MPEG-7 Multimedia Description Schemes

Philippe Salembier, *Member, IEEE,* and John R. Smith, *Member, IEEE*

***Abstract*—MPEG-7 Multimedia Description Schemes (MDSs) are metadata structures for describing and annotating audio-visual (AV) content. The Description Schemes (DSs) provide a standardized way of describing in XML the important concepts related to AV content description and content management in order to facilitate searching, indexing, filtering, and access. The DSs are defined using the MPEG-7 Description Definition Language, which is based on the XML Schema Language, and are instantiated as documents or streams. The resulting descriptions can be expressed in a textual form (i.e., human readable XML for editing, searching, filtering) or compressed binary form (i.e., for storage or transmission). In this paper, we provide an overview of the MPEG-7 MDSs and describe their targeted functionality and use in multimedia applications.**

***Index Terms*—Browsing, content-based retrieval, digital libraries, indexing, metadata, MPEG-7, multimedia databases, multimedia description schemes (MDS), retrieval, search, XML.**

I. Introduction

THE goal of the MPEG-7 standard [1], [8] is to allow interoperable searching, indexing, filtering, and access of audio-visual (AV) content by enabling interoperability among devices and applications that deal with AV content description. MPEG-7 describes specific features of AV content, as well as information related to AV content management. MPEG-7 descriptions take two possible forms: 1) a textual XML form suitable for editing, searching, and filtering, and 2) a binary form suitable for storage, transmission, and streaming delivery. Overall, the standard specifies four types of normative elements: Descriptors, Description Schemes (DSs), a Description Definition Language (DDL), and coding schemes.

The MPEG-7 Descriptors are designed primarily to describe low-level audio or visual features such as color, texture, motion, audio energy, etc., as well as attributes of AV content such as location, time, quality, etc. It is expected that most Descriptors for low-level features shall be extracted automatically in applications. On the other hand, the MPEG-7 DSs are designed primarily to describe higher level AV features such as regions, segments, objects, events, and other immutable metadata related to creation and production, usage, and so forth. The DSs produce more complex descriptions by integrating together multiple Descriptors and DSs, and by declaring relationships among the description components. In MPEG-7, the DSs are categorized as pertaining to the multimedia, audio, or visual domain. Typically, the MDSs describe content consisting of a combination of audio, visual data, and possibly textual data, whereas the audio or visual DSs refer specifically to features unique to the audio or visual domain, respectively. In some cases, automatic tools can be used for instantiating the DSs, but in many cases instantiating DSs requires human assisted extraction or authoring tools.

The MPEG-7 DDL is a language for specifying the syntax of the DSs and Descriptors. The DDL also allows the MPEG-7 standardized DSs and Descriptors to be extended for specialized applications. The DDL is based on the W3C XML Schema Language [9]. An MPEG-7 description is produced for a particular piece of AV content by instantiating the MPEG-7 DSs or Descriptors as defined by the DDL. The MPEG-7 coding schemes produce a binary form of description that is compact, easy to stream, and resilient to errors during transmission.

The objective of this paper is to provide an overview of the MPEG-7 MDSs. Fig. 1 provides an overview of the organization of the MDSs into different functional areas: basic elements, content management, content description, navigation and access, content organization, and user Interaction. The MPEG-7 DSs can be considered as a library of description tools and in practice, an application selects an appropriate subset of relevant DSs. This paper discusses each of the different functional areas of MDSs. At present, more detailed information can be found in the *MPEG-7 Experimentation Model (XM)* [6] and *MPEG-7 Committee Draft (CD)* [7] documents developed by the MPEG MDS Group. In addition, many of the relevant concepts of the multimedia domain are defined in the MPEG-7 conceptual model, which is included as an annex of the *MPEG-7 Requirements* document [8]. The paper is organized as follows. Section II describes the Basic Elements of the MDSs. Section III describes the DSs for content management, and Section IV describes the DSs for content description. Section V describes the DSs for navigation and access and Section VI describes the DSs for content organization. Finally, Section VII describes the DSs for user interaction.

II. Basic Elements

MPEG-7 provides a number of Schema Tools that assist in the formation, packaging, and annotation of MPEG-7 descriptions. An MPEG-7 description begins with a root element that signifies whether the description is complete or partial. A complete description provides a complete and standalone description of AV content for an application. On the other hand, a description unit carries only partial or incremental information that possibly adds to an existing description. In the case of a complete description, an MPEG-7 top-level element follows the root element. The top-level element orients the description around a specific description task, such as the description of a particular type of AV content, for instance an image, video, audio, or multimedia, or a particular function related to content management, such as creation, usage, summarization, and so forth. The

Manuscript received September 2000; revised March 2001.

P. Salembier is with the Department of Signal Theory and Communication, Universitat Politecnica de Catalunya, 08034 Barcelona, Spain (e-mail: philippe@gps.tsc.upc.es).

J. R. Smith is with IBM T. J. Watson Research Center, Hawthorne, NY 10532 USA (e-mail: jsmith@us.ibm.com).

Publisher Item Identifier S 1051-8215(01)04993-X.

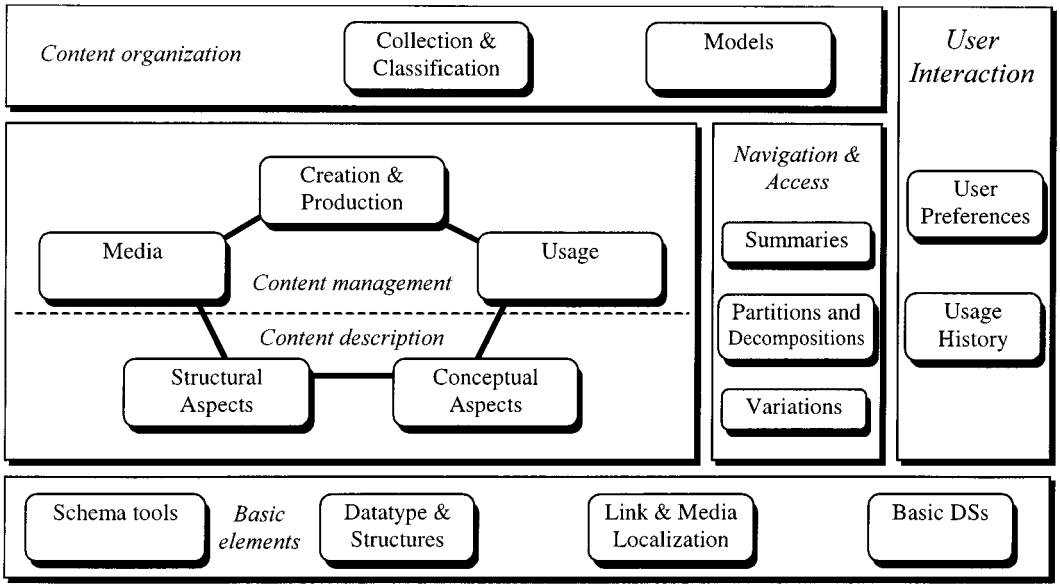

Fig. 1. Overview of the MPEG-7 MDSs.

top-level types collect together the appropriate tools for carrying out the specific description task. In the case of description units, the root element can be followed by an arbitrary instance of an MPEG-7 DS or Descriptor. Unlike a complete description which usually contains a "semantically complete" MPEG-7 description, a description unit can be used to send a partial description as required by an application—such as a description of a place, a shape and texture descriptor and so on. The *Package* DS describes a user-defined organization of MPEG-7 DSs and Ds into a package, which allows the organized selection of MPEG-7 tools to be communicated to a search engine or user. Furthermore, the *DescriptionMetadata* DS describes metadata about the description, such as creation time, extraction instrument, version, confidence, and so forth.

A number of basic elements are used throughout the MDS specification as fundamental constructs in defining the MPEG-7 DSs. The basic data types provide a set of extended data types and mathematical structures, such as vectors and matrices, which are needed by the DSs for describing AV content. The basic elements include also constructs for linking media files, localizing pieces of content, and describing time, places, persons, individuals, groups, organizations, and other textual annotation. We briefly discuss the MPEG-7 approaches for describing time and textual annotations.

Temporal Information: The DSs for describing time are based on the ISO 8601 standard, which has also been adopted by the XML Schema Language. The *Time* DS and *MediaTime* DS describe time information in the real world and in media streams, respectively. Both follow the same strategy described in Fig. 2. Fig. 2(a) illustrates the simplest way to describe a temporal instant and a temporal interval. A time instant t_1 can be described by a lexical representation using the Time Point. An interval $[t_1, t_2]$ can be described by its starting point t_1 (using the Time Point) and a Duration, $t_2 - t_1$. An alternative way to describe a time instant is shown in Fig. 2(b). It relies on Relative Time Point. The instant t_1 is described by a temporal offset with respect to a reference t_0, called Time Base. Note that the goal of the Relative Time Point is to define a temporal instant t_1, and not an interval as the Duration in Fig. 2(a). Finally, Fig. 2(c) illustrates the specification of time using a predefined interval called Time Unit and counting the number of intervals. This specification is particularly efficient for periodic or sampled temporal signals. Since the strategy consists of counting Time Units, the specification of a time instant has to be done relative to a Time Base (or temporal origin). In Fig. 2(c), t_1 is defined with a Relative Incremental Time Point by counting 8 Time Units (starting from t_0). An interval $[t_1, t_2]$ can also be defined by counting Time Units. In Fig. 2(c), Incremental Duration is used to count 13 Time Units to define the interval $[t_1, t_2]$.

Textual Annotation: Text annotation is an important component of many DSs. MPEG-7 provides a number of different basic constructs for textual annotation. The most flexible text annotation construct is the data type for free text. Free text allows the formation of an arbitrary string of text, which optionally includes information about the language of the text. However, MPEG-7 provides also a tool for more structured textual annotation by including specific fields corresponding to the questions "Who? What object? What action? Where? When? Why? and How?". Moreover, more complex textual annotations can also be defined by describing explicitly the syntactic dependency between the grammatical elements forming sentences (e.g., the relation between a verb and a subject, etc.). This latter type of textual annotation is particularly useful for applications where the annotation will be processed automatically. Lastly, MPEG-7 provides constructs for classification schemes and controlled terms. The classification schemes provide a language independent set of terms that form a vocabulary for a particular application or domain. Controlled terms are used in descriptions to make reference to the entries in the classification schemes. Allowing controlled terms to be described by classification schemes offers advantages over the

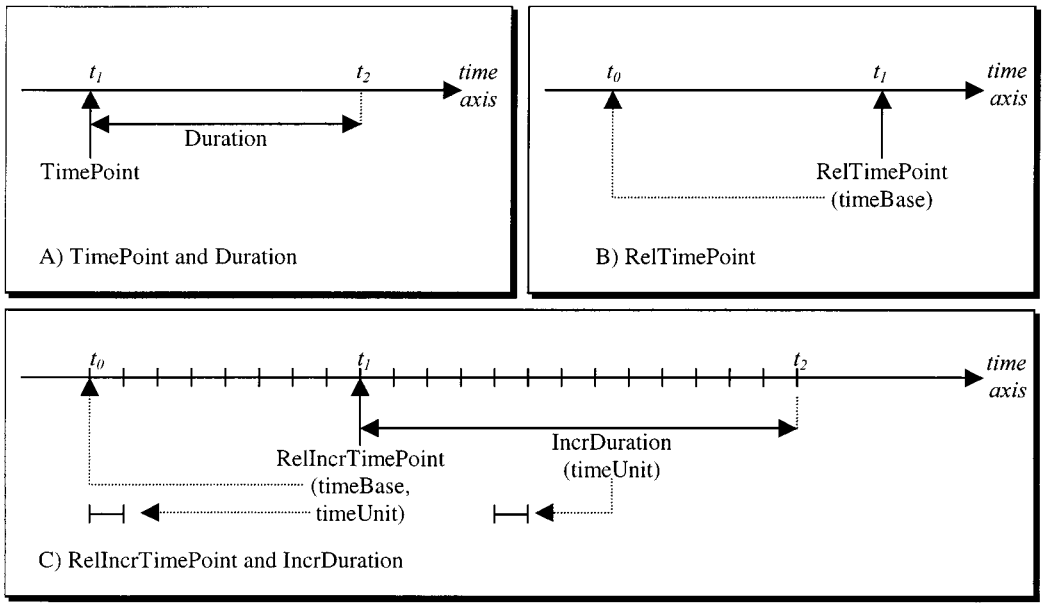

Fig. 2. Overview of the Time DSs.

standardization of fixed vocabularies for different applications and domains, since it is likely that the vocabularies for multimedia applications will evolve over time.

III. Content Management

MPEG-7 provides DSs for AV content management. These tools describe the following information: 1) creation and production; 2) media coding, storage, and file formats; and 3) content usage. More details about the MPEG-7 tools for content management are described as follows.[1]

- The *Creation Information* describes the creation and classification of the AV content and of other related materials. The creation information provides a title (which may itself be textual or another piece of AV content), textual annotation, and information such as creators, creation locations, and dates. The classification information describes how the AV material is classified into categories such as genre, subject, purpose, language, and so forth. It provides also review and guidance information such as age classification, parental guidance, and subjective review. Finally, the related material information describes whether there exists other AV materials that are related to the content being described.
- The *Media Information* describes the storage media such as the format, compression, and coding of the AV content. The media information DS identifies the master media, which is the original source from which different instances of the AV content are produced. The instances of the AV content are referred to as media profiles, which are versions of the master obtained perhaps by using different encodings, or storage and delivery formats. Each media profile is described individually in terms of the encoding parameters, storage media information, and location.
- The *Usage Information* describes the usage information related to the AV content, such as usage rights, usage record, and financial information. The rights information is not explicitly included in the MPEG-7 description; instead, links are provided to the rights holders and other information related to rights management and protection. The rights DS provides these references in the form of unique identifiers that are under management by external authorities. The underlying strategy is to enable MPEG-7 descriptions to provide access to current rights owner information without dealing with information and negotiation directly. The usage record DS and availability DSs provide information related to the use of the content such as broadcasting, on demand delivery, CD sales, and so forth. Finally, the financial DS provides information related to the cost of production and the income resulting from content use. The usage information is typically dynamic in that it is subject to change during the lifetime of the AV content.

[1]Many of the components of the content management DSs are optional. The instantiation of the optional components is often decided in view of the specific multimedia application.

IV. Content Description

MPEG-7 provides DSs for description of the structure and semantics of AV content. The structural tools describe the structure of the AV content in terms of video segments, frames, still and moving regions, and audio segments. The semantic tools describe the objects, events, and notions from the real world that are captured by the AV content.

A. Description of the Structural Aspects of Content

The *Segment* DS describes the result of a spatial, temporal, or spatio-temporal partitioning of the AV content. The *Segment* DS can describe a recursive or hierarchical decomposition of the AV content into segments that form a segment tree. The *SegmentRelation* DS describes additional spatio-temporal relationships among segments.

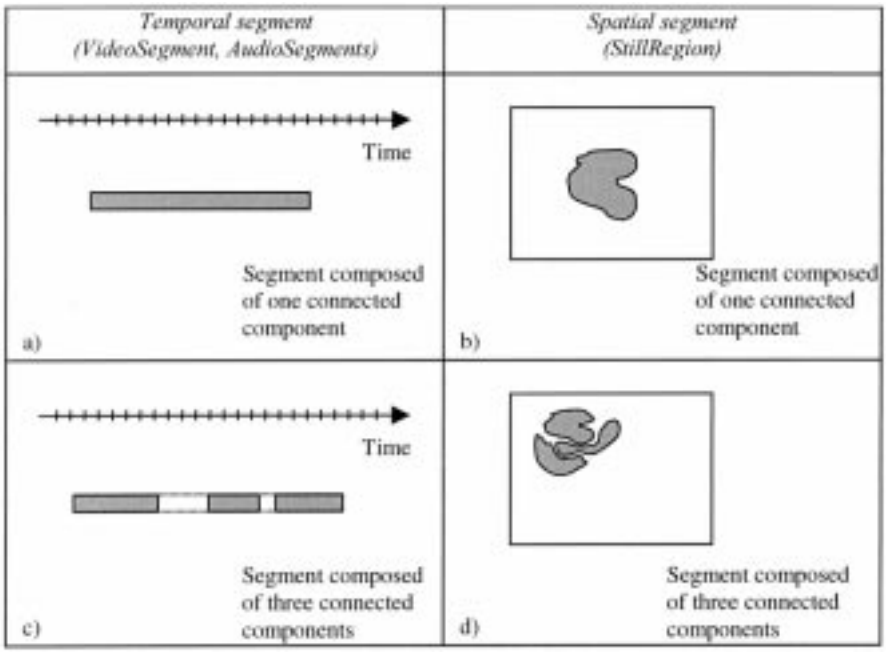

Fig. 3. Examples of segments. (a), (b) Segments composed of one connected component. (c), (d) Segments composed of three connected components.

The *Segment* DS forms the base abstract type of the different specialized segment types: audio segments, video segments, AV segments, moving regions, and still regions. As a result, a segment may have spatial and/or temporal properties. For example, the *AudioSegment* DS can describe a temporal audio segment corresponding to a temporal period of an audio sequence. The *VideoSegment* DS can describe a set of frames of a video sequence. The *AudioVisualSegment* DS can describe a combination of audio and visual information such as a video with synchronized audio. The *StillRegion* DS can describe a spatial segment or region of an image or a frame in a video. Finally, the *MovingRegion DS* can describe a spatio-temporal segment or moving region of a video sequence.

There exists also a set of specialized segments for specific type of AV content. For example, the *Mosaic DS* is a specialized type of *StillRegion*. It describes a mosaic or panoramic view of a video segment constructed by aligning together and warping the frames of a *VideoSegment* upon each other using a common spatial reference system. The *VideoText* and the *InkSegment* DSs are two subclasses of the *MovingRegion* DS. The VideoText DS describes a region of video content corresponding to text or captions. This includes superimposed text as well as text appearing in scene as well as. The InkSegment DS describes a segment of an electronic ink document created by a pen-based system or an electronic whiteboard.

Since the *Segment* DS is abstract, it cannot be instantiated on its own. However, the *Segment* DS contains elements and attributes that are common to the different segment types. Among the common properties of segments is information related to creation, usage, media location, and text annotation.

The *Segment* DS can be used to describe segments that are not necessarily connected, but composed of several nonconnected components. Connectivity refers here to both spatial and temporal domains. A temporal segment (*Video Segment, Audio Segment* and *AudioVisual Segment*) is said to be temporally connected if it is a sequence of continuous video frames or audio samples. A spatial segment (*Still Region*) is said spatially connected if it is a group of connected pixels. A spatio-temporal segment (*Moving Region*) is said spatially and temporally connected if the temporal segment where it is instantiated is temporally connected and if each one of its temporal instantiations in frames is spatially connected (Note that this is not the classical connectivity in a 3-D space).

Fig. 3 illustrates several examples of temporal or spatial segments and their connectivity. Fig. 3(a) and (b) illustrate a temporal and a spatial segment composed of a single connected component. Fig. 3(c) and (d) illustrate a temporal and a spatial segment composed of three connected components. Fig. 4 shows examples of connected and nonconnected moving regions. In this last case, the segment is not connected because it is not instantiated in all frames and, furthermore, it involves several spatial connected components in some of the frames.

Note that, in all cases, the Descriptors and DSs attached to the segment are global to the union of the connected components building the segment. At this level, it is not possible to describe individually the connected components of the segment. If connected components have to be described individually, then the segment has to be decomposed into various sub-segments corresponding to its individual connected components.

The *Segment* DS is recursive, i.e., it may be subdivided into sub-segments, and thus may form a hierarchy (tree). The resulting segment tree is used to describe the media source, the temporal and/or spatial structure of the AV content. For example, a video program may be temporally segmented into various levels of scenes, shots, and micro-segments; a table of contents may thus be generated based on this structure. Similar strategies can be used for spatial and spatio-temporal segments.

A segment may also be decomposed into various media sources such as various audio tracks or viewpoints from several cameras. The hierarchical decomposition is useful to design

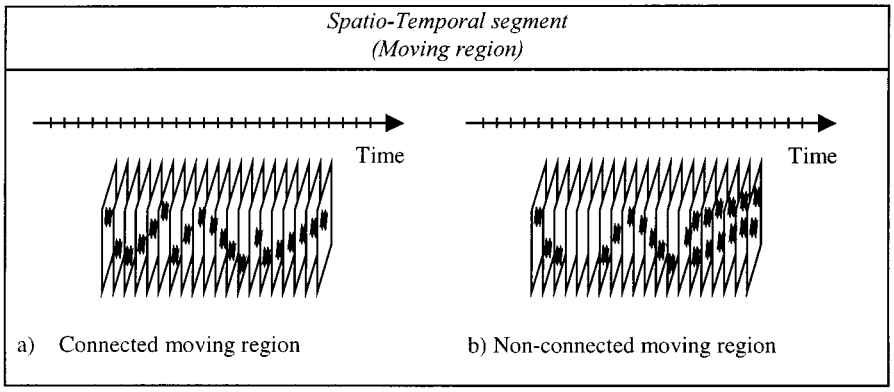

Fig. 4. Examples of (a) connected and (b) nonconnected moving region.

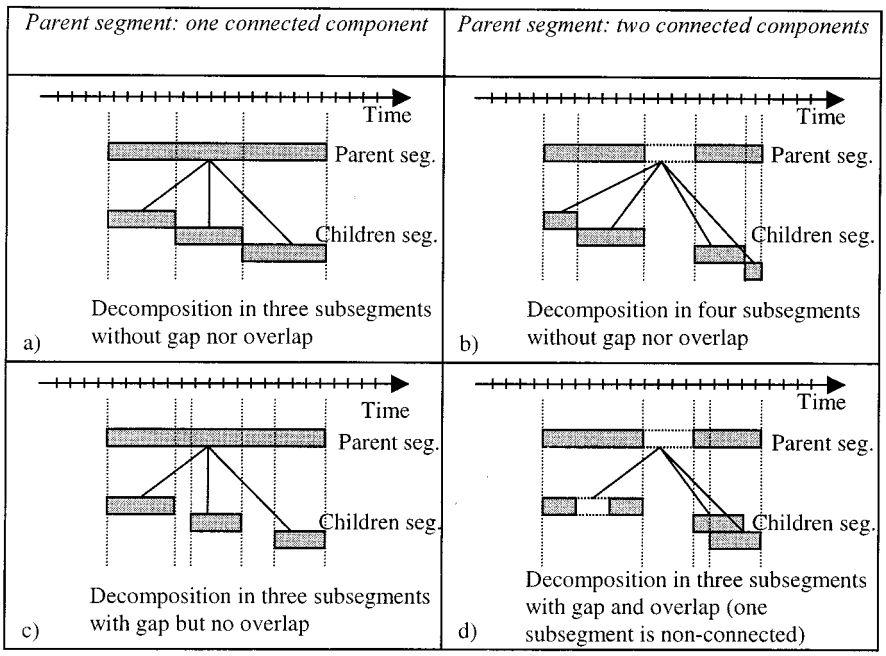

Fig. 5. Examples of segment decomposition. (a), (b) Segment Decompositions without gap nor overlap. (c), (d) Segment decompositions with gap or overlap.

efficient search strategies (global search to local search). It also allows the description to be scalable: a segment may be described by its direct set of Descriptors and DSs, but it may also be described by the union of the Descriptors and DSs that are related to its sub-segments. Note that a segment may be subdivided into sub-segments of different types, e.g., a video segment may be decomposed in moving regions that are themselves decomposed in still regions.

As it is done in a spatio-temporal space, the decomposition is described by a set of attributes defining the type of sub-division: temporal, spatial or spatio-temporal. Moreover, the spatial and temporal subdivisions may leave gaps and overlaps between the sub-segments. Several examples of decompositions are described for temporal segments in Fig. 5. Fig. 5(a) and (b) describe two examples of decompositions without gaps nor overlaps (partition in the mathematical sense). In both cases the union of the children corresponds exactly to the temporal extension of the parent, even if the parent is itself non connected [see the example of Fig. 5(b)]. Fig. 5(c) shows an example of decomposition with gaps but no overlaps. Finally, Fig. 5(d) illustrates a more complex case where the parent is composed of two connected components and its decomposition creates three children: the first one is itself composed of two connected components, whereas the two remaining children are composed of a single connected component. The decomposition allows gap and overlap. Note that, in any case, the decomposition implies that the union of the spatio-temporal space defined by the

TABLE I
SPECIFIC FEATURES FOR SEGMENT DESCRIPTION

Feature	*Video Segment*	*Still Region*	*Moving Region*	*Audio Segment*
Time	X	.	X	X
Shape	.	X	X	.
Color	X	X	X	.
Texture	.	X	.	.
Motion	X	.	X	.
Camera motion	X	.	.	.
Audio features	.	.	X	X

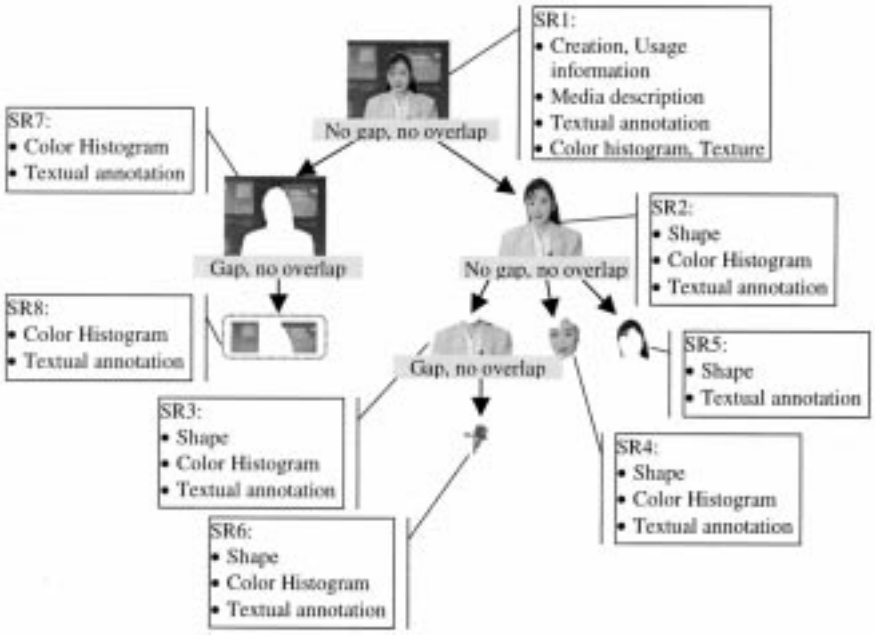

Fig. 6. Examples of image description with still regions.

children segments is included in the spatio-temporal space defined by their ancestor segment (children are contained in their ancestors).

As described above, any segment may be described by creation information, usage information, media information, and textual annotation. However, specific features depending on the segment type are also allowed. These specific features are reported in Table I. Most of the Descriptors corresponding to these features can be extracted automatically from the original content. For this purpose, a large number of tools have been reported in the literature (see this issue as well as [2]–[5], and reference herein). The instantiation of the decomposition involved in the *Segment* DS can be viewed as a hierarchical segmentation problem where elementary entities (region, video segment, and so forth) have to be defined and structured by inclusion relationship within a tree.

An example of image description is illustrated in Fig. 6. The original image is described as a Still Region *SR1*, which is described by creation (title, creator), usage information (copyright), and media information (file format), as well as a textual annotation (summarizing the image content), a color histogram, and a texture descriptor. This initial region can be further decomposed into individual regions. For each decomposition step, we indicate if Gaps and Overlaps are allowed. The segment tree is composed of eight still regions (note that SR8 is a single segment made of two connected components). For each region, Fig. 6 shows the type of feature that is instantiated. Note that it is not necessary to repeat in the tree hierarchy the creation, usage information, and media information, since the children segments are assumed to inherit their parent value (unless re-instantiated).

The description of the content structure is not constrained to rely on trees. Although, hierarchical structures such as trees are adequate for efficient access, retrieval and scalable description, they imply constraints that may make them inappropriate for certain applications. In such cases, the *SegmentRelation* DS has to be used. The graph structure is defined very simply by a set of nodes, each corresponding to a segment, and a set of edges, each corresponding to a relationship between two nodes. To illustrate the use of graphs, consider the example shown in Fig. 7.

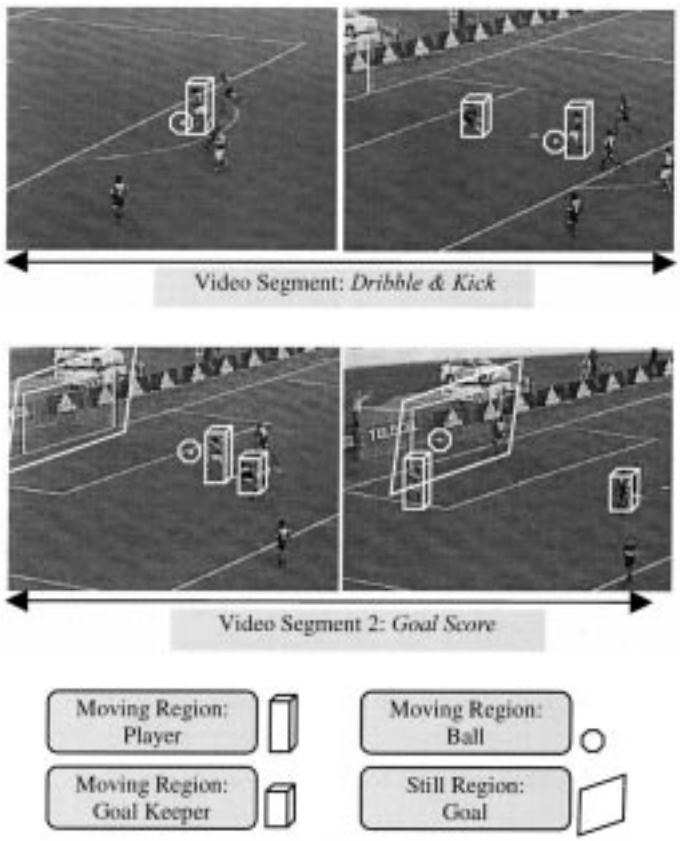

Fig. 7. Example of video segments and regions for the Segment Relationship Graph of Fig. 8.

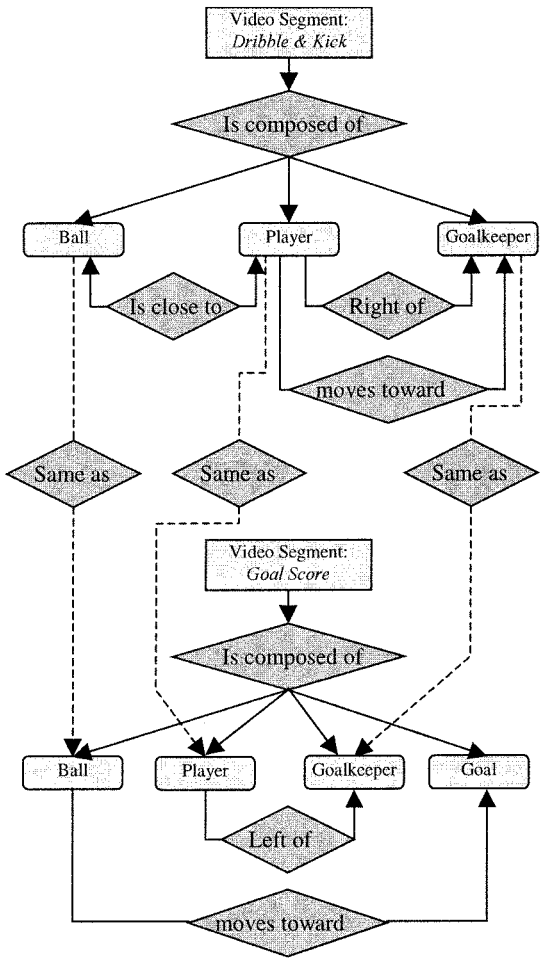

Fig. 8. Example of Segment Relationship Graph.

This example shows an excerpt from a soccer match. Two video segments, one Still Region and three Moving Regions, are considered. A possible graph describing the structure of the content is shown in Fig. 8. The video segment *Dribble & Kick* involves the *ball*, the *goalkeeper*, and the *player*. The *ball* remains *close to* the *player* who is *moving toward* the *goalkeeper*. The *player* appears on the *right of* the *goalkeeper*. The *goal score* video segment involves the same moving regions plus the still region called *goal*. In this part of the sequence, the *player* is on the *left of* the *goalkeeper* and the *ball* moves toward the *goal*. This very simple example illustrates the flexibility of this kind of representation. Note that this description is mainly structural because the relations specified in the graph edges are purely physical and the nodes represent segments (still and moving regions in this example). The only explicit semantic information is available from the textual annotation (where keywords such as ball, player, or goalkeeper can be specified).

B. Description of the Conceptual Aspects of Content

For some applications, the viewpoint described in the previous section is not appropriate because it highlights the structural aspects of the content. For applications where the structure is of no real use, but where the user is mainly interested in the semantic of the content, an alternative approach is provided by the *Semantic* DS. In this approach, the emphasis is not on segments but on *events, objects, places, time* in narrative worlds, and *abstraction*.

The narrative world refers to the context for a semantic description, that is, it is the "reality" in which the description makes sense. This notion covers the world depicted in the specific instances of AV content, as well as more abstract descriptions representing the possible worlds described in the possible media occurrences. A description may involve multiple narrative worlds depicted in multiple instances of AV content.

As shown in Fig. 9, the SemanticBase DS describes narrative worlds and semantic entities in a narrative world. In addition, a number of specialized DSs are derived from the generic SemanticBase DS, which describe specific types of semantic entities, such as narrative worlds, objects, agent objects, events, places, and time, as follows. The Semantic DS describes narrative worlds that are depicted by or related to the AV content. It may also be used to describe a template for AV content. In practice, the Semantic DS is intended to encapsulate the description of a narrative world. The Object DS describes a perceivable or abstract object. A perceivable object is an entity that exists, i.e., has temporal and spatial extent, in a narrative world (e.g., "Tom's piano"). An abstract object is the result of applying abstraction to a perceivable object (e.g., "any piano"). Essentially, this generates an object template. The AgentObject DS extends from the Object DS. It describes a person, an organization, a group of people, or personalized objects (e.g., "a talking cup in an animated movie"). The Event DS describes a perceivable or abstract event. A perceivable event is a dynamic relation involving one or more objects occurring in a region in time and space of a narrative world (e.g., "Tom playing the piano"). An abstract event is the result of applying abstraction to a perceivable event (e.g., "anyone playing the piano"). Here also, this generates a template of the event. Finally, SemanticPlace and SemanticTime DSs describe, respectively, a place and a time in a narrative world.

As in the case of the Segment DS, the conceptual aspect of description can be organized in a tree or in a graph. The graph structure is defined by a set of nodes, representing semantic notions, and a set of edges specifying the relationship between the nodes. Edges are described by the Semantic Relation DSs.

Beside the semantic description of individual instances in AV content, MPEG-7 Semantic DSs also allow the description of abstractions. Abstraction refers to the process of taking a description from a specific instance of AV content and generalizing it to a set of multiple instances of AV content or to a set of specific descriptions. Two types of abstraction, called *media abstraction* and *standard abstraction*, are considered.

A media abstraction is a description that has been separated from a specific instance of AV content, and can describe all instances of AV content that are sufficiently similar (similarity depends on the application and on the detail of the description). A typical example is that of a news event which can be applied to the description of multiple programs that may have been broadcasted on different channels.

A standard abstraction is the generalization of a media abstraction to describe a general class of semantic entities or de-

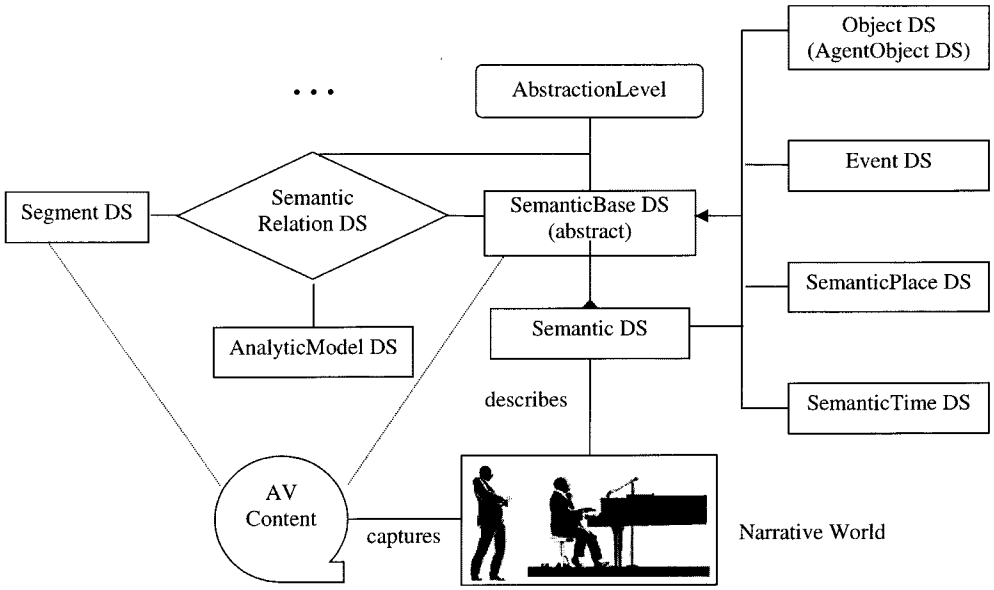

Fig. 9. Tools for the description of conceptual aspects.

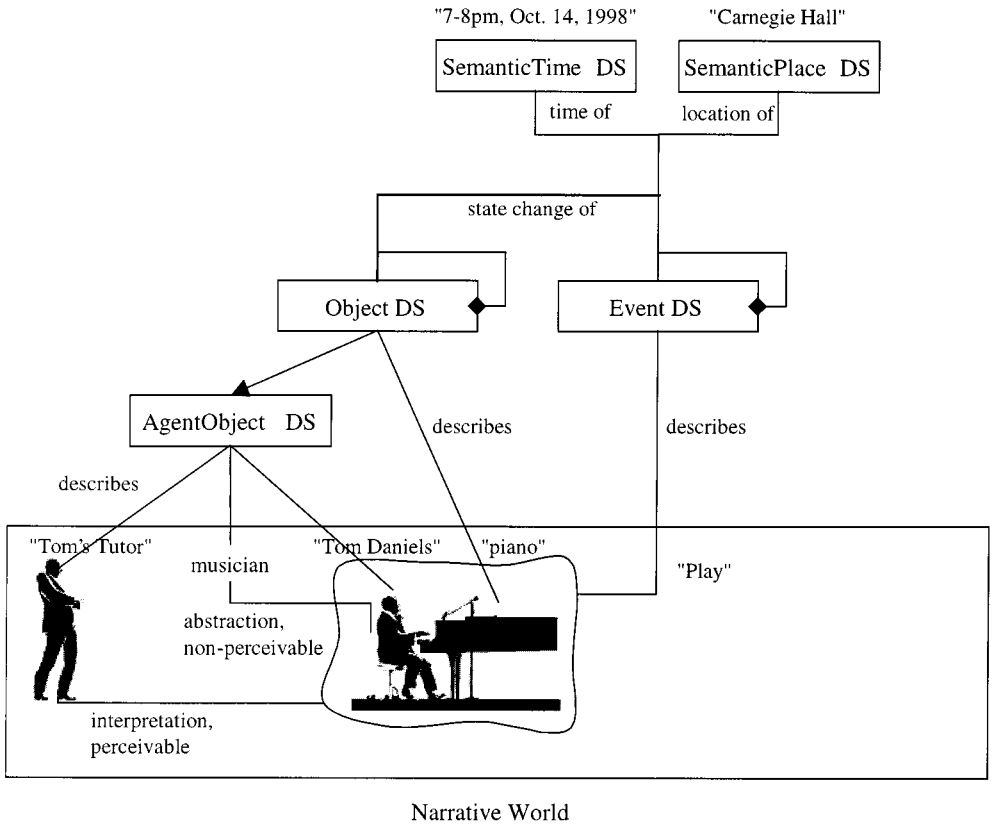

Fig. 10. Example of conceptual aspects description.

scriptions. In general, the standard abstraction is obtained by replacing the specific objects, events or other semantic entities by classes. For instance, if "Tom playing piano" is replaced by "a man playing piano," the description is now a standard abstraction. Standard abstractions can also be recursive, that is one can define abstraction of abstractions. Typically, a standard abstraction is intended for reuse, or to be used by reference in a description.

A simple example of conceptual aspects description is illustrated in Fig. 10. The narrative world involves Tom Daniels playing the Piano and his tutor. The event is characterized by a semantic time description: "7–8 PM on the 14th of October 1998," and a semantic place: "Carnegie Hall." The description involves one event, to play, and four objects: piano, Tom Daniels, his tutor, and the abstract notion of musicians. The last three objects belong to the class of Agent.

V. Navigation and Access

MPEG-7 facilitates navigation and access of AV content by describing summaries, views and partitions, and variations. The *Summary* DS describes semantically meaningful summaries and abstracts of AV content in order to enable efficient browsing and navigation. The *Space and Frequency View* DS describes structural views of the AV signals in the space or frequency domain in order to enable multiresolution access and progressive retrieval. The *Variation* DS describes relationships between different variations of AV programs in order to enable adaptive selection under different terminal, delivery, and user preference conditions. These tools are described in more detail as follows.

A. Summaries

The *Summarization* DS describes different compact summaries of the AV content that facilitate discovery, browsing, and navigation of the AV content. The summary descriptions allow the AV content to be navigated in either a hierarchical or sequential fashion. The hierarchical summary organizes the content into successive levels of detail. The sequential summary composes sequences of images, possibly synchronized with audio, to describe a slide-show or AV skim.

1) *Summarization DS*: the MPEG-7 summaries enable fast and effective browsing and navigation by abstracting out the salient information from the AV content. The *Summarization* DS contains links to the AV content, at the level of segments and frames. Given an MPEG-7 summarization description, a terminal device, such as a digital television set-top box, accesses the AV material composing the summary and renders the result for subsequent interaction with the user. The *Summarization* DS can describe multiple summaries of the same AV content, such as to provide different levels of detail or highlight specific features, objects, events, or semantics. By including links to the AV content in the summaries, it is possible to generate and store multiple summaries without storing multiple versions of the summary AV content
2) *HierarchicalSummary DS*: The *HierarchicalSummary* DS describes the organization of summaries into multiple levels in order to describe different levels of temporal detail. The *HierarchicalSummary* DS is constructed around the generic notion of temporal segments of AV content, described by *HighlightSegment* DSs. Each *HighlightSegment* contains locators to the AV content that provide access to the associated key-videoclips, key-audioclips, key-frames and key-sounds. Each may also contain textual annotations that describe the key-themes. These AV segments are grouped into summaries, or highlights, using the *HighlightSummary* DS. For example, in Fig. 11, the *HierarchicalSummary* contains two summaries, where the first summary consists of four highlight segments and the second summary consists of three highlight segments. The summaries could correspond to two different themes and could provide alternative views on the original AV content. The *HighlightSummary* DS is recursive in nature, enabling summaries to contain other summaries. This capability can be used to build a variety of hierarchical summaries, i.e., to describe content at different granularities. Additionally, multiple summaries may be grouped together using the *HierarchicalSummary* DS.

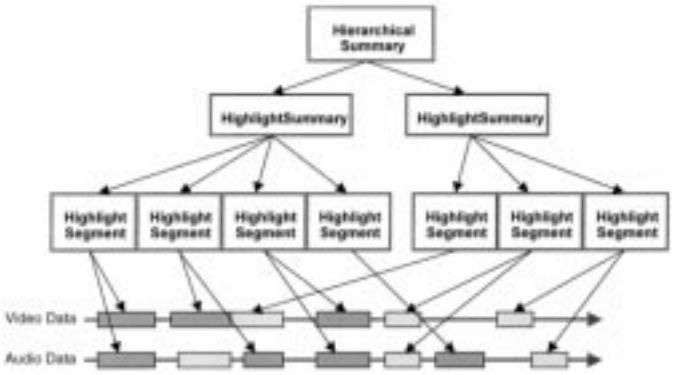

Fig. 11. Illustration of HierarchicalSummary DS containing two summaries.

Fig. 12 shows an example of a hierarchical summary of a soccer video. The *HierarchicalSummary* description gives three levels of detail. In this example, the video of the soccer game is summarized into a single frame at the root. The next level of the hierarchy provides three frames that summarize different segments of the video. Finally, the bottom level provides additional frames, depicting in more detail the scenes depicted in the segments.

3) *SequentialSummary* DS: the *SequentialSummary* DS describes a summary consisting of a sequence of images or video frames, which is possibly synchronized with audio. The *SequentialSummary* may also contain a sequence of audio clips. The AV content that makes up the *SequentialSummary* may be stored separately from the original AV content to allow fast navigation and access. Alternatively, the *SequentialSummary* may link directly to the original AV content in order to reduce storage.

B. Partitions and Decompositions

The *View DS* describes a structural view, partition, or decomposition of an audio or visual signal in space, time, and frequency. In general, the views of the signals correspond to low-resolution views, or spatial or temporal segments, or frequency subbands. The *Space and Frequency View* DS describes a view in terms of its corresponding partition in the space or frequency plane. In addition, the *Decomposition* DS describes a tree- or graph-based decomposition of an audio or visual signal or organization of views. In the tree- or graph-based decompositions, a node corresponds to a view, and a transition corresponds to an analysis and synthesis signal processing dependency among the connected views.

1) *View DS*: the *View* DS describes a space or frequency view of an audio or visual signal. The *SpaceView* DS describes a spatial view of an audio or visual signal, for example, a spatial segment of an image. The *FrequencyView* DS describes a view of an audio or visual signal within a particular frequency band, for example, a wavelet subband of an audio signal. The *SpaceFrequencyView* DS describes

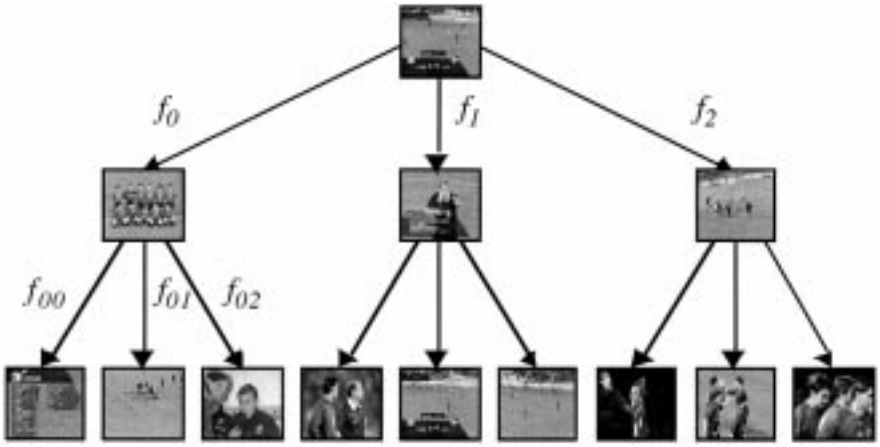

Fig. 12. Example of a hierarchical summary of a video of a soccer game providing a multiple level key-frame hierarchy. The hierarchical summary denotes the fidelity (i.e., f_0, f_1) of each key-frame with respect to the video segment referred to by the key-frames at the next lower level.

a multi-dimensional view of an audio or visual signal simultaneously in space and frequency, for example, a wavelet subband of a spatial segment of an image. The *ResolutionView* DS describes a low-resolution view of an audio or visual signal, such as a thumbnail view of an image. Conceptually, a resolution view is a special case of a frequency view that corresponds to a low-frequency subband of the signal. A *SpaceResolutionView* DS describes a view simultaneously in space and resolution of an audio or visual signal, for example, a low-resolution view of a spatial segment of an image.

2) *View Decompositions*: the *ViewDecomposition* DS describes a space and frequency decomposition or organization of views of an audio or visual signal. The *ViewSet* DS describes a set of views, which can have different properties of completeness and redundancy. For example, the set of wavelet subbands of an audio signal forms a complete and nonredundant set of views. The *SpaceTree* DS describes a spatial-tree decomposition of an audio or visual signal, for example, a spatial quad-tree image decomposition. The *FrequencyTree* DS describes a frequency-tree decomposition of an audio or visual signal, e.g., a wavelet packet-tree image decomposition. The *SpaceFrequencyGraph* DS describes a decomposition of an audio or visual signal simultaneously in space and frequency in which the views are organized using a space and frequency graph. The *VideoViewGraph* DS describes a specific type of decomposition of a video signal in both spatial- and temporal-frequency that corresponds to a 3-D subband decomposition. Finally, a *MultiResolutionPyramid* DS describes a hierarchy of multi-resolution views of an audio or visual signal, such as an image pyramid.

Fig. 13 shows an example Space and Frequency Graph decomposition of an image. The Space and Frequency Graph structure contains nodes that correspond to the different space and frequency views of the image. The views correspond to partitions of the 2-D image signal in space (spatial segments), frequency (wavelet subbands), and space and frequency (wavelet subbands of spatial segments). The space and frequency graph contains also transitions that correspond to the analysis and synthesis dependencies among the views. For example, in Fig. 13, each "S" transition indicates spatial decomposition while each "F" transition indicates frequency or subband decomposition.

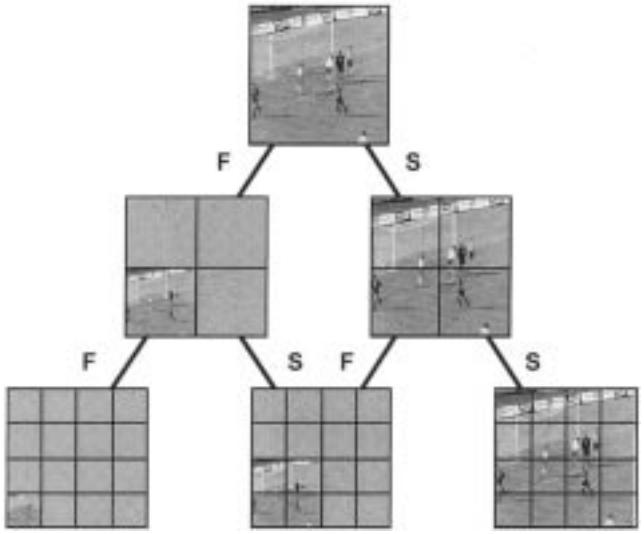

Fig. 13. The Space and Frequency Graph describes the decomposition of an audio or visual signal in space (time) and frequency.

C. Variations of the Content

The *Variation* DS describes variations of the AV content, such as compressed or low-resolution versions, summaries, different languages, and different modalities, such as audio, video, image, text, and so forth. One of the targeted functionalities of the *Variation* DS is to allow a server or proxy to select the most suitable variation of the AV content for delivery according to the capabilities of terminal devices, network conditions, or user preferences. The *Variations* DS describes the different alternative variations. The variations may refer to newly authored AV content, or correspond to AV content derived from another source. A variation fidelity value gives the quality of the variation compared to the original. The variation type attribute indicates the type of variation, such as summary, abstract, extract, modality

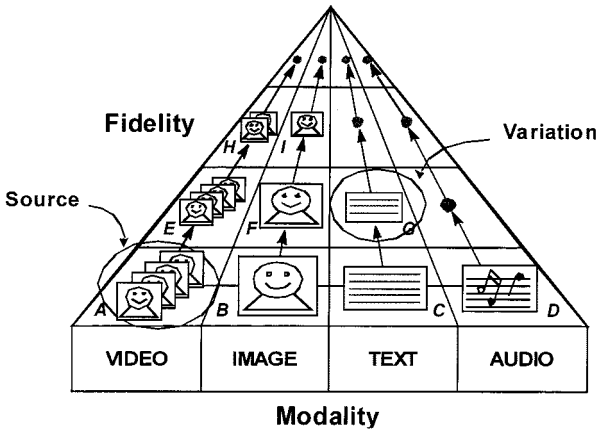

Fig. 14. Illustration of variations of a source AV program.

translation, language translation, color reduction, spatial reduction, rate reduction, compression, and so forth.

Fig. 14 illustrates a set of variations of an AV program. The example shows the source video program in the lower left corner (A) and eight variation programs. The variations have different modalities: two variations are video programs (E, H), three are images (B, F, I), two are text (C, G), and one is audio (D). Each of the variation programs has a specified fidelity value that indicates the fidelity of the variation program with respect to the source program.

VI. Content Organization

The *Collection* DS describes collections of AV content, descriptor instances, concepts, or mixed content. The collections can be used for tasks such as describing an album of songs, a group of objects, or a cluster of color feature descriptors. The *Model* DS describes parameterized models of AV content or a collection. The models can be expressed in terms of statistics or probabilities associated with the attributes of collections of AV content, or can be expressed through examples or exemplars of the AV content classes.

A. Collections

The *Collection* DS describes a collection related to AV content. The *Collection* DS includes tools for describing collections of AV material, collections of AV content descriptions, collections of semantic concepts, mixed collections of content, descriptions, and concepts, and collection structures in terms of the relationships among collections. Fig. 15 shows one organization of collections within a collection structure. In this example, each collection consists of a set of images with common properties, for example, each depicting similar events in a soccer game. Within each collection, the relationships among the images can be described, such as the degree of similarity of the images in the cluster. Across the collections, additional relationships can be described, such as the degree of similarity of the collections.

B. Models

The *Model* DS describes parameterized models of AV content, descriptors, or collections. The *ProbabilityModel* DS describes different statistical functions and probabilistic structures, which can be used to describe samples of AV content and classes of Descriptors using statistical approximation. The *AnalyticModel* DS describes a collection of examples of AV content or clusters of Descriptors that are used to provide a model for a particular semantic class. For example, a collection of art images labeled with tag indicating that the paintings are examples of the Impressionist period forms an analytic model. The *AnalyticModel* DS also optionally describes the confidence in which the semantic labels are assigned. The *Classifier* DS describes different types of classifiers that are used to assign the semantic labels to AV content or collections.

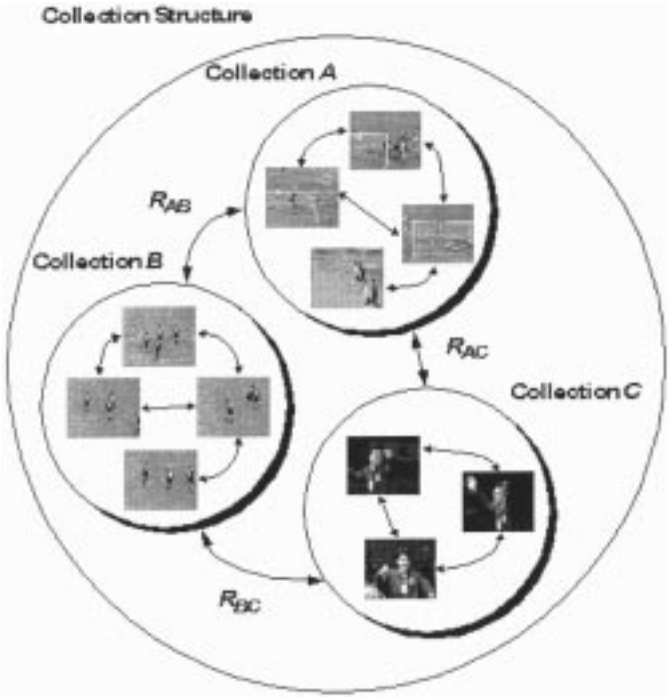

Fig. 15. The Collection Structure DS describes collections of AV content and the relationships (i.e., R_{AB}, R_{BC}, R_{AC}) within and across the collections.

VII. User Interaction

The *UserInteraction* DS describes preferences of users pertaining to the consumption of the AV content, as well as usage history. The MPEG-7 AV content descriptions can be matched to the preference descriptions in order to select and personalize AV content for more efficient and effective access, presentation and consumption. The *UserPreference* DS describes preferences for different types of content and modes of browsing, including context dependency in terms of time and place. The *UserPreference* DS describes also the weighting of the relative importance of different preferences, the privacy characteristics of the preferences and whether preferences are subject to update, such as by an agent that automatically learns through interaction with the user. The *UsageHistory* DS describes the history of actions carried out by a user of a multimedia system. The usage history descriptions can be exchanged between consumers, their agents, content providers, and devices, and may in turn be used to determine the user's preferences with regard to AV content.

VIII. Conclusion

In this paper, we presented an overview of the MDSs being developed by MPEG as part of the MPEG-7 standard. The MDSs are metadata structures for describing and annotating AV content and provide a way to describe in XML the important concepts related to AV content. The objective is to allow interoperable searching, indexing, filtering and access by enabling interoperability among devices that deal with AV content description.

We presented the organization of the MPEG-7 MDSs into different areas of basic elements, schema tools, content description, content management, content organization, navigation and access, and user interaction. In the process of developing the MPEG-7 MDSs, efforts have been made to harmonize with other multimedia metadata standards and existing practices. Beyond the standardized set of MPEG-7 MDSs, the MPEG-7 DDL also allows the extension of the MPEG-7 standard to accommodate specific AV content domains and applications.

Acknowledgment

The MPEG MDSs presented in this paper result from the contributions and collaborative efforts of many people. The authors are particularly grateful to the other editors of the *MPEG-7 MDS Experimentation Model (XM)* and *Committee Draft (CD) documents* [6], [7], including P. van Beek, A. Benitez, J. Heuer, J. Martinez, Y. Shibata and T. Walker. The authors also thank all members of the MPEG MDS Group for their many contributions toward the development of the MPEG-7 standard.[2]

References

[1] *IEEE Trans. Circuits Syst. Video Technol*, vol. 11, June 2001.
[2] *IEEE Trans. Image Processing*, vol. 9, Jan. 2000.
[3] *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, Sept. 1998.
[4] *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, Dec. 1999.
[5] *Signal Processing: Image Commun.*, vol. 16, Sept. 2000.
[6] *MPEG-7 Multimedia Description Schemes XM (v6.0)*, ISO/IEC JTC1/SC29/WG11/N3815, Jan. 2001.
[7] *MPEG ISO/IEC CD 15938-5 Information Technology—Multimedia Content Description Interface: Multimedia Description Schemes*, ISO/IEC JTC1/SC29/WG11/N3705, Oct. 2000.
[8] *MPEG. MPEG-7 Requirements Document V.13*, ISO/IEC JTC1/SC29/WG11/N3933, Jan. 2001.

[2]The public MPEG documents referenced in this paper can be downloaded from: http://www.cselt.it/mpeg/working_documents.htm

[9] XML Schema Part 0: Primer, Part 1: Structures, Part 2: Datatypes, W3C Candidate Recommendation (2000, Oct.). [Online]. Available: http://www.w3.org/TR/

Philippe Salembier (M'96) received the M.S. degrees from the Ecole Polytechnique, Paris, France, in 1983, and from the Ecole Nationale Superieure des Telecommunications, Paris, France, in 1985. He received the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1991.

From 1985 to 1989, he was with Laboratoires d'Electronique Philips, Limeil-Brevannes, France, working in the fields of digital communications and signal processing for HDTV. In 1989, he joined the Swiss Federal Institute of Technology, Lausanne, Switzerland, to work on image processing. He was a Postdoctoral Fellow at the Harvard Robotics Laboratory, Cambridge, MA, during 1991, then joined the Polytechnic University of Catalonia, Barcelona, Spain, where he is currently an Associate Professor, lecturing in the area of digital signal and image processing. His current research interests include image and sequence coding, compression and indexing, image modeling, segmentation problems, video sequence analysis, mathematical morphology and nonlinear filtering. In terms of standardization activities, he has been involved in the definition of the MPEG-7 standard ("Multimedia Content Description Interface") as chair of the "Multimedia Description Scheme" group until March 2001.

Dr. Salembier served as an Area Editor of the *Journal of Visual Communication and Image Representation* from 1995 to 1998 and as an AdCom Officer of the European Association for Signal Processing (EURASIP), in charge of the edition of the Newsletter, from 1994 to 1999. He is currently Deputy Editor of *Signal Processing*, and was previously a Guest Editor for the *Signal Processing* Special Issue on Mathematical Morphology (1994) and Special Issue on Video Sequence Analysis (1998), and was Co-Editor of a Special Issue on MPEG-7 proposals (2000). Finally, he is member of the Image and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society.

John R. Smith (S'93–M'97) received the M. Phil and Ph.D. degrees in electrical engineering from Columbia University, New York, in 1994 and 1997, respectively.

He is currently Manager of the Pervasive Media Management Group, IBM T. J. Watson Research Center, Hawthorne, NY, where he has lead the development of the SFGraph and VideoZoom systems for progressive retrieval of high-resolution images and video, respectively. He is also an Adjunct Professor at Columbia University. In MPEG, he has lead the effort to create the MPEG-7 Conceptual Model, and since March 2001, has served as Chair of the MPEG Multimedia Description Schemes (MDS) sub-group. Previously, at Columbia, he developed the WebSEEk image and video search engine and the VisualSEEk content-based retrieval system, His research interests include multimedia content analysis and retrieval.

Dr. Smith received the Eliahu I. Jury Award from Columbia University for outstanding achievement as a graduate student in the areas of systems communication or signal processing.