# MULTIMEDIA CONTENT DESCRIPTION IN THE INFOPYRAMID

*Chung-Sheng Li, Rakesh Mohan and John R. Smith*

IBM T.J. Watson Research Center
PO Box 704 , Yorktown Heights, NY 10598
{*csli, rakesh, jrsmith*}*@watson.ibm.com*

## ABSTRACT

There is a growing need for developing a content description language for multimedia that improves searching, indexing and managing of the multimedia content. The MPEG group recently established the MPEG-7 effort to standardize the multimedia content interface. The proposed interface will bridge the gap between various types of content meta-data, such as content features, annotations, relationships, and the search engines. In this paper, we develop a method of handling multimedia content description in a new multi-abstraction, multi-modal content representation framework called the InfoPyramid. The InfoPyramid facilitates the search, retrieval, manipulation, and transmission of multimedia data by providing a hierarchy for content descriptors. We illustrate the suitability of the InfoPyramid multimedia content description to MPEG-7 by examining four multimedia retrieval applications: a Web-image search engine, a satellite image retrieval system, an Internet content delivery system, and a TV news storage and retrieval system.

**Keywords:** MPEG-7, multimedia representation, XML.

## 1. INTRODUCTION

The recent rapid proliferation of multimedia content is increasing the need for efficient methods for searching, indexing and managing multimedia. Multimedia combines data from various modalities (video, images, audio, multilingual text), multiple abstraction levels (raw data, features, semantics, and meta-data), and multiple resolutions (image thumbnails, video key-frames). Consequently, it has become apparent that an interoperable, content-neutral description interface is needed to facilitate the searching and indexing of the growing amounts of multimedia content.

The MPEG standards group has recently begun to formulate the MPEG-7 multimedia content description interface. MPEG-7 will specify a standard set of content descriptors, description schemes, and an extensible content description framework that can be used, in general, to describe multimedia content. By developing standard processes for developing and publishing the content descriptions, it will be become easier for various search engines, agents, user applications and users to search, index and filter the multimedia content based on the content descriptions.

In this paper, we present a novel multimedia content description scheme developed in the InfoPyramid. The InfoPyramid is a framework for representing content in multiple modalities, levels of abstraction and resolutions. The InfoPyramid also aggregates description data, such as content features, annotations, and meta-data, with methods for analyzing, manipulating, and synthesizing the content and content descriptions, and rules for processing, and delivering the content. In short, the InfoPyramid facilitates the search, retrieval, manipulation, and transmission of multimedia data by providing a hierarchy for content descriptors.

We present the InfoPyramid multimedia content description system and describe an extensible content description data model. We then describe a system for representing the InfoPyramid content descriptions using XML. We examine the suitability of the InfoPyramid content description scheme to MPEG-7 by analyzing four multimedia search and retrieval applications that we are developing using the InfoPyramid,. In particular, we demonstrate the InfoPyramid multimedia content description framework in a Web-image search engine, a satellite image retrieval system, an Internet content delivery system, and a TV news storage and retrieval system.

## 2. INFOPYRAMIDS

The InfoPyramid is a framework for aggregating the individual components of multimedia content with content-descriptions, methods and rules for handling the content and content descriptions. The InfoPyramid describes content in different modalities, at different resolutions and at multiple abstractions. In addition, it may define methods for manipulating, translating, transcoding, and generating the content. The primary objective of the InfoPyramid is to provide a hierarchy for content descriptors in order to guide search and retrieval.

**Multi-modal**: Multimedia content is usually not in a single media format, or modality. A video clip can contain raw data from video, audio in two or more languages, and closed captions. In medical arena, MRI, CT, PET, and ultrasound can be captured for the same patient, resulting in multiple 3D scans of the same content.

For certain query and retrieval tasks, the appropriate content modality may not be available. The required modality may be generated by transforming other modalities. For example, a video clip can be transformed into images showing key-frames [5,6], while text can be synthesized into speech.

**Multi-resolution**: Each content component can also be described at multiple resolutions. Numerous resolution reduction techniques exist for constructing image and video pyramids. For example, Flashpix - that provides mechanisms for storing and retrieval of still images at multiple resolutions. Features and semantics at different resolutions are obtained from raw data or transformed data at different resolutions, thus resulting a feature or semantic pyramid.

**Multiple-abstraction levels**: The abstraction levels describe features and data in a hierarchical fashion to facilitate progressive search and navigation. For example, one hierarchy could be features, semantics and object descriptions, and annotations and meta-data. A hierarchy within the color histogram feature could be a color histogram with 512 bins, one with 16 bins and the average color value.

**Methods and Rules**: Methods generate content descriptors from the features of the data, or analyze, manipulate, provide modality translation, or process the data in various ways. In addition, the InfoPyramid may have rules to provide flexible application of the methods. Methods and rules provide linkage between different modalities, resolutions and abstractions.

Rather than forcing a strong separation between the data and the content description meta-data, the InfoPyramid offers a continuum between the data, various abstractions of the data, and content description data.

# 3. DESCRIPTION DATA MODEL

In order for the content descriptions to be usable by outside search engines and applications, we develop a standard set of content descriptor primitives, a framework for extending the content descriptors, and a process for representing, publishing, and transmitting the content descriptors using XML. The content descriptions are defined, in general, from an extensible set of content description data types. In the InfoPyramid, the content descriptor schemes are defined by explicitly specifying content descriptors and methods for comparing instances of the content.

## 3.1 Description data types

The multimedia content descriptors are defined from fundamental description data types and derived types derived by extending the fundamental data types.

In creating content-description instances, many of the types T utilize modifiers of the form T(t)[l], which specifies that T contains l elements of type t. For example, vector(integer)[256] defines a 256-dimensional vector in integer space.

Derived types D are created by derivation from the fundamental types. For example, D:[T] defines a derived type D which is derived from type T. In general, derived types are most useful when combining fundamental types as follows: D:[T1,T2,T3,...]. For example, consider the definition of the derived type "deformation": deformation:[sequence(shape)[N],path].

## 3.2 Standard Descriptors

In the Web image search engine and content-based satellite image retrieval system, we have developed a set of standard descriptors for images and videos which describe various visual features of the content, such as the color, texture and motion. For example, we define a standard descriptor for color histograms as

HVShist:histogram(real)[166], which corresponds to a particular definition of a 166-bin color histogram derived from HVS color space [1]. Similarly, we define a standard descriptor for texture, QMFtexture:vector(real)[9], which correspond to texture descriptions are defined by the spatial-frequency energies of nine subbands of the QMF wavelet transform of the image [4].

## 3.3 Description methods

The system also defines a fundamental set of description functions that operate on the description. The primary purpose of the description functions is to facilitate the comparison of description values, which allows searching, indexing and retrieval of the multimedia content.

The fundamental description functions comprise several classes: logic, similarity and transform. The logic functions perform binary evaluations. The similarity functions return a score. They define standard mathematical formulas for computing distances. Finally, the transform functions define operations on the description which transform it in some way. For example, the transform functions can define the relationship between one description type and another standard description type. For example, given the following standard type: rgbhist:histogram(integer)[512], which defines rgbhist to as a 512-bin histogram in RGB color space, another derived type may be declared such as

myhist:histogram(integer)[512],

which defines a color histogram in a different color space. Given that the new color space may be derived from the RGB color space, then myhist is obtained via transformation F of rgbhist, as follows

myhist = F(rgbhist).

The importance of this is clear in conducting queries across multiple archives of multimedia, as illustrated in Figure 1. For example, each archive may use a different color histogram description. In order for the search engine to query the multiple archives given a single query color histogram Q, the search engine must transform that query histogram into the appropriate histogram spaces of the specific archives [3], i.e., F1(Q), F2(Q), and F3(Q).

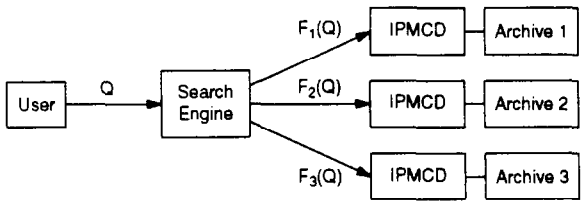


**Figure 1**: Content-based search in multiple archives requires transformations of the query Q to be compatible with the specific content-descriptions in each archive.

# 4. REPRESENTATION IN XML

We use the eXtensible Markup Language, or XML[8], as the representation language for InfoPyramids. This representation of InfoPyramids in XML can be viewed as a "serialization" of the InfoPyramid abstract data model, useful. XML is a tagged markup language for representing hierarchical, structured data.

XML is easily readable by both machines and humans. Additionally, XML is portable and extensible.

### 4.1 Descriptor Extensibility

New descriptors can be defined in XML by specifying the base class types and compare methods. For example, consider the following specification of a new color histogram description class:

*<HIST classname="myhist"*
*baseclass = "histogram(real)[64]" compare = "Euclidean" />*

which defines the descriptor class "myhist" which corresponds to a 64-bin histogram which utilizes the Euclidean distance metric to compare myhist descriptions.

The myhist content description instances are specified as follows:

*<HIST id="9991" myhist="83203411242342342..." />*

### 4.2 Descriptor Schemes

The multimedia content description language allows the development of descriptor schemes in which a set of content descriptors and their relationships are specified.

Consider the following example for new descriptors for color regions:

*<COLOR-REGION classname="colorregion"*
*baseclass="myhist.shape" compare="0.6*myhist.Euclidean + 0.4*shape.walk" />*
*<REGION-SET classname="regionset"*
*baseclass="set(colorregion)[N]" compare="sum(n=0;N-1)(colorregion.compare)" />*

### 4.3 Document Type Definitions: DTD

The domain of MPEG-7 descriptors is very large. A look through the early drafts of MPEG-7 [9] show that a large number of features and meta-data have already been proposed, and this list is only going to increase. Most of these are specific to particular media objects or application domain. XML includes an excellent mechanism, the Document Type Definition or DTDs[8] which make it possible to manage the plethora of meta-data and feature descriptors by DTDs, which support the subset for a particular media or application. The DTDs also makes it easy for a particular community to share and conform to a specific set of MPEG-7 descriptors by subscribing to a common set of DTDs.

## 5. APPLICATIONS

We examine four multimedia retrieval applications in which we utilize the InfoPyramid multimedia content description system to facilitate the searching, indexing and filtering of multimedia content.

### 5.1 Web image search engines

We are utilizing the multimedia content description system in the development of a Web image search engine [1]. The objective of the Web image search engine is to catalog the images and videos on the World-Wide Web and allow users to search the catalog. The Web image search engine uses content descriptors to index the images and videos by visual features, text, and semantic concepts.

The Web image search engine uses a set of descriptors which are automatically and semi-automatically generated. The visual features of the images are defined by a color histogram and a texture vector which are automatically computed. The system assigns each image and video a set of terms which are automatically extracted from the parent Web document and Web address. The Web image search engine also assigns various concept labels to each image and video by looking-up the terms in a term-concept dictionary. This process is semi-automatic in the sense that the concept labels may be later verified manually. Each concept class belongs to a concept ontology that is also developed manually.

The content descriptions in the Web image search engine are represented using the InfoPyramid multimedia content description language. The content descriptions types are defined as follows:

*<HIST classname="color" baseclass="histogram(real)[166]" compare="Euclidean" />*
*<TEXTURE classname="texture" baseclass="histogram(real)[9]" compare="Euclidean" />*
*<TEXT classname="text" baseclass="set(term)" compare="String"/>*
*<CONCEPT classname="concepts" baseclass="set(concept)" compare="String" />* .

The Web image search engine specifies the content description instances as follows:

*<FEATURES>*
*<HIST color="83203411242342342..." />*
*<TEXTURE texture="284..." />*
*<TEXT text="term1/term2/term3/... " />*
*<CONCEPT concepts="concept1/concept2/concept3/..." />*
*</FEATURES>*

In this way, any search engine may search the catalog of image and video content descriptions.

### 5.2 Satellite image retrieval systems

A content-based retrieval system of satellite images have been reported in [4]. In this system, image content is represented as an InfoPyramid with four modalities: (1) Pixel, or the original image (2) Feature (3) Semantic and (4) Metadata.

We distinguish between simple and composite objects. A simple object is defined as a region of an image that is homogeneous with respect to an appropriate descriptive quantity, or attribute. A composite object consists of multiple simple objects with pairwise spatial (e.g. adjacent, next to, west of), temporal (e.g. before, after) relationships. A simple object can be defined at any of the modalities.

This system can answer queries such as "find all the regions of cauliflower fields that have clubroot disease." Here, the search target is specified by a composite object containing cauliflower field regions and a clubroot disease regions.

## 5.3 Internet Content Customization

We are using the InfoPyramid to allow content providers to represent Internet content in a form that allows its customized delivery according to client device characteristics and user preferences. We are also developing the InfoPyramid to be a transient structure that facilitates the transcoding of Internet content on the fly to customize the retrieval and display of Internet content [2].

## 5.4 TV News Application

We have also implemented a video database application using the InfoPyramid model. This application automatically captures and indexes television news stories and makes them available for search over the Internet. The details can be found in [6,7], we will give a summary here.

The system captures TV news broadcasts and the closed caption stream. It then segments the news program into individual news stories. The text transcript of the each story, contained in the close caption is fed to a text indexer. A user can query this database of news stories over the Internet using text queries.

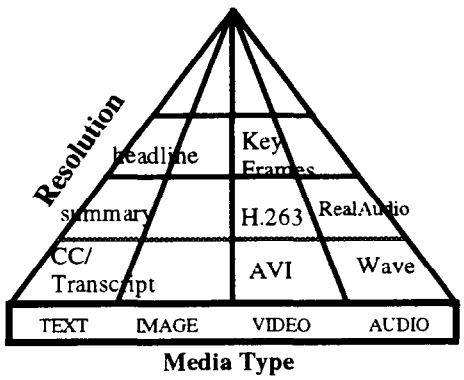


**Figure 2**: TV News Story InfoPyramid

Figure 2 shows the InfoPyramid for a news story. The InfoPyramid has different modalities of content, namely video, audio and text at various resolutions.

The news story is thus represented at different resolutions from full AVI video through key-frames with text, to audio alone down to the level of just a text title for the story.

This InfoPyramid can be represented in XML as:

```
<NEWS-STORY>
 <Program> ABC Evening News</Program>
 <date>11/2/97</date>
.....
  <Video>
   <Component>
    <Content-type>video/avi</Content-type>
    <URL>http://foo.com/news1.avi</URL>
    <bit-rate>1Mbs</bit-rate> .....
   </Component>
   <Component>
    <Content-type>video/Bamba</Content-type>
    .....
   </Component>
  .....
  </Video>
  <transcript>.....</transcript>
</NEWS-STORY>
```

As a news story's content is multi-modal, depending on the query, the right modality has to be exposed to the search mechanism. For example, text based queries require access to the textual transcript, while a visual search may make use of the key-frames. Similarly, in reply to a query, different modalities and/or resolutions may have to be returned. When summaries are requested for browsing, the InfoPyramid returns the key frames of the video and a summary of the transcript. When a specific story is requested, the InfoPyramid returns the appropriate video. Thus, the InfoPyramid provides a uniform representation both for search and for retrieval.

# 6. CONCLUSIONS

We have presented a new content description scheme, the InfoPyramid and its representation in XML, which can be used in MPEG-7. We have demonstrated the usefulness and flexiblity of this scheme through four diverse applications covering a spectrum of MPEG-7 issues.

# 7. REFERENCES


[1] J. R. Smith and S.-F. Chang, Visually searching the web for content, *IEEE Multimedia*, 1997, Vol. 4, No. 3, pp. 12 -- 20.

[2] J. R. Smith, R. Mohan, C.-S. Li, Transcoding Internet content for heterogenous client devices, *IEEE ISCAS-98, Special session on Next Generation Internet*, June, 1998, to appear.

[3] S.-F. Chang, J. R. Smith, M. Beigi, A. Benitez, Visual information retrieval from large distributed on-line repositories, *Communications of the ACM*, December 1997.

[4] L. Bergman, V. Castelli and C.-S. Li, Progressive Content Based Retrieval from Large Satellite Image Archives, *DLIB Magazine*, http://www.dlib.org, October, 1997.

[5]L. Teodosio and W. Bender, "Salient video stills: content and context preserved," in *Proceeding ACM Multimedia 93*, Anaheim, CA, 1993, pp. 39-46.

[6] R. Mohan, Text Based Browsing of TV News, *SPIE 2916, Multimedia Storage and Archiving Systems*, Boston, Nov. 1996.

[7] R. Mohan, Indexing Television News, *Visual'97*, San Diego, CA, Dec. 1997.

[8] Extensible Markup Language (XML), W3C Working Draft, http://www.w3.org/TR/WD-xml, November, 1997.

[9] MPEG-7 Requirements, ISO/IEC JTC1 /SC29/ WG11/ N1921, October 1997.