IN-DEPTH ARTICLES: INTELLECTUAL FOUNDATIONS AND DESCRIPTIONS OF MPEG-7 TOOLS FOR MULTIMEDIA DESCRIPTION

Structure Description Tools

Philippe Salembier
Universitat Politècnica de Catalunya, Campus Nord, D5, Jordi Girona, 1–3, 08034 Barcelona, Spain.
E-mail: philippe@gps.tsc.upc.es

Ana B. Benitez
Thomson Corporate Research, 2233 North Ontario Street, Suite 100, Burbank, CA 91504.
E-mail: ana.benitez@thomson.net

This article provides an overview of the tools specified by the MPEG-7 standard for describing the structure of multimedia content. In particular, it focuses on tools that represent segments resulting from a spatial and/or temporal partitioning of multimedia content. The segments are described in terms of their decomposition and the general relations among them as well as attributes or features of segments. Decomposition efficiently represents segment hierarchies and can be used to create tables of contents or indexes. More general graph representations are handled by the various standard spatial and temporal relations. A segment can be described by a large number of features ranging from those targeting the life cycle of the content (e.g., creation and usage) to those addressing signal characteristics such as audio, color, shape, or motion properties.

Introduction

The goal of the MPEG-7 standard (ISO/IEC, 2002–2004; Manjunath et al., 2002) is to allow interoperable indexing, filtering, searching, and access of multimedia content by enabling interoperability among devices and applications that deal with multimedia content description. MPEG-7 specifies tools for the description of features related to the content as well as information related to its management. The scope of the standard is to define the representation of the description, that is, the syntax and the semantics of the structures used to create MPEG-7 descriptions. For most description tools, the standard does not provide normative tools for the generation or the consumption of the description. Their inclusion is not necessary to guarantee interoperability. Moreover, their omission allows future improvements to be included in MPEG-7-compliant applications.

Overall, the standard specifies four types of normative elements: descriptors (Ds), description schemes (DSs), a Description Definition Language (DDL), and coding schemes. In MPEG-7, a descriptor defines the syntax and the semantics of an elementary feature. A descriptor can deal with low-level features, which represent signal characteristics, such as color, texture, shape, motion, audio energy, or audio spectrum, as well as higher-level features, such as the title or the author. In general, the description of a piece of content involves a large number of descriptors. The descriptors are structured and related within a common framework based on description schemes (DSs). The DSs define a model of the description using the descriptors as building blocks. In MPEG-7, the syntax of descriptors and description schemes is defined by the Description Definition Language (DDL), which is an extension of the Extensible Markup Language (XML) Schema language (W3C, 2001). The DDL is used not only to define the syntax of MPEG-7 description tools but also to allow the declaration of new description tools that are related to specific applications.

Structure Description Tools

The objective of this article is to provide an overview of the MPEG-7 DSs and Ds that target the description of the structural aspects of the content. More information about describing other aspects of multimedia content can be found in other articles of this *Perspectives* or in (ISO/IEC, 2002–2004; Manjunath, Salembier, & Sikora, 2002). The description of the

© 2007 Wiley Periodicals, Inc. • Published online 21 May 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20572

structure of multimedia content relies on the notion of *segments*. The structure description tools represent the structure of multimedia data in space and/or time by describing general and application-specific segments of multimedia content together with their attributes, hierarchical decompositions, and relations. Segments are the result of a spatial, temporal, or spatiotemporal partitioning of the multimedia content. Decomposition efficiently represents segment hierarchies and can be used to create tables of contents or indexes. More general graph representations are handled by the various standard spatial and temporal relations. A segment can be described by a large number of features ranging from those targeting the life cycle or management of the content (e.g., creation and usage) to those addressing the signal characteristics such as audio, color, shape, or motion properties.

Figure 1 shows an example of how the structure description tools (see Description of the Content Structure) together with content management description tools (see Content Management) and feature description tools (see Audio and Visual Features) can be used to describe an image. In this example, the entire image is described as a still region (SR1), which is a group of pixels in a two-dimensional (2D) image or a video frame. This figure also exemplifies the spatial decomposition of a still region (SR1) into two still regions (SR2 and SR3), and the description of the spatial relation left between two still regions (SR2 and SR3). Various properties dealing with creation information, textual annotation, or color features are described for each still region in addition to creation, media, and usage information for the entire image.

Description of the Content Structure

As noted, the description of the structure of multimedia content relies on the notion of segments. The Segment DS describes the result of a spatial, temporal, or spatiotemporal partitioning of the multimedia content. In this section, we describe in more detail the MPEG-7 Segment DSs, their attributes, as well as their related decomposition and relation tools.

Segment Entities

The MPEG-7 *Segment* DS describes segments of multimedia content in space, time, and/or media source. The *Segment* DS is an abstract type (which cannot be instantiated on its own) that defines the set of attributes and properties that are common to all types of segments. The description of specific segment types is handled by DSs that are derived from the *Segment* DS. There are 23 segment description tools in all, including the abstract *Segment* DS.

The most basic *Segment* DS dealing with pure visual information is the *Still Region* DS (a still region is a group of pixels in the digital case); it describes a spatial region of a 2D image or a video frame. Other purely visual segments are the *Video Segment* DS and the *Moving Region* DS, which describe, respectively, a temporal interval (a group of frames in a video) and a spatiotemporal region (a group of pixels in a group of video frames) of a video sequence. The *Audio Segment* DS represents a temporal interval of an audio sequence (a group of samples in the digital case).

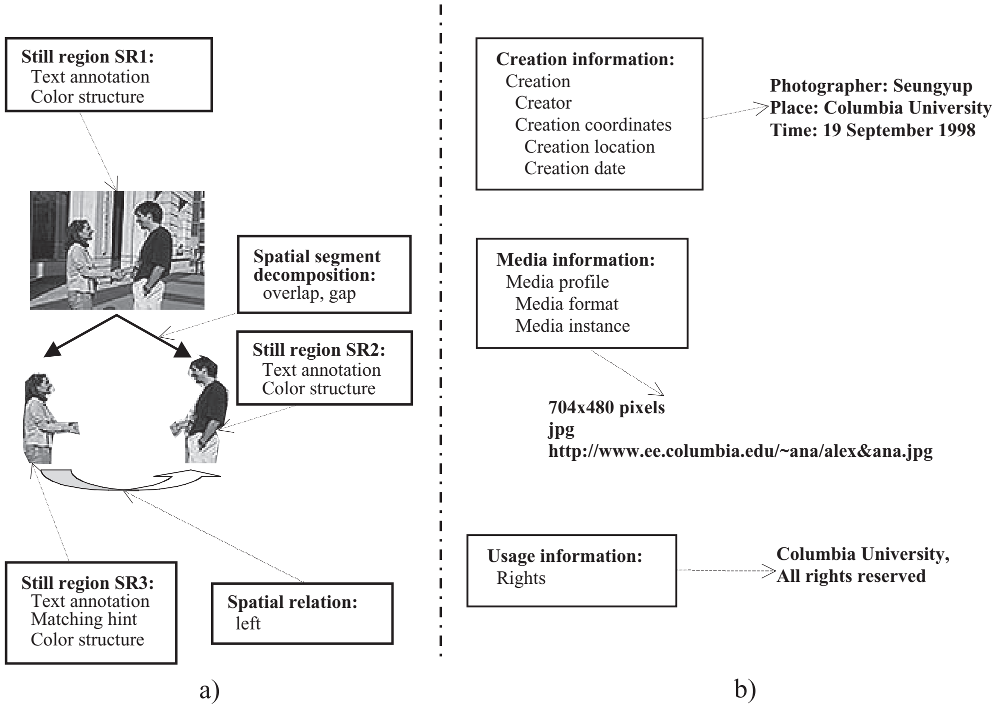

FIG. 1. Illustration of the description of an image in MPEG-7 using (a) structure description tools (Description of Content Structure; e.g., still region and spatial relation) and feature description tools (Audio and Visual Features; e.g., color structure); and (b) content management description tools (Content Management; creation, media, and usage information).

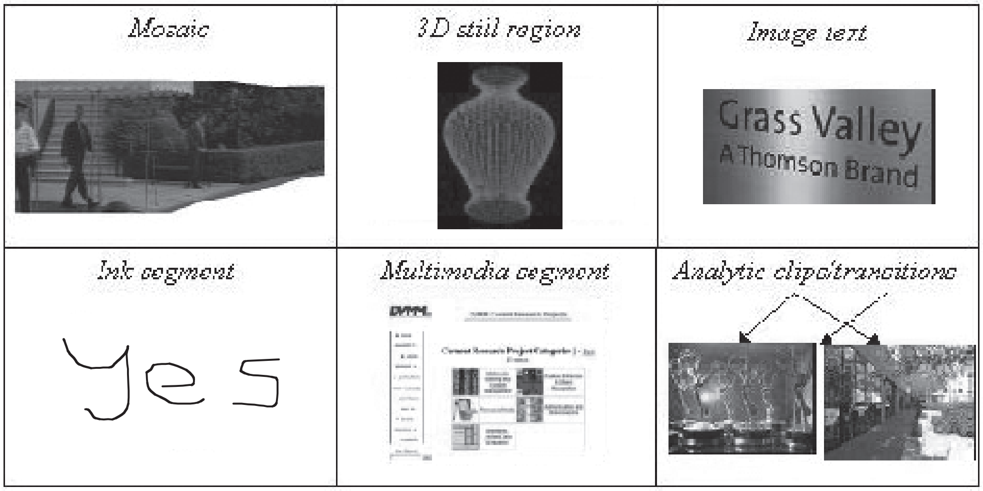

FIG. 2. Examples of a mosaic, a 3D still region, an image text, an ink segment, a multimedia segment, two analytic clips, and an analytic transition.

Both the *Audio-Visual Segment* DS and the *Audio-Visual Region* DS describe audio and visual data in an audiovisual sequence. In particular, the *Audio-Visual Segment* DS describes a temporal interval of audiovisual data, which corresponds to both the audio and the video in the same temporal interval, whereas the *Audio-Visual Region* DS describes an arbitrary spatiotemporal segment of audiovisual data, which corresponds to the audio in an arbitrary temporal interval and the video in an arbitrary spatiotemporal interval.

There is also a set of specific *Segment DSs* that describe specific content types such as mosaic, 3D, image/video text, ink content, multimedia content, and video editing (see Figure 2). For example, the *Mosaic* DS extends from the *Still Region* DS and describes a mosaic or panoramic view of a video segment. A mosaic is usually constructed by aligning and blending together the frames of a video segment on each other using a common spatial reference system. The *Still Region 3D* DS represents a 3D spatial region of a 3D image. The *Image Text* DS and the *Video Text* DS extend, respectively, the *Still Region* DS and the *Moving Region* DS. They describe a still region and a moving region that correspond to text, respectively. The *Ink Segment* DS represents a spatiotemporal segment of ink content created by a pen-based system or an electronic whiteboard. The *Multimedia Segment* DS represents a composite of segments forming a multimedia presentation such as an MPEG-4 presentation or a Web page. Finally, the edited video segment DSs such as the *Analytic Clip* DS and the *Analytic Transition* DS extend originally from the *Video Segment* DS and describe, respectively, different types of shots and transitions between shots resulting from video editing work. The video editing description is "analytic" in the sense that it is made *a posteriori* on the final edited video content.

The *Segment* DS can describe segments that are not connected but composed of several separated connected components. *Connectivity* refers here to both spatial and temporal dimensions. A temporal segment (instance of the *Video Segment, Audio Segment, Audio-Visual Segment, Ink Segment DSs*, and the audio part of the *Audio-Visual Region* DS) is said to be temporally connected if it is a sequence of continuous video frames and/or audio samples, in the digital case. A spatial segment (instance of the *Still Region* DS) is said to be spatially connected if it is a group of connected pixels, in the digital case. A spatiotemporal segment (instance of the *Moving Region* DS and the video part of the *Audio-Visual Region* DS) is said to be spatially and temporally connected if the temporal segment where it is instantiated is temporally connected and if each one of its spatial instantiations in frames is spatially connected. Note that this definition of connectivity in a 3D space is not the classic one. Figure 3 illustrates several examples of temporal, spatial, and spatiotemporal segments that are either connected or composed of several connected components.

The *Segment* DS is abstract and, therefore, cannot be instantiated on its own. However, the *Segment* DS contains elements and attributes that are common to all segment types. Among the most important common properties of segments, there is information related to the creation, the usage, and the media. The corresponding Ds and DSs are discussed in Segment Attributes and Content Management. Note that, in all cases, the Ds and DSs attached to the segment are global to the full extent of the segment, that is, the union of the connected components composing the segment being described. At this level, it is not possible to describe the individual connected components of the segment. If connected components have to be described individually, then the segment has to be decomposed into various subsegments corresponding to its individual connected components using the segment decomposition tools (see section Segment Decompositions).

Segment Attributes

As mentioned, any kind of segment can be described in terms of its media, creation, and usage information. It can be further described by textual annotations, visual features, audio features, and other segment attributes. Specific features such as the characterization of the connected components of

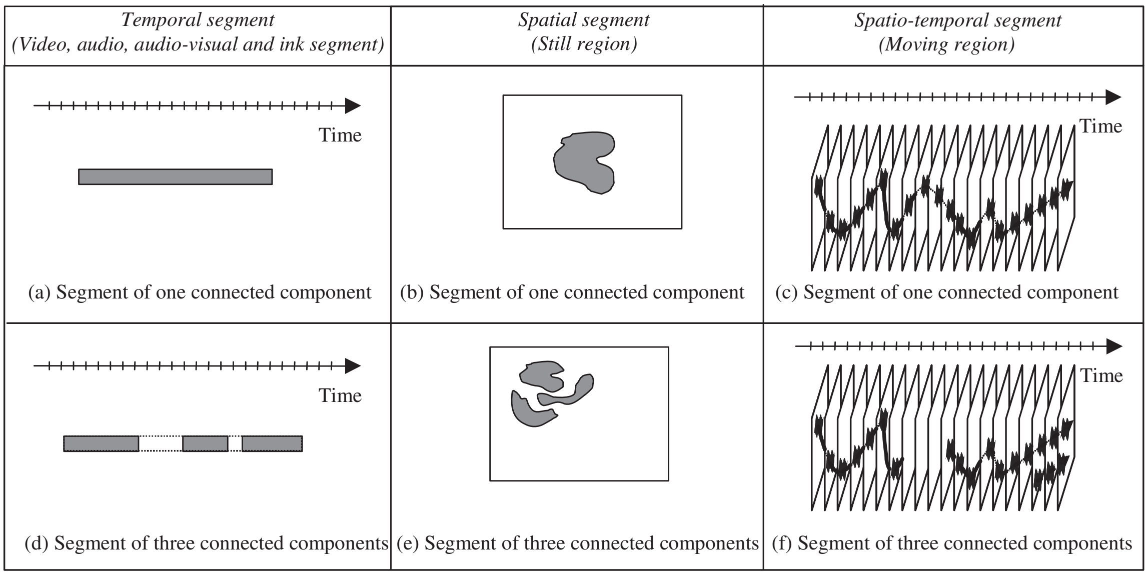
Temporal segment
(Video, audio, audio-visual and ink segment)
Spatial segment
(Still region)
Spatio-temporal segment
(Moving region)
Time
Time
(a) Segment of one connected component
(b) Segment of one connected component
(c) Segment of one connected component
Time
Time
(d) Segment of three connected components
(e) Segment of three connected components
(f) Segment of three connected components

FIG. 3. Examples of segments: (a), (b), (c) segments composed of a single connected component; (d), (e), (f) segments composed of three connected components.

the segment, the segment importance, and the relevance of some of its descriptors can also be described. Some of these features are briefly discussed here.

The Spatial Mask, Temporal Mask, and Spatiotemporal Mask descriptors describe the localization of separated connected components (subregions or subintervals). They are used to describe the spatial and temporal location of the subregions or subintervals of nonconnected segments (e.g., the start time and duration for each of the three intervals of the temporal segment in Figure 3(d)). For example, the *Spatial Mask* D describes the localization of the spatial connected components of a still region. Similarly, the *Temporal Mask* D describes the localization of the temporal connected components of a video segment, an ink segment, an audio segment, an audiovisual segment, or the audio part of an audiovisual region. Finally, the *Spatiotemporal Mask* D describes the localization of the spatiotemporal components of a moving region or of the visual part of an audiovisual region.

The importance of segments and segment descriptors is described by the Matching Hint and Point of View descriptors, respectively. The *Matching Hint* D describes the relative importance of instances of audio or visual Ds/DSs (e.g., the color structure of a still region) or parts of audio or visual Ds/DSs (e.g., the fourth value element of a color structure description of a still region) in instances of segments. For specific applications, the *Matching Hint* D can improve the retrieval performance by specifying the most relevant descriptors for matching, which may depend on the application and also may vary from segment to segment. On the other side, the Point of View describes the relative importance of segments given a specific viewpoint. The Point of View assigns values between 0 and 1 to segments on the basis of a viewpoint specified by a string (e.g., "Home team" for a soccer game).

Other segment attribute tools describe specific media, creation, and handwriting recognition information related to the ink segments. These tools are the *Ink Media Info, Hand Writing Recognition Information* and *Hand Writing Recognition Result* DSs. The *Ink Media Information* DS describes parameters of the input device of the ink content (e.g., width, height, and lines of the writing field's bounding box), writer's handedness (e.g., right or left), and ink data style (e.g., "cursive" or "drawing"). The *Hand Writing Recognition Information* DS describes the handwriting recognizer and the ink lexicon used by the recognizer. The *Hand Writing Recognition Result* DS describes the results of a handwriting recognizer, in particular, the overall quality and some results of the recognition process with corresponding accuracy scores.

Finally, depending on the nature of the segment, visual and audio Ds/DSs can be used to describe specific features related to the segment. Examples of visual features are color, shape, texture, and motion, whereas audio features involve audio spectrum, audio power, fundamental frequency, harmonicity, timbre, melody, and spoken content, among others. The corresponding tools are reviewed in Audio and Visual Features.

Segment Decompositions

An important part of the structural description of content addresses the decomposition or subdivision of segments into subsegments. Decomposition supports the creation of segment hierarchies to generate, for example, tables of contents and indexes. The decomposition is described by the *Segment Decomposition* DS, which is an abstract type that represents an arbitrary decomposition of a segment. It is important to note that MPEG-7 does not specify the way a segment should be segmented into subsegments nor describe the segmentation process; however, it can describe the criteria used during the segmentation and the dimensions affected by the segmentation: space, time, and/or media source.

The basic DSs derived from the *Segment Decomposition* DS describe specific types of decomposition: the *Spatial Segment Decomposition*, *the Temporal Segment Decomposition*, *the Spatio-Temporal Segment Decomposition*, and the *Media Source Decomposition DSs.* For example, an image can be decomposed spatially into a set of still regions corresponding to the objects in an image, which, at the same time, can be decomposed into other still regions. Similar decompositions can be generated in time and/or space for video and other multimedia content. Media source decompositions divide segments into their media constituents such as audio tracks or viewpoints from several cameras. The subsegments resulting from a decomposition may overlap in time, space, and/or media source. Furthermore, their union may not cover the full time, space, and media extents of the parent segment, thus leaving gaps. Two attributes in the *Segment Decomposition* DS indicate whether a decomposition leaves gaps or overlaps. Note that, in any case, the segment decomposition implies that the union of the spatiotemporal and media spaces defined by the child segments is included in the spatiotemporal and media spaces defined by their parent segment (i.e., children are included within their parents).

Several examples of decompositions for temporal segments are included in Figure 4. Figure 4(a) and 4(b) show two examples of segment decompositions with neither gaps nor overlaps (i.e., a partition in the mathematical sense). In both cases, the union of the temporal extents of the children corresponds exactly to the temporal extent of the parent, even if the parent is itself nonconnected. Figure 4(c) shows an example of decomposition with gaps but no overlaps. Finally, Figure 4(d) illustrates a more complex case in which the parent is composed of two connected components and its decomposition generates three children with gaps and overlaps: the first child is itself nonconnected and composed of two connected components; the two remaining children are composed of a single connected component.

The decomposition of a segment may result in segments of a different nature. For example, a video segment may be decomposed into other video segments as well as into still or moving regions. However, not all the combinations of parent segment type, segment decomposition type, and child segment type are valid. For example, an audio segment can only be decomposed in time or media into other audio segments; the spatial decomposition of an audio segment into *still regions* is not valid. The valid decompositions among specific types of segments are specified in the MPEG-7 standard.

Describing the hierarchical decomposition of segments is useful to design 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 Ds and DSs, but it may also be described by the union of the Ds and DSs that are related to its subsegments.

An example of image description with several spatial decompositions is illustrated in Figure 5. The full image (whose ID is “SR1”) is described using the *Still Region* DS whose creation (title, creator), usage (copyright), media (file format), text annotation (summary of image content), and color properties are described using description tools that are discussed in Description of the Content Structure and Content Management. This first still region is decomposed into two still regions, which are further decomposed into other still regions. For each decomposition step, Figure 5 indicates whether gaps and overlaps are generated. The complete segment hierarchy is composed of nine still regions (note that “SR9” and “SR7” are single segments composed of two separated connected components). For each region, Figure 5 shows the type of feature that is described (interestingly, text annotation is the only one applicable to all).

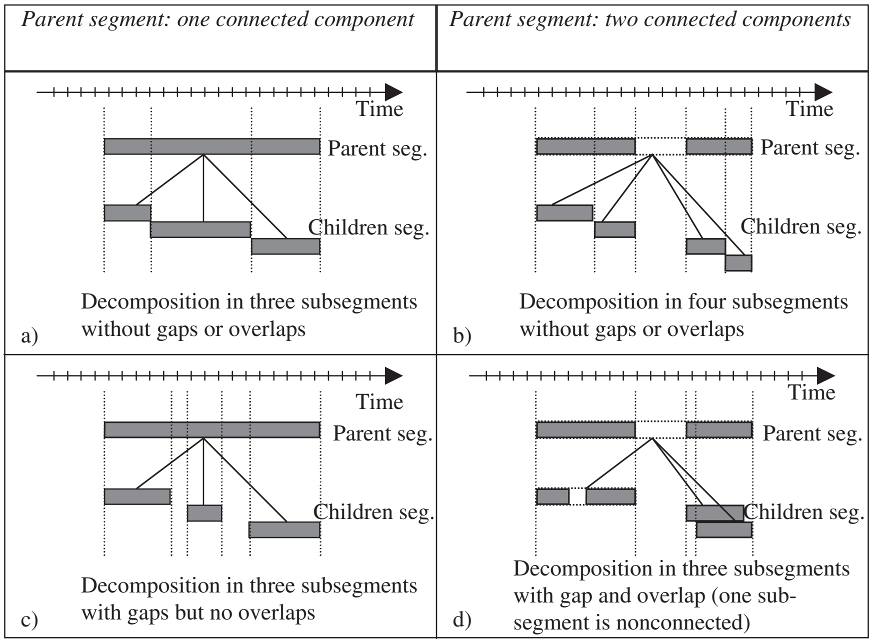
Parent segment: one connected component
Parent segment: two connected components
Time
Parent seg.
Children seg.
a)
Decomposition in three subsegments without gaps or overlaps
b)
Decomposition in four subsegments without gaps or overlaps
c)
Decomposition in three subsegments with gaps but no overlaps
d)
Decomposition in three subsegments with gap and overlap (one subsegment is nonconnected)

FIG. 4. Examples of segment decompositions: (a) and (b) decompositions with neither gaps nor overlaps; (c) and (d) decompositions with gaps and/or overlaps.

DOI: 10.1002/asi

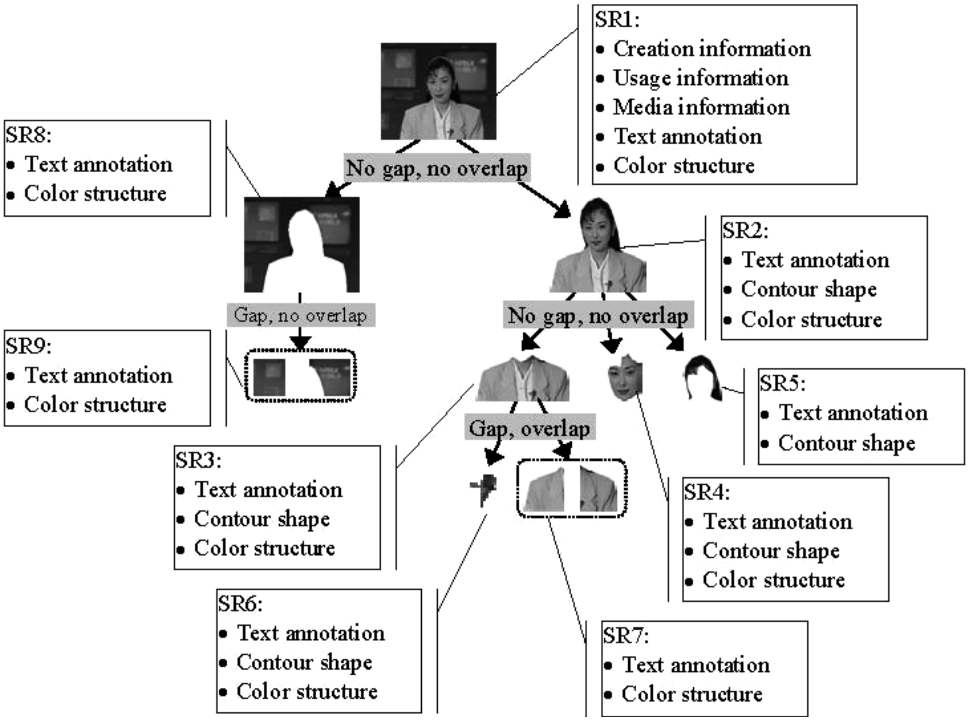

FIG. 5. Example of an image description as a hierarchy of still regions and associated features.

Structural Relations

The description of the content structure in MPEG-7 is not constrained to rely on hierarchies. Although hierarchical structures provided by the segment decomposition tools are adequate for efficient access, retrieval, and scalable description, they imply constraints that may make them inappropriate for certain applications. In such cases, the MPEG-7 graph and segment relation tools can be used to describe more general structures. MPEG-7 has standardized a set of common structural relations, but it also allows the description of nonnormative relations (discussed in the soccer example later).

The *Spatial Relation* and the *Temporal Relation* classification schemes specify spatial and temporal relations, respectively. The normative structural relations in MPEG-7 are listed by type in Table 1. For each normative segment relation, MPEG-7 has also standardized the inverse relation.

TABLE 1. Normative segment relations in MPEG-7 listed by type.

Type	Normative relations
Spatial	South, north, west, east, northwest, northeast, southwest, southeast, left, right, below, above
Temporal	Precedes, follows, meets, metBy, overlaps, overlappedBy, during, contains, strictDuring, strictContains, starts, startedBy, finishes, finishedBy, coOccurs, contiguous, sequential, coBegin, coEnd, parallel, overlapping

Note that the inverse relation is defined as follows: “A InverseRelation B” ↔ “B Relation A.”

To illustrate the use of the segment entity and relation tools, consider the examples in Figure 6 and Figure 7. These examples show an excerpt from a soccer match and its description. One video segment and four moving regions are described. A graph of structural relationships describing the

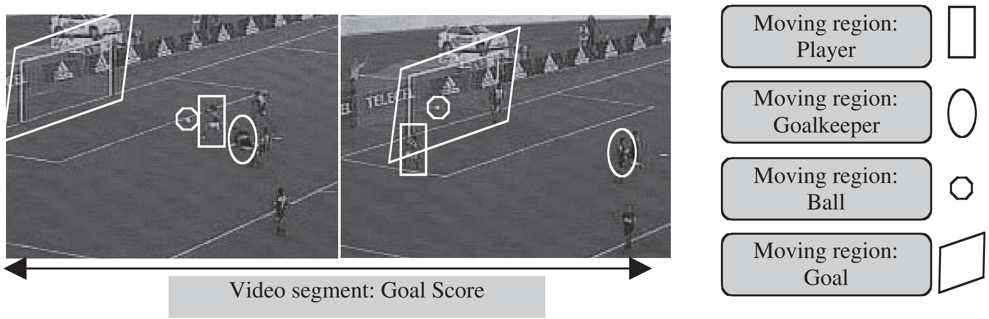

FIG. 6. Excerpt from a soccer match with the corresponding video segment and moving regions that participate in the graph in Figure 7.

DOI: 10.1002/asi

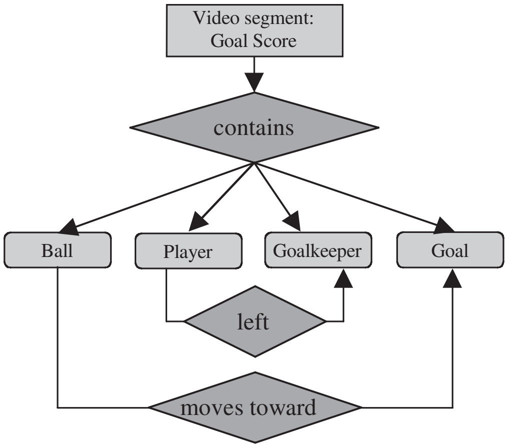

FIG. 7. Example of a graph of structural relationships for the video segment and moving regions in Figure 6.

relations between the segments is shown in Figure 6. The video segment "Goal Score" involves moving regions "Ball," "Goalkeeper," "Player," and "Goal." The moving region "Player" is on the left of the moving region "Goalkeeper," and the moving region "Ball" moves toward the moving region "Goal." This example illustrates the flexibility of this kind of representation. Note that this description is mainly structural because the relationships specified in the graph edges are purely spatiotemporal and visual and the nodes represent segments. The only explicit semantic information is available from the textual annotation (in which keywords such as "Ball," "Player," or "Goalkeeper" are used). All the relations in this example are normative except "moves toward."

Content Management

As mentioned previously, any type of segment can be described by a set of DSs dealing with the description of the life cycle of multimedia content including its creation, usage, or media encoding. They are organized in three main areas dealing with description of the creation process of the media and of the usage.

Content Creation

This area contains author-generated information about the content creation process. This information cannot usually be extracted from the entity itself. That is, the information is related to, but not explicitly depicted in, the actual entity content. The creation information is "wrapped" in the *Creation Information* DS, which is composed of one *Creation* DS, an optional *Classification* DS, and an arbitrary number of *Related Material* DSs.

- The *Creation* DS describes the creation of the entity content, including places, dates, actions, materials, staff (technical and artistic, e.g., directors and actors), and organizations involved. Titles and abstracts can also be described using the *Creation* DS.
- The *Classification* DS classifies the content for searching and filtering. It includes user-oriented classifications such as language, subject, genre, and production type (e.g., documentary and news), as well as service-oriented classifications such as purpose, parental guidance, market segmentation, and media review.
- The *Related Material* DS specifies additional information about the entity content available in other materials.

Media Information

In the simplest case, the content is created or recorded only once. In this situation, the content can be represented as a unique medium with its associated format, coding scheme, creation information, and usage information. A typical example of this simplest case is a picture from a consumer digital photograph album. Nevertheless, more complex scenarios have to be considered, for example, in which a single event, called a *reality* (e.g., a sports event), is recorded in different modalities or with different encoding formats. In this case, different "content entities" are produced. The *Media Information* DS describes each content entity.

Furthermore, each content entity can be encoded using various coding schemes or parameters and stored in various formats. The combination of a coding scheme with its associated parameters and formats determines a media profile. Finally, within the same media profile, several copies or instances of the same content can be created. The original media profile is called the *master profile*. Note that the quality of audio or visual content may decrease when the signal goes through compression, transmission, or signal conversion. As a result, MPEG-7 provides a set of DSs/Ds for describing these profiles and instances as well as their corresponding quality.

Media information can also provide transcoding hints that allow the creation of additional media variations for applications that need to adapt the content for transmission or delivery under specific network conditions or to terminals with specific characteristics. The transcoding hints provide information on ways to generate new content from the current piece of content. They do not describe the relation between two existing pieces of content.

Usage Information

The usage information describes rights, financial aspects, and availability of the content. It is wrapped in the *Usage Information* DS, which contains one *Rights* descriptor, an optional *Financial* descriptor, and zero or more *Availability* and *Usage Record* DSs.

- The *Rights* D provides access to information about the rights holders and the access rights. No rights information is explicitly described; MPEG-7 simply provides links to information related to rights management and protection. It provides these references in the form of unique identifiers that are managed by external authorities. The underlying strategy is to enable MPEG-7 descriptions to provide access

DOI: 10.1002/asi

to current rights owner information without dealing with the information and negotiation directly.

- The *Financial* D contains information related to the costs generated and incomes produced by multimedia content.
- The *Availability* DS describes the availability for access of the content entity's media instances. It contains tools for referencing the associated media instance and describing, among other aspects, the type of publication medium, the disseminator, additional financial information such as publication costs and price of use, and the availability period.
- The *Usage Record* DS describes the past use of the content. It contains a reference to the associated instances of the *Availability* DS.

Note that the *Usage Information* DS may be updated or extended each time the content is used or when there are new ways to access to the content.

Audio and Visual Features

Segments can be further described by their audio or visual properties. The MPEG-7 standard has specified a set of DSs and descriptors for this purpose. Note that, in general, each audio or visual descriptor can be used only for a limited set of segment types specified by the standard.

Visual Description Tools

The main features characterized by the visual descriptors are color, texture, shape, and motion.

Color features. MPEG-7 has standardized eight color descriptors: Color Space, Color Quantization, Dominant Color, Scalable Color Histogram, Color Structure, Color Layout, and GoF/GoP Color. The Color Space descriptor specifies a color space; whereas the Color Quantization descriptor specifies the partitioning of the color space into discrete bins. The first two descriptors are intended to be used in conjunction with the remaining descriptors. The Dominant Color descriptor is suitable for representing local features where a small number of colors are enough to characterize the color information in the region of interest. It defines the set of dominant colors, the percentage of each color in the region of interest, and, optionally, the spatial coherence. This descriptor is mainly used in retrieval by similarity. The Scalable Color Histogram descriptor represents a color histogram encoded with a Haar transform to provide scalability in terms of bin numbers and accuracy. It is particularly attractive for image-to-image matching and color-based retrieval. The Color Structure descriptor captures both color content and its structure. Its main functionality is image-to-image matching. The extraction method essentially computes the relative frequency of 8 × 8 windows that contain a particular color. The Color Layout descriptor specifies the spatial distribution of colors for high-speed retrieval and browsing. It targets not only image-to-image matching and video-clip-to-video-clip matching, but also layout-based retrieval for color, such as sketch-to-image matching that is not supported by other color descriptors. The descriptor represents the DCT values of an image or a region that has been previously partitioned into 8 × 8 blocks and where each block is represented by its dominant color. The last color descriptor is the GoF/GoP Color descriptor. It extends the Scalable Color Histogram descriptor defined for still images to video sequences or collection of still images.

Texture features. There are three texture descriptors: *Homogeneous Texture, Texture Browsing* (more similar to human description), and *Edge Histogram. Homogeneous Texture* has emerged as an important visual primitive for automated searching and browsing through large collections of similar-looking patterns. The Homogeneous Texture descriptor relies on a frequency decomposition with a Gabor filter bank. The frequency bands are defined by a scale parameter and an orientation parameter. The first and second moments of the energy in the frequency bands are then used as the components of the descriptor. The *Texture Browsing* descriptor provides a qualitative representation of the texture similar to a human characterization, in terms of perceivable qualities such as dominant direction, regularity, and coarseness. It is useful for texture-based browsing applications. Finally, the Edge Histogram descriptor represents the histogram of five possible types of edges, namely, four directional edges and one nondirectional edge. The descriptor primarily targets image-to-image matching (query by example or by sketch).

Shape. There are three main shape descriptors: *Region Shape, Contour Shape,* and *3D Shape*. The *Region Shape* and *Contour Shape* descriptors are intended for shape matching. They do not provide enough information to reconstruct the shape nor to define its position in the image. Two shape descriptors have been defined because, in terms of applications, there are at least two major interpretations of shape similarity, either by contour similarity or by pixel-distribution similarity. The *Contour Shape* descriptor essentially captures the points of high curvature along the contour (position of the point and value of the curvature). This representation has a number of important properties: It captures characteristic features of the shape, allowing efficient similarity-based retrieval. It is robust to nonrigid deformation and partial occlusion. The *Region Shape* descriptor captures the distribution of all pixels within a region. Note that this descriptor can deal with regions made of several connected components or including holes. The 3D shape information can also be described in MPEG-7. Most of the time, 3D information is represented by polygonal meshes. The *3D Shape* descriptor provides an intrinsic shape description of 3D mesh models through their shape spectrum.

Motion features. There are four motion descriptors: *Camera Motion, Motion Trajectory, Parametric Motion,* and *Motion Activity.* The *Camera Motion* descriptor characterizes 3D camera motion parameters. It supports the following basic camera operations: fixed, tracking (horizontal transverse movement, also called "traveling "in the film industry), booming (vertical transverse movement), dollying (translation along the optical axis), panning (horizontal rotation), tilting (vertical rotation), rolling (rotation around the optical axis), and zooming (change of the focal length). The *Motion Trajectory* descriptor characterizes the temporal evolution of key points. It is composed of a list of key points (x, y, z, t) along with a set of optional interpolating functions that describe the trajectory between key points. The *Parametric Motion* descriptor defines the motion of regions in video sequences as a 2D parametric model. Specifically, affine models include translations, rotations, scaling, and combination of them. Planar perspective models make it possible to take into account global deformations associated with perspective projections. Finally, quadratic models make it possible to describe more complex movements. The parametric model is associated with arbitrary regions over a specified time interval. Finally, the *Motion Activity* descriptor captures the intuitive notion of "intensity of action" or "pace of action" in a video segment. It is based on five main features: the intensity of the motion activity (value between 1 and 5), the direction of the activity (optional), the spatial localization, and the spatial and the temporal distributions of the activity.

Audio Features

Most audio description tools are based on audio features that permit similarity in sounds (such as music and speech) to be assessed. The similarity is based on characteristics such as spectrum, harmony, timbre, and melody contained in the *Audio Segment* DS. The set of audio descriptors can be grouped in four basic categories on the basis of on the functionality they support:

- Robust audio matching, supported by the *Audio Signature* DSs, which describes spectral flatness of sounds
- Timbre matching (identification, search, and filtering), supported by the *Harmonic Instrument Timbre* and the *Percussive Instrument Timbre* descriptors
- Melodic search, supported by the *Melody Contour* DS (efficient melody description) and the *Melody Sequence* DS (complete melody description)
- Sound recognition and indexing, supported by the *Sound Model DS,* the *Sound Classification Model DS,* the *Sound Model State Path descriptor,* and the *Sound Model State Histogram* descriptor

Finally, a very important piece of information about audio segments relies on the description of spoken content. In MPEG-7, this feature is handled by the *Spoken Content Lattice* DS and the *Spoken Content Header* descriptor.

Conclusions

This article has reviewed the major set of MPEG-7 tools devoted to the description of the structural aspects of multimedia content. The basic description scheme supporting this functionality is the *Segment* DS. It describes the outcome of a spatial, temporal, or spatiotemporal partitioning of the multimedia content. The structural aspects of the content are represented through decomposition of the content and relations among the components produced by the partitioning. Decomposition efficiently represents hierarchies and can be used, for example, to create tables of contents or indexes. More general graph representations are handled by the various standard spatial and temporal relations. Each segment can be described by a large number of features ranging from those targeting the life cycle of the content (its creation, its various versions and copies, its usage) to those addressing the signal characteristics such as audio, color, shape, or motion properties. Moreover, MPEG-7 has specified a fairly large number of description schemes (DSs) or descriptors (Ds) related to specific functionalities used across many applications (for example, the *Spoken Content* or the *Camera Motion* DSs).

Disclaimer

The views expressed herein are those of the authors and are not necessarily those of Thomson Inc., or its affiliates.

References

International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) (2002, 2003, 2004). 15938. Part 1 to 8. Information Technology—Multimedia Content Description Interface (MPEG-7).

Manjunath, B.S., Salembier, P., & Sikora, T. (Eds.). (2002). Introduction to MPEG-7: Multimedia Content Description Interface. Chichester, UK: John Wiley & Sons.

W3C (2001). XML Schema, W3C recommendation. Retrieved May 2, 2001, from http://www.w3.org/XML/Schema

Bios

Philippe Salembier received degrees from the Ecole Polytechnique and the Ecole Nationale Supérieure des Télécommunications, both in Paris, France, and a Ph.D. from the Swiss Federal Institute of Technology (EPFL) in 1991. He was a Postdoctoral Fellow at the Harvard Robotics Laboratory, Cambridge, MA. He is currently a professor at the Technical University of Catalonia, Barcelona, lecturing in the area of digital signal and image processing. He was chair of the MPEG-7 "Multimedia Description Scheme" group between 1999 and 2001.

Ana B. Benitez has been a Senior Member of the Technical Staff at Thomson Corporate Research, Burbank, CA, USA, since 2004. She received her telecommunications engineering degree from the Polytechnic University of Catalonia (UPC) in Barcelona, Spain, in 1996 and her Ph.D. from Columbia University in 2005 in the Department of Electrical Engineering, where she was a Kodak Graduate Fellow and a member of the ADVENT Consortium and the Digital Video Multimedia (DVMM) Lab. She chaired the Ad Hoc Group on MPEG-7 Multimedia Description Schemes Core Experiments and she was an Editor of the Multimedia Description Schemes part of the MPEG-7 standard.

DOI: 10.1002/asi