Extraction, Description and Application of Multimedia Using MPEG-7

Ana B. Benitez
Depart. of Electrical Engineering, Columbia University,
1312 Mudd, 500 W 120th Street, MC 4712,
New York, NY 10027, USA
email: ana@ee.columbia.edu

Shih-Fu Chang
Depart. of Electrical Engineering, Columbia University,
1312 Mudd, 500 W 120th Street, MC 4712,
New York, NY 10027, USA
email: sfchang@ee.columbia.edu

Abstract - **In this paper, we present the multimedia description scheme tools specified by the MPEG-7 standard for describing multimedia data such as images and video. In particular, we focus on the description tools that represent the structure and semantics of multimedia data to whose development we have actively contributed. We also describe some of our research prototype systems dealing with the extraction and application of MPEG-7 structural and semantic descriptions. These systems are AMOS, a video object segmentation and retrieval system, and IMKA, an intelligent multimedia knowledge application using the MediaNet knowledge representation framework.**

I. INTRODUCTION

In recent years, there has been an important increase of available multimedia data in both the scientific and consumer domains and facilities to access the multimedia data. However, the extraction of useful information from the multimedia data and the application of this information in practical systems such as multimedia search engines are still open problems. The most important barrier has been the lack of a comprehensive, simple, and flexible representation of multimedia data that enables interoperable, scalable, and efficient multimedia applications.

MPEG-7 aims at standardizing tools for describing multimedia data to tear this barrier. The MPEG-7 framework consists of Descriptors (Ds), Description Schemes (DSs), a Description Definition Language (DDL), and coding schemes. Descriptors represent features or attributes of multimedia data such as color, texture, textual annotations, and media format. Description schemes specify more complex structures and semantics grouping descriptors and other description schemes such as segments of, and objects depicted in, multimedia data. The description definition language allows defining and extending descriptors and description schemes. Finally, coding schemes to compress descriptions and tools are needed to satisfy the storage and transmission requirements. MPEG-7 has been an international standard since 2001.

The specification of the MPEG-7 standard is divided into several parts, which correspond to the different MPEG groups working on it [5][8]. The parts of the MPEG-7 standard are systems - transmission and encoding of MPEG-7 descriptions and tools -, DDL - description definition language based on XML-Schema -, video - Ds and DSs for video data -, audio - Ds and DSs for audio data -, Multimedia Description Scheme (MDS) - generic Ds and DSs for any media-, reference software - reference implementation of the standard -, and conformance - guidelines and procedures for testing conformance to MPEG-7.

This paper provides an overview of the MPEG-7 MDS tools [4], which are organized on the basis of their functionality into basic elements, content description, content management, content organization, navigation and access, and user interaction. We have actively contributed to the development of these tools, especially to the content description tools, which are the focus of this paper. Content description tools describe the perceivable content of multimedia data including its structure and semantics.

The creation and application of MPEG-7 descriptions are outside the scope of the MPEG-7 standard. However, MPEG-7 is becoming a significant driver of new research for multimedia analysis, storage, searching, and filtering [6], among others. In this paper, we also present two of our research prototypes systems, AMOS and IMKA, which demonstrate the extraction and application of MPEG-7 structure and semantic descriptions, respectively. AMOS is a video object segmentation and retrieval system [7]. IMKA is an intelligent multimedia knowledge application using the MediaNet knowledge representation framework [2] [3].

This paper is organized as follows. In section I, we provide an overview of the MPEG-7 MDS tools. We describe the content description tools in detail providing examples in section II. Section III presents the research prototype systems AMOS and IMKA. Finally, we conclude with a summary in section IV.

II. OVERVIEW OF MPEG-7 MDS TOOLS

The MPEG-7 MDS tools [4] are organized on the basis of their functionality into basic elements, content description, content management, content organization, navigation and access, and user interaction, as shown in Figure 1. We shall provide an overview of these tools in this section.

The *basic elements* are the schema tools, basic datatypes, media localization tools, and basic tools repeatedly used to define other MPEG-7 MDS tools. The schema tools define the root and top-level elements for MPEG-7 descriptions, and the base types and packaging of MPEG-7 tools. The root element is the starting element of complete or partial MPEG-7 descriptions to allow both the complete and incremental transmission of MPEG-7 descriptions. MPEG-7 descriptions can be associ-

ated metadata such as such as version, creator, and rights. Top-level elements are the elements that immediately follow the root element in a description such as multimedia content entity (e.g., image, video, and multimedia collection.), abstraction (e.g., world, model, and summary), and management (e.g., user and creation) elements. The base types are the tools from which the MPEG-7 tools (e.g., Ds and DSs) are derived. Finally, package tools describe the packaging of MPEG-7 tools into hierarchical folders.

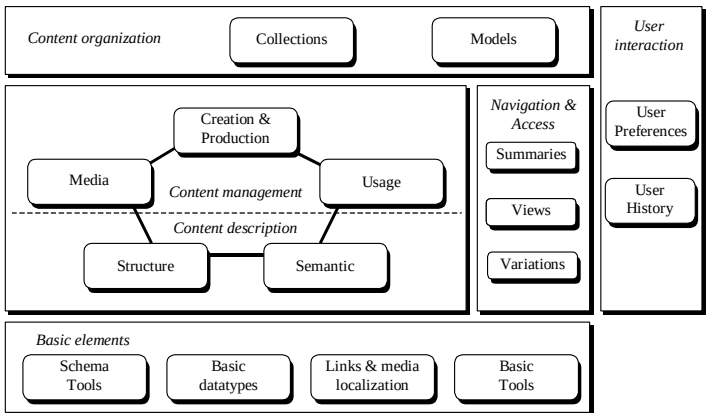

Fig. 1: Organization of the MPEG-7 MDS tools.

Basic datatypes represent constrained integer and real values, vectors and matrices, country and region codes, references, unique identifiers, time points and duration, etc. Media localization tools represent links to multimedia data using media URIs, inline media, and spatio-temporal locators. Some basic tools are text annotation tools, classification schemes and controlled term tools. Text annotation tools represent unstructured and structured textual annotations. Textual annotations can be structured by questions such as who, where, and when, by keywords, or by syntactic dependency relations. Classification scheme and controlled term tools describe hierarchical classifications of textual and graphical terms in subject-like areas, relations among the terms, and references to terms in classification schemes. Other basic tools describe persons, group of persons, organizations, geographical locations either real or fictional, relations and graph structures, criteria for ordering descriptions, audience's affective response to multimedia data, and the pronunciation of a set of words.

The *content management* tools relate to the media, creation and usage of the multimedia data. This information is not usually perceived in the media. Media tools describe media information of the multimedia data as a unique identifier and a set of media profiles. A media profile describes a coded version of the multimedia data in terms of format, quality, and transcoding hints, among others. Creation tools describe the creation and production of the multimedia data including the creation process (e.g., title, abstract, creator, creation place, time, and tools), the classification (e.g., target audience, genre, and rating), and related material (e.g., trailers and web pages for movies). Usage tools describe the usage process of the multimedia data including the rights, the publication (e.g., emission and edition), audience, and financial results.

The *content description* tools describe the perceivable content of multimedia data in terms of the structure and semantics of multimedia data. Structure description tools represent the structure of multimedia data in space, time, and media by describing segments such as still regions, video segments, and audio segments, and attributes, hierarchical decompositions, and relations of segments. Correspondingly, the structure description tools are segment entity, attribute, decomposition, and relation tools. Similarly, semantic description tools represent the narrative world depicted, or related to multimedia data by describing semantic entities in the narrative world such as objects, agent objects, events, concepts, semantic states, semantic places, and semantic times, and attributes and relations of semantic entities. Correspondingly, the semantic description tools are semantic entity, attribute, and relation tools.

The *content organization* tools address the collection, modeling, and classification of multimedia data. Collection tools describe unordered sets of multimedia data, segments, semantic entities, descriptors, or mixed sets of the above. These tools can also describe relations among collections. There are three types of model tools: probability, analytical, and cluster model tools. Probability model tools associate statistics or probabilities to collections; analytical model tools associate labels or semantics to collection; finally, the cluster model tools can associate not only statistics or probabilities but also labels or semantics to collections. Finally, classification model tools describe known collections in order to classify unknown ones.

The *navigation and access* tools describe summaries, partitions and decompositions, and variations of multimedia data. Hierarchical and sequential summary tools describe hierarchical or sequential summaries of multimedia data. Hierarchy summaries are composed of highlight segments of the multimedia data that are organized in hierarchies. Sequential summaries are composed of images and audio clips that can be synchronized and presented to the user at different speeds. Partitions and decomposition tools describe decomposition of images, video, or audio-visual data in space and/or frequency partitions. Variation tools describe relations between different variations of multimedia data (e.g., summary, revision, and modality conversion).

Finally, the *user interaction* tools describe user preferences and history in the consumption of multimedia data. User preference tools describe preferences of users pertaining to the filtering, search, and browsing of multimedia data (e.g., creation and classification preferences). The user history tools describe history of users in consuming the multimedia data as set of user actions (e.g., record, pause, and play) with associated temporal information.

III. CONTENT DESCRIPTION TOOLS

We have actively contributed to the development of the MDS descriptors and description schemes, in particular, to the content description tools, which are described in more detail in this section. As mentioned in the previous section, content description tools describe the structure and semantics of mul-

timedia data. Figure 2 shows the structure and semantic descriptions of an image.

A. Structure Description Tools

The structure description tools represent the structure of multimedia data in space, time, and media by describing general and application-specific *segments* of multimedia data together with their attributes, hierarchical decompositions, and relations.

The most basic visual segment is the *still region*, which is a group of pixels in a 2D image or a video frame. Figure 2 shows three examples of still regions corresponding to the full image and the two regions depicting the persons in the image (SR1, SR2, and SR3). Other strictly visual segments are the *video segment*, defined as a group of pixels in a video, and the *moving region*, defined as a set of pixels in a group of frames in a video. The *audio segment* is defined as a group of samples in an audio sequence. Both *audio-visual segment* and the *audio-visual region* contain audio and visual data in an audio-visual sequence corresponding to a group of frames and audio samples synchronized in time, and an arbitrary group of pixels and audio samples, accordingly.

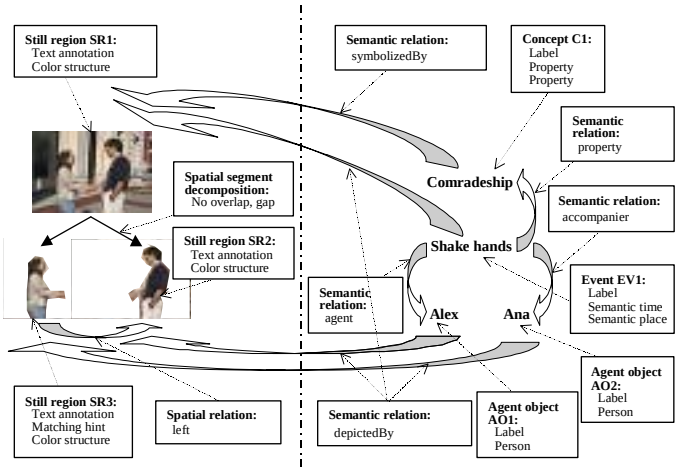

Fig. 2: Structure (left) and semantic (right) description of an image.

The application-specific segments are *mosaics* (a panoramic view of a video segment), *3D still regions* (3D spatial region of a 3D image), *image* and *video texts* (still region and moving region corresponding to text, respectively), *ink segments* (temporal segment of electronic ink data), *multimedia segments* (composite of segments forming a multimedia presentation such as an MPEG-4 presentation or a web page), and edited video segments (video segments resulting from editing work).

Visual and audio features, and media, creation and usage information of the multimedia data in segments can be described using the visual and audio Ds and DSs, and the media, creation and usage tools, respectively. *Segment attribute* description tools describe other attributes of segments related to spatio-temporal, media, and graph masks; importance for matching and point of view; creation and media of ink segments; and hand writing recognition (e.g., lexicon and results).

The *segment decomposition* tools describe the decomposition of segments in hierarchies of segments to form, for example, tables of contents of multimedia data. MPEG-7 has defined four types of segment decompositions: spatial, temporal, spatio-temporal, and media decompositions. An image can be decomposed spatially into a set of still regions corresponding to objects in the image, which, at the same time, can be decomposed into other still regions. Figure 2 exemplifies the spatial decomposition of a still region into two still regions. Similar decompositions can be generated in time and/or space for video and other multimedia data. A media decomposition divides segments into its media constituents such as audio and video tracks. The sub-segments resulting from a decomposition may overlap in time, space, and/or media; furthermore, their union may not cover the full time, space, and media extents of the parent segment, thus leaving gaps. The spatial decomposition shown in Figure 2 has gaps but no overlaps.

The *structural relation* description tools can describe general relations among segments. Figure 2 shows an example of the spatial relation *left* between two still regions (SR2 and SR3). Current normative structural relations in MPEG-7 are spatial relations (e.g., “left” “north”, “above” and “under”), and temporal relations (e.g., “before”, “sequential” and “parallel”). Other relations that can be described among segments are basic relationships (“union” and “intersection”), and semantic relations (e.g., “keyFor” -a segment is key for another- and “annotates” - a segment describes another-).

B. Semantic Description Tools

The semantic description tools represent narrative worlds depicted by or related to multimedia data by describing *semantic entities* in the narrative world such as objects, agent objects, events, concepts, semantic states, semantic places, and semantic times, together with their attributes and relations.

A *narrative world* refers to the reality in which the description makes sense, which may be the world depicted in the multimedia data (e.g., earth world and Roman mythology world). *Objects* and *events* are perceivable entities that exist or take place in time and space in the narrative world, respectively. *Agent objects* are objects that are persons, group of persons, or organizations. As example, the semantic description in Figure 2 shows an event and two agent objects corresponding to "Shake hands" (EV1), and the two persons "Alex" and "Ana" (AO1 and AO2). *Concepts* are entities that cannot be perceived in the narrative world or described as the generalization of perceivable semantic entities. "Comradeship" (C1) in Figure 2 is a concept. *Semantic states* are parametric attributes of semantic entities and semantic relations at a specific point in time and space (e.g., weight and height of person). Finally, *semantic places and times* are locations and times in the narrative world, respectively. The event "Shake hands" have associated a semantic place and a semantic time.

Semantics entities can be described in terms of labels, a textual definition, properties, and features of the segments where they appear. *Semantic attribute* description tools describe other attributes of semantic entities related to abstraction levels and semantic measurements in time and space. Abstraction

refers to the process of taking a concrete semantic description of multimedia data (e.g., "Alex is shaking hands with Ana" in the image shown in Figure 2) and generalizing it to a set of multimedia data (media abstraction, e.g., "Alex is shaking hands with Ana" in any picture) or a set of concrete semantic descriptions (formal abstraction, e.g., "Any man is shaking hands with any woman").

The *semantic relation* description tools describe general relations among semantic entities and other entities. Current normative semantic relations in MPEG-7 can describe how semantic entities relate in a narrative (e.g., “agent”-an object that initiates the action of an event- and “accompanier”-an object that is a join agent in an event-,). Semantics relations can also describe how the definitions of semantics entities relate (e.g., “generalizes”–a semantic entity has a more general meaning than another and “exemplifies"-a model that is an example of a semantic entity-) and the localization of semantic entities in space, time and media (“depictedBy”-a segment that depicts a semantic entity- and “symbolizedBy”-a segment that symbolizes a semantic entity-). Figure 2 includes examples of some semantic relations.

IV. RESEARCH PROTOTYPE SYSTEMS

The creation and application of MPEG-7 descriptions are outside the scope of the MPEG-7 standard. However, MPEG-7 is becoming a significant driver of new research for multimedia analysis, storage, searching, and filtering, [6], among others. In this section, we present two of our research prototypes systems, AMOS and IMKA, which demonstrate the generation and application of MPEG-7 structure and semantic descriptions, respectively, in a retrieval application.

A. AMOS: Video Object Segmentation and Search

AMOS is a video object segmentation and retrieval system [7]. In this framework, a video object (e.g. person, car) is modeled and tracked as a set of regions with corresponding visual features and spatio-temporal relations (see Figure 3.a). The region-based model also provides an effective base for similarity retrieval of video objects.

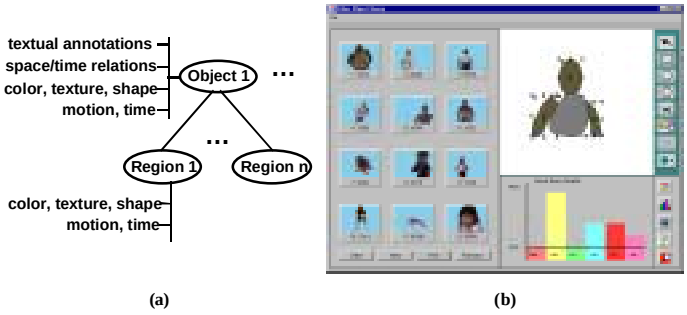

Fig. 3. (a) Video object representation in AMOS. (b) Query interface of AMOS: query results (left), query canvas (top right), and feature weights (bottom right).

AMOS effectively combines user input and automatic region segmentation for defining and tracking video objects at a semantic level. First, the user roughly outlines the contour of an object at the starting frame, which is used to create a video object with underlying homogeneous regions. This process is based on a region segmentation method that involves color and edge features and a region aggregation method that classifies regions into foreground and background. Then, the object and the homogeneous regions are tracked through successive frames. This process uses affine motion models to project regions from frame to frame and a color-based region growing to determine the final projected regions. Users can stop the segmentation at any time to correct the contour of video objects. Extensive experimental results have demonstrated excellent results. Most tracking errors are caused by uncovered regions and can be corrected with a few user inputs.

AMOS also extracts salient regions within video objects that users can interactively create and manipulate. Visual features and spatio-temporal relations are computed for video objects and salient regions and stored in a database for similarity matching. The features include motion trajectory, dominant color, texture, shape, and time descriptors. Currently three types of relations among the regions of a video object are supported: orientation spatial (angle between two regions), topological spatial (contains, does not contain, or inside), and directional temporal (start before, at the same time, or after). Users can enter textual annotations for the objects.

AMOS accepts queries in the form of sketches or examples and returns similar video objects based on different features and relations. Figure 3.b shows the query interface of AMOS. The query process of finding candidate video objects for a query uses a filtering together with a joining scheme. The first step is to find a list candidate regions from the database for each query region based on the visual features. Then, the region lists are joined to obtain candidate objects and their total distance to the query is computed by matching the spatio-temporal relations.

The mapping of the video object representation of AMOS to MPEG-7 Ds and DSs in straightforward. In fact, the AMOS system has been used to generate descriptions to evaluate parts of MPEG-7 [1]. A video can be described as a video segment that is spatio-temporally decomposed into moving regions corresponding to the segmented video objects. In the same way, these moving regions can be decomposed into other moving regions corresponding to the salient regions of the objects. The visual features, textual annotations, and spatio-temporal relations can be described using the normative visual, textual annotation, and segment relation tools.

B. IMKA: Intelligent Multimedia Knowledge Application

IMKA, is an intelligent multimedia knowledge application using the MediaNet knowledge representation framework [2][3]. MediaNet uses multimedia information for exemplifying semantic and perceptual information about the world [3]. MediaNet knowledge bases can be built from collections of annotated images [2] and used to enhance the retrieval of multimedia data [3].

MediaNet represents the world using concepts and relationships between the concepts that are defined and exemplified by multimedia information such as text, images, video sequences,

and audio-visual descriptors. In MediaNet, concepts can represent either semantically meaningful objects or perceptual patterns in the world. MediaNet models the traditional semantic relationship types such as generalization and aggregation but adds additional functionality by modeling perceptual relationships based on feature descriptor similarity and constraints. Figure 4 shows an example of a MediaNet knowledge base illustrating the concepts Human and Hominid.

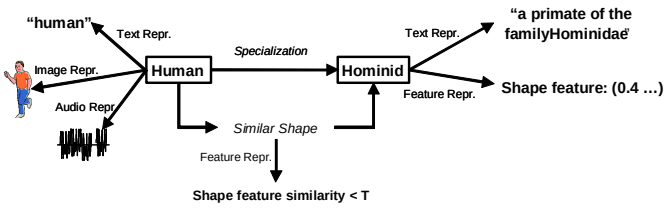

Fig. 4. MediaNet knowledge base illustrating the concepts Human and Hominid.

We construct a MediaNet knowledge base automatically from a collection of annotated images as follows [2]. First, we discover low-level or perceptual knowledge by clustering the images based on visual and text feature descriptors, and discovering similarity and statistical relationships between the clusters. We extract high-level or semantic knowledge by disambiguating the senses of words in the annotations using WordNet and the images clusters, and finding semantic relations between the senses in WordNet. Then, interrelations among concepts (e.g., clusters and senses) are discovered using classifiers and probabilistic Bayesian networks. Finally, we can reduce the size of, or summarize multimedia knowledge by clustering the concepts based on their distances. We have also proposed techniques for evaluating the quality of the extracted knowledge using information and graph theory notions.

An intelligent image retrieval system for images has been implemented by extending a typical image retrieval system with a MediaNet knowledge base and a query processor that translates and expands queries across multiple media modalities. First, the query processor classifies each incoming query into a set of relevant concepts based on features extracted from the query and the image examples of the concepts. The initial set of relevant concepts is then extended with other semantically similar concepts. A visual query is issued to the descriptor similarity-based search engine for the incoming query and each relevant concept. Finally, the results of all the queries are merged into a unique list based on how similar the concepts that generated those results were to the initial user query. We have found that MediaNet can improve the performance of image retrieval applications for semantic queries (e.g., find Tapirs and other animals); however, additional experiments are needed to further demonstrate the performance gain.

MediaNet knowledge bases can be encoded using MPEG-7 semantic and model tools, which could greatly benefit the exchange and re-use of knowledge and intelligence among multimedia applications. Each concept in MediaNet can be a semantic entity. The textual examples of a concept can be textual labels of the semantic entity. Other media examples of concepts can be described using probability models and audio-visual examples of semantic entities. Relationships among concepts in MediaNet can be encoded as relationships among the corresponding semantic entities.

V. SUMMARY

MPEG-7 provides a comprehensive suite of tools for describing multimedia data that has the potential to revolutionize current multimedia applications. In particular, MPEG-7 includes tools for describing the structure, semantics, media, creation, usage, collections, models, summaries and views of multimedia data that allow efficiently searching, filtering, browsing, and access of multimedia. In addition, MPEG-7 can also describe user preferences and history in consuming multimedia data that allow personalized multimedia services and devices. However, MPEG-7 has also introduced many new research problems related to the extraction and application of MPEG-7 descriptions. This paper has provided an overview of these tools focusing on the tools for describing the structure and semantics of multimedia data. It has also presented two research prototypes systems that can extract and use semantic and structure descriptions in a retrieval scenario.

VI. ACKNOWLEDGMENTS

The MPEG-7 tools presented in this paper are the result of the contributions and collaborative efforts of many people. The authors are particularly grateful to the members of the MPEG MDS Group for their many contributions in the development of MPEG-7.

VII. REFERENCES

[1] A. B. Benitez and S.-F. Chang, "Validation Experiments on Structural, Conceptual, Collection, and Access Description Schemes for MPEG-7", Digest of the IEEE 2000 International Conference on Consumer Electronics (ICCE-2000), Los Angeles, CA, June 13-15, 2000.

[2] A. B. Benitez and S.-F. Chang, "Automatic Multimedia Knowledge Discovery, Summarization and Evaluation", IEEE Trans. on Multimedia, 2003 (submitted); available at http://www.ee.columbia.edu/dvmm/newPublication.htm.

[3] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation", Proceedings of IS&T/SPIE 2000 Conference on Internet Multimedia Management Systems, Vol. 4210, Boston, MA, Nov. 6-8, 2000.

[4] MPEG MDS Group, "Text of ISO/IEC 15938-5 FDIS Information Technology –Multimedia Content Description Interface – Part 5 Multimedia Description Schemes", ISO/IEC JTC1/SC29/WG11 MPEG01/N4242, Sydney, July 2001.

[5] MPEG, Working Documents for MPEG-7 Standard, http://www.cselt.it/mpeg/working_ documents.htm.

[6] Y. C. Chang, M. L. Lo, J. R. Smith, "Issues and solutions for storage, retrieval, and search of MPEG-7 documents", Proceedings of IS&T/SPIE 2000 Conference on Internet Multimedia Management Systems, Vol. 4210, Boston, MA, Nov. 6-8, 2000.

[7] D. Zhong and S.-F. Chang, "An Integrated System for Content-Based Video Object Segmentation and Retrieval", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp.1259-1268, 1999.

[8] B. S. Manjunath, P. Salembier and T. Sikora, "Introduction to MPEG 7: Multimedia Content Description Language", Wiley, 2002.