*Journal of Universal Computer Science, vol. 7, no. 6 (2001), 530-547*
*submitted: 1/5/01, accepted: 1/6/01, appeared: 28/6/01 © Springer Pub. Co.*

# MPEG and its Relevance for Content-based Multimedia Retrieval

**Werner Haas**
(Institute of Information Systems & Information Management
JOANNEUM RESEARCH
Graz, Austria
Werner.Haas@joanneum.at)

**Harald Mayer**
(Institute of Information Systems & Information Management
JOANNEUM RESEARCH
Graz, Austria
Harald.Mayer@joanneum.at)

**Abstract:** The utilization of new emerging standards such as MPEG-7 is expected to be a major breakthrough for content-based multimedia data retrieval. The main features of the MPEG standards series and of related standards, formats and protocols are presented. It is discussed, how they, despite their partially early and immature stage, can best be utilized to yield effective results in the context of a knowledge management environment. Complementary to that, the current status and state of the art in content-based retrieval for images, video and audio content is briefly presented. In the context of the KNOW-Center we are developing a prototype platform to implement a user friendly and highly informative access to audiovisual content as a potential component for a future knowledge management system. The technical requirements and the system architecture for the prototype platform are described.

**Key Words:** content-based search and retrieval, MPEG, knowledge management, databases

**Category:** H.3.1, H.3, K.1

## 1 Introduction

The technology developments of recent years, most prominent among them the success of the Internet together with the transition to digital technologies for broadcasting has resulted in an enormous increase in digital audiovisual content available for private and commercial use. The “lost in Hyperspace-syndrome” has thus been drastically extended into the multimedia area.

This situation has challenged research communities and industry to answer with appropriate solutions for quickly searching, filtering and retrieving relevant multimedia material. Examples are search requests that are formulated by spoken queries, hand-drawn sketches, similar images or text based formulations on a high semantic level in order to collect material for a new program from a TV archive.

Relevant research has concentrated onto two areas: *content-based retrieval* of audiovisual data and new emerging related *standards*. Among the standardisation efforts, MPEG-7 addresses the description of multimedia content on a metadata level, that gives major focus to the semantic information level, defining what most users would desire to retrieve. Whereas MPEG-7 does not specify any methods, however, research in multimedia indexing and retrieval has become a popular and successful area, that has produced a high number of partially very powerful methodologies and algorithms. Generally applicable results are only rarely available, however. An overview on standards and content-based state of the art will be given in this paper, while later on a prototype with selected functionality will be described.

## 2 The Family of MPEG Standards

### 2.1 MPEG overview

The Moving Picture Experts Group (MPEG) is a working group of ISO/IEC (International Standards Organisation/International Electrotechnical Committee) in charge of the development of standards for coded representation of digital audio and video. Since 1988 the group has produced a series of standards, which were initially focussed onto bit-efficient representation of audio-visual content, i.e. compression, decompression, processing and coded representation of moving pictures, audio and combinations of the two.

Besides standards strictly related to bit-efficient representation of audio-visual content, MPEG has soon started producing other standards that relate to describing content and to the practical use of those standards. An example is given by Intellectual Property Management and Protection (IPMP) [see IPMP].

*MPEG-1* is the standard for the storage and retrieval of moving pictures and audio on storage media. On it such products as Video CD and MP3 are based.

*MPEG-2* is the standard for digital television. It supports the transition from analogue to digital format for satellite broadcasting and cable television. Products as Digital Television, set top boxes and DVD are based on this standard.

*MPEG-4* enables to code content as objects. Those objects can be manipulated individually or collectively on an audiovisual scene. It is supposed to be the standard for multimedia for interactive TV, the Web and mobility.

*MPEG-7* is formally named “Multimedia Content Description Interface”. It is the standard that describes multimedia content such that users can search, browse and retrieve content more effectively and efficiently than today’s search engines. Whereas MPEG-1, MPEG-2 and MPEG-4 are already accepted standards, completion of MPEG-7 is officially scheduled for September 2001. This may however be further postponed.

*MPEG-21* as the recent activity of the group aims at defining a “Multimedia Framework”. Work has started in June 2000 and has already produced some reports. A first Working Draft for 'Digital Item Identification and Description' (DIID) was issued. This part of the MPEG-21 standard will uniquely identify multimedia content and elements within that content according to international standards for identifiers (ISAN, the International Standard Audiovisual Number).

## 2.2 The role of MPEG-4, MPEG-7 and XML

For the goal of efficient content-based search and retrieval – keeping in mind the final goal of integrating this into a knowledge management system – description of content is as important as the ability to package annotations and content together and to transport it. This is where MPEG-4, -7 and in the future probably also MPEG-21 will have their role in knowledge management systems.

MPEG-4 [see MPEG-4] provides technologies to satisfy the needs of authors, service providers and end-users. It does so by standardizing *coding* (representation of media objects, that may be generated by conventional means like cameras, microphones or synthetically by computer), *composition* (creation of compound media objects that appear as audiovisual scenes), *multiplexing* (for transport over networks, taking into account necessary QoS for each part of the media) and *interaction* (providing interactivity between receiver and transmitter).

One important – and for video annotation very interesting – feature of MPEG-4 is the concept of “video object” and “video object plane”. This allows separate handling and annotation of those objects. Another feature is, that media objects may have 2D and 3D dimensionality and as also audio may have spatial distribution.

MPEG-7 (Multimedia Content Description Interface) [see MPEG-7] provides a standardised content description for various types of audio/visual material (audio, speech, video, pictures…). The objective is to quickly and efficiently search and retrieve audiovisual material. To allow interoperability, the standard adopts normative elements, such as Descriptors (D’s), Description Schemes (DS’s), the Description Definition Language (DDL) [see ISO/IEC JTC1/SC29/WG11 N3702] as well as Coding and System Tools. The Descriptors define the syntax and the semantics of the representation of features, while the Description Schemes specify the structure and semantics of the relationships between Descriptors or other Descriptions. Many descriptors have been submitted for MPEG-7 [see ISO/IEC JTC 1/SC 29/WG 11 N3705], some of which either accepted and included in the eXperimental Model (XM), which is a platform and tool set to evaluate and improve the tools of MPEG-7 [see ISO/IEC JTC 1/SC 29/WG 11 N3815], or are in the experimentation (Core Experiments, CE) phase. MPEG-7 has adopted XML Schema as its DDL [see Nack and Lindsay 1999a] and [Nack and Lindsay 1999b].

Two parallel levels of descriptors are defined: the *syntactic* one, which describes the perceptual properties of the content, such as colour, texture, shape, layout and motion, in the visual data case, or pitch and energy level in the audio data case and the *semantic* one, which describes the meaning of content, in terms of semantic objects and events.

As an example, among the over 100 MPEG-7 descriptions currently being developed [see Day 2000] the MovieRegion Description Scheme allows to see content from predefined viewpoints: creation and production, usage, media, structural aspects and conceptual aspects. The standard is not restricted to the views mentioned here but can be also used to describe other aspects (e.g. user preferences…).

The fact has to be stressed, that MPEG-7 does neither deal with the description generation (e.g. automatic extraction, indexing) nor with the description consumption (e.g. search, retrieval, …). This is left completely to the creativity of researchers and to appropriate applications.

# 3 Related Standards, Initiatives and Projects

There is a number of activities in progress, which to some extent overlap with standards defined within the MPEG series. Those which seem to be of importance for the issues covered in this paper are briefly discussed and their main relationship to the respective member of the MPEG family is explained.

## 3.1 General Multimedia Standards

### 3.1.1 X3D – eXtensible 3D

This is an effort taken by the former VRML-, now Web3D-consortium. X3D overlaps to some extent with MPEG-4, that has borrowed much of its 3D representation from VRML. One main difference lies in the fact, that in VRML the browsers assume, that all audiovisual content is downloaded first to the client and then played, while MPEG-4 has the concept of embedding scenes and media into the stream.

### 3.1.2 SMIL – Synchronized Multimedia Integration Language

With SMIL, the W3C (W3 consortium) has specified a format for integrating independent multimedia objects into a synchronized multimedia presentation. Using SMIL, an author can describe the temporal behaviour of the presentation, describe the layout of the presentation on a screen and associate hyperlinks with media [see SMIL]. Syntax follows strongly an HTML/XML approach with extensions for the presentation of independent media objects. As compared to MPEG-4, SMIL does not specify fine-grain synchronization and does not provide explicit 3D support.

### 3.1.3 BHTML

The committee for digital broadcasting of the United States Federal Communication Commission has specified BHTML as an extension to HTML. It has been designed primarily for digital TV with browsing functionality and is therefore a direct competitor for MPEG-4. In comparison to SMIL, BHTML is downsized in other functionalities which are not so important for digital TV applications.

An extensive survey about MPEG-4 and related standards may be found in [see Battista et al. 1999], [see Battista et al. 2000].

### 3.1.4 HyTime – Hypermedia/Time-Based Structuring Language

HyTime is a standard framework for integrated hypermedia, based on SGML technology and documents. It extends SGML in a large number of functionalities [see DeRose and Durand 1994]. It allows to define element types or classes, called architectural forms. As a consequence, hyperlinks and event schedules may be specified with great flexibility. Multimedia documents may be linked in time and space with different types of links (contextual, independent, aggregate, query). It is a powerful and very general standard, and thus has relationship and influence on MPEG-7, in particular on the linking mechanisms for MPEG-7 DDL [see Nack and Lindsay 99a] and [Nack and Lindsay 99b].

### 3.1.5 MHEG-5

In the MHEG standards series, the Multimedia and Hypermedia Information Encoding Expert Group (MHEG) within ISO/IEC specifies the coded representation and the interchange of multimedia and hypermedia information objects. This ranges from storage devices over local networks to telecommunication or broadband networks. MHEG focuses on the interchange of a *final-form representation* of multimedia objects which retain spatial and temporal relationships. This includes interaction objects such as buttons, text entry, and scrolling areas where selection and modification are possible. Other components are regular content objects and composite objects. Behavioural objects deal with action, linking, and scripting. MHEG-5 is the fifth part of the MHEG suite, aimed at interactive client/server applications.

As an interesting fact, MHEG is one of the very few instances, where SGML/XML or derivatives are not used. The text form of MHEG code is written in ASN.1 (Abstract Syntax Notation version 1), also an ISO standard. The final form of MHEG code is binary, not textual, and this binary form must be common to all hardware platforms for the standard to work.

MHEG-5, with respect to its description of content is related to work done in MPEG-7. On the other hand, MHEG-5 was built with special emphasis on interactive TV and set top boxes, an area that is also directly covered by MPEG-4.

## 3.2 Standards in the broadcast communities

In particular the audiovisual archives of broadcasters have had the problem of defining common metadata and common content exchange formats. It is obvious, that the content of those archives will have tremendous impact on the consumer side not only for entertainment, but also for educational purposes. Efficient knowledge management for this content is a goal even above the currently desired goals of interoperability, standardization and efficient search and retrieval of material. Therefore in this paper attention is paid also to developments going on in this area. Most of the activities described in this chapter are highly related to MPEG-7 and in less intensity with MPEG-4 and the upcoming MPEG-21 efforts.

### 3.2.1 SMPTE

The European Broadcasting Union (EBU) and the Society of Motion Picture and Television Engineers (SMPTE) has formed the “Joint EBU/SMPTE Task Force for the Harmonisation of Standards for the Exchange of Television Programme Material as Bit Streams”. In 1997 and 1998 they have produced reports, one on “User Requirements”, and a second one on “Systems, Compression Issues, Wrappers and Metadata and Networks and Transfer Protocols”, respectively.

The SMPTE Metadata Dictionary (SMPTE 335M-2000) is a reference book of audio-visual descriptors. These descriptors cover the entire production chain (pre-production, postproduction, acquisition, distribution, transmission, storage and archiving). A hierarchical registration of metadata items is possible through a general scheme. Different description sets from other activities were combined into one common set.

The dictionary is made up of 10 categories (extensible to 255) dealing with the different aspects to be described. The data are encoded in the KLV (Key-Length-Value) protocol. The SMPTE Universal Label is taken as the key. The automatically created length is according to ISO standards and the value is taken from the metadata dictionary [see SMPTE standards].

The Unique Material Identifier (SMPTE 330M-2000) describes the format of a unique identifier for material like video, audio and data. The identifiers referring to that standard are created locally (thus not asking a general database for a registration) but are still globally unique. This is a major difference to other identification methods. The reason why this uniqueness is possible lies in the fact that the identifier is made up of 2 parts: a Basic UMID and the Signature metadata. The Basic UMID contains the universal label, the length, the instance number and the material number. The Signature metadata is made up of time/date information, spatial coordinates, country and organisation codes and the name of the creator.

### 3.2.2 SMEF

The SMEF™ (Standard Media Exchange Framework) [see SMEF™ DATA MODEL] is a data model, which allows the description of all information related to the production, development, use and management of media assets. The model offers a semantic and a logical view on the items, logical clusters of items and the relationships in between the clusters. The model consists of two parts: a data dictionary defining the entities and attributes and a number of entity relationship diagrams (ERDs), which show the structure in the form of relations between the entities and also the cardinalities in these relations.

The Model has been developed within the BBC’s Technology Directorate. It has been compiled taking into account the work of relevant standards bodies (e.g. SMPTE and MPEG-7). It is intended that further development of SMEF will continue to incorporate the standards developed by these bodies.

### 3.2.3 EBU P/META

Another initiative of the EBU (European Broadcast Union) led by BBC is the project P/META (Metadata exchange standards). The objective is to standardise the structuring of media related information, which may be carried separately or embedded in the media itself. This project can be seen as complementary to other activities of EBU and SMPTE (e.g. Metadata Dictionary, UMID) as well. The main goals of this project [see EBU P/META] are to use the BBC Standard Media Exchange Framework (SMEF) as the core information architecture and to validate and extend the SMEF model. It also wants to establish understanding of the use of unique identifiers in metadata e.g. the SMPTE UMID, and to develop protocols for their management between members. Furthermore it wants to co-operate with standards bodies in related industries such as music and print publishing, to collate all relevant unique identifier schemes and map them against each other, e.g. with the EU INDECS project [see INDECS] and the DOI Foundation [see DOI].

### 3.2.4 AAF – Advanced Authoring Format

AAF is a software implementation of SMPTE metadata and SMPTE labels, designed particularly to make it easy to work with large collections of interrelated sets of metadata and essence. Besides the ability to format and manipulate metadata itself, the AAF software toolkit provides added capabilities for management of metadata sets, user extensions, and plug-in modules.

AAF is moving through committees in SMPTE. Some elements of AAF have been incorporated into MPEG-4, and SMPTE and MPEG-7 are harmonising their metadata descriptions. The Pro-MPEG forum is studying AAF compatibility [see The AAF Association]. The role and position of the respective standards and initiatives is visualized in Figure 1.

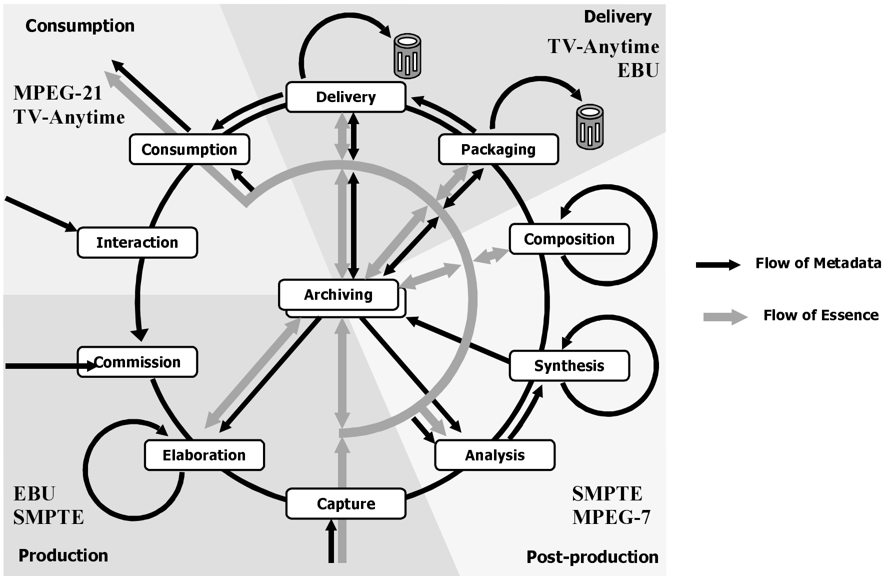


Figure 1: Broadcast workflow and positioning of standards
(after [see AAF 2000])

### 3.2 5 MXF – Material Exchange Format

The main objective of the Material Exchange Format is to exchange programme material together with attached metadata information about the material body. The MXF format is specifically targeted at professional video and broadcast applications, which is a major differentiator from consumer applications at the one end and complex content authoring at the other.

The Pro-MPEG Forum is an association of broadcasters and programme makers with strong participation of equipment manufacturers and component suppliers [see Pro-MPEG Forum]. Also due to composition of the Forum, there is relatively close co-ordination with SMPTE and compliance with AAF.

## 4 Content-based Multimedia Retrieval

Many professional groups share the need for content-based retrieval systems. The requirements of these groups and application areas, e.g. crime prevention, medicine and publishing can vary considerably. In this section we will describe how the current state of the art Content-based Retrieval (CBR) methods can be characterised, independent from their application area. We concentrate on image and video retrieval and give a brief overview on audio retrieval methods [see also Eakins and Graham 1999] and [Aigrain 1996].

### 4.1 Image retrieval techniques: current practice and state-of-the-art

First of all it is useful to characterise image queries into 3 levels of abstraction of increasing complexity [see Eakins 1996]. Please note that these levels do not contain queries by associated (administrative) metadata, such as who created the media object, where and when, because this is primarily a text indexing and retrieval issue.

Level 1: retrieval by *primitive features* such as colour, texture, shape or the spatial location of image elements. In case of video also motion information is an primitive feature. This level of retrieval uses features which are directly derivable from the images and video themselves, without the need to refer to any external knowledge base.

Level 2: retrieval by *derived* or *logical features,* involving some degree of logical inference about the identity of the objects depicted in the image. To extract such logical features usually some reference to external knowledge is needed. E.g. to answer queries like “find pictures with the Grazer Uhrturm” or “find pictures with Tiger Woods” one needs the knowledge that certain structures have been named “Grazer Uhrturm” or identify persons such as “Tiger Woods”. However, these criteria are still reasonably objective.

Level 3: retrieval by *abstract attributes*, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. Example for such a query could be “find pictures depicting happiness”, which could be issued by artists for newspapers or magazines. Complex reasoning and often subjective judgement are required to successfully handle this type of queries.

Level 2 and level 3 are often referred together as *semantic retrieval* [see Gudivada and Raghavan 1995], hence the gap between level 1 and 2 is named *semantic gap.*

Video queries are categorised in the same way as image queries, as they mainly consist of image data. A common way of how to organise video for archiving and retrieval is to prepare a storyboard of annotated still images (*keyframes*) representing each scene. However, there is one major difference, as video usually also has a soundtrack, containing music, speech and other sounds. Sometimes there is also text appearing in the video (trailer) or even closed-caption text used to provide subtitles. All of this information can provide additional cues for retrieval.

#### 4.1.1 Current image retrieval techniques

Almost all current content-based image retrieval systems, commercial and experimental, operate at level 1 of the query categories described above. Automatically extracted features like measures of colour, texture or shape are used to describe images and stored with the images in a database. A typical system allows the user to formulate queries by submitting an example, some offer the possibility to submit sketches of the sought-after images. Some of the commonly used feature types are described below:

Colour: a colour histogram, which shows the proportion of pixels of each colour within the image, is calculated for each image and stored in a database. The user can search either by describing the desired amount of particular colours or by submitting an example, which histogram is calculated and compared to those in the database. The most commonly used matching method was first developed by Swain and Ballard [see Swain and Ballard 1991] and is called histogram intersection.

Texture: texture can be used to distinguish between areas with similar colour, such as sky and sea. Essentially these calculate the relative brightness of selected pairs of pixels from each image. From these it is possible to calculate measures of image texture such as degree of *contrast*, *coarseness*, *directionality* and *regularity* [see Tamura et al. 1978]. Queries can be formulated in the same way as above, by supplying an image example or selecting from a given palette of known textures.

Shape: retrieval by shape is one of the most obvious requirements at the primitive level. There is considerable evidence that most natural objects are primarily recognised by their shape. Queries are formulated either by example images or as user-drawn sketches. Two main types of shape features are commonly used: global features like aspect ratio, circularity and moment invariants [see Niblack et al. 1993] and local features such as sets of consecutive boundary segments [see Mehrotra and Gary 1995]. Shape matching of three-dimensional objects is more challenging, especially where only a single 2-D view of the object is available. One approach is to generate a series of alternative 2-D views of a 3-D model and match them with the query image.

Other types: several other types of features are used in content-based retrieval, which rely on complex transformation of the pixel intensities and have no direct counterpart in human descriptions of images. One of the mainly used techniques is the wavelet transformation to model an image in several different resolutions. Promising results have been reported by matching these wavelet features from sample and stored images [see Jacobs et al. 1995] and [see Liang and Kuo 1998].

As mentioned before, methods for retrieval of videos rely on adaptations of techniques developed for image retrieval. Usually a video is first divided into shots, i.e. scenes without changes in main content, camera position or angle. Such changes can be detected by the analysis of the motion vector field (which is also part of the

MPEG compression scheme) and colour histograms. From each shot a keyframe can be extracted and the standard image retrieval methods can be applied.

#### 4.1.2 Existing systems

There are several image retrieval systems available as commercial packages, including QBIC from IBM [see Flickner et al. 1995], the VIR Image Engine from Virage Inc. [see Gupta et al. 1996] and Visual RetrievalWare from Excalibur Technologies [see Feder 1996]. The European companies LookThatUp (www.lookthatup.fr) and Cobion (www.cobion.de) also offer very powerful products and services for content-based image recognition. These products are in use within video archives, Web search engines for finding images on the Web and professional image stock databases.

### 4.2 Audio retrieval techniques

There are several approaches on content-based identification and search of audio material. Due to the smaller complexity of the problem (only one-dimensional signal as compared to the two-dimensional images) there are more mature research results available.

Audio retrieval techniques have to be divided into two categories: *Speech recognition* and *general audio* or *music recognition*. The first one is nowadays widely available and in use in office applications of personal computers and in integrated telephony applications.

For music recognition the retrieval queries are formulated either by humming or whistling a melody or by giving a music example. [See Wold et al. 1996] from Muscle Fish describe a system for finding similar sounds to a given example. This system extracts time-varying properties from sampled sound files and for each property the mean, variance and autocorrelation over the entire file is recorded. At the time of their publication the system was used for comparison of noises, like scratches, bells and laughing but is nowadays extended for whole song identification.

The IST project RAA (Recognition and Analysis of Audio) develops a system for identifying songs considerably faster than real-time, which is robust against transmission and compression effects and highly scalable in terms of the amount of original titles in the audio database [see Neuschmied et al. 2001].

### 4.3 Application of MPEG-7

#### 4.3.1 Application fields foreseen by MPEG-7

MPEG-7 addresses and supports a broad range of application areas, e.g. multimedia digital libraries, broadcast media selection, multimedia editing, home entertainment devices and so on. It also wants to contribute to making the Web searchable for multimedia objects as it is today for text.

It is outside the scope of the MPEG-7 standard to define the way how data is used to answer particular queries, but the authors of the MPEG-7 overview give the following sophisticated examples of some query scenarios [see ISO/IEC JTC 1/SC 29/WG 11 N4031, 2001]:

- Play a few notes on a keyboard and retrieve a list of musical pieces similar to the required tune, or images matching the notes in a certain way, e.g. in terms of emotion.
- Draw a few lines on a screen and find a set of images containing similar graphics, logos, ideograms, …
- Define objects, including colour patches or textures and retrieve examples among which you select the interesting objects to compose your design.
- On a given set of multimedia objects, describe movements and relations between objects and so search for animations fulfilling the described temporal and spatial relations.
- Describe actions and get a list of scenarios containing such actions.
- Using an excerpt of Pavarotti's voice, obtaining a list of Pavarotti's records, video clips where Pavarotti is singing and photographic material portraying Pavarotti.

Obviously these examples involve to a good amount level 2 and even level 3 query mechanisms, which are currently still under research and it is unclear when such technologies will be available for general purpose applications.

However, MPEG-7 provides with its descriptors (D), description schemes (DS) and the description definition language (DDL) a lot of elements which enable content-based search and retrieval applications.

### 4.3.2 MPEG-7 high-level audio description tools

The MPEG-7 Audio standard defines an *audio description framework*, which contains low-level tools designed to provide a basis for higher level audio applications. In addition to that the following four sets of audio description tools are integrated in the final committee draft:

- *Musical timbre description tools:* describing the perceptual features of instrument sounds.
- *Sound recognition description tools:* a collection of tools for indexing and categorization of general sounds, with immediate application to sound effects.
- *Spoken content description tools:* detailed description of words spoken within an audio stream.
- *Melody description tools:* a compact representation for melodic information, which allows for efficient and robust melodic similarity matching, e.g. in query by humming.

#### 4.3.3 MPEG-7 visual description tools

The MPEG-7 visual description tools consist of the following basic structures and basic visual features, each category consists of elementary and sophisticated descriptors:

- *Basic Structures:* this includes the *Grid Layout*, the *Time Series*, *Multi View*, the *Spatial 2D Coordinates* and *Temporal Interpolation*.
- *Colour Descriptors:* there are eight descriptors: *Colour Space*, *Dominant Colour*, *Colour Quantisation*, *Group of Frames Colour*, *Colour-Structure* and *Scalable Colour*. All these descriptors allow for the detailed description of colour features in visual content.
- *Texture Descriptors:* consist of *Homogenous Texture*, an important primitive for searching and browsing through large collections of similar looking patterns; *Texture Browsing*, which provides a perceptual characterization of texture, similar to a human characterization, in terms of regularity, coarseness and directionality; *Edge Histogram*, representing the spatial distribution of five types of edges, namely four directional edges and one non-directional edge.
- *Shape Descriptors:* consist of *Region-Based Shape*, which can describe any shapes, including complex shapes that consists of holes in the object or several disjoint regions; *Contour-Based Shape*, uses so-called Curvature Scale-Space representation, which captures perceptually meaningful features of the shape; *3D Shape*, aims at providing an intrinsic shape description of 3D mesh models, targeted at search and retrieval of 3D model databases.
- *Motion Descriptors:* consist of *Camera Motion*, characterizing 3D camera motion parameters, which can be automatically generated by the capture device; *Motion Trajectory*, which is defined as the localization, in time and space, of one representative point of an object; *Parametric Motion*, describing motion of objects as a 2D parametric model; *Motion Activity*, capturing the intuitive notion of ‘intensity of action’ or ‘pace of action’ in a video segment.
- *Localization:* the *Region Locator* enables localization of regions within images by specifying them with a brief and scalable representation of a box or polygon; the *Spatio Temporal Locator* describes spatio-temporal regions in a video sequence, such as moving object regions, and provides localization features.
- *Others:* Currently this includes the *Face Recognition* descriptor, which can be used to retrieve face images which match a query face image. It represents the projection of a face vector onto a set of basis vectors which span the space of possible face vectors. This feature set is extracted from a normalized face image, containing 56 lines with 46 intensity values in each line.

#### 4.3.4 Sample applications and projects

Various European R&D projects are trying to provide audiovisual archiving systems in the philosophy of MPEG-7, partially starting from databases with proprietary documentation and user access interfaces, handling metadata information. Among the first approaches are those of VICAR (Esprit-24916, http://iis.joanneum.ac.at/vicar) and DiVan (Esprit-24956, http://divan.intranet.gr) projects. The AVIR project (ESPRIT-28798, http://www.extra.research.philips.com/euprojects/avir), in turn, has proposed a language for expressing metadata information and description schemes, following up the developments towards the MPEG-7 standard. The ACTS DICEMAN project (ACTS308, http://www.teltec.dcu.ie/diceman) has been developing an MPEG-7 database implementation.

FAETHON (IST-1999-20502) [see Delopoulos and Haas 2001] has the goal to extract high level semantic information out of existing syntactic or (lower level) semantic data like those encapsulated in MPEG-7 structures (descriptors and description schemes). It will concentrate on the subjective extraction of semantic information, depending on users' profile by applying interpretation rules.

The TV-Anytime Forum [see TV-Anytime Forum] is a group of organisations and industry partners. In their TV Anytime project they are developing a framework of tools and technologies for movies on demand, broadcast recording, searching and filtering, for retrieving information from the web, together with e-commerce and remote education. It aims at the mass-market, high volume storage for home consumers. An implementation will probably be located in a future combination of VCR and set-top-box. In the context of this project, MPEG-7 capabilities will be utilized for the "metadata standard" audiovisual descriptors, for content referencing and rights management. The typical application scenario is the so called Electronic Program Guide (EPG), that will enable users to discover and access (parts of) programs and documents from digital broadcast or the web [see Pfeiffer 2000].

# 5 Application Architecture for Knowledge Management Prototype

## 5.1 Management of multimedia data

Current knowledge management systems concentrate mainly on knowledge contained in text-based information types. For this data type there exist several low and high-level search and retrieval methods which are well described in literature and are also available in commercial products and are therefore not discussed in this article.

The digital management of multimedia content including video, audio, still images, animations, 3D- and various other types of objects and documents has been recognized as the major challenge for a major part of future knowledge management systems for search, retrieval, preview and partial distribution of these assets. The usage of standards, particularly of MPEG-7 for describing multimedia data is a feasible approach to create a unique management system for storing and retrieving of different multimedia data with different data formats.

In such a system a database is required that can manage MPEG-7 and therefore XML-Schema documents. MPEG-7 documents have to be saved, such that is possible to search for individual metadata. The extent of the description of multimedia data and therefore which metadata are specified in MPEG-7 documents is highly variable, depending on media, application area and user's semantic view.

There are also extremely high storage and performance requirements for the content database. For video, audio and image data huge data sets have to be managed. To support content-based retrieval (e.g. similarity search for images or audio), features and feature vectors need to be extracted. Based on these features a similar or equal content can be found in the database. After finding the content, not only the whole video or piece of music should be retrieved. It is also necessary to get for example specific images of a video or parts of audio data as a response to the query from the database.

### 5.2 Prototype application architecture

The goal of a basic research project carried out by Joanneum Research for the KNOW Center is to develop a concept for a system for annotation, storage and retrieval of multimedia data. By the implementation of a prototype for such a system experience will be gained about required software technologies and in particular about the MPEG-7 standard and performance of the DB technology required.

For the metadata we differentiate between *general metadata* (already standardized) and *retrieval data* (colour histogram, texture features, object contours, etc) which may also be standardized, but are delivered by specialized applications in a non MPEG-7 compliant format. They will later be used together for content-based search (similarity search in the prototype). The extraction of metadata at a low semantic level will be done automatically for a few selected functions. This annotation will be enriched manually in areas relevant for the users' requirements. Examples of low-level metadata are the retrieval data or the data which describe the structure of the multimedia data (shots, key-frames, etc.).

The architecture of this system can be seen in Figure 2. The main components have the following tasks:

- Automatic and manual acquisition of the metadata from different multimedia data like image, audio, video and animation data (PowerPoint, SMIL, MPEG-4, etc.).
- Storage of multimedia data and MPEG-7 documents in databases (XML for MPEG-7 compliant annotation, relational or object oriented databases for content and retrieval data.
- Metadata and content-based retrieval of the multimedia data.

The multimedia database will separately save the multimedia data (content), the MPEG-7 documents and the retrieval data. For the retrieval data specific search algorithms and index structures are necessary.

A Web-based retrieval tool is used to query the multimedia database. Search results contain extracts of the annotations, parts of the content (e.g. key-frame of a video). and references for downloading the multimedia data and the appropriate MPEG-7 document.

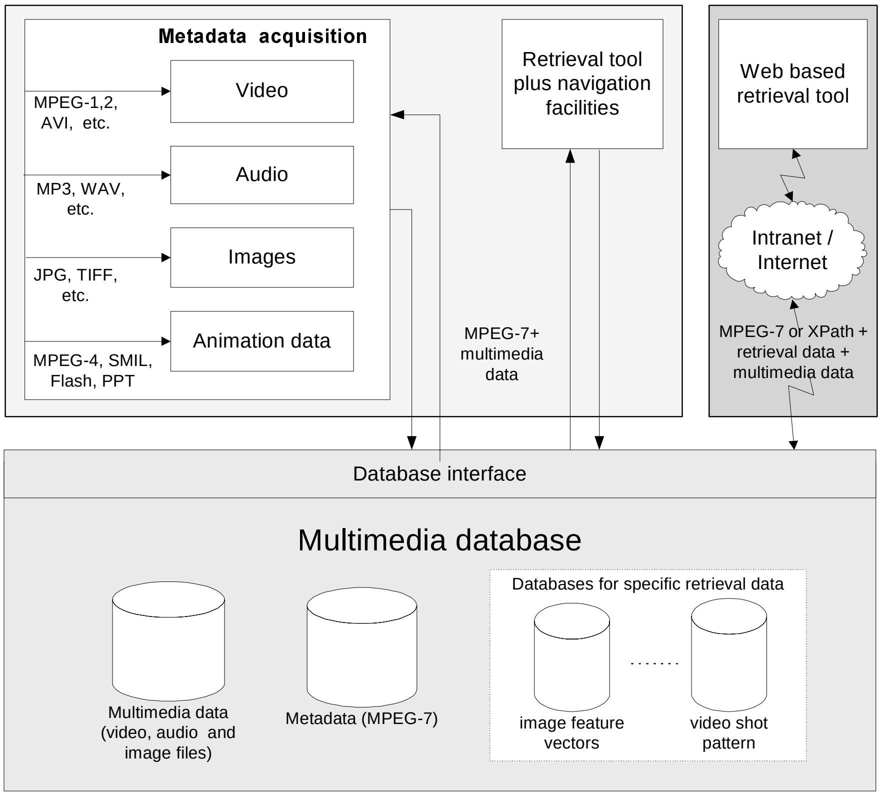


Figure 2: System Architecture for content-based retrieval prototype

## 6 Conclusion

For the efficient search and retrieval of audiovisual data, a high number of methodologies and algorithms is available in the research field. The MPEG-7 standard is describing extensively the metadata of this content, geared towards efficient search and retrieval. Thus, it seems straightforward to implement a content-based indexing and retrieval system for audiovisual data based on these technologies.

Unfortunately, MPEG-7 has not yet achieved a completely stable stage yet, and there are no fully developed applications yet, based on this standard. Storage of MPEG-7 based metadata requires handling of XML-based documents. Native XML-databases seem to be the obvious choice, but also in this field there is lack of experience.

The multimedia indexing and retrieval algorithms – and where available, applications – are based on proprietary data structures and are not yet based on MPEG-7 compliant descriptors. They are also not general purpose, but mostly very specialized for specific application areas.

We have chosen an approach, that – despite the early stage of standardisation and immature development of general algorithms for CBR – is based on these modern standards and methodologies. We are implementing a prototype that covers the complete workflow, from (automatic) annotation over storage to later search and retrieval. We will however limit the number of applications for automatic search and retrieval for a few selected ones. This is done in order to get experience with all components and their integration. The inclusion of additional and better search and retrieval methodologies and of more automatic tools will then be an easier task, based on a solid and proven concept.

**Acknowledgements**

This work has been motivated and initiated through participation in the KNOW-Center Graz (Competence Center for knowledge-based Applications and Systems), funded through the support within the Austrian Competence Center program K plus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.kplus.at) and the industrial partners of the center. This support is gratefully acknowledged by the authors.

Much of the background and complementary know-how has been and is being acquired through participation in EU funded projects. Most prominent among them are VICAR (Esprit-24916, http://iis.joanneum.at/vicar), VIZARD (IST-2000-26354), RAA (IST-1999-12585, http://raa.joanneum.at), PRESTO (IST-1999-20013, http://presto.joanneum.at) and FAETHON (IST-1999-20502). The support of the EC is gratefully acknowledged.

## References

[AAF 2000] AAF Technical information, http://www.aafassociation.org/html/techinfo/index.html.

[Aigrain 1996] Aigrain, P. et al.: “Content-based representation and retrieval of visual media – a state-of-the-art review”; Multimedia Tools and Applications 3(3), pp. 179-202.

[Battista et al. 1999] Battista, S., Casalino, F., Lande, C.: “MPEG-4: A Multimedia Standard for the Third Millenium, Part 1”; IEEE Multimedia, 6, 4, October-December 1999, pp. 74-83.

[Battista et al. 2000] Battista, S., Casalino, F., Lande, C.: “MPEG-4: A Multimedia Standard for the Third Millenium, Part 2”; IEEE Multimedia, 7, 1, January-March 2000, pp. 76-84.

[Day 2000] Day, N.: ”MPEG-7 Daring To Describe Multimedia Content”; XML-Journal, 1,6, (2000), pp. 24-27.

[Delopoulos and Haas 2001] Delopoulos, A., Haas, W. et al.: “Unified Access to Heterogenous Audiovisual Content”; to be published in Proc. of CBMI’01, Brescia (2001).

[DeRose and Durand 1994] DeRose, S., Durand, D.: “Making Hypermedia Work – A User’s Guide to HyTime”; Kluwer Academic Publishers, Boston, 1994.

[DOI] http://www.doi.org/.

[Eakins 1996] Eakins, J.P.: "Automatic image content retrieval – are we getting anywhere?"; Proceedings of Third International Conference on Electronic Library and Visual Information Research (ELVIRA3), De Montfort University, Milton Keynes, pp. 123-135.

[Eakins and Graham 1999] Eakins, J.P. and Graham, M.: "Content-based image retrieval"; JISC Technology Application Programme, Report No. 39. http://www.jtap.ac.uk.

[EBU P/META] European Broadcasting Union: PMC Project P/META (Metadata exchange standards): http://www.ebu.ch/pmc_meta.html.

[Feder 1996] Feder, J.: "Towards image content-based retrieval for the World-Wide Web"; in Advanced Imaging 11(1), pp. 26-32.

[Flickner et al. 1995] Flickner M. et al.: "Query by image and video content: the QBIC system"; IEEE Computer 28(9), pp. 23-32.

[Gudivada and Raghavan 1995], Gudivada, V.N. and Raghavan, V.V.: "Content-based image retrieval systems"; IEEE Computer 28(9), pp. 18-22.

[Gupta et al. 1996] Gupta, A. et al.: "The Virage image search engine: an open framework for image management"; in Storage and Retrieval for Image and Video Databases IV, Proceedings SPIE 2670, pp. 76-87.

[INDECS] http://www.indecs.org/.

[IPMP] http://www.mpeg.org/MPEG/.

[ISO/IEC JTC1/SC29/WG11 N3747] "MPEG-4 Overview"; v 16, International Organization for Standardisation, October 2000, La Baule, France.

[ISO/IEC JTC1/SC29/WG11 N3702] "Multimedia content description interface – Part 2 Description definition language"; International Organization for Standardisation, October 2000 ,La Baule, France.

[ISO/IEC JTC1/SC 29/WG 11 N3705] "Multimedia Content Description Interface – Part 5 Multimedia Description Schemes"; v 1.0, International Organization for Standardisation, October 2000, La Baule, France.

[ISO/IEC JTC1/SC 29/WG 11 N3815] "Multimedia Description Schemes XM"; v 6.0, International Organization for Standardisation, January 2001, Pisa, Italy.

[ISO/IEC JTC1/SC 29/WG 11 N4031] "Overview of the MPEG-7 Standard"; v 5.0, International Organization for Standardisation, March 2001, Singapore.

[Jacobs et al. 1995] Jacobs, C.E. et al.: "Fast multiresolution image querying"; Proceedings of SIGGRAPH 1995, Los Angeles, CA, pp. 277-286.

[Liang and Kuo 1998] Liang, K.C. and Kuo, C.C.J.: "Implementation and performance evaluation of a progressive image retrieval system"; in Storage and Retrieval for Image and Video Databases VI, Proceedings SPIE 3312, pp. 37-48.

[Mehrota and Gary 1995] Mehrota, R. and Gary, J.E.: "Similar-shape retrieval in shape data management"; IEEE Computer 28(9), pp. 57-62.

[MPEG-4] http://www.cselt.it/mpeg/standards/MPEG-4/MPEG-4.htm.

[MPEG-7] http://www.cselt.it/mpeg/standards/MPEG-7/MPEG-7.htm.

[MPEG-7 main page] GMD - Forschungszentrum Informationstechnik GmbH, http://www.darmstadt.gmd.de/mobile/MPEG-7/index.html.

[Nack and Lindsay 1999a] Nack, F., Lindsay, A.: "Everything You Wanted to Know About MPEG-7: Part 1"; IEEE Multimedia, 6(3), July-September 1999, 65-77.

[Nack and Lindsay 1999b] Nack, F., Lindsay, A.: "Everything You Wanted to Know About MPEG-7: Part 2"; IEEE Multimedia, 6(4), October-December 1999, 64-73.

[Neuschmied et al. 2001] Neuschmied, H., Mayer, H. and Batlle, E.: "Content-based Identification of Audio Titles on the Internet"; to be published at Wedelmusic 2001, Florence.

[Niblack et al. 1993] Niblack, W. et al.: "The QBIC project: querying images by color, texture and shape"; IBM Research Report RJ-9203.

[Pfeiffer 2000] Pfeiffer, S., Srinivasan, U.: "TV Anytime as an application scenario for MPEG-7"; http://woodworm.cs.uml.edu/~rprice/ep/pfeiffer/index.html, Copyright ACM, 2000.

[Pro-MPEG Forum] http://www.pro-mpeg.org/.

[SMEF™ DATA MODEL] SMEF™ DATA MODEL v 1.5: British Broadcasting Corporation (2000).

[SMIL] http://www.w3.org/TR/REC-smil/.

[SMPTE standards] http://www.smpte.org/stds/s336m.pdf.

[Swain and Ballard 1991] Swain, M.J. and Ballard, D.H.: "Color indexing"; International Journal of Computer Vision 7(1), pp. 11-32.

[Tamura et al. 1978], Tamura, H. et al.: "Textural features corresponding to visual perception"; IEEE Transactions on Systems, Man and Cybernetics 8(6), pp. 460-472.

[The AAF Association] The AAF Association, http://www.aafassociation.org/.

[TV-Anytime Forum] http://www.tv-anytime.org/.

[Wold et al. 1996] Wold, E., Blum, T., Keislar, D., and Wheaton, J.: "Content-Based Classification, Search, and Retrieval of Audio"; IEEE Multimedia, Vol. 3., No. 3, 1996, pp. 27-36.