International Journal of Image and Graphics, Vol. 1, No. 3 (2001) 527–546

MPEG-7: A STANDARD FOR MULTIMEDIA CONTENT DESCRIPTION

FERNANDO PEREIRA*

Instituto Superior Técnico - Instituto de Telecomunicações, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ROB KOENEN†

InterTrust Technologies Corporation, 4750 Patrick Henry Drive, Santa Clara, CA 95054, USA

Multimedia information is getting more abundant and the means to produce it are becoming a commodity, but finding and managing multimedia content is getting harder and harder. In 1996, MPEG (Moving Picture Experts Group) started a project formally named "Multimedia Content Description Interface", but better known as MPEG-7, acknowledging the need to efficiently and effectively describe and retrieve multimedia information and recognizing the substantial technological developments in the area of multimedia content description.

MPEG-7 sets a standard for multimedia description tools, notably so-called descriptors, description schemes, systems tools and a description definition language. MPEG-7 is generic in the sense that it is not specially designed or optimized for a particular application domain. It is however clear that image and video database applications are among its most important application domains.

This paper intends to overview the context, objectives, technical approach, workplan and achievements of the MPEG-7 standard. The first version and main body of this standard should be ready by July 2001.

Keywords: Multimedia Content Description; Standardization; MPEG-7.

1. Motivation and Goals

The amount of digital multimedia information accessible to the masses is growing everyday, not only in terms of consumption, but also in terms of production. Digital still cameras directly storing in JPEG format have hit the mass market and digital video cameras directly recording in MPEG-1 format are also available. This transforms every one of us in a potential content producer, capable of creating content that can be easily distributed and published using the Internet. But, if it is today easier and easier to acquire, process and distribute multimedia content, it should also be equally easy to access the available information, because huge amounts of

*E-mail: Fernando.Pereira@lx.it.pt
†E-mail: rkoenen@intertrust.com

digital multimedia information are being generated, all over the world, everyday. In fact, there is no use in making available multimedia information that can only be found by chance. Unfortunately, the more information gets available, the harder it becomes to identify and find what you want, and the more difficult it becomes to manage the information.

The anticipated need to efficiently manage and retrieve multimedia content, and the foreseeable increase in the difficulty of doing so was recognized by MPEG (Moving Picture Experts Group) in July 1996. At the Tampere meeting, MPEG[1] stated its intention to provide a solution in the form of a "generally agreed-upon framework for the description of audio-visual content". To this end, MPEG initiated a new work item, formally called "Multimedia Content Description Interface", generally known as MPEG-7.[2] MPEG-7 will specify a standard way of describing various types of multimedia information, irrespective of their representation format, e.g., analog or digital, and storage support, e.g., paper, film or tape. Participants in the development of MPEG-7 represent broadcasters, equipment and software manufacturers, digital content creators, owners and managers, telecommunication service providers, publishers and intellectual property rights managers, as well as university researchers. MPEG-7 is quite a different standard than its predecessors. MPEG-1, -2 and -4, all represent the content itself — "the bits" — while MPEG-7 represents information about the content — "the bits about the bits". But, there is some overlap and sometimes the frontiers are not that sharp.

1.1. *Why a standard?*

There are many ways to describe multimedia content, and indeed many proprietary ways are in use in various digital asset management systems today. Such systems, however, do not allow a search across different repositories for a certain piece of content, and do not facilitate content exchange between different databases using different systems. These are interoperability issues, and creating a standard is an appropriate way to address them.

The MPEG-7 standard addresses this kind of "interoperability", offers the prospect of lowering product cost through the creation of mass markets, and the possibility to make new, standards-based services "explode" in terms of number of users. To end users, the standard will enable tools allowing them to easily "surf on the seas and filter the floods of multimedia information". To consumer and professional users alike, MPEG-7 will facilitate management of multimedia content. Of course, in order to be adopted, the standard needs to be technically sound. Matching the needs and the technologies in multimedia content description was thus the task of MPEG in the MPEG-7 standardization process.

1.2. *The objectives*

Like the other members of the MPEG family, MPEG-7 will be a standard representation of multimedia information satisfying a set of well-defined requirements,

which in this case, relate to the description of multimedia content. "Multimedia information" includes still pictures, video, speech, audio, graphics, 3D models and synthetic audio. The emphasis is on audio-visual content, and the standard will not specify new description tools for describing and annotating text itself, but will rather consider existing solutions for describing text documents, e.g., HTML, SGML and RDF,[3] supporting them as appropriate. While MPEG-7 includes statistical and signal processing tools, using textual descriptors to describe multimedia content is essential for information that cannot be derived by automatic analysis or human viewing the content. Examples include name of a place, date of acquisition, as well as more subjective annotations. Moreover, MPEG-7 will allow linking multimedia descriptions to any relevant data, notably the described content itself.

MPEG-7 is being designed as a generic standard in the sense that it will not be especially tuned to any specific application. MPEG-7 addresses content usage in storage, online and offline, or streamed, e.g., broadcast and (Internet) streaming. MPEG-7 supports applications operating in both real-time and nonreal-time environments. In this context, a "real-time environment" means that the description information is associated with the content while that content is being captured.

MPEG-7 descriptions will often be useful stand-alone, e.g., if only a summary of the multimedia information is needed. More often, however, they will be used to locate and retrieve the same multimedia content represented in a format suitable for reproducing the content: digital (and coded) or even analog. In fact, as mentioned above, MPEG-7 data is intended for content *identification* purposes, while other representation formats, such as MPEG-2 and MPEG-4, are mainly intended for content *reproduction* purposes. The boundaries may be not so sharp, but the different standards fulfill different requirements. MPEG-7 descriptions may be physically co-located with the corresponding "reproduction data", in the same data stream or in the same storage system. The descriptions may also live somewhere else on the globe. When the various multimedia representation formats are not co-located, mechanisms linking them are needed. These links should be able to work in both directions: from the "description data" to the "reproduction data" and vice versa.

Since MPEG-7 intends to describe multimedia content regardless of the way the content is available, it will neither depend on the reproduction format or on the form of storage. Video information could, for instance, be available as MPEG-4, -2, or -1, JPEG, or any other coded form — or not even be coded at all: it is entirely possible to generate an MPEG-7 description for an analogue movie or for a picture that is printed on paper. There is a special relationship between MPEG-7 and MPEG-4, however, as MPEG-7 is grounded on an object-based data model, which is also used by MPEG-4.[4] Like MPEG-4, MPEG-7 can describe the world as a composition of multimedia objects with spatial and temporal behavior, allowing object-based multimedia descriptions. As a matter of fact, each object in an MPEG-4 scene can have an MPEG-7 description (stream) associated with it; this description can be accessed independently.

1.3. *Normative versus non-normative*

A standard should seek to provide interoperability while trying to keep the constraints on the freedom of the user to a minimum. To MPEG, this means that a standard must offer the maximum of advantages by specifying the minimum necessary, thus allowing for competing implementations and for evolution of the technology in the so-called "non-normative" areas. MPEG-7 will only prescribe the multimedia description format (syntax and semantics) and usually not the extraction and encoding processes. Certainly, any part of the search process is outside the realm of a standard. Although good analysis and retrieval tools will be essential for a successful MPEG-7 application, their standardization is not required for interoperability.[5] In the same way, the specification of motion estimation and rate control is not essential for MPEG-1 and MPEG-2 applications, and the specification of segmentation is not essential for MPEG-4 applications.

Following the principle of "specifying the minimum for maximum usability", MPEG will concentrate on standardizing the tools to express the multimedia description. The development of multimedia analysis tools — automatic or semi-automatic — as well as of the tools that will use the MPEG-7 descriptions — search engines and filters — will be a task for the industries that will build and sell MPEG-7 enabled products. This strategy ensures that good use can be made of the continuous improvements in the relevant technical areas. New automatic analysis tools can always be used, also after the standard is finalized, and it is possible to rely on competition for obtaining ever better results. In fact, it will be these very non-normative tools that products will use to distinguish themselves, which only reinforces their importance.

1.4. *Low-level versus high-level*

The description of content may typically be done using two broadly defined types of features: low-level and high-level. The so-called low-level features are those like color and shape for images, and pitch and timbre for speech. High-level features typically have a semantic value associated to what the content means to humans, e.g., genre classification and rating. Low-level features have three important characteristics:

- can be extracted automatically: not the specialists but the machinery will worry about the great amount of information to describe,
- are objective: problems such as subjectivity and specialization are eliminated,
- are native to the audio-visual content: allows queries to be formulated in a way more adequate to the content in question, e.g., using colors, shapes and motion.

Although low-level features are easier to extract (they can typically be extracted fully automatically), most (nonprofessional) consumers would like to express their queries at the semantic level, where automatic extraction is rather difficult. One of MPEG-7's main strengths is that it provides a description framework that supports

the combination of low-level and high-level features in a single description. In combination with the highly structured nature of MPEG-7 descriptions, this capability constitutes one of the major differences between MPEG-7 and other available or emerging multimedia description solutions.

1.5. *Extensibility*

There is no single "right" description for a piece of multimedia content. What is right strongly depends on the application domain. MPEG-7 defines a rich set of core description tools. However, it is impossible to have MPEG-7 specifically addressing every single application. Therefore, it is essential that MPEG-7 be an open standard, extensible in a normative way to address description needs, and thus application domains, that are not fully addressed by the core description tools. The power to build new description tools (possibly based on the standard ones) is achieved through a standard description language, the Description Definition Language (DDL).

2. MPEG-7 Basic Description Elements

MPEG-7 specifies the following types of tools[3]:

- ***Descriptors* (*D*)** — *A Descriptor* (*D*) *is a representation of a Feature*; *a Feature is a distinctive characteristic of the data that signifies something to somebody. A Descriptor defines the syntax and the semantics of the Feature representation. A Descriptor allows an evaluation of the corresponding feature via the descriptor value. It is possible to have several descriptors representing a single feature, i.e., to address different relevant requirements/functionalities. Examples are*: *a timecode for representing duration, color moments and histograms for representing color, and a character string for representing a title.*
- ***Description Schemes* (*DS*)** — *A Description Scheme* (*DS*) *specifies the structure and semantics of the relationships between its components, which may be both Descriptors and Description Schemes. A DS provides a solution to model and describe multimedia content in terms of structure and semantics. A simple example is*: *a movie, temporally structured as scenes and shots, including some textual descriptors at the scene level, and color, motion and audio amplitude descriptors at the shot level.*
- ***Description Definition Language* (*DDL*)** — *The Description Definition Language is a language that allows the creation of new Description Schemes and, possibly* (*although not in MPEG-7 version* 1), *Descriptors. It also allows the extension and modification of existing Description Schemes.*
- ***Systems Tools*** — *Tools related to the binarization, synchronization, transport and storage of descriptions, as well as to the management and protection of intellectual property.*

These are the "normative elements" of the standard. "Normative" means that if these elements are implemented, they must be implemented according to the standardized specification since they are essential to guarantee interoperability. Feature extraction, similarity measures and search engines are also relevant, but will not be standardized.

For the sake of legibility and organization, the MPEG-7 standard is structured in seven parts[6]:

- ***Part 1 – Systems*** — *Specifies the tools that are needed to prepare MPEG-7 descriptions for efficient transport and storage, to allow synchronization between content and descriptions, and the tools related to managing and protecting intellectual property*[7];
- ***Part 2 – Description Definition Language*** — *Specifies the language for defining new description schemes and perhaps also new descriptors*[8];
- ***Part 3 – Visual*** — *Specifies the descriptors and description schemes dealing only with visual information*[9];
- ***Part 4 – Audio*** — *Specifies the descriptors and description schemes dealing only with audio information*[10];
- ***Part 5 – Generic Entities and Multimedia Description Schemes*** — *Specifies the descriptors and description schemes dealing with generic (non-audio or video specific) and multimedia features*[11];
- ***Part 6 – Reference Software*** — *Includes software corresponding to the specified MPEG-7 tools*[12];
- ***Part 7 – Conformance Testing*** — *Defines guidelines and procedures for testing conformance of MPEG-7 descriptions and terminals.*[5]

Parts 1 to 5 specify the core MPEG-7 technology, while Parts 6 and 7 are "supporting parts". While the various MPEG-7 parts are rather independent and thus can be used by themselves, or in combination with proprietary technologies, they were developed in order that the maximum benefit results when they are used together.

3. MPEG Standardization Process

Two foundations of the success of the MPEG standards so far are the toolkit approach and the "one functionality, one tool" principle.[13] The toolkit approach means setting a horizontal standard that can be integrated with, for example, different kinds of transmission solutions. MPEG does not set vertical standards across many layers in the ISO stack. The "one functionality, one tool" principle implies that no two tools will be included in the standard if they provide essentially the same functionality. To apply this approach, the standard development process is organized as follows:

Table 1. MPEG-7 workplan.

October 16, 1998	Call for proposals Final version of the MPEG-7 Proposal Package Description (PPD)
December 1, 1988	Pre-registration of proposals
February 1, 1999	Proposals due
February 15—19, 1999	Evaluation of proposals (in an Ad Hoc Group meeting hold in Lancaster, UK)
March 1999	First version of the MPEG-7 eXperimentation Model (XM)
December 1999	Working Draft stage (WD)
October 2000	Committee Draft stage (CD)
March 2001	Final Committee Draft stage (FCD) after ballot with comments
July 2001	Final Draft International Standard stage (FDIS) after ballot with comments (after this step, the text of the standard is set in stone)
September 2001	International Standard (IS) after yes/no ballot

(i) identification of the relevant applications and extraction of relevant requirements;
(ii) open Call for Proposals on the basis of these requirements;
(iii) evaluation of Proposals against the requirements;
(iv) collaborative specification of tools to fulfill the requirements;
(v) verification that the developed tools fulfill the identified requirements.

Because MPEG always operates in new fields, the requirements landscape will keep moving and the above process is not applied rigidly. Some steps may be taken more than once and iterations are sometimes needed. The time schedule, however, is always closely observed by MPEG. Although all decisions are taken by consensus, the process keeps a high pace, allowing MPEG to timely provide technical solutions. For MPEG-7, this process translates to the workplan presented in Table 1.

After an initial period dedicated to the specification of objectives and the identification of applications and requirements, MPEG-7 issued, in October 1998, a Call for Proposals[14] to gather the best available technology fitting the MPEG-7 requirements. Six hundred and sixty-five proposal pre-registrations were received by the December 1, 1998 deadline.[15] Of these, 390 (59%) were actually submitted as proposals by the February 1st, 1999 deadline. Out of these 390 proposals, there were 231 descriptors and 116 description schemes. The proposals for normative elements were evaluated by MPEG experts, in February 1999, in Lancaster (UK), following the procedures defined in the MPEG-7 Evaluation documents.[16,17] A special set of audio-visual content was provided to the proposers for usage in the evaluation process; this content has also being used in the collaborative phase. The content set consists of 32 Compact Discs with sound tracks, pictures and moving video.[18] It has been made available to MPEG under the licensing conditions defined in Ref. 19. Broadly, these licensing terms permit usage of the content exclusively for MPEG-7 standard development purposes. While fairly straightforward methodologies were

used for the evaluation of the audio-visual description tools in the MPEG-7 competitive phase, more powerful methodologies were developed during the collaborative phase in the context of the tens of so-called "core experiments". A core experiment is a well-defined experiment carried out by two or more independent parties. In the collaborative development phase, technical choices for the standard are made on the basis of such core experiments. After the evaluation of the technology received, choices and recommendations were made and the collaborative phase started with the most promising tools.[20]

In the course of developing the standard, additional calls may be issued when not enough technology is available within MPEG to meet the requirements, but there must be indications that the technology does indeed exist. The "collaboration after competition" approach concentrates the efforts of many research teams throughout the world on further improving the technology that was already demonstrated to be top-ranking.

4. Applications and Requirements

MPEG-7 targets a wide range of application environments and it will offer different levels of granularity in its descriptions, along axes such as time, space and accuracy. Descriptive features must be meaningful in the context of an application, so the descriptions for the same content can differ according to the user domain and application. This implies that the same material can be described in various ways, using different features, and with different levels of abstraction for those features. It is thus the task of the content description generator to choose the right features and corresponding granularity. It may be clear then that no single "right" description exists for any piece of content; all descriptions may be equally valid from their own usage point of view. The strength of MPEG-7 is that these descriptions will all be based on the same description tools, and can be exchanged in a meaningful way.

MPEG-7 requirements are application driven. The relevant applications are all those that should be enabled by the MPEG-7 toolbox. Addressing new applications, i.e., those that do not exist yet but will be enabled by the standard, has the same priority as improving the functionality of existing ones. There are many application domains that should benefit from the MPEG-7 standard, and no application list drawn up today can be exhaustive.

The MPEG-7 Applications document includes examples of both improved existing applications as well as new ones that may benefit from the MPEG-7 standard, and organizes the example applications into three sets, as follows[21]:

- **Pull applications** — Applications such as storage and retrieval in audio-visual databases, delivery of pictures and video for professional media production, commercial musical applications, sound effects libraries, historical speech database, movie scene retrieval by memorable auditory events, and registration and retrieval of trademarks.

- **Push applications** — Applications such as user agent driven media selection and filtering, personalized television services, intelligent multimedia presentations, and information access facilities for people with special needs.
- **Specialized professional applications** — Applications that are particularly related to a specific professional environment, notably tele-shopping, bio-medical, remote sensing, educational and surveillance applications.

For each listed application, the MPEG-7 Applications document gives a description of the application, the corresponding requirements, and a list of relevant work and references. The set of applications in the MPEG-7 Applications document[21] is a living set, which will be augmented in the future, intended to give the industry — clients of the MPEG work — some hints about the application domains addressed. If MPEG-7 will enable new and "unforeseen" applications to emerge, this will show the strength of the toolkit approach.

Although MPEG-7 intends to address as many as possible application domains, it is clear that some applications are more important than others due to their relevance in terms of foreseen business, research investment, number of MPEG experts, etc. This is clearly the case for storage and retrieval in audio-visual databases, which since the very beginning has been the leading application, sometimes even wrongly understood as the only MPEG-7 application.

In order to develop useful tools for the MPEG-7 toolkit, functionality requirements have been extracted from the identified applications. The MPEG-7 requirements[3] are currently divided into five sections: descriptors, description schemes, Description Definition Language, descriptions and systems requirements. Whenever applicable, visual and audio requirements are considered separately. The requirements apply, in principle, to both real-time and nonreal-time systems as well as to offline and streaming applications, and they should be meaningful to as many applications as possible.

5. MPEG-7 Description Tools

Since March 1999, MPEG has developed a set of multimedia description tools addressing the identified MPEG-7 requirements. These tools are specified in the various parts of the MPEG-7 standard.

5.1. *Systems and DDL tools*

The MPEG Systems Group is in charge of developing a set of "systems tools" and the Description Definition Language (Parts 1 and 2 of the standard). The systems tools allow preparing the MPEG-7 descriptions for efficient transport and storage (binarization), to synchronize content and the corresponding descriptions, and to manage and protect intellectual property. The DDL is the language for defining new Description Schemes and extending existing ones.

MPEG-7 descriptions may be delivered independently or together with the content they describe. The MPEG-7 architecture allows conveying data back from the terminal to the transmitter or server, such as queries. The Systems layer encompasses mechanisms allowing synchronization, framing and multiplexing of MPEG-7 descriptions and may also be capable of providing the multimedia content data if requested. The delivery of MPEG-7 content on particular systems is outside the scope of the Systems specification.[6] Existing delivery tools, such as TCP/IP, MPEG-2 Transport Stream (TS) or even a CD-ROM may be used for this purpose.

MPEG-7 elementary streams consist in consecutive individually accessible portions of data named *Access Units*; an access unit is the smallest data entity to which timing information can be attributed. MPEG-7 elementary streams contain information of different nature: (i) description schema information defining the structure of the MPEG-7 description, and (ii) description information that is either the complete description of the multimedia content or fragments of the description.

MPEG-7 data can be represented either in textual format, in binary format or a mixture of the two formats, depending on the application.[6] MPEG-7 defines a unique bi-directional mapping between the binary format and the textual format. This mapping can be lossless either way.

The syntax of the textual format is defined in Part 2: DDL — Description Definition Language[8] of the standard. The syntax of the binary format — BiM (Binary format for MPEG-7 data) — is defined in Part 1: Systems[7] of the standard. Description schemes are defined in Parts 3, 4 and 5 (Visual, Audio and Multimedia Description Schemes) of the standard.[9–11]

The Description Definition Language (the textual format) is based on W3C's XML (eXtensible Markup Language) Schema Language[22]; however, some extensions to XML Schema are being developed in order that all the DDL requirements[3] are fulfilled by the MPEG-7 DDL.[8] In this context, the DDL can be broken down into the following logical normative components[6]:

- **XML Schema structural language components** — These components correspond to Part 1 of the XML Schema specification[22] and provide facilities for describing and constraining the content of XML 1.0 documents.
- **XML Schema datatype language components** — These components correspond to Part 2 of the XML Schema specification[22] and provide facilities for defining datatypes to be used to constrain the datatypes of elements and attributes within XML Schemas.
- **MPEG-7 specific extensions** — These extensions correspond to the features added to the XML Schema Language to fulfill MPEG-7 specific requirements, notably new datatypes.[8]

MPEG-7 DDL specific parsers will be developed by adding validation of the MPEG-7 additional constructs to standard XML Schema parsers. In fact, while a DDL parser is able to parse a regular XML Schema file, a regular XML Schema

parser may parse an MPEG-7 textual description although with a reduced level of validation due to the MPEG-7 specific datatypes that cannot be recognized.

The BiM format[7] is a DDL compression tool, which can also be seen as a general XML compression tool since it also allows to efficiently binary encode XML files. There are two major reasons for having a binary format for MPEG-7 data. First, in general, the transmission or storage of the textual format requires a much higher bandwidth than necessary from a theoretical point of view (compression gains of 98% may be obtained for certain cases); an efficient compression of the textual format is applied when converting it to the binary format. Second, the textual format is not very appropriate for streaming applications since it only allows the transmission of a description tree in the so-called depth-first tree order. However, for streaming applications, more flexibility is required with respect to the transmission order of the elements: the BiM provides this flexibility. The BiM allows randomly searching or accessing elements of a binary MPEG-7 description directly on the bitstream, without parsing the complete bitstream before these elements. At the description-consuming terminal, the binary description can either be converted to the textual format or directly parsed.

5.2. *Visual tools*

The MPEG Video Group is responsible for the development of the MPEG-7 visual description tools (Part 4 of the standard).[9] MPEG-7 visual description tools include basic structures and descriptors or description schemes enabling the description of some visual features of the visual material, such as color, texture, shape and motion, as well as the localization of the described objects in the image or video sequence. These tools are defined by their syntax in DDL and binary representations and semantics associated with the syntactic elements.[9] For each tool, there are normative and non-normative parts: the normative parts specify the textual and binary syntax and semantics of the structures, while the non-normative ones propose relevant associated methods such as extraction and matching.

5.2.1. *Basic structures: elements and containers*

The MPEG-7 visual basic elements are[9]:

- **Spatial 2D coordinates** — This descriptor defines a 2D spatial coordinate system to be used by reference in other Ds/DSs, when relevant. It supports two kinds of coordinate systems: local and integrated. In a local coordinate system, the coordinates used for the creation of the description are mapped to the current coordinate system applicable; in an integrated coordinate system, each image (frame) of e.g., a video sequence may be mapped to different areas with respect to the first frame of a shot or video.
- **Temporal interpolation** — The Temporal_Interpolation descriptor characterizes temporal interpolation using connected polynomials, to approximate

multi-dimensional variable values that change with time, such as object position in a video sequence. The descriptor size is usually much smaller than describing all position values.

The MPEG-7 visual containers — structures allowing the combination of visual descriptors according to some spatial/temporal organization — are[9]:

- **Grid layout** — The grid layout is a splitting of the image into a set of rectangular regions, so that each region can be described separately. Each region of the grid can be described in terms of descriptors such as color or texture.
- **Time series** — The TimeSeries structure describes a temporal series of descriptors in a video segment and provides image to video frame and video frame to video frame matching functionalities. Two types of TimeSeries are defined: RegularTimeSeries and IrregularTimeSeries; in the RegularTimeSeries, descriptors are located regularly (with constant intervals) within a given time span; alternatively, in the IrregularTimeSeries, descriptors are located irregularly along time.
- **Multiple view** — The MultipleView descriptor specifies a structure combining 2D descriptors representing a visual feature of a 3D object seen from different view angles. The descriptor forms a complete 3D view-based representation of the object, using any 2D visual descriptor, such as shape, color or texture.

5.2.2. *Descriptors and description schemes*

The MPEG-7 visual descriptors cover five basic visual features: color, texture, shape, motion and localization. These descriptors and description schemes are essential tools to describe and retrieve images and videos stored in databases.

(a) **Color**

There are seven MPEG-7 **color** descriptors[9]:

- **Color space** — Defines the color space used in MPEG-7 color based descriptions; the following color spaces are supported: RGB, YCbCr, HSV, HMMD, linear transformation matrix with reference to RGB, monochrome.
- **Color quantization** — Defines the uniform quantization of a color space.
- **Dominant color** — Specifies a set of dominant colors in an arbitrarily shaped region.
- **Scalable color** — Defines a color histogram in the HSV color space, encoded by a Haar transform; its binary representation is scalable in terms of bin numbers and bit representation accuracy over a broad range of data rates.
- **Color layout** — Specifies the spatial distribution of colors for high-speed retrieval and browsing; it can be applied either to a whole image or to any part of an image, including arbitrarily shaped regions.

- **Color structure** — Captures both color content (similar to that of a color histogram) and the structure of this content via the use of a structuring element composed of several image samples.
- **GoF/GoP color** — Defines a structure required for representing the color features of a collection of (similar) images or video frames by means of the scalable color descriptor defined above. The collection of video frames can be a contiguous video segment or a noncontiguous collection of similar video frames.

(b) **Texture**

There are three MPEG-7 **texture** descriptors[9]; texture represents the amount of structure in an image such as directionality, coarseness, regularity of patterns, etc.:

- **Homogeneous texture** — Represents the energy and energy deviation values extracted from a frequency layout.
- **Texture browsing** — Relates to a perceptual characterization of the texture, similar to a human characterization, in terms of regularity (irregular, slightly regular, regular, highly regular) (Fig. 1), directionality (0°, 30°, 60°, 90°, 120°, 150°) and coarseness (fine, medium, coarse, very coarse), referred as Perceptual Browsing Component (PBC).
- **Edge histogram** — Represents the spatial distribution of five types of edges in local image regions as shown in Fig. 2 (four directional edges and one nondirectional edge in each local region called sub-image).

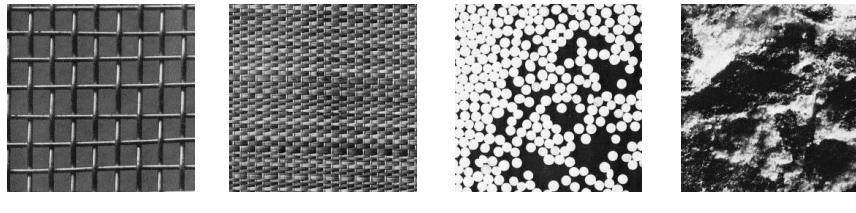
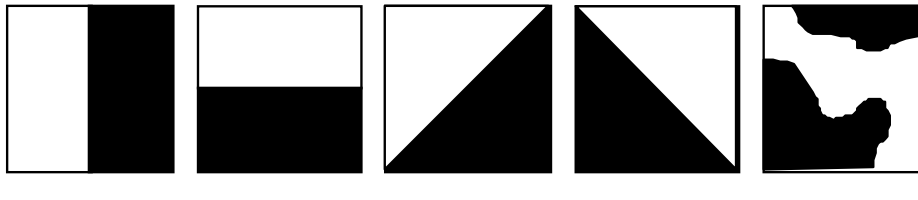


Fig. 1. Examples of texture regularity from highly regular to irregular.[9]



(a) vertical edge (b) horizontal edge (c) 45° edge (d) 135° edge (e) nondirectional edge

Fig. 2. The five type of MPEG-7 edges used in the edge histogram descriptor.[9]

(c) **Shape**

There are three MPEG-7 **shape** descriptors[9]:

- **Region-based shape: Angular Radial Transform** — Makes use of all pixels constituting the shape and can describe any shape, i.e., not only a simple shape with a single connected region but also a complex shape consisting of several disjoint regions. The region-based shape descriptor is a set of ART (Angular Radial Transform) coefficients. The ART is a 2D complex transform defined on a unit disk in polar coordinates,

$$F_{nm} = \langle V_{nm}(\rho,\theta), f(\rho,\theta)\rangle = \int_0^{2\pi}\int_0^1 V^*_{nm}(\rho,\theta), f(\rho,\theta)\rho d\rho d\theta\,,$$

  where F_{nm} is an ART coefficient of order n and m, $f(\rho,\theta)$ is an image function in polar coordinates, and $V_{nm}(\rho,\theta)$ is the ART basis function (Fig. 3).

- **Contour-based shape (Curvature Scale-Space representation)** — Describes a closed contour of a 2D object or region in an image or video sequence. This descriptor is based on the Curvature Scale Space (CSS)[9] representation of the contour, typically making very compact descriptions (below 14 bytes in size, on average) (Fig. 4).

- **Shape 3D** — Provides an intrinsic shape description of 3D mesh models, based on the histogram of 3D shape indexes representing local curvature properties of the 3D surface.

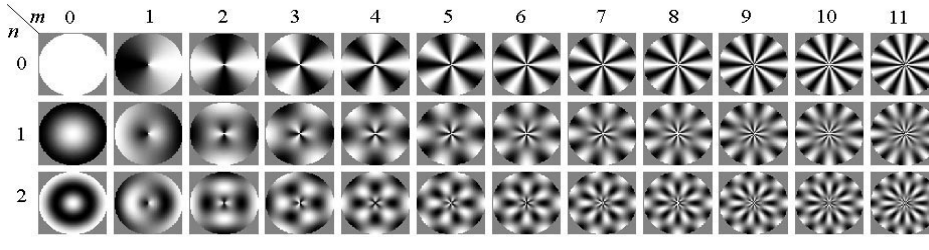


Fig. 3. Real parts of the ART basis functions.[9]

Fig. 4. Example shapes: The ART descriptor can describe all these shapes but the CSS descriptor can only describe the shapes with one single connected region.[9]

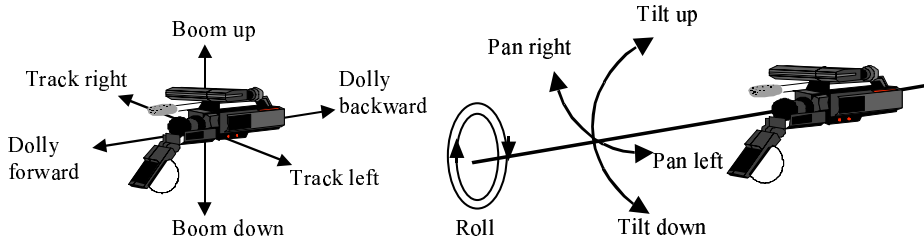


Fig. 5. Types of MPEG-7 camera motion.[9]

(d) **Motion**

There are four MPEG-7 **motion** descriptors[9]:

- **Camera motion** — Characterizes 3D camera motion based on 3D camera motion parameter information, which can be automatically extracted or generated by capture devices. This descriptor supports the following basic camera operations (see Fig. 5): fixed, panning (horizontal rotation), tracking (horizontal transverse movement, also called traveling in the film industry), tilting (vertical rotation), booming (vertical transverse movement), zooming (change of the focal length), dollying (translation along the optical axis) and rolling (rotation around the optical axis).
- **Motion trajectory** — Defines a spatio-temporal localization of a representative point of a moving region, for example the centroid.
- **Parametric motion** — Characterizes the evolution of arbitrarily shaped regions over time in terms of a 2D geometric transform (translational, rotation/scaling, affine, perspective, quadratic). This descriptor addresses the motion of objects in video sequences, as well as global motion; if it is associated with a region, it can be used to specify the relationship between two or more feature point motion trajectories, according to the underlying motion model.
- **Motion activity** — Captures the intuitive notion of "intensity of action" or "pace of action" in a video segment. The activity descriptor includes the following five attributes: intensity of activity, direction of activity, spatial distribution of activity, spatial localization of activity and temporal distribution of activity.

(e) **Localization**

There are two MPEG-7 **localization** descriptors[9]:

- **Region locator** — 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** — Describes spatio-temporal regions in a video sequence by one or several sets of reference regions and motion using two

description schemes: FigureTrajectory and ParameterTrajectory. These two description schemes are selected according to moving object conditions: if a moving object region is rigid and the motion model is known, ParameterTrajectory is appropriate; if a moving region is nonrigid, FigureTrajectory is appropriate.

Finally, there is also a face recognition descriptor[9] based on eigenfaces, which represents the projection of a face vector onto a set of 49 basis vectors that span the space of possible face vectors.

The set of MPEG-7 visual description tools presented above provides a very powerful solution in terms of describing visual content for a large range of applications, and specially for image and video database storage and retrieval.

5.3. *Audio Tools*

The MPEG Audio Group had the task to specify the MPEG-7 audio description tools (Part 4 of the standard). MPEG-7 audio description tools are organized in the following areas[10]:

- **Scalable series** — Efficient representation for series of feature values; this is a core part of MPEG-7 audio.
- **(Low-level) Audio framework** — Collection of low-level audio descriptors, many built upon the scalable series, e.g., waveform, spectral envelope, loudness, spectral centroid, spectral spread, fundamental frequency, harmonicity, attack time, spectral basis.
- **Silence** — Descriptor identifying silence.
- **Spoken content** — Set of description schemes representing the output of Automatic Speech Recognition (ASR).
- **Timbre description** — Collection of descriptors and description schemes describing the perceptual features of instrument sounds.
- **Sound effects** — Collection of descriptors and description schemes defining a general mechanism suitable for handling sound effects.
- **MelodyContour** — Description scheme allowing retrieval of musical data.
- **Melody** — More general description framework for melody.

5.4. *MDS Tools*

The MPEG Multimedia Description Schemes (MDS) Group had the task of specifying a set of description tools, dealing with generic as well as multimedia entities (Part 5 of the standard).[11] Generic entities are those that can be used in audio, visual, and textual descriptions, and therefore are "generic" to all media, e.g., vector, histogram, time, etc. Multimedia entities are those that deal with more than

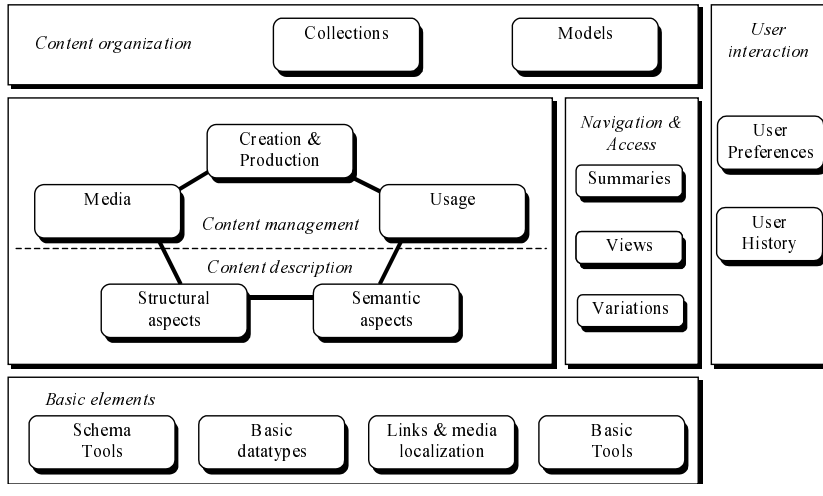


Fig. 6. Overview of the MDS description tools.[11]

one medium, e.g., audio and video. MDS description tools (Fig. 6) can be grouped into six different classes according to their functionality[11]:

- **Content description: Structural and Conceptual Aspects** — Representation of perceivable information in terms of structural and conceptual aspects; the visual and audio descriptors and descriptions schemes presented above allow the structural description of the content and thus these tools fit in the Structure DS.[11] The structural and semantic trees may be linked together leading to a graph, which provides a powerful description of the content since it combines a more low-level description with a more abstract description.
- **Content management: Media, Usage, Creation and Production** — Information related to the management of the content, notably information that cannot be automatically extracted from the content, e.g., rights holders (content usage), author's name (creation and production).
- **Content organization: Collection and Classification, and Model** — Organization and classification of content as well as analytic modeling.
- **Navigation and access: Summary and Variation** — Summaries, which enable fast visualization and sonification as well as to generate variations of the content.
- **User: User Preferences** — Limited set of descriptors for user preferences pertaining to multimedia content.
- **Basic Elements: Datatype and Structures, Schema Tools, Link and Media Localization, and Basic DSs** — Tools such as basic datatypes, mathematical structures, schema tools (root element, top-level element and packages), etc.,

which are found as elementary components of more complex DSs. The root element is a wrapper of the description that may be a complete description (in the sense of an application) or a partial/incremental description.

Since the overall value of the MPEG-7 standard will strongly depend not only on the quality of each set of tools but also on the quality of their interaction and complementary capabilities, the standard was developed in an intense collaborative framework, where synergies could be fully exploited. The best example of this practice is the MPEG-7 eXperimentation Model (XM), which integrates the systems, DDL, audio, visual and MDS tools into one comprehensive experimental Framework. This will ultimately become Part 6 of MPEG-7: Reference Software.[12]

6. Conclusion

MPEG-1 and MPEG-2 have been successful standards, and it is expected that MPEG-4 will also set new frontiers in terms of multimedia representation. Following these projects, MPEG has moved on to address the problem of identifying, selecting and managing various types of multimedia material. MPEG-7 offers a standardized way of describing multimedia. In comparison with other available or emerging audio-visual description frameworks, MPEG-7 can be characterised by its:

(i) Genericity: the capability to describe multimedia content from many application environments;
(ii) Object-based data model: the capability to independently describe individual objects, e.g., scenes, takes, objects within a scene — be they MPEG-4 or not;
(iii) Integration of low-level and high-level features/descriptors into a single architecture, allowing to combine the power of both types of descriptors;
(iv) Extensibility, provided by the Description Definition Language, which allows MPEG-7 to keep growing, to be extended to new application areas, to answer to newly emerging needs and to integrate novel description tools.

It is expected that the emerging MPEG-7 standard will constitute another great step in the area of multimedia content representation.

Acknowledgments

The authors would like to thank all the MPEG-7 members for the interesting and fruitful discussions in meetings and by e-mail, which substantially enriched their technical knowledge.

References

1. MPEG Home Page, http://www.cselt.it/mpeg/.
2. MPEG Requirements Group, "Introduction to MPEG-7," Doc. ISO/MPEG N4032, Singapore MPEG Meeting, March 2001.

3. MPEG Requirements Group, "MPEG-7 requirements," Doc. ISO/MPEG N4035, Singapore MPEG Meeting, March 2001.
4. F. Pereira, "MPEG-4: why, what, how and when?" *Tutorial Issue on the MPEG-4 Standard*, *Signal Processing*: *Image Communication* **15**(4–5) (1999).
5. MPEG Requirements Group, "MPEG-7 interoperability, conformance testing and profiling," Doc. ISO/MPEG N4039, Singapore MPEG Meeting, March 2001.
6. MPEG Requirements Group, "MPEG-7 overview," Doc. ISO/MPEG N4031, Singapore MPEG Meeting, March 2001.
7. MPEG Systems Group, "MPEG-7 systems final committee draft," Doc. ISO/MPEG N4001, Singapore MPEG Meeting, March 2001.
8. MPEG Systems Group, "MPEG-7 DDL final committee draft," Doc. ISO/MPEG N4002, Singapore MPEG Meeting, March 2001.
9. MPEG Video Group, "MPEG-7 visual final committee draft," Doc. ISO/MPEG N4062, Singapore MPEG Meeting, March 2001.
10. MPEG Audio Group, "MPEG-7 audio final committee draft," Doc. ISO/MPEG N4004, Singapore MPEG Meeting, March 2001.
11. MPEG Multimedia Description Schemes Group, "MPEG-7 multimedia description schemes final committee draf," Doc. ISO/MPEG N3966, Singapore MPEG Meeting, March 2001.
12. MPEG Implementation Studies Group, "MPEG-7 reference software final committee draft," Doc. ISO/MPEG N4006, Singapore MPEG Meeting, March 2001.
13. L. Chiariglione, "The challenge of multimedia standardization," *IEEE Multimedia* **4**(2), 1997.
14. MPEG Requirements Group, "MPEG-7 call for proposals," Doc. ISO/MPEG N2469, Atlantic City MPEG Meeting, October 1998.
15. MPEG Requirements Group, "MPEG-7 list of proposal pre-registrations," Doc. ISO/MPEG N2567, Rome MPEG Meeting, December 1998.
16. MPEG Requirements Group, "MPEG-7 evaluation process," Doc. ISO/MPEG N2463, Atlantic City MPEG Meeting, October 1998.
17. MPEG Requirements Group, "MPEG-7 Proposal Package Description (PPD)," Doc. ISO/MPEG N2464, Atlantic City MPEG Meeting, October 1998.
18. MPEG Requirements Group, "Description of MPEG-7 content set," Doc. ISO/MPEG N2467, Atlantic City MPEG Meeting, October 1998.
19. MPEG Requirements Group, "Licensing agreement for MPEG-7 content set," Doc. ISO/MPEG N2466, Atlantic City MPEG Meeting, October 1998.
20. MPEG Requirements Group, "Results of MPEG-7 technology proposal evaluations and recommendations," Doc. ISO/MPEG N2730, Seoul MPEG Meeting, March 1999.
21. MPEG Requirements Group, "MPEG-7 applications," Doc. ISO/MPEG N3934, Pisa MPEG Meeting, January 2001.
22. W3C, "XML Schema (Primer, Structures, Datatypes)," W3C Working Draft, April 2000.

Fernando Pereira was born in Vermelha, Portugal in October 1962. He was graduated in Electrical and Computers Engineering by Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Portugal, in 1985. He received his M.Sc. and Ph.D. in Electrical and Computers Engineering from IST, in 1988 and 1991, respectively. He is currently an Associate Professor at the Electrical and Computers Engineering Department of IST. He is responsible for the participation of IST in many national and international research projects. He is an Area Editor of *Signal Processing*: *Image Communication Journal* and an Associate Editor of *IEEE Transactions of Circuits and Systems for Video Technology*. He is a Member of the Scientific and Program Committees of several international conferences and workshops. He is the author of more than one hundred journal and conference papers. He has participated in the work of ISO/MPEG for many years, acting as the Head of the Portuguese delegation, and chairing many Ad Hoc Groups related to the MPEG-4 and MPEG-7 standards. His current areas of interest are video analysis, processing, coding and description, and the development of interactive multimedia services, notably for the Internet and mobile environments.

Rob Koenen received his "Ingenieur" (MSEE) degree from Delft University of Technology, the Netherlands, in 1989. He studied Electrical Engineering, specializing in Information Theory. In 1990, he joined KPN Research, where he has researched various aspects of audio-visual communication, working as a research group and Programme Manager. His projects have addressed: image coding research, audio-visual communication for people with special needs, interactive broadband multimedia for residential users, mobile multimedia, the strategic deployment of new multimedia services, audio-visual quality assessment and multimedia standardization. As an MPEG delegate, he has played a key role in the development of the MPEG-4 standard since 1993, and in defining the upcoming MPEG-7 standard since 1995. Ir. Koenen now works as a Director of Product Development for InterTrust Technologies Corporation in Santa Clara, CA, USA. He chairs the MPEG Requirements subgroup and he is the Initiator and President of the MPEG-4 Industry Forum (M4IF). He is also an Associate Editor of *IEEE Transactions on Circuits and Systems for Video Technology*.