The MPEG-7 Description Standard

Nina Jaunsen
Dept of Information and Media Science
University of Bergen, Norway
September 2004

The increasing use of multimedia in the general society and the need for user-friendly effective multimedia retrieval systems are indisputable. The fact that effective retrieval is reliant on complete and thorough description (among other things) is also fairly agreed upon. Various standards for the description of multimedia are being developed and some exist already. Typically such standards tend to be specialised for certain applications or domains leaving only a few representing general-purpose multimedia description. As of today, the MPEG-7 seems to be recognised as the most complete general-purpose description standard for multimedia. Whether the MPEG-7 multimedia description standard qualifies as an appropriate general-purpose description standard and is compliant with the requirements of such a standard is a discussion beyond the scope of this thesis.

MPEG-7 is an ISO/IEC (International Standards Organization/International Electro-technical Committee) approved standard developed by MPEG (Moving Picture Expert Group), a working group developing international standards for compression, decompression, processing, and coded representation of audio-visual data. The standard was initialised in 1996 and it represents a continuously evolving framework for standardising multimedia content description. In the context of this thesis the proposed MPEG-7 standard represents an assessment and basis for evaluation of a general MIRS’ ability to adequately, according to the standard, describe image-media content applied in a digital museum context.

The following review of the MPEG-7 standard is based on MPEG-7 documentation (2003) and is merely intended to provide an overview of the standard and to emphasise the image-media specific descriptive elements.

---

1 This paper is part of a master’s thesis in Information Science at the University of Bergen, Norway titled: Improving Image Retrieval through Enhanced Image Description.
An overview

MPEG-7 offers a comprehensive set of audiovisual description tools including metadata elements and their structures and relationships defined by the standard in the form of Descriptors and Description Schemes.

The main elements of the MPEG-7 standard:

- Description Tools: Descriptors (D), defining the syntax and semantics of each feature (metadata element), and Description Schemes (DS), specifying the structure and semantics of the relationships between elements, which may be both Descriptors and Description Schemes.
- A Description Definition Language (DDL) to define the syntax of the MPEG-7 Description Tools and to allow the creation of modified/extended/new Description Schemes and/or Descriptors.
- System tools, to support binary coded representation for efficient storage and transmission, transmission mechanism (both for textual and binary formats), multiplexing of descriptions, synchronization of descriptions with content, management and protection of intellectual properties in MPEG-7 descriptions etc.

The overall main tools used to develop MPEG-7 descriptions are the description definition language (DDL), the Description Schemes (DSs) and the Descriptors (Ds). Descriptors typically represent quantitative measures binding a specific feature to a set of values. Description Schemes can be interpreted as models of the real world object and the environments they represent. A DS is a model of the description itself, specifying the specific Descriptors to be used in a description, and the relationships between the Descriptors or between other Description Schemes. The DDL defines the syntactic rules to define, express and combine Description Schemes and Descriptors.

XML Scheme has been the general DDL (Description Definition Language) used for the syntactic definition of MPEG-7 Description Tools and is also proposed as the preferred language for the textual representation of content description. However, the MPEG-7 standard is not in itself dependent of XML as its DDL as long as the language of choice complies with the standard’s requirements regarding flexibility and extensibility.

The MPEG-7 Description Tools

The Description Tools consist of the Visual Description Tools dealing with visual descriptions only, Audio Description Tools dealing with audio descriptions and the Multimedia Description Schemes handling generic characteristics and compound multimedia descriptions. Generic characteristics refer to features and attributes applicable to all media types.

The Descriptors are designed primarily to describe features usually referred to as low-level audiovisual features such as colour, texture, motion, audio energy etc, as well as attributes
associated with AV content such as location, time, quality and so forth. It is expected that most low-level Descriptors can be extracted automatically in the various applications.

The MPEG-7 Description Schemes are primarily designed to express and describe higher-level AV features such as objects, events, regions, segments and other metadata related to the creation, production and usage of the media objects. They produce more complex descriptions by integrating together multiple Descriptors and Description Schemes, and by declaring relationships among the description components. Typically, the multimedia DSs 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 Description Schemes, but in many cases instantiating DSs requires human assisted extraction or authorizing tools.

The collected Description Tools allow and provide for content description from various viewpoints. Though often presented as separate entities/approaches, they are interrelated and can be combined in many ways. Depending on the application (context), some approaches will be more appropriate and thus more present and some less appropriate and perhaps absent or only partly present in a content description.

**Multimedia Description Schemes**

The MPEG-7 Multimedia Description Schemes represent metadata structures dealing with generic and multimedia entities. They can be organized into the following areas and presented in a graphical model. Although all areas more or less (explicitly or implicitly) contribute to image description, the areas presumed to be most relevant and essential for image description are highlighted in grey.

1. Content Organization
2. Navigation and Access
3. User Interaction
4. Basic Elements
5. Content Management
6. Content Description
Content Organization refer to structures for organizing and modelling collections of AV content, segments, events and/or objects and describing the common properties provided for by the Collection Structure DS and various Model DSs. A series of images depicting the same whale could represent such a collection and be described using the Collection Structure DS. Various relationships between the images within a collection and even across different collections such as temporal order, placement and degree of similarity can be specified as well. The collections can be further described using different models and statistics in order to characterize the common attributes of the collection members.

To support Navigation and Access of AV content MPEG-7 provides DSs describing summaries, views, partitions and variations of the AV content. The summary descriptions allow the AV content to be navigated either hierarchical or sequential. The hierarchical summary organizes the content into successive levels of detail and the sequential summary provides a sequence of images composing a kind of slideshow or AV skim. The View DS describes a structural view, partition, or decomposition of an AV signal in space, time, and frequency. In general, views of signals correspond to low-resolution views, spatial or temporal segments or frequency sub-bands. The Variation DS describes variations of AV content, such as summaries and abstracts; scaled, compressed and low-resolution versions; different languages and modalities such as audio, video, image, text etc.

User Interaction structures describe user preferences and usage history pertaining to the use and general consumption of multimedia material. MPEG-7 content descriptions can be matched against user preferences to personalize the AV content access, presentation and consumption. The Usage History DS describes the history of actions carried out by an end-user and can be exchanged.
between consumers, their agents, content providers and perhaps in turn be used to determine the user’s preferences with regard to AV data.

The Basic Elements include components and structures necessary for the development of complex and compound description schemes. Schema Tools assist in the formation, packaging, and annotation of MPEG-7 descriptions. The Basic data types provide a set of extended data types and mathematical structures such as vectors and matrices, needed by the DSs for describing some AV content. Links and media localization represent constructs for linking media files and localizing pieces of content. The Basic Tools provide constructs for describing time, place, individuals, groups, organizations, and other textual annotation. In addition, constructs for classification schemes and controlled terms are provided by the Basic Elements.

Content Management includes tools describing the life cycle of the content, from creation and production, media coding (storage and file formats), to consumption and usage. “Content” is here referred to as the specific structure representing a real world object. A content described by MPEG-7 descriptions can be available in different modalities, formats, coding schemes and there can be several instances.

1. The Creation and Production Description Tools describe author-generated information regarding the creation and production process of the AV content. The Creation Information DS is composed of information regarding the creation and classification of the AV content and of information regarding related material. The creation can describe titles, textual annotations, creators, creation locations and associated dates. Classification describes how the AV material is classified into categories such as genre, subject, purpose and language etc. Related material describes other existing AV objects that are related to the content in question. Because this information is merely associated with the content and not explicitly depicted in the actual content this information cannot be automatically extracted and must usually be added manually.

2. The Media Description Tools describe the storage features of the media such as the format, compression and coding of the AV content. It identifies the master media (original source) of the AV content and from which instances or versions can be produced. Possible instances of AV content are referred to as Media Profiles representing versions of the original source obtained by using different formats or encoding.

3. The Content Usage Description Tools describe usage information related to the AV content such as usage rights, availability, usage record and financial information. The usage information is typically dynamic indicating that it is likely to change during the lifetime of the AV content. For this reason it may be preferable to not include this information explicitly in the MPEG-7 description, but to rather provide links to the right holders and/or other information regarding rights management, availability and content usage.
**Content Description** provides Description Schemes for describing AV content from the viewpoint of its physical and logical structure and from the viewpoint of real-world semantics and conceptual notions. The structural tools describe AV content in terms of video segments, frames, still and moving regions and/or audio segments. The semantic tools describe objects, events and conceptual notions from the real world that are captured by the AV content.

1. Structural aspects of the AV content are based on segments (Segment DS) representing spatial, temporal or spatial-temporal structures of the content. A segment is a section of an AV content item and a segment can be decomposed into sub-segments corresponding to the required level of detail, often forming a hierarchical segment tree. Spatial and temporal segmenting may leave gaps and overlaps between the sub-segments. Each segment is typically further described using Visual (media specific) description tools usually based on low-level features of the content such as colours, textures and shapes. The Segment DS is an abstract class and merely used to define the common properties of its subclasses. Among such common properties (elements and attributes) of segments is information related to creation, usage, media location and textual annotation. The Segment DS forms the base abstract type of specialized segment types (segment subclasses) such as audio segments, video segments, audio-visual segments, moving region/still region segments. Structural descriptions may also include a Graph DS that allows the representation of other and perhaps more complex spatial and/or temporal relations, used to describe relationships between segments not presented by the hierarchical/tree structures alone.

Figure 2 illustrates a segment hierarchy based on a root image (still region). The root segment is shown together with some common segment properties and the sub-segments are shown together with a suggested set of more media specific Description Schemes/Descriptors, which could be suitable for proper segment description and thus image description.
Figure 2 Image descriptions by segmenting
Figure 3 illustrates a possible graph (Graph DS) describing the structure of the image content by revealing other existing relationships between the segments.
2. Conceptual aspects of the AV content are revealed based on Semantic DSs typically involving entities such as objects, events, abstract concepts, places and time, all in narrative worlds. A narrative world refers to the context for a semantic description, the “reality” in which the description makes sense. The semantic description is based on the generic SemanticBase DS and the several derived and specialized DSs, describing the specific types of semantic entities, such as narrative worlds, objects, events, places and time. As in the case of the Segment DS, the conceptual aspects of the 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 relationships between the nodes. Figure 3.4 illustrates some of the different tools (DS) typically used to describe the semantic AV content and how these can be linked to segments of images.

Figure 3.4 Some tools for describing conceptual aspects
Beside the semantic description of individual instances (specific image) of AV content, the Semantic DSs also allow 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. Figure 3.5 illustrates possible conceptual aspects and abstractions of a specific instance (image) of AV content.

Figure 3.5 Example of Conceptual aspects

The Structure DSs and Semantic DSs can be related by a set of links allowing the AV content to be described on the basis of both content structure and semantic structures together. The links relate different Semantic concepts to the instances within the AV content described by the Segments. Furthermore, most of the MPEG-7 Content Description and Content Management DSs are linked together and in practice, also often included within each other in the MPEG-7 descriptions. Depending on the requirements of the application, some aspects of the AV content description can be emphasised such as semantic description and/or creation description, while other aspects can be minimised or ignored, such as the media or structure description.
The MPEG-7 Visual Description Tools

The MPEG-7 Visual Description Tools intend to support the description of elements specific for visual content such as still images and video. The visual description tools consist of some defined basic structures and descriptors covering the essential, according to MPEG-7, visual features.

The five visual related basic structures represent structural methods for decomposing the image or image segments for visual specific description.

- The **Grid Layout** splits the image into a set of equally sized rectangular regions in order to describe each region in terms of other Descriptors such as colour/texture separately. Regions may further be split into sub-regions etc.

- The **2D-3D Multiple Views** specify a structure combining 2D Descriptors representing a visual feature of a 3D object seen from different view angles. The descriptors form a complete 3D view representation of the object. The 2D-3D descriptor supports integration of the 2D Descriptors used in the image to describe features of the 3D (real world) objects.

- The **Spatial 2D Coordinates** define a 2D spatial coordinate system and a unit to be used by reference in other Ds/DSs when relevant. The coordinate system is defined by a mapping between image and coordinate system. One of the advantages using this descriptor is that MPEG-7 descriptions need not be modified even if the image size is changed or a part of the image is removed.

The two last basic visual structures, Time Series and Temporal Interpolation, are based on temporal aspects for video media and are not regarded relevant for the description of image data.

MPEG-7’s Visual Descriptors cover the common basic visual features of colour, texture and shape and also the aspects of Motion, Localization and Face recognition.

The Colour Descriptors

The seven colour descriptors are briefly presented to indicate various methods for colour representation in colour descriptions.

**Colour space Descriptor** – A colour space is based on a colour model (mathematical model) describing the way colours can be represented as tuples of numbers, typically as three or four values, also referred to as colour components. The colour space defines the total range of colours, which can be described using a particular colour model. The Descriptor indicates the colour space used in a specific colour-based description.

**Colour Quantisation Descriptor** – a uniform quantification of a colour space. Quantisation refers to the reduction of the numbers of unique colours in an image. The number of bins that the quantiser produces is configurable to support greater flexibility.
Dominant Colour Descriptors – suitable for representing local (object or image region) features where a small number of colours are sufficient to characterize the colour information in the region of interest.

Scalable Colour Descriptor - the Scalable Colour Descriptor is a Colour Histogram based on the Hue Saturation Value (HSV) model (colour space). HSV defines the colour space in terms of three constituent components: hue (a colour type such as red, blue etc), saturation (the intensity of the colour) and value (the brightness of the colour). Its binary representation is scalable in terms of bin numbers and bit representation accuracy over a broad range of data rates.

Colour Layout Descriptor– represents the spatial distribution of colour as visual signals in a very compact form. This compactness allows visual signal matching functionalities with high retrieval efficiency at a very low computational cost.

Colour-Structure Descriptor – captures both colour content (similar to colour histogram) and information about the structure of this content.

GoF/GoP Colour Descriptor– the Group of Frames/Group of Pictures colour descriptor extends the ScalableColour descriptor that is defined for a still image to colour description of a video segment or a collection of still images.

The Texture Descriptors
Texture is defined as the arrangement of particles of any material (wood, metal, fabric etc.) as it affects the appearance or feel of the surface. Can also be interpreted as structure or composition.

Homogenous Texture Descriptor– provides a precise quantitative description of a texture and is appropriate for identifying similar textures (patterns/structures) in search and retrieval. A parking lot with cars parked at regular intervals is a good example of a homogenous pattern when viewed from a distance. Agricultural and vegetation areas are other examples of homogenous patterns useful to identify especially when browsing aerial and satellite imagery. The computation of this descriptor is based on orientation- and scale-tuned filters. Homogenous texture has emerged as an important visual descriptor for searching and browsing large collections of similar looking patterns.

Texture Browsing Descriptor– is useful for representing homogenous texture for browsing type applications. Provides a perceptual characterization of texture, similar to human characterization, in terms of regularity, coarseness and directionality. The texture computation proceeds similarly as the Homogeneous Texture Descriptor, filtering the images with a bank of orientation- and scale-tuned filters.

Edge Histogram Descriptor- represents the spatial distribution of five types of edges, namely four directional edges (vertical, horizontal, 45 degrees and 135 degrees diagonal) and one non-directional edge (isotropic). Since edges play an important role for image perception, it can retrieve images with similar semantic meaning. Thus, it primarily targets image-to-image matching (by example or by sketch), especially for natural images with non-uniform edge distribution.
The Shape Descriptors

Region Shape Descriptor— The shape of an object may consist of a single connected region, a set of regions or regions consisting of holes and disjointed areas. The Region Shape Descriptor makes use of all pixels constituting the shape within a frame, and can thus allegedly efficiently describe shapes consisting of one single connected region as well as more complex shapes regardless of minor deformation along the boundaries of the shape (object). It should be capable of recognising shape-based similarity in spite of disjointed regions and other minor deviations of the shape.

![Figure 6](image)

Figure 6 a, b and d represent single connected region shapes while c represents shape consisting of holes and minor deformations along the boundaries.

The figure above represents two sets of images the Region Shape Descriptor should be able to recognise as similar. The two beluga whales constituting as single connected regions (a. and b.) and a sei-whale skeleton presented as both a single connected region (d.) and a more complex shape consisting of holes and disjoint areas (c.).

Contour Shape Descriptor - captures characteristic shape-features of an object or region based on its contour. The Descriptor uses a so-called Curvature Scale-Space representation capturing perceptually meaningful features of the shape (contours). Some important properties of the Curvature Scale-Space representation of contours and thus the Contour Shape Descriptor are:

- Shape generalisation, seeks to identify perceptual similarity among different shapes
- Non-rigid motion robustness, attempting to recognise movement variations of the shape
- Partial occlusion robustness, recognising similar shapes despite partial occlusions (non-concluded shapes)
- Invariant to certain perspective transformations such as camera angels and parameters
The characteristic shape features of the whales illustrated in figure 3.7 should typically be recognised as whales (or similar) by their contours and perceived to be similar though slightly different in their shapes (generalisation).

The Contour Shape Descriptor can prove to be a very useful tool if actually able to generalise the shapes and recognise the typical human perceivable shape similarity of semantically meaningful objects.

*Shape 3D Descriptor*- 3D information is usually represented as polygon meshes or rather 3D mesh model coding (MPEG-4). The Shape Descriptor described in detail provides an intrinsic shape description of 3D mesh models representing the 3D information.

**The Motion Descriptors**
MPEG-7 recognises four different Motion Descriptors including Camera Motion, Motion Trajectory, Parametric Motion and Motion Activity. As the Motion Descriptors consider aspects of motion they are not applicable to images and thus not within the scope of this thesis.

**The Localization Descriptors**
- *Region Locator Descriptor* – enables localization of regions within images or frames by specifying them with a brief and scalable representation of a Box or a Polygon.

- *Spatio Temporal Locator Descriptor* - describes spatial-temporal regions in a video sequence, such as moving object regions, and provides localization functionality. Not applicable to images.

- *Face Recognition Descriptor* - can be used to retrieve face images matching a query face image. The Face Recognition feature set is extracted from a normalized face image containing 56 lines with 46 intensity values in each line. The centre of the two eyes in each face image are located on the 24th row and the 16th and 31st column for the right and left eye respectively. The normalised face image is used to extract a one dimensional face-vector consisting of the luminance pixel values from the normalised face arranged into the vector by using a raster scan starting at the top-left corner of the image and finishing at the bottom-right corner of the image. The face recognition feature set is then calculated by projecting the one-dimensional face-vector onto a space defined by a set of basis vectors (spanning the space of possible face-vectors). Basically, the descriptor represents the projection of a face-vector onto a set of basis vectors that span the space of possible face-vectors.