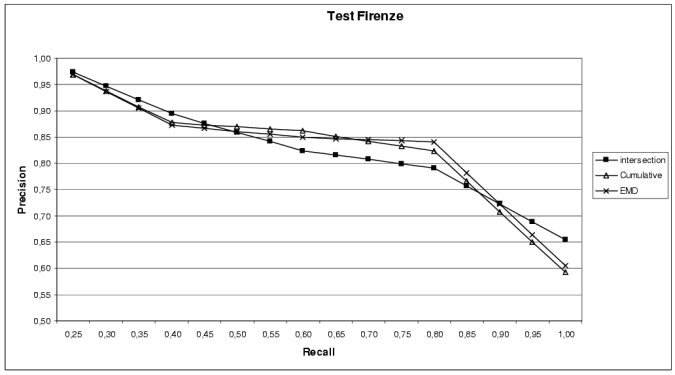

Fig. 4. Results obtained with a first *easy* database, provided by the University of Florence.

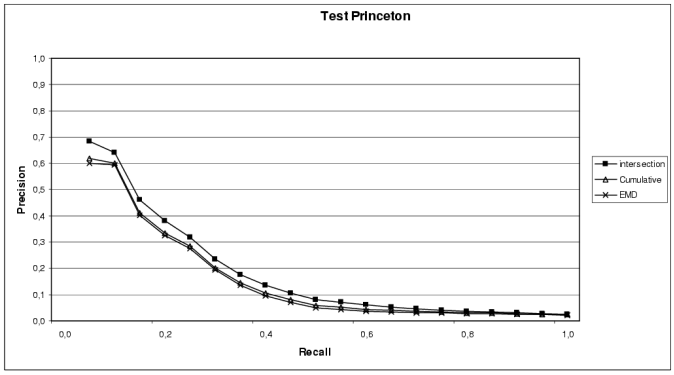

Fig. 5. Results obtained with the Princeton database. It is interesting noting that the results are better that those obtained in Fig. 6

vertices with corresponding normals $\mathbf{N}_i$ and $\mathbf{N}_j$. Let us work in the tangent space of $\mathbf{P}_i$ to make measurements about how the normal vector changes. The tangent space is the plane containing $\mathbf{P}_i$ and that is perpendicular to $\mathbf{N}_i$. The not necessarily unit-length direction to the adjacent vertex is $\mathbf{D}_{ji} = \mathbf{P}_j - \mathbf{P}_i$, but is not necessarily in the tangent plane. We can project it onto the plane and normalize it by

$$\begin{array}{l} \mathbf{P}_{ji} = \mathbf{D}_{ji} - (\mathbf{N}_i \cdot \mathbf{D}_{ji})\mathbf{N}_i \\ \mathbf{W}_{ji} = \frac{\mathbf{P}_{ji}}{|\mathbf{P}_{ji}|} \end{array} \tag{6}$$

The difference of normals is

$$\mathbf{R}_{ji} = \mathbf{N}_j - \mathbf{N}_i \tag{7}$$

The approximation to the rate of change of vertex normals is

$$\mathbf{R}_{ji} = (D\mathbf{N})\mathbf{W}_{ji} \tag{8}$$

where $D\mathbf{N}$ is the derivative matrix of the normal vector field at the position $\mathbf{V}_i$. The problem is still that we do not know what is $D\mathbf{N}$. Think of $D\mathbf{N}$ as a collection of 9 unknown values. Any approximate rate of change may be used to help establish the values of the unknowns. For a closed manifold triangle mesh, each vertex has at least 3 edges emanating from it. Therefore the corresponding approximate rate of change equations give us at least 9 pieces of information about the unknowns. We can formulate construction of the unknowns as a least-squares problem. Suppose that the vertex $\mathbf{P}$ has n adjacent vertices. Each adjacent vertex leads to a unit-length direction vector $\mathbf{W}_j$ in the tangent plane of $\mathbf{P}$ and each adjacent vertex normal leads to a difference normal $\mathbf{R}_j$, $1 \leq j \leq n$. The approximate rate of change equations can be written jointly as

$$(D\mathbf{N})W = D\mathbf{N}[\mathbf{W}_1|\ldots|\mathbf{W}_n] = [\mathbf{R}_1|\ldots|\mathbf{R}_n] = R \tag{9}$$

where W is a $3 \times n$ matrix whose columns are the specified unit-length direction vectors and R is a $3 \times n$ matrix whose columns are the specified normal differences. Multiplying by W^{T} leads to

$$(D\mathbf{N})(WW^{\mathrm{T}}) = RW^{\mathrm{T}} \tag{10}$$

The matrix WW^{T} is a 3×3 matrix. If W were to have rank 3, then WW^{T} would be invertible in which case we could solve for $D\mathbf{N}$ directly. The problem is that the columns of W are all tangent vectors at $\mathbf{P}$ and lie in the same plane; consequently W has rank 2 and we cannot invert WW^{T}.

To remedy the problem, we can additionally stipulate that the normal vectors do not change as you move in the normal direction. This is an artificial constraint because you cannot leave the mesh when making surface measurements, but it makes intuitive sense. The mathematical condition for this is

$$(D\mathbf{N})\mathbf{N} = 0 \tag{11}$$

Note that this is not the condition implied by having unit-length normals. That is, since $\mathbf{N} \cdot \mathbf{N} = 1$ for all $\mathbf{N}$, we can take the derivative and obtain

$$0 = D(\mathbf{N} \cdot \mathbf{N}) = \mathbf{N}^{\mathrm{T}} D\mathbf{N} + (D\mathbf{N})^{\mathrm{T}}\mathbf{N} = 2\mathbf{N}^{\mathrm{T}} D\mathbf{N} \tag{12}$$

This equation is different than the previous one since $D\mathbf{N}$ is not a symmetric matrix. The artificial constraint can be included in the set of equations that determine $D\mathbf{N}$ as

$$(D\mathbf{N})W = D\mathbf{N}[\mathbf{W}_1|\ldots|\mathbf{W}_n|\mathbf{N}] = [\mathbf{R}_1|\ldots|\mathbf{R}_n|0] = R \tag{13}$$

where W and R are now $3 \times (n+1)$ matrices. The matrix W now has rank 3, so W is invertible. We can solve for the derivative matrix:

$$D\mathbf{N} = RW^{\mathrm{T}}(WW^{\mathrm{T}}) - 1 \tag{14}$$

and use any mathematical library approach to ensure numerical stability.

III. 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 University of Florence. The second test was performed on the Princeton Shape Benchmark [3] which contains a database of 1,814 classified 3D models collected from 293 different web domains. For each 3D model, there is an Object File Format (.off) file with the polygonal surface

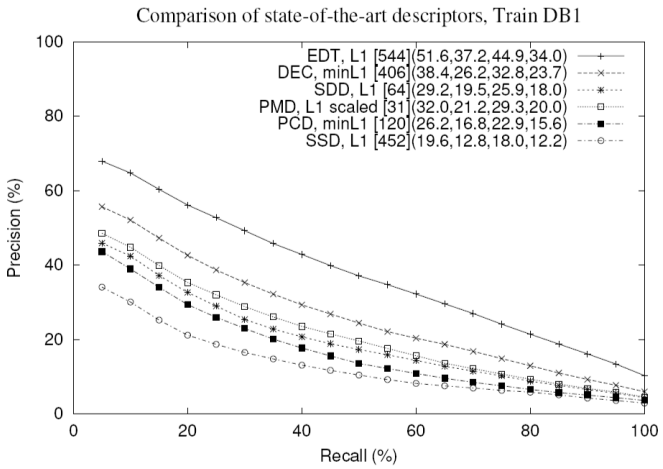

Fig. 6. Results presented in literature [5]. It is possible to see that the MPEG-7 descriptor (SSD) provides the worst results with respect to all other proposals.

geometry of the model, a textual information file containing meta-data for the model, and a JPEG image file containing a thumbnail view of the model.

For comparing the Shape Spectrum we employed three different histogram matching metrics: the histogram intersection, the cumulative histogram and the Earth Movers Distance [4]. This last approach sees the comparison as a Transportation Problem: it tries to find the least expansive flow of goods from supplier to consumer (each consumer-supplier chain has an associated cost to deliver one unit of goods). For every distance measure we employed a Training Set to obtain the best parameters and the results are shown on the Test Set. In particular for the SSD granularity the best results (according with [5]) are provided by 64 bins (the default value in MPEG-7 is 100) and no significant performance loss is observed by reducing the number of bits per bin up from 12 up to 8.

In our tests we found the surprising result that the simple Histogram Intersection provides the best results, with respect to both other measures. This is not apparent from the Florence Test (Fig. 4), but becomes evident in the Princeton Test (Fig. 5). Another interesting outcome was the comparison of the MPEG-7 descriptor results with those reported in literature: the SSD always provided the worst performances, but on the same database we obtained much better precision at low recall rates. Fig. 6 shows the results obtained in [5] on the Princeton Database, after optimization of the different parameters.

IV. Conclusion

In this paper, an in depth analisis of the MPEG-7 Shape Spectrum Descriptor was performed. As a first conclusion, it is possible to say that our implementation of the SSD gives slightly better results with respect to the one of [5], which uses the implementation provided in [6] known as the MPEG-7 eXperimentation Model (XM), in particular at low recall rates. Nevertheless the results do not stand the comparison to other state of the art descriptors.

A second conclusion is that even an accurate and definitely very complex histogram comparison metric cannot discriminate effectively different SSDs. The use of histogram intersection, the standard L_1 norm, is perfectly adequate to the retrival aims.

The use of the MPEG-7 SSD is definitely a bad choice, and should be avoided by all means. The main problem here is that no other possibility are offered by the standard, neither it is possible to provide a personalized descriptor within the MPEG-7 framework. So we are left with the problem of which descriptor to use and how to communicate our choice/choices to the other systems.

Acknowledgment

The authors would like to thank Davolio Matteo for his contribution to the code fixing and the Visual Information Processing Lab of the University of Florence for providing their 3D model database.

References

[1] D. V. Vranic, “3d-shape descriptors: An overview,” in *Proceedings of Graphiktag 2003*, Frankfurt am Main, Germany, Sept. 29, 2003, pp. 57–66. [Online]. Available: http://www.gdv.informatik.uni-frankfurt.de/events/Proceeding_gesamt.pdf

[2] *Information Technology Multimedia Content Description Interface Part 5 Multimedia Description Schemes*, ISO/IEC Std. 15 938-5, July 2001.

[3] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, “The princeton shape benchmark,” in *Proceedings of International Conference on Shape Modeling and Applications 2004*, Genova, Italy, June 7–9, 2004, pp. 167–178. [Online]. Available: http://shape.cs.princeton.edu/benchmark/benchmark.pdf

[4] Y. Rubner, C. Tomasi, and L. J. Guibas, “A metric for distributions with applications to image databases,” in *Proceedings of the IEEE International Conference on Computer Vision (ICIP98)*, Bombay, India, Jan. 1998, pp. 59–66. [Online]. Available: http://vision.stanford.edu/public/publication/rubner/rubnerIccv98.ps.gz

[5] D. V. Vranic, “3d model retrieval,” Ph.D. dissertation, University of Leipzig, 2004. [Online]. Available: http://www.informatik.uni-leipzig.de/~vranic/my_papers/dv-phd04.pdf

[6] MPEG-7 Implementation Studies Group, *Information Technology - Multimedia Content Description Interface - part 6: Reference Software*, ISO/IEC Std. 15 938-6 / N4006, Mar. 2001.