Management of Multimedia Semantics Using MPEG-7

Uma Srinivasan and Ajay Divakaran

TR2004-116 October 2004

Abstract

This chapter presents the ISO/IEC MPEG-7 Multimedia Content description Interface Standard from the point of view of managing semantics in the context of multimedia applications. We describe the organisation and structure of the MPEG-7 Multimedia Description schemes, which are metadata structures for describing and annotating multimedia content at several levels of granularity and abstraction. As we look at MPEG-7 semantic descriptions, we realise they provide a rich framework for static descriptions of content semantics. As content semantics evolves with interaction, the human user will have to compensate for the absence of detailed semantics that cannot be specified in advance. We explore the practical aspects of using these descriptions in the context of different applications and present some pros and cons from the point of view of managing multimedia semantics.

Management of Multimedia Semantics using MPEG-7

Uma Srinivasan
CSIRO ICT Centre
Locked Bag 17, North Ryde NSW 1670
Australia
Ph: +612 9325 3100
Fax: +612 9325 3200
Email: Uma.Srinivasan@csiro.au

Ajay Divakaran
Mitsubishi Electric Research Laboratories
201 Broadway, Cambridge, MA 02139
Tel: 617-621-7521
Fax: 617-621-7550
Email: ajayd@merl.com

ABSTRACT

This chapter presents the ISO/IEC MPEG-7 Multimedia Content description Interface Standard from the point of view of managing semantics in the context of multimedia applications. We describe the organisation and structure of the MPEG-7 Multimedia Description schemes, which are metadata structures for describing and annotating multimedia content at several levels of granularity and abstraction. As we look at MPEG-7 semantic descriptions, we realise they provide a rich framework for static descriptions of content semantics. As content semantics evolves with interaction, the human user will have to compensate for the absence of detailed semantics that cannot be specified in advance. We explore the practical aspects of using these descriptions in the context of different applications and present some pros and cons from the point of view of managing multimedia semantics.

1. Introduction

MPEG-7 is an ISO/IEC Standard that aims at providing a standard way to describe multimedia content, to enable fast and efficient searching and filtering of audiovisual content. MPEG-7 has a broad scope to facilitate functions such as indexing, management, filtering, authoring, editing, browsing, navigation, and searching content descriptions. The purpose of the standard is to describe the content in a machine-readable format for further processing determined by the application requirements.

Multimedia content can be described in many different ways depending on the context, the user, the purpose of use and the application domain. In order to address the description requirement of a wide range of applications, MPEG-7 aims to describe content at several levels of granularity and abstraction to include description of features, structure, semantics, models, collections and metadata about the content.

Initial research focus on feature extraction techniques influenced the description of content at the perceptual feature level. Examples of visual features that can be extracted using image processing techniques are colour, shape and texture. Accordingly, there are several MPEG-7 Descriptors (Ds) to describe visual features. Similarly there is a number of low level Descriptors to describe audio content at the level of spectral, parametric and temporal features of an audio signal. While these Descriptors describe objective measures of audio and visual features, they are inadequate for describing content at a higher level of semantics to describe relationships among audio and visual descriptors within an image or over a video segment. This need is addressed through the construct called Multimedia Descriptions Scheme (MDS), also referred to simply as Description Scheme (DS). Description schemes are designed to describe higher-level content features such as regions, segments, objects and events, as well as metadata about the content, its usage, etc. Accordingly there are several groups or categories of MDS tools.

An important factor that needs to be considered while describing audiovisual content is the recognition that humans start to interpret and describe the meaning of the content that goes far beyond visual features and cinematic constructs introduced in films. While such meanings and interpretations cannot be extracted automatically, as they are contextual, they can be described using free text descriptions. MPEG-7 handles this aspect through several description schemes that are based on structured free text descriptions.

As our focus is on management of multimedia semantics, we look at MPEG-7 MDS constructs from two perspectives: (a) the level of granularity offered while describing content and (b) the level of abstraction available to describe multimedia semantics. Section 2 provides an overview of the MPEG-7 constructs and how they hang together. Section 3 looks at MDS tools to manage multimedia semantics at multiple levels of granularity and abstraction. Section 4 takes a look at the whole framework form the perspective of different applications. Section 5 presents some discussions and conclusions.

2. MPEG-7 Content Description and Organisation

The main elements of MPEG-7 as described in the MPEG-7 Overview document [2] are a set of tools to describe the content, a language to define the syntax of the descriptions, and system tools to support efficient storage and transmission, execution and synchronization of binary encoded descriptions.

The Description Tools provide a set of Descriptors (D) that define the syntax and the semantics of each feature, and a library of Description Schemes (DS) that specify the structure and semantics of the relationships between their components, that may be both Descriptors and Description Schemes. A description of a piece of audiovisual content is made up of a number of D’s and DS’s determined by the application. The description tools can be used to create such descriptions, which form the basis for search and retrieval. A Description Definition Language (DDL) is used to create and represent the descriptions. DDL is based on XML and hence allows the processing of descriptions in a machine-readable format. Content descriptions created using these tools could be stored in variety of ways. The descriptions could be physically located with the content in the same data stream or the same storage system, allowing efficient storage and retrieval. However, there could be instances where content and its descriptions may not be co-located. In such cases, we need effective ways to synchronise the content and its Descriptions. System tools support multiplexing of description, synchronization issues, transmission mechanisms, file format, etc. Figure 1 [2] shows the main MPEG-7 elements and their relationships.

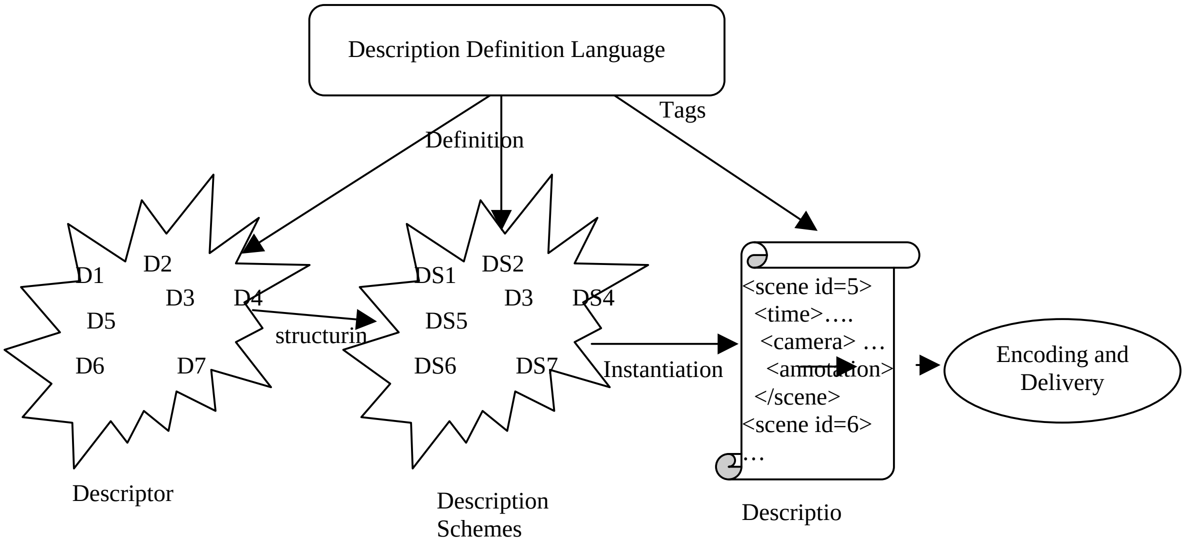

Figure 1: MPEG-7 Elements

MPEG-7 has a broad scope and aims to address the needs of several types of applications [3]. MPEG-7 descriptions of content could include:

- Information describing the creation and production processes of the content (director, title, short feature movie).
- Information related to the usage of the content (copyright pointers, usage history, and broadcast schedule).
- Information of the storage features of the content (storage format, encoding).
- Structural information on spatial, temporal or spatio-temporal components of the content (example: scene cuts, segmentation in regions, region motion tracking).
- Information about low-level audio and visual features in the content (example: colors, textures, sound timbres, melody description).
- Conceptual information of the reality captured by the content (example: objects and events, interactions among objects).

- Information about how to browse the content in an efficient way (example: summaries, variations, spatial and frequency sub-bands).
- Information about collections of objects.
- Information about the interaction of the user with the content (user preferences, usage history).

MPEG-7 Multimedia Description Schemes (MDS) are metadata structures for describing and annotating audiovisual content at several levels of granularity and abstraction (to describe what's in the content) and metadata - a description about the content. These Multimedia Descriptions Schemes are described using XML to support readability at the human level and processing capability at the machine level.

MPEG-7 Multimedia DS's are categorised and organised into the following groups: Basic Elements, Content Description, Content Management, Content Organization, Navigation and Access, and User Interaction. Figure 2 shows the different categories and presents a big picture view of Multimedia DS's.

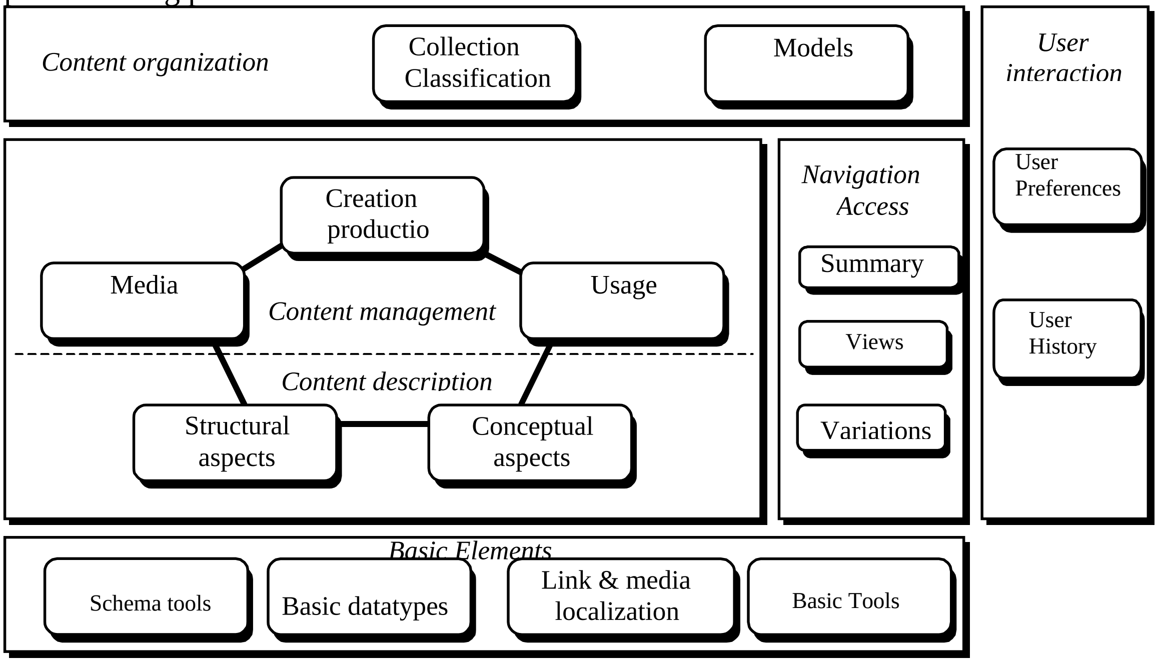

Figure 2: Overview of MPEG-7 Multimedia Description Schemes [2].

Basic Elements

Basic elements provide the fundamental constructs in defining MPEG-7 DS's. This includes basic data types and a set of extended data types such as vectors, matrices to describe the features and structural aspects of the content. The basic elements also include constructs for linking media files, localising specific segments, describing time and temporal information, place, individual/s, groups, organizations, and other textual annotations.

Content Description

MPEG-7 DS's for content description are organised into two categories: DS's for describing structural aspects and DS's for describing conceptual aspects of the content. The structural DS's describe audiovisual content at a structural level organised around a segment. The *Segment* DS represents the spatial, temporal or spatio-temporal structure of an audiovisual segment. The *Segment* DS can be organised into a hierarchical structure to produce a table of contents for indexing and searching audiovisual content in a structured way. The segments can be described at different perceptual levels using Descriptors for colour, texture, shape, motion, and so on. The conceptual aspects are described using *Semantic* DS, to describe objects, events and abstract concepts. The structure DS's and semantic DS's are related by a set of links that relate different semantic concepts to content structure. The links relate semantic concepts to instances within the content described by the segments. Many of the content description DS's are linked to D's are linked to DS's in content management group.

Content Management

MPEG-7 DSs for content management include tools to describe information pertaining to creation and production, media coding, storage and file formats, and content usage.

Creation information provides information related to the creators of the content, creation locations, dates, other related material, etc. These could be textual annotations or other multimedia content such an image such as a logo. This also includes information related to classification of the content from a viewer's point of view.

Media information describes information including location, storage and delivery formats, compression and coding schemes, and version history based on media profiles.

Usage information describes information related to usage rights, usage record, and related financial information. While rights management is not handled explicitly, the Rights DS provides references in the form unique identifiers to external rights owners and regulatory authorities.

Navigation and Access

The DS's under this category facilitate browsing and retrieval of audiovisual content. There are DS's that facilitate browsing in different ways based on summaries, partitions and decompositions and other variations. The *Summary* DS's support hierarchical and sequential navigation modes. Hierarchical summaries can be described at different levels of granularity, moving from coarse high-level descriptions to more detailed summaries of audiovisual content. Sequential summaries provide a sequence of images and frames synchronised with audio and facilitate a slide show style of browsing and navigation.

Content Organisation

The DS's under this category facilitate organising and modelling collections of audiovisual content descriptions. The *Collection* DS helps to describe collections at the level of objects, segments, and events, based on common properties of the elements in the collection.

User Interaction
This set of DS's describes user and usage preferences, usage history to facilitate personalization of content access, presentation and consumption.
For more details of the full list of DS's, the reader is referred to the MPEG-7 URL at: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm and [1].

3. Representation of Multimedia semantics

In the previous section we described the MPEG-7 constructs and the method of organising the MDS from a functional perspective, as presented in various official MPEG-7 documents. In this section we look at the Ds and DS's from the perspective of addressing multimedia semantics and its management. We look at the levels of granularity and abstraction that MPEG-7 Ds and DS's are able to support. The structural aspects of content description are meant to describe content at different levels of granularity ranging from visual descriptors to temporal segments. The semantic DS's are developed for the purpose of describing content at several abstract levels in a free text, but in a structured form.

MPEG-7 deals with content semantics by considering "narrative worlds." Since MPEG-7 targets description of multimedia content, which is mostly narrative in nature, it is reasonable for it to view the participants, background, context, and all the other constituents of a narrative as the narrative world. Each narrative world can exist as a distinct semantic description. The components of the semantic descriptions broadly consist of entities that inhabit the narrative worlds, their attributes and their relationships with each other.

3.1 Levels of Granularity

Let us consider a video of a play that consists of four acts. Then we can segment the video temporally into four parts corresponding to the acts. Each act can be further segmented into scenes. Each scene can be segmented into shots while a shot is defined as a temporally continuous segment of video captured by a single camera. The shots can in turn be segmented into frames. Finally, each frame can be segmented into Spatial Regions. Note that each level of the hierarchy lends itself to meaningful semantic description. Each level of granularity lends itself to distinctive D's. For instance, we could use the *texture* descriptor to describe the texture of spatial regions. Such a description is clearly confined to the lowest level of the hierarchy we just described. The 2-D shape descriptors are similarly confined by definition. Each frame can also be described using the scalable *color* descriptor, which is essentially a color histogram. A shot consisting of several frames however has to be described using the *Group of Frames color* descriptor, which aggregates the histograms of all the constituent shots using for instance, the median. Note that while it is possible to extend the *Color* description to a video segment of any length of time, it is most meaningful at the shot level and below. The *MotionActivity* descriptor can be used to meaningfully describe any length of video, since it merely captures the "pace" or action in the video. Thus, a talking head segment would be described as “low action” while a “car chase” scene would be described as “high action.” A one-hour movie that mostly consists of car chases could reasonably be described as “high action.” The motion trajectory descriptor on the other hand is meaningful only at the shot level and meaningless at any lower or higher level. In other words, each level of granularity has its own set of appropriate descriptors that may or may not be appropriate at all levels of the hierarchy. The aim of such description is to enable content retrieval at any desired level of granularity.

3.2 Levels of Abstraction

Note that in the previous section, the hierarchy stemmed from the temporal and spatial segmentation, but not from any conceptual point of view. Therefore such a description does not let us browse the content at varying levels of semantic abstraction that may exist at a given constant level of temporal granularity. For instance, we may be only interested in dramatic dialogues between characters A and B in one case, and in any interactions between character A and character B in another. Note that the former is an instance of the latter and therefore is at a lower level of abstraction. In the absence of multi-layered abstraction, our content browsing would have to be either excessively general through restriction to the highest level of abstraction, or excessively particular through restriction to the lowest level of abstraction. Note that to a human being, the definition of “too general” and “too specific” depends completely on the need of the moment, and therefore is subject to wide variation. Any useful representation of the content semantics has to therefore be at as many levels of abstraction as possible.

Returning to the example of interactions between the characters A and B, we can see that the semantics consists of the entities A and B, with their names being their attributes and whose relationship with each other consists of the various interactions they have with each other. MPEG-7 considers two types of abstraction. The first is media abstraction, i.e. a description that can describe more than one instance of similar content. We can see that the description “all interactions between characters A and B,” is an example of media abstraction since it describes all instances of media in which A and B interact. The second type of abstraction is formal abstraction, in which the pattern common to a set of multimedia examples contains placeholders. The description “interaction between any two of the characters in the play” is an example of such formal abstraction. Since the definition of similarity depends on the level of detail of the description and the application, we can see that these two forms of abstraction allow us to accommodate a wide range of abstraction from the highly abstract to the highly concrete and detailed.

Furthermore, MPEG-7 also provides ways to describe abstract quantities such as properties, through the Property element, and concepts, through the *Concept* DS. Such quantities do not result from an abstraction of an entity, and so are treated separately. For instance, the beauty of a painting is a property and is not the result of somehow generalizing its constituents. Concepts are defined as collections of properties that define a category of entities but do not completely characterize it.

Semantic entities in MPEG-7 mostly consist of narrative worlds, objects, events, concepts, states, places and times. The objects and events are represented by the *Object* and *Event* DS's respectively. The *Object* DS and *Event* DS provide abstraction through a recursive definition that allows for example sub-categorization of objects into sub-objects. In that way, an object can be represented at multiple levels of abstraction. For instance, a continent could be broken down into continent-country-state-district etc., so that it can be described at varying levels of semantic granularity. Note that the *Object* DS accommodates attributes so as to allow for the abstraction we mentioned earlier, i.e. abstraction that is related to properties rather than generalization of constituents such as districts. The "hospitable nature of the continent's inhabitants" for instance cannot result from abstraction of districts to states to countries etc.

Semantic entities can be described by labels, by a textual definition, or in terms of properties or of features of the media or segments in which they occur. The *SemanticBase* DS contains such descriptive elements. The AbstractionLevel data type in the *SemanticBase* DS describes the kind of abstraction that has been performed in the description of the entity. If it is not present, then the description is considered concrete. If the abstraction is a media abstraction, then the dimension of the AbstractionLevel element is set to zero. If a formal abstraction is present, the dimension of the element is set to one or higher. The higher the value is, the higher is the abstraction. Thus, a value of 2 would indicate an abstraction of an abstraction.

The *Relation* DS rounds off the collection of representation tools for content semantics. Relations capture how semantic entities are connected with each other. Thus examples of a relation is "doctor-patient," "student-teacher" etc. Note that since each of the entities in the relation lends itself to multiple levels of abstraction and the relations in turn have properties, there is further abstraction that results from relations.

4. Applications

As we cover MPEG-7 semantic descriptions, we realize that they provide a rich framework for static description of content semantics. Such a framework has the inherent problem of providing an embarrassment of riches, which makes the management of the browsing very difficult. Since MPEG-7 content semantics is very graph-oriented, it is clear that it does not scale well as the number of concepts/events/objects goes up. Creation of a deep hierarchy through very fine semantic subdivision of the objects would result in the same problem of computational intractability. As the content semantic representation is pushed more and more towards a natural language representation, evidence from natural language processing research indicates that the computational intractability will be exacerbated. In our view therefore, the practical utility of such representation is restricted to cases in which either the concept hierarchies are not unmanageably broad, or the concept hierarchies are not unmanageably deep or both.

Our view is that in interactive systems, the human uses will compensate for the shallowness or narrowness of the concept hierarchies through their domain knowledge. Since humans are known to be quick at sophisticated processing of data sets of small size, the semantic descriptions should be at a broad scale to help narrow down the search space. Thereafter, the human can compensate for the absence of detailed semantics through use of low-level feature based video browsing techniques such as video summarization. Therefore, MPEG-7 semantic representations would be best used in applications in which a modest hierarchy can help narrow down the search space considerably. Let us consider some candidate applications:

Educational Applications

At first glance, since education is after all intended to be systematic acquisition of knowledge, a semantics-based description of all the content seems reasonable. Our experience indicates that restriction of the description to a narrow topic allows for a rich description within the topic of research and makes for a successful learning experience for the student. Any application in which the intention is to learn abstract concepts, an overly shallow concept hierarchy will be a hindrance. Hence, our preference for narrowing the topic itself to limit the breadth of the representation so as to buy some space for a deeper representation. The so called "edutainment" systems fall in the same general category with varying degrees of compromise between the richness of the descriptions and the size of the database. Such applications include tourist information, cultural services, shopping, social, film and radio archives etc.

Information Retrieval Applications

Applications that require retrieval from an archive based on a specific query rather than a top-down immersion in the content, typically consist of very large databases in which even a small increase in the breadth and depth of the representation would lead to an unacceptable increase in computation. Such applications include journalism, investigation services, professional film and radio archives, surveillance, remote sensing, etc. Furthermore, in such applications, the accuracy requirements are much more stringent. Our view is that only a modest MPEG-7 content semantics representation would be feasible for such applications. However, even a modest semantic representation would be a vast improvement over current retrieval.

Generation of MPEG-7 Semantic Meta-Data

It is also important to consider how the descriptions would be generated in the first place. Given the state of the art, the semantic meta-data would have to be manually generated. That is yet another challenge posed by large-scale systems. Once again, the same strategy of either tackling modest databases, or creating modest representations or a combination of both would be reasonable. Once again, if the generation of the meta-data is integrated with its consumption in an interactive application, the user could enhance the meta-data over time. This is perhaps a challenge for future researchers.

5. Discussion and Conclusion

Managing multimedia content has evolved around textual descriptions and/or processing audiovisual information and indexing them using features that can be automatically extracted. The question is how do we retrieve the content in a meaningful way. How can we correlate user's semantics with archivist's semantics? Even though MPEG-7 DS's provide a framework to support such descriptions, MPEG-7 is still a standard for describing features of multimedia content. Although there are DS's to describe the metadata related to the content, there is still a gap in describing semantics that evolves with interaction and user's context. There is a static aspect to the descriptions, which limits adaptive flexibility needed for different types of applications. Nevertheless, a standard way to describe the relatively unambiguous aspects of content does provide a starting point for many applications where the focus is content management.

The generic nature of MPEG-7 descriptions can be both strength and a weakness. The comprehensive library of DS's is aimed to support a large number of applications, and there are several tools to support the development of descriptions required fro a particular application. However, this requires a deep knowledge of MPEG-7, and the large scope becomes a weakness, as it becomes impossible to pick and choose from a huge library, without understanding the implications of the choices made. As discussed in section 4, often a modest set of content descriptions, DS's and elements may suffice for a given application. This requires an application developer to first develop the descriptions in the context of the application domain, determine the DS's to support the descriptions, and then identify the required elements in the DS's. This is an involved process and cannot be viewed in isolation of the domain and application context. As MPEG-7 compliant applications start to be developed, it is possible, that there could be context-dependent elements and DS's that are essential to the application, but not described in the standard as the application context cannot be predetermined during the definition stage.

In conclusion, it is still early days for MPEG-7 and its deployment in managing the semantic aspects of multimedia applications. As the saying goes 'the proof of the pudding lies in the eating', and the success of the applications will determine the success of the standard.

6. References

1. Introduction to MPEG-7: Multimedia Content Description Interface, B. S. Manjunath (Editor), Philippe Salembier (Editor), Thomas Sikora (Editor), John Wiley and Sons.
2. MPEG-7 Overview (version 9), ISO/IEC JTC1/SC29/WG11N5525, Editor José M. Martínez , March 2003.
3. MPEG-7 Applications Document v.10, Editor Anthony Vetro, Editor, ISO/IEC JTC1/SC29/WG11/N3934, January 2001.