The MPEG-7 Standard and the Content-based Management of Three-dimensional Data: A Case Study

Eric Paquet and Marc Rioux
Visual Information Technology
Institute for Information Technology
National Research Council
Ottawa (Ontario) K1A 0R6 Canada
E-mail: eric.paquet@iit.nrc.ca, marc.rioux@iit.nrc.ca

Abstract

Content-based management of three-dimensional data in the framework of the MPEG-7 standard is considered. An overview on MPEG-7 and three-dimensional data is presented. Descriptors and description schemes suitable for three-dimensional data are introduced and their integration in the framework of MPEG-7 is detailed. The shape descriptors are based on cords, wavelet transform and three-dimensional statistical moments. Implications of the normative and the non-normative parts of the standard are analyzed and their relations are considered. The descriptors and description schemes are evaluated according to MPEG-7 criteria. Experimental results indicating the potential of the descriptors are presented. Some considerations on dynamic scenes are introduced.

1. Introduction

More and more multimedia information is available around the world. The number of users and the amount of information are progressing at a very fast rate. To be useful that information has to be filtered and retrieved. In order to perform a search it is necessary to have a description of the content. That description can be a set of keywords, a structured text or some more abstract representation.

If the description of the content is not standardized it is very difficult to retrieve the content because the descriptions involved are not necessarily compatible and their distribution can be limited to a restricted subset of users. In order to overcome this problem most search engines retrieve the documents and describe them with keywords by analyzing the textual information that surrounds the multimedia information. That procedure has two major drawbacks: it is time-consuming and the keywords extracted do not necessarily match the multimedia content. Actually the most qualified person to describe the content is the provider. In order to share its knowledge the provider needs a representation that can be understood by anybody accessing the content. This is why a standard representation of the content is needed. The aim of MPEG-7 is to become such a standard.

2. MPEG-7 and Three-dimensional Data

MPEG-7 [1] is a member of the MPEG family which itself is joint technical committee of the ISO and IEC: ISO/IEC JTC1/SC29/WG11. Contrary to MPEG 1, 2 and 4, MPEG-7 does not address coding but content description. As a matter of fact MPEG-7 is formally known as Multimedia Content Description Interface. What is standardized by MPEG-7 is called the normative part while what is not standardized is known as the non-normative part. The normative part of MPEG-7 is the description. The feature extraction and the search engine belong to the non-normative part. They are not standardized for two main raisons: they are not necessary for interoperability and MPEG-7 does not want to close the door to future technical improvements.

MPEG-7 is made out of three main elements: the descriptor D, the description scheme DS and the description definition language DDL. The descriptor is the representation of a feature: a feature being a distinctive or characteristic part of the data. The descriptor may be composite. The description scheme is formed by the combination of a plurality of descriptors and description schemes. The description scheme specified the structure and the relation between the descriptors. Finally the DDL is the language used to specify the description scheme.

They are many three-dimensional objects and scenes on the Web. Most of them are in the VRML format including the ISO standard VRML 97. Along with VRML other formats are emerging or are under development like Metastream from Metacreation, VRML C from the VRML Consortium, Fahrenheit and AAF from Microsoft. As a matter of fact there are more than 40 common commercial formats in use. Some of them like DXF and ACIS are mostly used for CAD. Among the most important applications of three-dimensional data are industrial design, inspection, virtual reality, catalogues, collections and medical applications. In most applications an important number of files is involved. In order to search efficiently those data it is important to have a description suited for them. Interesting works on this subject can be found in [2-6]. The present paper addresses the description of the scale, shape and color of three-dimensional objects in the context of MPEG-7.

3. Description of Three-dimensional Objects in the Framework of MPEG-7

One of the main characteristics of three-dimensional data is their physical dimensions or scale. In order to describe the scale it is proposed to utilize three descriptors: the volume, a bounding box and a bounding sphere. The volume is simply a float value representing the volume of the object. The bounding box is set of three floats representing the dimensions of the smallest box that can contain the object. The bounding sphere is a float value corresponding to the radius of the smallest sphere that can contain the object. The three descriptors form a DS describing the scale of the object. The scale is feature that can be easily used to rapidly filter objects for a specific query. In the DS a field is used to specify the unit used for the measurement: meters, inches, etc.

The color distribution is represented by a set of three histograms corresponding to the red, green and blue component of the color. This representation has been retained because commonly used in three-dimensional formats. The histograms form a DS for the color distribution. It has to be pointed out that each histogram is normalized in order to make their comparison easier. Each histogram has a header made out of a single number representing the number of channels.

Three-dimensional shape is of particular interest. Since the extraction is not part of the MPEG-7 standard, a special care must be taken in order that the description depends as little as possible on the extraction. In order to describe the geometrical shape we use the concept of a cord. A cord is a vector that goes from the center of mass of the object to the center of mass of a given triangle belonging to the surface of the object. In order to define the orientation of a cord we use a reference frame that does not depend on the orientation of the object. The reference frame is defined as the eigen vectors of the tensor of inertia of the object. Assuming a triangular decomposition, the tensor of inertia is defined by:

$$I = [I_{qr}] = \left[\frac{1}{n} \sum_{i=1}^{n} \left[m_i (q_i - q_{CM})(r_i - r_{CM}) \right] \right] \quad (1)$$

Where m_i is the *mass* of a triangle and q and r are the Cartesian coordinates. Each axis is identified by its eigen value. The axes are labeled one, two and three by descending order of their corresponding eigen values.

$$\left[I\vec{a}_i = \lambda_i \vec{a}_i \right]_{i=1,2,3} \quad (2)$$

The orientation of the cord is completely determined by the two angles between the cord and the first two reference axes. The norms of the cords are normalized: that means that the cord distribution does not depend on the scale of the object. The angles and the norm define uniquely the cord which, is our MPEG-7 feature. Now we are interested in the distribution of those cords so we define the cord distribution as a set of three histograms: the first histogram represents the distribution of the first angle, the second histograms the distribution of the second angle and the third histogram the distribution of the norms. Each histogram is a descriptor and the set formed by them is a DS. Each histogram has a header made out of a single number representing the number of channels. Depending on the number of channels the resulting DS can be very compact. The size of the histogram is also dictated by the precision of the representation and by the discrimination capability that is needed.

The behavior of a cord can be better understood by considering a regular pyramid and a step pyramid. Most people agree that they belong to the same category. If normal vectors would be used to represent the pyramids, five directions would characterize the regular pyramid while six directions would characterize the step pyramid. So they would be classified as two distinct objects. If a cord representation is used the histograms corresponding to the regular pyramid and to the step pyramid are much more similar. Consequently a cord can be understood as a slow varying normal vector. Some results have been published recently [7] and some are provided here. For detailed results the reader is invited to try our demo located at http://cook.iitsg.nrc.ca:8800/Nefertiti.

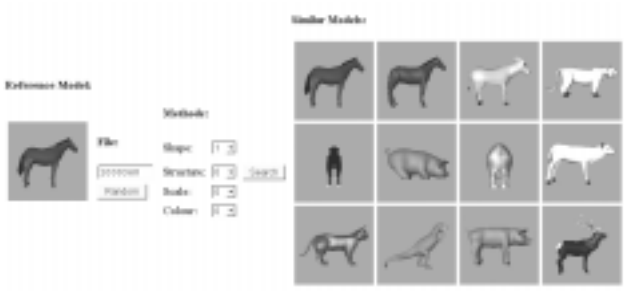

Figure 1. Query for a horse.

In addition to the cord, we propose another descriptors based on a wavelet representation. This is because in addition to be bounded by a surface a three-dimensional object is also a volume. In order to analyzed the volume we use a three-dimensional wavelet transform. Let us review the procedure. For our purpose we use DAU4 wavelets which have two vanishing moments. The $N \times N$ matrix corresponding to that transform is

$$W = \begin{bmatrix} c_0 & c_1 & c_2 & c_3 & & & & & & & & \\ c_3 & -c_2 & c_1 & -c_0 & & & & & & & & \\ & & c_0 & c_1 & c_2 & c_3 & & & & & & \\ & & c_3 & -c_2 & c_1 & -c_0 & & & & & & \\ & & & & & & \ddots & & & & & \\ & & & & & & & c_0 & c_1 & c_2 & c_3 & \\ & & & & & & & c_3 & -c_2 & c_1 & -c_0 & \\ c_2 & c_3 & & & & & & & & c_0 & c_1 & \\ c_1 & -c_0 & & & & & & & & c_3 & -c_2 & \end{bmatrix} \quad (3)$$

The wavelet coefficients are obtained by applying the matrix W on the three axes defined by the tensor of inertia. We use those axes because the wavelet transform is not translation and rotation invariant. In order to apply the transform to the object the latter has to be binarized by using a voxel representation. The wavelet coefficients represent a tremendous amount of information. In order to reduce it, we compute a set of statistical moments for each level of resolution of the transform. For each moment, a histogram of the distribution of the moment is built. The channels correspond to the level of detail and the amplitude to the normalized moments. This histogram is the descriptor of the object. The DS is formed by the encapsulation of the histograms. A field in the DS indicates the number of moments used which is also the number of histograms. The DS has also a field to indicate the number of levels of detail corresponding to the transform.

The last descriptor is based on three-dimensional statistical moments. The three-dimensional statistical moment M_{qrs} is defined as

$$M_{qrs} = \sum_{i=1}^{n} m_i \left(x_i - x_{cm}\right)^q \left(y_i - y_{cm}\right)^r \left(z_i - z_{cm}\right)^s \quad (4)$$

The statistical moments are not rotation invariant. In order to solve this problem they are computed in the same reference of the wavelet descriptors. The order of the moments is related to the level of detail. Three numbers indicating the order and a float indicating the value of the moment for the descriptor. The DS is made out of a field indicating the number of moments and the corresponding moment descriptors. Overviews of all the descriptors and DS that we have presented can be seen in the following figure; a pseudo-code is used for the DDL.

```
Object
{
  Histogram
     {int numberOfChannels;
      float histogram[numberOfChannels];}

  Scale
    {String unit,searchCriterion;
     float volume,sphereRadium,x,y,z;}

 Color
    {int numberOfChannels;
     String searchCriterion;
     Histogram red[numberOfChannels];
     Histogram green[numberOfChannels];
     Histogram blue[numberOf channels];}

  Shape
  {
    Cord
      {int numberOfChannels;
       String searchCriterion;
       Histogram cordAngle_1[numberOfChannels];
       Histogram cordAngle_2[numberOfChannels];
       Histogram cordNorm[numberOfChannels];}

    Wavelet
      {int numberOfLevels;
       int numberOfMoments;
       String searchCriterion;
       Histogram histWavelet[numberOfMoments];}

    Moments
    {
      int numberOfMoments;
      String search Criterion;

      Moment
        {int q,r,s;
         float value;}

      Moment moments[numberOfMoments];
    }
  }
}
```

Figure 2: OO representation of D and DS.

Even if the non-normative part does not belong to the standard it may have some serious impact on it. As stated earlier the non-normative part includes the extraction of the descriptors and the search engine. Let us start with the search engine. The search engine is concerned with the comparison of the descriptors and the formulation of the query. The comparison is the most critical part. It is related to the criterion that is used to compare the descriptors. That criterion can be a metric like the Hamming distance or the inner product or a non-metric likes clustering. The fact that the comparison criterion is not part of the standard may be problematic. As a matter of fact it is possible to optimize a descriptor for a given comparison criterion. Our experiments have indicated that the Hamming distance seems to be the most appropriate metric for the cord and wavelet descriptors.

Knowing that the descriptors and the comparison criterion may be related to each other and that a bad criterion may significantly deteriorate the results we propose to add a field in the DS in order to indicate the type of comparison criterion that is recommended to use with the descriptors. It would be possible to standardize the nomenclature used in that field. The parameter would indicate which comparison criterion should be used in the search. This parameter would not be mandatory and could be overruled by the search engine. In any case the ultimate decision would be left to the search engine.

As stated earlier the extraction of the descriptors is not part of the standard. We are going to review the extraction process in order to determine what are the implications. We have proposed DS for the scale, color distribution and shape. The extraction of the scale does not pose any particular problem because the volume, the dimensions of the smallest bounding box and the radius of the smallest sphere are not ambiguous: the details of the implementation should not modify significantly their values.

The case of the color distribution is more complicated. The color is defined in a non-unique way. Among the most common methods are texture mapping and vertices. A texture map is a color picture. By associating a set of two texture coordinates with each geometrical vertex it is possible to map the texture on the surface of the object. The texture element is deformed in order to fit the geometrical element on which it is mapped. That means that the color is defined both on the vertices and the surface surrounding them. In the second case a color is associated with each vertex. There is no color associated directly with the surrounding surface. In the simplest case the color is defined as a triplet while in more sophisticated models diffuse and specular color are taken into account.

For all those models the color representation can be reduced to a triplet and the corresponding histograms can be computed. Even if the resulting color could vary from one implementation to the next the variations are usually small. In the worst case it is always possible to reduce the number of channel in the histograms. Actually we could modify our descriptors and DS for the color by defining a set of three histograms for the diffuse color, a set of three histograms for the specular color and one histogram for the specularity. If a unique triplet defines the color it is possible to describe the color distribution by only using the histograms corresponding to the diffuse color leaving the other histograms equal to zero. Such a DS would allow a more detailed description of the color distribution. A red, green and blue triplet is used for the color. This is not the best representation from a human point of view. The hue, saturation and intensity representation and many others are more appropriated for that purpose. As a matter of fact the well-known JPEG is based on a luminance-chrominance representation. Nevertheless the red, green and blue representation is used because any representation can be converted to it. If require another representation can be used in the comparison process by converting the reference triplet to the appropriate system.

For the shape analysis two aspects are involved: data modeling and descriptors computation. In order to compute the cord descriptors a triangular mesh of the object is needed. Many common three-dimensional formats like VRML use triangular meshes as their standard representation. In many applications the object are represented with parametric surfaces like NURBS or non-uniform rational beta splines. Those representations can always be reduced to a triangular mesh. As a mater of fact most if not all rendering software and hardware need to reduce them to triangular meshes in order to perform the rendering. This is the reason why we use triangular mesh as our standard representation.

The next step is to compute the tensor of inertia. In this case the specific implementation of the computation may be a problem. The reason is that the distribution of the vertices is not uniform on the surface of the object. A usual case is a planar surface that is represented by a few vertices while a highly curved surface is represented by many of them. Consequently the area or mass around the vertices must be taken into account. That can be done by using the center of mass of each triangle of the mesh and by attributing them a mass corresponding to the area of their surface. Of course they are many other ways: as long as the *mass* is taken into account all methodology should provide a satisfactory result for the tensor of inertia. The computation of the eigen vectors and values is a well-known problem and should not be of concern. The

computation of the orientations and modulus of the cords should provide results that are consistent because those values are defined in an unambiguous way. Thus despite the fact that the extraction and the search engine are not normative it is possible to implement the descriptors and DS in a consistent way.

The wavelet descriptors can be easily extracted. As explained before the reference frame can be calculated without particular problems and there is a fast implementation of the wavelet transform. The computation of the binary model is a well-known problem. Actually the voxel representation is often used in order to generate a triangular mesh for three-dimensional raw data: the two representations are complimentary [8]. The moments-based descriptor is as simple to calculate as the tensor of inertia. Consequently the remarks are the same.

4. Evaluation of D and DS

MPEG-7 has issued a set of criterion for evaluating D and DS. The most important criteria used by MPEG-7 for the descriptors are the effectiveness, the application domain, the expression efficiency or value calculation, the processing efficiency or matching, scalability and multi-level representation [1]. Let us review those criteria in order to determine if our descriptors are suitable for MPEG-7.

The effectiveness determine if the descriptor capture an important characteristic of the content. The scale is a very important characteristic because it captures the physical dimensions of the object. Only three-dimensional objects can provide that information directly. The color is not specific to three-dimensional objects but in many applications it is an important criteria for filtering and retrieval. The shape is also a criterion that is unique to three-dimensional data in the sense that this is the only representation that can provide a complete and unambiguous geometrical information as opposed to pictures.

The application domain refers to the range of application domains. Today three-dimensional data are used in CAD applications, inspection, visualization, virtual reality, medical applications and the Web. A large number of scanners are available and important collections of three-dimensional objects already exist or are under construction like the CESAR project for human modeling. Those databases are confronted to the same problems: finding an efficient way to filter and retrieve useful information. We believe that our descriptors can help to provide an answer to that problem.

The expression efficiency determines if the descriptor expresses the features precisely and completely. The cord representation tends to capture the global and regional characteristics while minimizing the importance of the local features. If local characteristics are needed it is possible to increase the sensibility of the descriptors by adding more channels to the histograms. The wavelet descriptor describes the object at different levels of resolution. That make it a good candidate for coarse to fine search. The same observation can be made for the moments-based descriptor.

The processing efficiency refers to the non-normative part of the standard: the extraction and the comparison of the descriptors. As we have shown in a previous section the descriptors can be easily extracted from most three-dimensional formats and the outcome of the process has almost no dependency on the particular implementation of the calculation.

An application is scalable if its performance does not deteriorate with larger amount of data. Our descriptors are scalable at the object level because the size of the cord descriptors is not a function of the size of the object file: it does not depend on the number of triangles. For the wavelet descriptors it depends logarithmically on the number of voxels. This is totally acceptable in most applications. Moments-based descriptors are also scalable: the size of the moment description does not depend on the size of the data set.

At the database level our descriptors are also scalable in the sense that the outcome of a query is determined by comparing the reference descriptor with all the other descriptors. Consequently the complexity of the search depends linearly on the number of objects. This is acceptable for most applications and quit comment in content-based applications.

Finally let us talk about the multi-level representation. That means that the descriptor represents the features at multiple level of abstraction. In the cord case the abstraction level is related to the number of channels in the histograms. If they are many channels the description is precise but the level of abstraction is lower while if the number of channels is limited the description is less precise but the level of abstraction is higher. In case of the wavelet-based description the abstraction is controlled by the number of voxels representing the object: the more voxels the more level of detail we have. Moments-based

descriptors also provide a multi-level description that corresponds to the orders of the moments.

According to MPEG-7 the DS should be evaluated according to the following criterion [1]: effectiveness, application domain, comprehensiveness, flexibility, extensibility, scalability, simplicity and abstraction at multiple hierarchical levels. Our DS is very simple: it encapsulates related descriptors in a single class. In the case of the scale the volume, the dimensions of the bounding box and the radius of the sphere are encapsulated in a single class call the Scale. The red, green and blue histograms are also encapsulated in a single class call the Color and the histogram describing the distribution of the cords, the wavelet coefficients and the moments are encapsulated in a single class called Shape.

A class called Object inherits those classes. The proposed scheme is simple and is scalable in the sense that it does not depend on the dimensions of the object. The scheme can be easily extended to three-dimensional scenes. A three-dimensional scene is made out of objects. Our DS can describe each object present in the scene. Another DS can handle the relations between the objects. Their relative position and orientation could be such a relation. The DS is also flexible because more descriptors can be added to the three basic classes: scale, shape and color. The DS does not have any multiple hierarchical levels but the descriptors already provide those characteristics.

The DS provides a comprehensive description in the sense that the object is composed of three classes that clearly define its major components: the scale, the color distribution and the shape. It is left to the search engine to combine those attributes in an appropriate fashion in order to handle a particular query. The object scheme that we have presented can also be extended to parts description. An object is made out of many parts. Each part can be considered as an object. So we can use our DS to describe a part. A DS specifying the relations between the parts can describe the resulting object. Among the relations we have relative positions and orientations, physical constraints and functionality. For the object parts and scenes we believe that the DS has to be application dependent in order to describe the relation between the components in a useful way.

5. Conclusions

Descriptors and DS suitable for describing three-dimensional data in terms of scale, color and shape have been introduced. Their implementation in the framework of MPEG-7 has been discussed as well as the implications of the normative and non-normative part of the standard. We have evaluated the descriptors and DS according to the criteria stated by MPEG-7.

In the case of static objects the descriptors do not need to be streamed with the data. They can remain be on the same or on a remote location as long as there is a link between them. Some formats like VRML 97 have engines and script nodes that can be used to introduce a temporal behavior. In that case the description has to be streamed with the content. Streaming the descriptors provides a frame by frame description of the object but does not provide any inter-frame description. As MPEG-1 and 2 are providing a coding for the inter-frame content we believe that MPEG-7 should address the description of inter-frame content. That description could include the dynamic of the action, the location of the action and a measurement of the importance of the action relatively to other simultaneous actions.

References

[1] MPEG-7: Evaluation Process Document, ISO/IEC JTC1/SC29/WG11 **N2463**, Atlantic City, October 1998.

[2] J. Rossignac, “Interactive Exploration of Distributed 3D Databases over the Internet”, *Proceedings Computer Graphics International Hanover*, pp. 324-335 (1997).

[3] Y. Liu and F Dellaert, “A Classification Based Similarity Metric for 3D Image Retrieval”, *CVPR*, pp. 800-805 (1998).

[4] Y. Gdalyahu and D. Weinshall, “Automatic Hierarchical Classification of Silhouettes of 3D Objects”, *CVPR*, pp. 787-793 (1998).

[5] J. H. Yi and D. M. Chelberg, “Model-Based 3D Object Recognition Using Bayesian Indexing”, *Computer Vision and Image Understanding* **69**, pp. 87-105 (1998).

[6] C. S. Chua and R Jarvis, “Point Signature: A New Representation for 3D Object Recognition”, *International Journal of Computer Vision* **25**, pp. 63-85 (1997).

[7] E. Paquet and M. Rioux, "Content-based Access of VRML Libraries", IAPR *International Workshop on Multimedia Information Analysis and Retrieval, Lecture Notes in Computer Sciences-Springer* **1464**, pp. 20-32 (1998).

[8] B. Curless and M. Levoy, “A Volumetric Method for Building Complex Models from Range Images”, SIGGRAPH, (1996).