Interoperable Content-based Access of Multimedia in Digital Libraries

John R. Smith
IBM T. J. Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532 USA

ABSTRACT

Recent academic and commercial efforts in digital libraries have demonstrated the significant potential for large-scale, on-line search and retrieval of cataloged multimedia content. By improving access to scientific, educational, and historical documents and information, digital libraries create powerful opportunities for revamping education, accelerating scientific discovery and technical advancement, and improving knowledge. Furthermore, digital multimedia libraries go well beyond traditional libraries in storing and indexing diverse and complex types of material consisting of images, video, graphics, audio and multimedia. In this paper, we examine several challenges addressed by MPEG-7 for enabling network-based digital multimedia libraries to interoperate with diverse networked multimedia systems and allow universal access to widely distributed multimedia content. MPEG-7 provides a framework for self-describing digital multimedia that uses a common vocabulary to describe the features, semantics, structure and models of audio-visual content. We describe how MPEG-7 can be used to enable interoperable content-based searching, indexing, filtering and browsing of multimedia in digital libraries. We also examine the MPEG-7 application of Universal Multimedia Access, which involves the scalable or adaptive delivery of multimedia content to patrons of digital libraries regardless of capabilities of terminal device, conditions of communication bandwidth, or client support for media formats and modalities.

Keywords: Digital libraries, content-based retrieval, multimedia databases, MPEG-7, Universal Multimedia Access

1. Introduction

The Moving Picture's Experts Group (MPEG) is developing a new standard called the "Multimedia Content Description Interface," also known as MPEG-7. The goal of MPEG-7 is to enable fast and efficient searching, filtering and adaptation of audio-visual content. The effort is being driven by specific requirements taken from a large number of applications related to image, video and audio databases, media filtering and interactive media services (radio, TV programs), image libraries, and so forth. One important application of MPEG-7 relates to the concept of "*Universal Multimedia Access*." Universal Multimedia Access (UMA) allows the scalable or adaptive delivery of multimedia to users and terminals regardless of conditions of communication bandwidth or capabilities of terminal devices and their support for media formats.

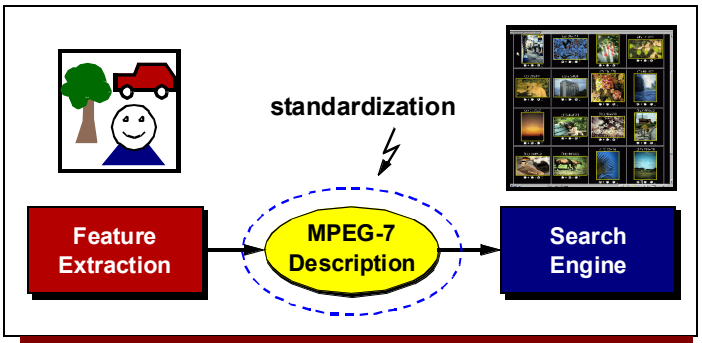
1.1. MPEG-7 Digital Libraries

With the tremendous growth in digital image and video data, it is becoming increasingly important to effectively store, search and retrieve such data. Recent advances in multimedia databases have resulted in technologies for managing a variety of multimedia formats including images, video, audio, and text [2]. In particular, developments in content-based retrieval have brought new capabilities for querying and accessing images and video by content [3]. Content-based access has been shown to improve the ability to effectively search, filter and access images and video [4]. Recent research has investigated content-based access along a number of different dimensions including: *features* – perceptual attributes of the image and video content, such as color, texture, shape; *structure* – spatio-temporal organization of image and video information content, such as segments and regions, models – collections, clusters and feature classes; and semantics – real world entities and scenes depicted in the images and video. Overall, content-based access of digital libraries requires the confluence of many diverse technologies related to content analysis, indexing, and querying at these different levels.

Traditional text-based indexing has been found to be insufficient for indexing multimedia data because of problems related to incompleteness, inconsistency, difficulty in automation, and subjectivity in assigning textual annotations. Other technical barriers for searching in multimedia digital libraries result from the problems in automatically attaining sufficient levels of image understanding, scene recognition and object extraction for unconstrained audio-visual data. Although much has been accomplished in computer vision and artificial

intelligence in the past decades, especially within specific problem domains, the unconstrained vision and understanding problems remain largely unsolved. A final technical barrier results from the expectations of the user for searching and filtering. Evidence has shown that users consider high-level semantic concepts when looking for specific multimedia content. Typical high-level semantics include objects, people, places, scenes, events, actions and so forth. Typically, it is difficult to derive this information automatically from the multimedia data.

Given today's state-of-the-art multimedia searching and filtering technology, the users of digital libraries are confronted with tools that rely on text-based, manually generated annotations or feature-based automatically generated descriptors of color, texture, shape, and so forth. Typically, the users cannot easily translate their high-level concepts into a sufficient set of feature descriptions, nor does the matching at the feature level consistently produce satisfying results [6].



Feature Extraction	MPEG-7	Search Engine
Tools for annotation and authoring	XML descriptions	Searching and Filtering
Content analysis	Description Schemes (DSs)	Classification
Feature extraction	Descriptors (Ds)	Manipulation and adaptation
Modeling of content-based description data	Language (DDL)	Summarization and visualization
	Description Coding Schemes	Indexing
	Principal Concepts	Complex querying

Figure 1 MPEG-7 standardizes the description structures, the description language and coding schemes. As a result, tools and methods are needed for feature extraction, searching and filtering to enable deployment of interoperable systems.

1.1.1. Multimedia Content Description

MPEG-7 is standardizing specific XML metadata structures called Description Schemes (DS) and binary Descriptors that are used to describe and annotate audio-visual data [1]. In MPEG-7, the DSs are categorized as pertaining to visual, audio, or generic description [7]. In general, the DSs can contain MPEG-7 Descriptors (D) and other DSs, and can be extended for domain-specific applications [10,11]. The MPEG-7 DSs provide a way to describe in XML the important concepts related to audio-visual data in order to facilitate the searching, indexing and filtering of audio-visual data. The DSs are designed to describe both the generic aspects of audio-visual data management and the specific content and features of the audio-visual material. The generic DSs provide a way to specify immutable meta-data related to the creation, production, usage and management of the audio-visual data. For example, these DSs can be used to describe the title and author of an audio-visual program. The audio-visual content-specific DSs provide a way to specify the content directly at a number of levels including signal structure, features, models and semantics. Other audio-visual specific DSs are designed to allow efficient navigation and access of the audio-visual data. The MPEG-7 DSs are defined using the MPEG-7 Description Definition Language (DDL), which is based on the XML-Schema Language. DDL allows the creation and extension of DSs. The DDL generates XML descriptions that are human-readable and can be searched, transmitted and filtered in applications that deal with audio-visual content.

1.1.2. Challenges for MPEG-7

Significant technical challenges remain in allowing the wide adoption of MPEG-7. First, MPEG-7 standardizes only the description structures (DS and Descriptors) and the description language (DDL). As a result, technologies for generating the using the descriptions, such as feature extraction, searching and filtering are not being developed or specified as part of the MPEG-7 Standard. Second, in order to be most effective for users, it is necessary to develop advanced analysis and classification methods that are capable of producing annotations that are semantically meaningful. This represents a significant improvement over the state-of-art in developing systems that enabling searching and filtering at the semantic level rather than requiring operations on low-level visual feature descriptions based on color, texture, shape and motion. This involves the development of new technology and methods for effectively deriving semantics from low-level feature information.

1.2. MPEG-7 Universal Multimedia Access

From the terminal device perspective, UMA is relevant for emerging applications that involve delivery of multimedia for pervasive computing (hand-held computers, palm pilots, portable media players), consumer electronics (television set-top boxes, digital video recorders, television browsers, Internet appliances) and mobile applications (cell-phones, wireless computers) [1]. From the content perspective, UMA involves important concepts of scalable or layered encoding, progressive data representation, and object- and scene-based encodings that inherently provide different embedded levels of content quality. From the network perspective, UMA involves important concepts related to the growing variety of communication channels, dynamic bandwidth variation and perceptual quality of service (QoS). Finally, from the human perspective, UMA involves different preferences of the user (recipients of the content) and the content publisher in choosing the form, quality and personalization of the content. UMA promises to integrate these different perspectives into a new class of content adaptive applications that allows users to access multimedia content without concern for specific encodings, terminal capabilities or network conditions.

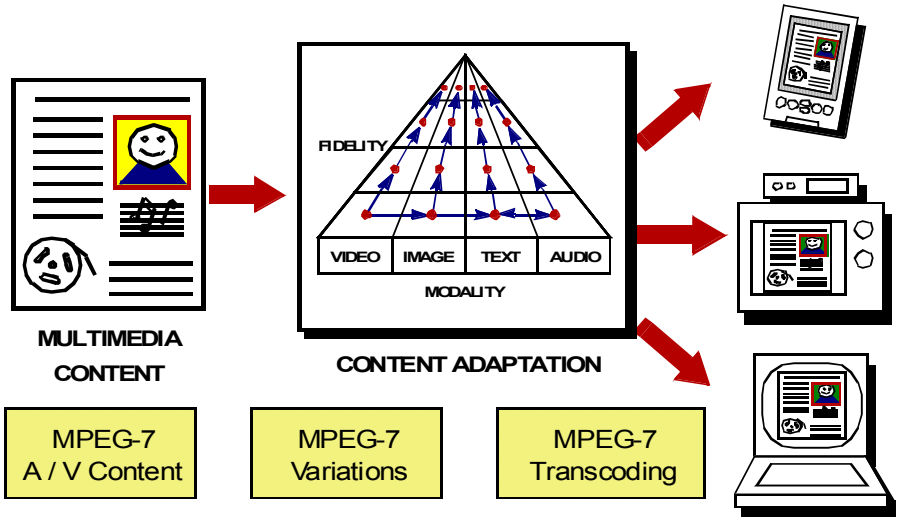


Figure 2 Figure 2: Illustration of Universal Multimedia Access (UMA) in which the appropriate variations of the audio-visual programs are selected according to the capabilities of the terminal devices. The MPEG-7 transcoding hints may also be used to further adapt the programs to the devices.

UMA can be enabled in a number of ways, as shown in Figure 2, which include (1) transcoding – manipulating the content on-the-fly in order to adapt to different terminal and network conditions, (2) selection – storing, selecting and delivering different variations of the content from a content server, (3) progressive retrieval – providing scalable representations of audio-visual signals that allow them to be retrieved progressively, and (4) summarization – providing personalized summaries of audio-visual programs that allow mobile users to efficiently browse and view the content. Transcoding has the advantage of not requiring the storage of excess data at the server. However, transcoding does require processing of the content, such as at a server or proxy, which typically introduces delay. On the other hand, pre-materialization of different variations of the content increases the amount of data that needs to be stored at the server, but it typically requires the different variations to be selected without processing of the content, thereby resulting in less delay.

MPEG-7 is addressing the requirements of UMA by providing a number of different tools as described above. For one, MPEG-7 has created specific requirements for Description Schemes that are part of the standard to describe different abstraction levels and variations of multimedia content [10,11]. For example, the different abstraction levels include the composition of objects from sub-objects, the extraction of plot structure of a video, summaries of audio-visual programs and different variations of the multimedia data with different resource requirements. In addition, MPEG-7 has created requirements that it shall support the transcoding, translation, summarization and adaptation of multimedia material according to the capabilities of the client devices, network resources, and user and author preferences. For example, adaptation hints may be provided that indicate how a photograph should be compressed to adapt it for a hand-held computer, or how a video should be summarized to speed-up browsing over a low-bandwidth network.

MPEG-7 has taken an important forward-looking approach in standardizing meta-data for Universal Multimedia Access. MPEG-7 is addressing UMA by providing tools for transcoding, managing variations of multimedia

content, representing multimedia data using scalable data representations, and specifying summaries. The corresponding Description Schemes have great utility in the deployment of multimedia applications that allow the scalable delivery of multimedia to users and terminals regardless of conditions of communication bandwidth, capabilities of terminal devices for processing, storage and display, and support for media formats. Based on these requirements, MPEG-7 is standardizing the following Description Schemes for UMA: Media Transcoding Hints, Variations, Space and Frequency Views, and Summaries.

1.2.1. Media Transcoding Hints

The MPEG-7 Media Transcoding Hints give information that can be used to guide the transcoding of multimedia, including object-, segment- and region-based description of the importance, priority, and content value, and description of transcoding behavior based on transcoding utility functions and network scaling profile. In addition, Media Coding tools give information about multimedia data including the image and video frame size (width and height), frame rates of video, data sizes for image, video and audio download and storage, formats and MIME-types. MPEG-7 provides Media Transcoding Hint tools for specifying transcoding hint information for Universal Multimedia Access (UMA). The primary motivation of the Media Transcoding Hints is to allow content servers, proxies or gateways to adapt image, video, audio and multimedia content to different network conditions, user and publisher preferences, and capabilities of terminal devices with limited communication, processing, storage and display capabilities. MPEG-7 provides the following types of transcoding hint information as part of the Media Transcoding Hints:

- **Importance** – specifies the relative importance of segments, regions, objects, or audio-visual programs. The importance takes values from 0.0 to 1.0, where 0.0 indicates the lowest importance and 1.0 indicates the highest importance.
- **Spatial resolution hint** – specifies the maximum allowable spatial resolution reduction factor for perceptibility. The SpatialResolutionHint takes values from 0.0 to 1.0, where 0.5 indicates that the resolution can be reduced by half, and 1.0 indicates the resolution cannot be reduced.
- **Shape hint** – specifies the amount of shape change in the media. The ShapeHint takes values from 0.0 to 1.0, where 0.0 indicates that no change has occurred and 1.0 indicates that all the pixels that define an object have been displaced.
- **Difficulty hint** – specifies the transcoding difficulty of the media. The DifficultyHint takes values from 0.0 to 1.0, where 0.0 indicates the lowest importance and 1.0 indicates the highest importance.
- **Motion hints** – specifies motion uncompensability and motion intensity information. Motion uncompensability specifies the amount of new content in a segment or region. Motion intensity specifies the motion intensity of a segment or region.

1.2.2. Content Management of Variations

The Variations describe different variations of multimedia data. The variations can be derived from the multimedia data, such as by applying methods for extraction, summarization or translation, or can simply represent alternative versions of multimedia data. In UMA applications, the variations can be selected and delivered as replacement, if necessary, to adapt to client terminal capabilities, such as display size, processing power, local storage, data format compatibility, or network conditions. The Variation DS represents the associations or relationships between different variations of audio-visual programs. The Variation DS serves important content management functionalities by tracking the variations of audio-visual content that result from various types of multimedia processing such as summarization, translation, reduction, revision and so forth. The Variation DS also serves an important role in applications such as Universal Multimedia Access by allowing the selection among the different variations of the audio-visual programs in order to select the most appropriate one in adapting to the specific capabilities of the terminal devices, network conditions or user preferences.

Figure 3 illustrates a set of Variations of an audio-visual program. The example shows the source video program in the lower left corner and eight variations, two others that are video programs, three that are images, two that are text, and one that is audio. Each of the variations has a specify fidelity value that indicates the fidelity of the variation program with respect to the source program. The following describes a sub-set of the Variations illustrated above using the VariationSet DS. The VariationSet DS describes first the source program giving its MediaInstance information and CreationInformation. Then following in the description is a set of three variation programs (sub-set of the eight illustrated above – C, E). For each variation program, the Variation DS gives MediaInstance information, fidelity, priority and VariationRelationship. The example shows in some cases that multiple VariationRelationships can be specified, such as in the case that the Variation program E is derived from the source program via both spatial reduction and compression.

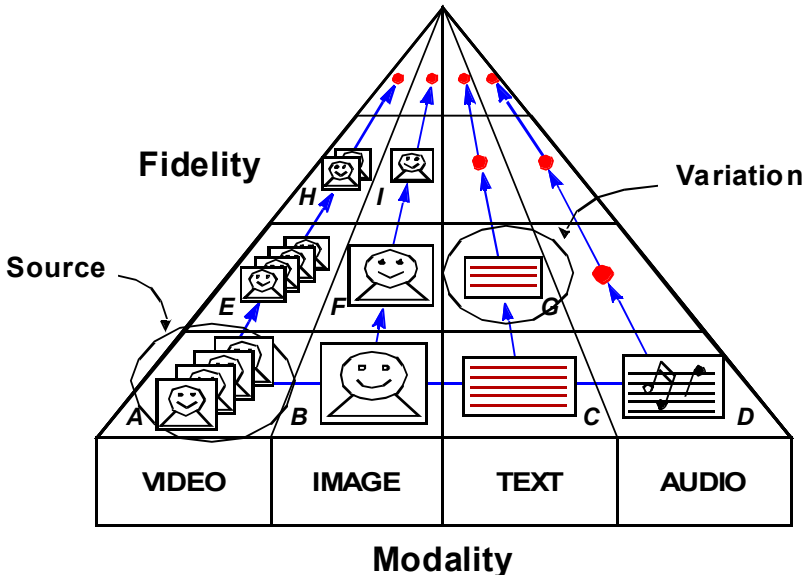


Figure 3 Figure 3: Illustration of different variations of a source audio-visual program. The variation programs have different modalities (video, image, text, audio) and fidelities with respect to the source program.

```
<VariationSet>
   <Source><Video>
          <MediaInformation>
               <MediaInstance>
                    <MediaLocator><MediaURI> A </MediaURI></MediaLocator>
               </MediaInstance>
          </MediaInformation>
          <CreationInformation>
               <Creation><Title><TitleText> Soccer video </TitleText></Title></Creation>
          </CreationInformation>
     </Video></Source>
   <Variation fidelity="0.75" priority="3">
          <VariationProgram>
               <MediaLocator><MediaURI> C </MediaURI></MediaLocator>
          </VariationProgram>
          <VariationRelationship> extract </VariationRelationship>
          <VariationRelationship> languageTranslation </VariationRelationship>
   </Variation>
   <Variation fidelity="0.5" priority="2">
          <VariationProgram>
               <MediaLocator><MediaURI> E </MediaURI></MediaLocator>
          </VariationProgram>
          <VariationRelationship> spatialReduction </VariationRelationship>
          <VariationRelationship> compression </VariationRelationship>
   </Variation>
</VariationSet>
```

1.2.3. Space and Frequency Views

The Space and Frequency Views provide a way to describe and manage progressive and layered representations of audio-visual data [5]. For example, the Space and Frequency Graph DS provides a data structure the embeds numerous important layered data representations such as those based on wavelet decompositions, multi-resolution pyramids, spatial quad-trees, spatial- and temporal-frequency subband decompositions, and so forth. The MPEG-7 View Partitions and Decompositions describe different partitions and decompositions of image, video and audio signals in the space and/or frequency domain. The MPEG-7 View Partitions describe different views of audio-visual content such as low-resolution views, spatial or temporal segments, or frequency subbands. Generally, the MPEG-7 Space and Frequency View DSs specify the views in terms of their corresponding partition in the space or frequency plane. The MPEG-7 View Decompositions describe different tree- and graph-based decompositions of the audio-visual signals and provide different organizations of views of

the audio-visual content. The decompositions specify node elements of the tree- and graph-based data structures that correspond to the Views and transition elements that correspond to the analysis and synthesis dependencies among the Views.

1.2.4. Summarization

The MPEG-7 Summaries facilitate discovery, browsing, navigation, visualization and sonification of audio-visual content. The descriptions provide compact summaries of the audio-visual content, which facilitate discovery, browsing, navigation, visualization and sonification. The Summary DSs allow the audio-visual content to be navigated in either a hierarchical or sequential fashion. The hierarchical summary decompositions organize the content into successive levels that describe the audio-visual content at different levels of detail from coarse to fine. The sequential summaries provide sequences of images or video frames, possibly synchronized with audio, that compose a slide-show or audio-visual skim. The Summarization DSs contain links to the audio-visual content, including the segments and frames. Given an MPEG-7 Summarization description, a terminal device, such as a digital television set-top box, can access the audio-visual material composing the summary and render the result for subsequent interaction with the user. The Summarization DS allow the creation of multiple summaries of the same content, which may be formed at different levels of detail. By including links to the audio-visual content in the summaries, it is possible to generate and store multiple summaries without storing multiple versions of the summary audio-visual content

2.Conclusions

In summary, MPEG-7 has taken an important forward-looking approach in standardizing meta-data for Universal Multimedia Access. MPEG-7 is addressing the management of variations of multimedia content, scalable representations of multimedia data and transcoding. These Description Schemes will have great utility in the deployment of multimedia applications that allow the scalable delivery of multimedia to users and terminals regardless of conditions of communication bandwidth, capabilities of terminal devices for processing, storage and display, and support for media formats. MPEG-7 enables UMA in a number of ways including: (1) transcoding – manipulating the content on-the-fly in order to adapt to different terminal and network conditions, (2) selection – storing, selecting and delivering different variations of the content from a content server, (3) progressive retrieval – providing scalable representations of audio-visual signals that allow them to be retrieved progressively, and (4) summarization – providing personalized summaries of audio-visual programs that allow mobile users to efficiently browse and view the content.

3. References

1. J. R. Smith, A. Puri, M. Tekalp, "MPEG-7 Multimedia Content Description Standard," *IEEE Intern. Conf. on Multimedia and Expo (ICME-2000),* July, 2000. Tutorial.
2. J. R. Smith, "Digital Video Libraries and the Internet," I*EEE Communications Mag*., Special issue on the Next Generation Internet, Vol. 37, No. 1, January, 1999, pp. 92 – 99.
3. J. R. Smith and S-F Chang, "Visually Searching the Web for Content," *IEEE Magazine*, Vol. 3, No. 4, July/September, 1997, pp. 12 – 20
4. J. R. Smith and S-F Chang, "VisualSEEk: a fully Automated Content-Based Retrieval System," *ACM Multimedia*, November, 1996, pp. 87 – 98.
5. J. R. Smith, "VideoZoom Spatio-temporal video browser," *IEEE Trans. Multimedia*, Vol. 1, No. 2, June, 1999, pp. 157 – 171
6. J. R. Smith and C.-S. Li, "Image classification and querying using composite region templates," *Journal of Computer Vision and Pattern Recognition*, Vol. 75, No. 1/2, July/August, 1999, pp. 165 – 174
7. P. Salembier and J. R. Smith, "MPEG-7 Multimedia Description Schemes," *IEEE Trans. Circuits and Systems for Video Technology,* 2001.
8. J. R. Smith, "Universal Multimedia Access," *SPIE Photonics East, Multimedia Systems and Applications III,* November 2000
9. J. R. Smith, R. Mohan, and C-S. Li. Scalable Multimedia Delivery for Pervasive Computing, *ACM Multimedia*, Orlando, FL, November 1999.
10. MPEG-7 Applications Document v.9.1, *ISO/IEC JTC1/SC29/WG11/N3548*, MPEG99, Beijing, CN, July 2000.
11. MPEG-7 Requirements Document V.10, *ISO/IEC JTC1/SC29/WG11/N2996*, MPEG99, Melbourne, Vic, October 1999.
12. MPEG-7 Content Description for Universal Multimedia Access, *ISO/IEC JTC1/SC29/WG11/M4749*, MPEG99, Vancouver, BC July 99.