ERIC7: An Experimental Tool for Content-Based Image Encoding and Retrieval under the MPEG-7 Standard

L. Gagnon, S. Foucher, V. Gouaillier

R&D Department, Computer Research Institute of Montreal
550 Sherbrooke Street West, Suite 100, Montreal (Quebec), H3A 1B9, Canada
email: {lgagnon, sfoucher, vgouaill}@crim.ca

Abstract

ERIC7 is a software test-bed that implements Content-Based Image Retrieval (CBIR) compatible with the MPEG-7 multimedia standard. In its current version, the system allows automatic MPEG-7/XML encoding of up to 15 color, texture and shape descriptors with two query modes: search from example images and database clustering. In addition, the system allows to navigate graphically among the various descriptors in the XML files in order to easily track encoding results and system performance. The system is being tested on various types of images, including images of man-made and natural objects, as well as "scape-type" scenes. The system architecture is expandable to video indexing and retrieval.

Introduction

The aim of this paper is to report about the development of a software test-bed for automatic content-based image encoding and retrieval that is compatible with the new MPEG-7 multimedia standard. ERIC7, a French acronym that stands for "Environnement générique de recherche d'images par contenu compatible MPEG-7", is being developed by the Vision and Imaging team at the Computer Research Institute of Montreal, a research center dedicated to R&D in Information and Communication Technology.

In its current version, ERIC7 uses various low-level image analysis tools (color, texture and shape descriptors) developed in-house or integrated from specialized libraries, and operates in order to experiment with the various aspects of the MPEG-7/XML schema. Query can be of two types: search by example and database clustering. In addition, the interface allows to navigate graphically among the various descriptors in the XML files in order to easily track the encoding results.

Content-Based Image Retrieval (CBIR) is a very active research topics; either within the MPEG-7 standard [1, 2] or not [3]. Many prototype systems have been developed. Some target audio, others video and multimedia (see references [1-3] for more details). To our knowledge, there is no commercial system available yet that is fully MPEG-7 compliant but a couple of interesting demos from IBM and CANON are available on the Internet that provide manual semantic encoding [4]. One particularity of ERIC7 with respect to the other systems is that it specifically targets automatic encoding as well as navigation within the MPEG-7 schema for research and analysis purposes.

The paper is organized as follows. Section 1 briefly presents an overview of the MPEG-7 standard, with emphasize on the visual part of the XML schema. Section 2 describes the software architecture of ERIC7 along with its graphical search functionality. Section 3 provides an application example of the use of ERIC7 for outdoor scenes encoding and

retrieval. The application shows search and clustering capabilities on a database of images of man-made and natural objects. We finally conclude and give a brief outline of an ongoing video project related to ERIC7.

1. MPEG-7 overview

MPEG-7, ISO/IEC International Standard 15938, was developed by the MPEG (Moving Picture Experts Group). Formally named "Multimedia Content Description Interface", MPEG-7 standard defines a normative indexing of multimedia content at many level ranging from low-level features to higher semantic description [5]. It also allows to record information on both content management (media description, navigation and access, user interaction) and content itself (structure and semantics). However, only the structure of the description is normative. Production or consumption of MPEG-7 descriptions are not within the scope of the standard.

MPEG-7 can address different types of media in various formats and it offers a framework sufficiently generic to support a broad range of applications that necessitate a degree of interpretation of the meaning of multimedia content such as Content Base Retrieval, content management, navigation, filtering, and automated processing. It enables interoperability between applications and systems which processes and manage multimedia content. The benefits of its use are numerous. MPEG-7 is harmonised with other standards that have demonstrated success and acceptance in both traditional media and new media businesses, e.g., W3C (XML, XML Schema), IETF (URI, URN, URL), Dublin Core, SMPTE Metadata Dictionary, etc. [4].

An MPEG-7 description is an XML file instantiating a subset of predefined normative tools. These tools are of two types: Descriptors (D), that define the XML syntax and the semantics of each feature of the multimedia content, and Description Schemes (DS), that assemble many Descriptors and Description Schemes by specifying the structure and semantics of the relationships between them. These tools are defined by the Description Definition Language (DDL), which is based on the W3C XML Schema and that allows creation of new DSs or extension of existing ones. A set of systems tools are also available to support binarisation, multiplexing, synchronisation, transport and storage of descriptions as well as the management and protection of intellectual property.

MPEG-7 comprises 25 tools specific to the description of visual content including still images, video and 3D models. Visual attributes such as color, texture, shape, motion, and even face features, can be represented by these tools.

The collaborative development of the standard has produced the eXperimentation Model (XM) software, which is a simulation platform for the MPEG-7 tools [6]. Its purpose is to provide a common framework to test MPEG-7 compatible applications. Besides the normative components as the encoding of the Descriptors and Description Schemes according to DDL, XM also implements non-normative components as the feature extraction and search capabilities. To each D or DS corresponds a set of applications divided in two types: the server (extraction) applications and the client (search, filtering and/or transcoding) applications. Only the results produced by XM are normative and every MPEG-7 compliant applications should match these results. However, the methods implemented in XM are not part of the standard.

2. System architecture

The system architecture of ERIC7 is depicted on Figure 1. It is composed of three main parts: the database, the server and the Web client. The database is where the raw images and their corresponding MPEG-7/XML files are stored.

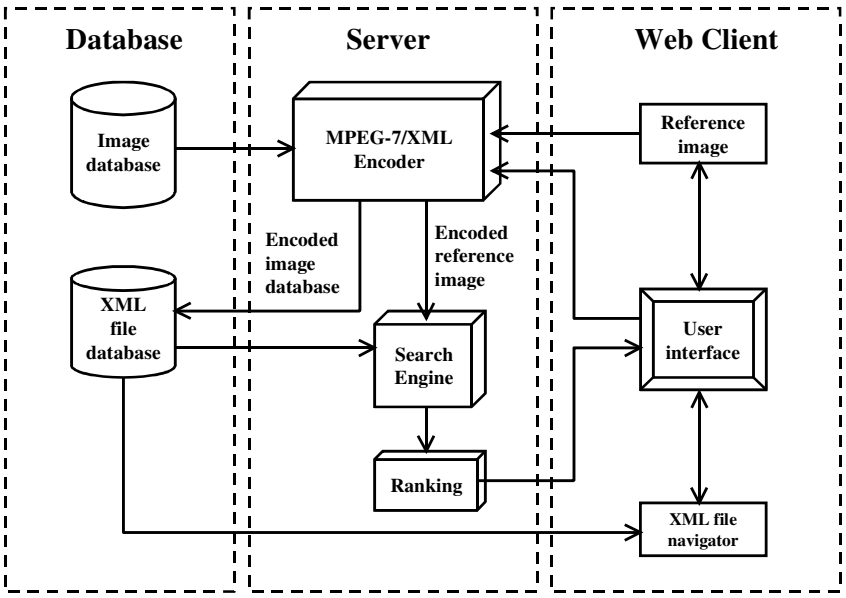

Figure 1: High-level architecture diagram of ERIC7

The main encoding and search applications are on the server part. The MPEG-7/XML files are generated by the encoder according to the features selected by the user. The encoding is done with our improved version of XM that allows multi-feature requests. In addition, the XM features have been extended with those of Virtual Retrieval Ware (VRW), a specialized library that provides low-level encoding for CBIR applications and sold by Convera (formerly Excalibur Technologies). This has been done by embedding VRW within XM using wrapping tools provided with XM.

In the current version of ERIC7, 15 features can be selected: 10 from XM (HomogeneousTexture, TextureBrowsing, EdgeHistogram, ColorLayout, ColorStructure, DominantColor, ScalableColor, ContourShape, RegionShape and RegionLocator), 4 from VRW (VRWColor, VRWRoughness, VRWTexture and VRWShape) and a last one based on Fourier spectrum shape [7]. Because Fourier spectrum shape reflects the overall structural organization of the scene content, it provides complementary low-level descriptors to color and textural descriptors. An application of a spectrum shape descriptor will be given in the next section.

The search engine calculates the distance between the query image and the search image set. The distances are all normalized between 0 and 1 to allow combinations.

The client part manages all the user requests through a HTML GUI which controls all the encoding, retrieval and visualization capabilities. The user can select two operation modes: search by example or clustering. He can also navigate within the XML files using a tool that generates UML diagrams. This is particularly useful to access encoding information in order to test system performance for a specific set of features.

3. Application example

A typical use case of ERIC7 is presented here in order to illustrate some of the main system's features. The chosen application is the discrimination between artificial and natural scenes, i.e. scenes containing large man-made objects (e.g. buildings) versus landscapes, using the spectral shape descriptor [7].

The spectrum shape encodes the overall structural organization of the scene content. It is implemented as follows. First, the image intensity is locally normalized in order to equalize the image contrast. Second, a polar fractional model is adjusted to the Fourier spectrum amplitude $A(f_x, f_y)$ using the model $A(f_x, f_y) \equiv A(f, \theta) \sim G(\theta) / f^{\alpha}$, where f and θ are the polar coordinates of the spatial frequencies f_x and f_y. The function $G(\theta)$ are obtained by linear fitting of the averaged energy spectrum on logarithmic units (see [7] for details). For natural images, value of α is around 2 (see also [7] for details). Third, a binary shape is formed based on the region delimited by $A(f=0.5, \theta)$. Finally, the binary shape is encoded using the RegionShape descriptor provided in XM. Figure 2 shows two examples of the above spectrum shape descriptor for a man-made and a natural scene. One can see that energy has a tendency to be isotropic for a natural scene and non-isotropic for an artificial scene.

Figure 2: Example of spectrum shape descriptor for man-made (left) and natural scene (right).

In ERIC7, the user has access to various windows that provides various graphical presentation of CBIR results. Figures 3 and 4 show examples of them. In Figure 3, the user have asked for the retrieval of images in the database that are similar to a city image found on the internet (top left image in Figure 3). The similarity is based only on the spectrum shape descriptor described above. Left-hand side of Figure 3 shows the result retrieval for the 10 best hits. Red color indicates good matching results. Right-hand side of Figure 3 is an example of hierarchical clustering done on the image database for the SpectralShape descriptor. The images are grouped into 12, 6, 3 and 2 clusters, according to their similarity content. This tool gives the user a way to find possible clusters of similar images in the database.

Left-hand side of Figure 4 shows another way to visualize CBIR results when two descriptors are used. Here the request is done with the SpectrumShape and EdgeHistogram descriptors on the same internet image as in Figure 3. The most similar images are arranged in a table as follows. The five first best matches of the SpectrumShape descriptor are sorted according to their EdgeHistogram distances and put in the first row. The same procedure is done for the five next best matches (second row) and so on. The background color indicates the similarity of the image with respect to the query image according to the SpectrumShape descriptor; the most similar image has a red background and the less similar has a blue background. For example, the first image on the first row on the left-hand side of Figure 4 is the one that is the closest to the query image with respect to EdgeHistogram; the second image on the first row

has a red background so it is the closest in terms of SpectralShape. A similar table is calculated by interchanging the two descriptors.

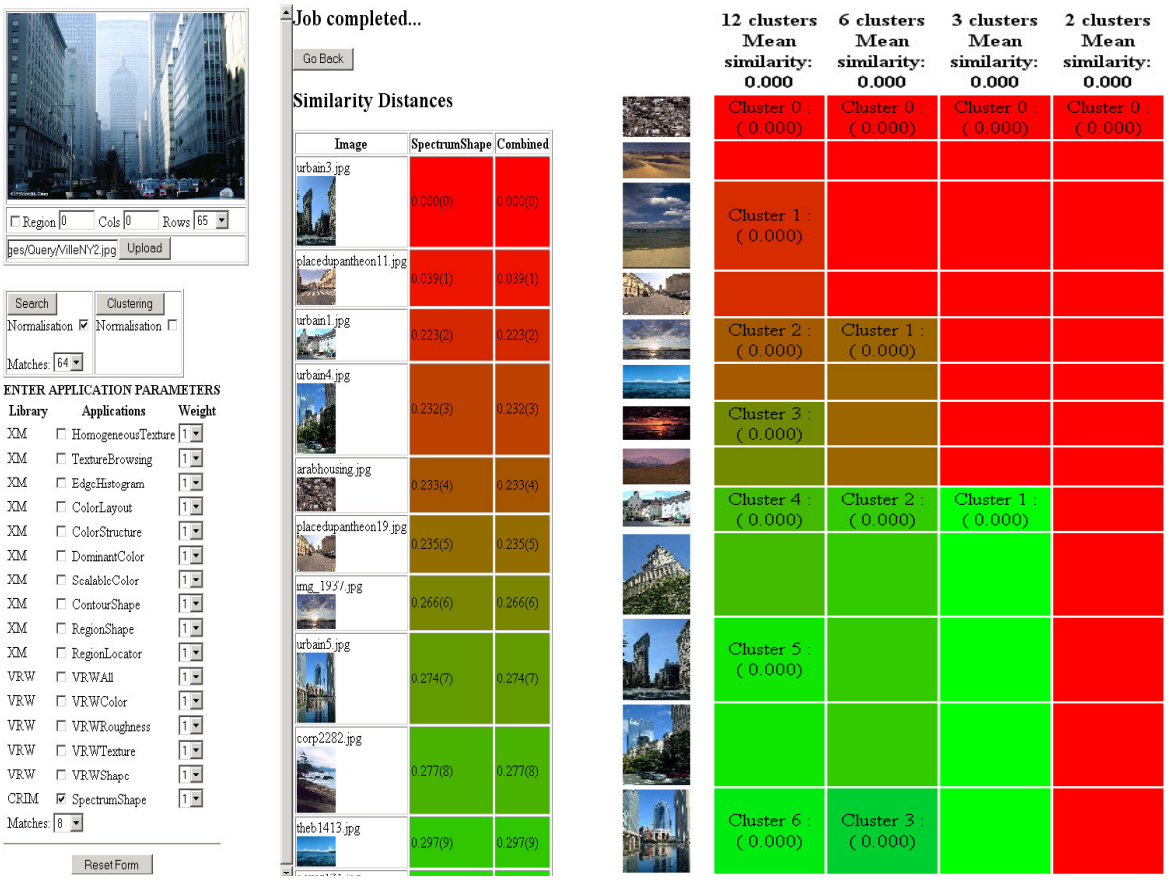

Figure 3: (left) Query example with an image from internet (image at the top left) and results of the query using the SpectrumShape descriptor; (right) Hierarchical clustering technique gives the user a way to find possible clusters of similar images in the database.

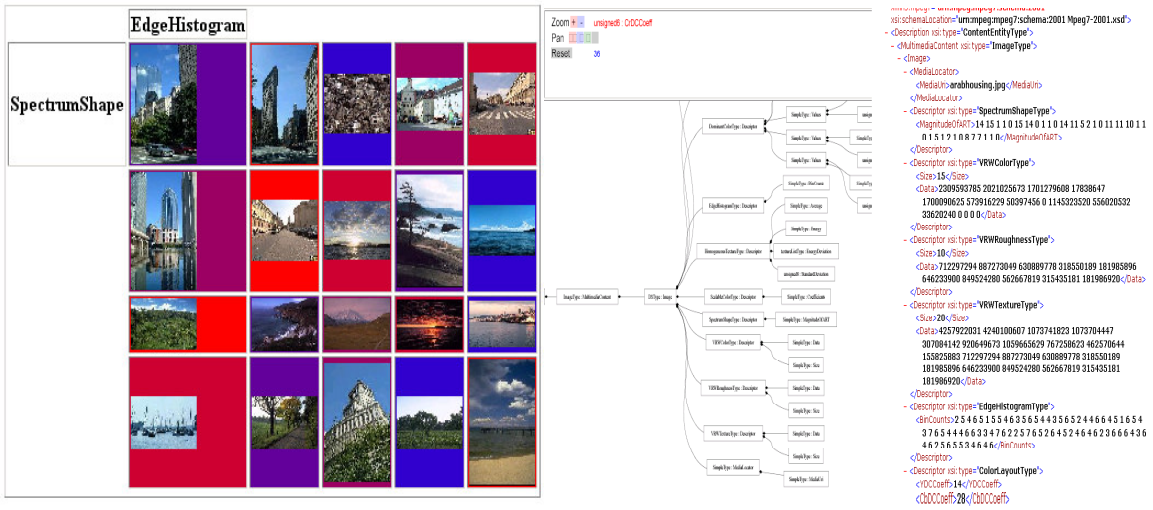

Figure 4: (left) Example of table arrangement of a query result based on two descriptors (see text for details); (right) Graphical UML representation of and MPEG-7/XML encoding file showing the hierarchical structure of the MPEG-7 tags generated after image encoding.

Finally, right-hand side of Figure 4 shows an example of automatic generation of a graphical UML representation for one of the MPEG-7/XML encoded file in the image data set. This turns out to be a very useful tool to interactively navigate among the various descriptors and tags in large MPEG-7/XML files. The user has access to all the tag values of each box by mouse pointing to it. He can also zoom-in or zoom-out within the block diagram at his convenience.

Conclusion and future works

We have briefly reported about the development of a new CBIR system fully compatible with MPEG-7. ERIC7 is based on image analysis tools developed in-house or integrated from specialized libraries. It is designed to experiment with the various aspects of the visual MPEG-7/XML schema. Query can be of two types: search by example and database clustering. The interface allows to navigate graphically among the various descriptors in the XML files and through interactive UML graphics. This is of particular importance for the tracking and analysis of intermediate encoding results as well as for the optimization of the system performance. Thus, ERIC7 provides a useful "laboratory" for MPEG-7-based CBIR with the help of analysis and exploration tools.

Works are under progress to extend ERIC7 to a fully MPEG-7 compliant audio-visual indexing and retrieval system. In collaboration with the National Film Board of Canada and the Canadian Internet development organization CANARIE, we are developing a modular and scalable MPEG-7 Audio-visual Documentation Indexing System (MADIS), packaged into a demonstrable test-bed, for research and development on content-based retrieval of films. The description decomposition we have adopted for MADIS is based on a temporal decomposition into visual shots. For each shot, key frames are identified to allow shot retrieval by examples. Specific visual and audio characteristics are extracted and encoded for each shot segment. Visual features include global texture and color content, indoor and outdoor scenes, face detection and motion activity. The audio band within a visual shot is segmented according to speech and non-speech segments. An automatic speech recognition system developed at CRIM and based on Hidden Markov Models is used to encode the speech segments in phone lattices in order to allow robust word recognition. Non-speech segments are further segmented into three classes: music, silence and other sounds. More details about MADIS will be reported soon.

References

[1] http://www.diffuse.org/meta.html
[2] International Organization for Standard, "MPEG-7 Projects and Demos", Draft document, March 2001
[3] B. Johansson, "A Survey on Contents Based Search in Image Database", December 2000
[4] http://mpeg-industry.com/
[5] "Introduction to MPEG-7: Multimedia Content Description Interface", Edited by B. S. Manjunath, P. Salembier, T. Sikora, John Wiley & Sons, 2002
[6] http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html
[7] A. Oliva, A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope", International Journal of Computer Vision, 42(3), p. 145-175, 2001.