Everything You Wanted to Know About MPEG-7: Part 1

Frank Nack
GMD-IPSI

Adam T. Lindsay
Starlab

Audio-visual information must allow some degree of interpretation, which can be passed onto, or accessed by a device or a computer code. MPEG-7 aims to create a standard for describing these operational requirements. We provide an overview on the development, functionality, and applicability of MPEG-7. We present the role of MPEG-7 and outline ideas for using MPEG-7 technology based on examples of improved versions of existing applications as well as completely new ones.

Part I of this article provides an overview of the development, functionality, and applicability of MPEG-7. We'll first present the role of MPEG-7 within the context of past MPEG standards. We then outline ideas of what should be possible using MPEG-7 technology. In Part II, we'll discuss the description of MPEG-7's concepts, terminology, and requirements. We'll then compare MPEG-7 to other approaches on multimedia content description.

Because of the Internet's popularity, the last decade has experienced a rapid increase of digital audio-visual information. Though the increasing availability of potentially interesting information has enriched our lives (for example, e-mail and the World Wide Web), the overwhelming amount of information also raises a fundamental problem: How fast and easily can desirable information be made available? The more interesting—that is, specific and useful—material available, the harder it is to locate.

A noticeable indicator of the existing tension between humans and the vast amounts of information available lies in the popularity of search engines available on the Web. Unfortunately, current solutions let users only search for textual information. Identifying audio-visual information proves difficult, as no generally recognized description of this material exists. In general, it's not possible to efficiently search the Web for, say, a picture of the motorbike from "Terminator II," or for a sequence where King Lear congratulates his assistants on the night after the battle, or for "twenty minutes of video according to my preferences of today." You can envisage similar examples for audio, in which you whistle a melody to find a song or quote a movie to find the context. It's true that in specific cases, solutions do exist. Multimedia databases on the market today let users search for pictures using characteristics like color, texture, and information about the shape of objects in the picture.

Furthermore, the question of identifying content isn't restricted to database retrieval applications—the problem applies equally to other areas. For instance, we're promised a world of 500-plus broadcast television channels, which will of course make it harder to select a potentially interesting channel. Domains other than search or filtering include image understanding (surveillance, intelligent vision, smart cameras, and so on) or media conversion (speech to text, picture to speech, visual transcoding, and so on).

In October 1996, the Moving Pictures Expert Group (MPEG) started a new work item to provide a solution to the questions described above. The newest member of the MPEG family, called the multimedia content description interface (MPEG-7), will extend the limited capabilities of proprietary solutions in identifying content that exists today, notably by including more data types. In other words, MPEG-7 aims to standardize a core set of quantitative measures of audio-visual features, called Descriptors (D), and structures of descriptors and their relationships, called Description Schemes (DS) in MPEG-7 parlance. MPEG-7 will also standardize a language—the Description Definition Language (DDL)—that specifies Description Schemes to ensure flexibility for wide adoption and a long life. You can index and search for audio-visual material that has MPEG-7 data associated with it. This material may include still pictures, graphics, 3D models, audio, speech, video, and information about how these elements combine in a multimedia presentation (for example, scenarios or composition information). We expect the standard core set of MPEG-7 functionality will facilitate those classes of applications that have widespread use and will provide interoperability.

The family of MPEG standards

MPEG is a working group of the International Organization for Standardization/International

1070-986X/99/$10.00 © 1999 IEEE

List of Standards

H.261

CCITT Recommendation H.261, Line Transmission on Non-telephone Signals: Video Codec for Audiovisual Services at p × 64 Kbps, Int'l Telegraph/Int'l Telecommunication Union and Telephone, Geneva, 1990.

JPEG

ISO/IEC IS 10918, Information Technology—Digital Compression and Coding of Continuous-Tone Still Images, Int'l Organization for Standardization, Geneva, 1993.

MPEG-1

Information Technology—Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbps, ISO, Geneva, 1993.

- *ISO/IEC 11172-1, Part 1: Systems*
- *ISO/IEC 11172-2, Part 2: Video*
- *ISO/IEC 1117-3, Part 3: Audio*
- *ISO/IEC 11172-4, Part 4: Conformance Testing*
- *ISO/IEC 11172-5, Part 5: Software Simulation*

MPEG-2

Information Technology—Generic Coding of Moving Pictures and Associated Audio Information (MPEG-2), ISO, Geneva, 1996.

- *ISO/IEC IS 13818-1, Part 1: Systems*
- *ISO/IEC IS 13818-2, Part 2: Video*
- *ISO/IEC IS 13818-3, Part 3: Audio*
- *ISO/IEC IS 13818-4, Part 4: Compliance testing*
- *ISO/IEC IS 13818-5, Part 5: Software simulation (Future TR)*
- *ISO/IEC IS 13818-6, Part 6: Extensions for Digital Storage Media Command and Control (DSM-CC)*
- *ISO/IEC IS 13818-9, Part 9: Extension for RealTime Interface for Systems Decoders*

MPEG-4, version 1

R. Koenen, *ISO/IEC JTC1/SC29/WG11 N2323, MPEG-4 Overview (Dublin version)*, ISO, Geneva, July 1998.

MPEG-7

MPEG home page, http://www.cselt.it/mpeg

ISO/MPEG N2728, Applications for MPEG-7, MPEG Requirements Group, ISO, Geneva, March 1999, http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/N2728.html.

ISO/MPEG N2729, MPEG-7 Context and Objectives, MPEG Requirements Group, Geneva, March 1999, http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/N2729.html.

Electronics Commission (ISO/IEC), in charge of developing international standards for compression, decompression, processing, and coded representation of moving pictures, audio, and their combination. So far, MPEG has produced MPEG-1, MPEG-2, MPEG-4 version 1, and is currently working on MPEG-4 version 2 and MPEG-7.

We'll now provide a brief overview of the different work items. We'll show how they're connected and in what ways they differ.

MPEG-1: Storage and retrieval

MPEG-1 is the standard for storage and retrieval of moving pictures and audio on storage media (see the sidebar "List of Standards" for this and other standards mentioned in this article). MPEG-1 provides a nominal data stream compression rate of about 1.2 Mbits per second—the typical CD-ROM data transfer rate—but can deliver data at a rate of up to 1,856,000 bps. MPEG-1 distinguishes four types of image coding for processing: I (intra-coded pictures), P (predictive coded pictures), B (bidirectionally predictive pictures), and D-Frame (coding based on discrete cosine only parameter) images.

To allow audio compression in acceptable quality, MPEG-1 enables audio data rates between 32 and 448 Kbps. MPEG-1 explicitly considers other standards and functionalities, such as JPEG and H.261, suitable for symmetric and asymmetric compression. It also provides a system definition to specify the combination of several individual data streams.

Note that MPEG-1 doesn't prescribe compression in real time. Furthermore, though MPEG-1 defines the process of decoding, it doesn't define the decoder itself. The quality of an MPEG-1 video without sound at roughly 1.2 Mbps (the single-speed CD-ROM transfer rate) is equivalent to a VHS recording.[1]

We should mention that MPEG-1 provides a means for transmitting metadata. In general, two mechanisms exist, the transimission of

- user data extensions within a video stream or
- data in a separated private data stream that gets multiplexed with the audio and video stream as part of the system stream.

Since both methods attach additional data into the MPEG-1 stream, they either increase the demand of bandwidth for transmission/storage or reduce the quality of the audio-visual streams for a given bandwidth.

No format for the coding of those extra streams was defined, which led to proprietary solutions. This might explain why these mechanisms aren't widely adopted.

MPEG-2: Digital televison

MPEG-2, the digital television standard, strives

for a higher resolution—up to 100 Mbps—that resembles the digital video studio standard CCIR 601 and the video quality needed in HDTV. As a compatible extension to MPEG-1, MPEG-2 supports interlaced video formats and a number of other advanced features, such as those to support HDTV. As a generic standard, MPEG-2 was defined in terms of extensible profiles, each supporting the feature required by an important application class. The Main Profile, for example, supports digital video transmission at a range of 2 to 80 Mbps over cable, satellite, and other broadcast channels. Furthermore, it supports digital storage and other communications applications. An essential extension from MPEG-1 to MPEG-2 is the ability to scale the compressed video, which allows the encoding of video at different qualities (spatial-, rate-, and amplitude-based scaling[2]).

The MPEG-2 audio coding was developed for low bit-rate coding of multichannel audio. MPEG-2 extends the MPEG-1 standard by providing five full bandwidth channels, two surround channels, one channel to improve low frequencies, and/or seven multilingual channels, and the coding of mono and stereo (at 16 kHz, 22.05 kHz, and 24 kHz). Nevertheless, MPEG-2 is still backward-compatible with MPEG-1.

MPEG-2 provides an MPEG-2 system with definitions of how video, audio, and other data combine into single or multiple streams suitable for storage and transmission. Furthermore, it provides syntactical and semantical rules that synchronize decoding and presentation of audio and video information.

With respect to transmission/storage, the same mechanisms developed for MPEG-1 were assigned to MPEG-2. Additionally, some of the MPEG-2 header contains a structured information block, covering such application-related information as copyright and conditional access. The amount of information is restricted to a number of bytes. Reimers[3] described an extensive structuring of content, coding, and access of such metadata within MPEG-2.

Originally, there were plans to specify MPEG-3 as a standard approaching HDTV. However, during the development of MPEG-2, researchers found that it scaled up adequately to meet HDTV requirements. Thus, MPEG-3 was dropped.

MPEG-4: Multimedia production, distribution, and content access

Though the results of MPEG-1 and MPEG-2 served well for wide-ranging developments in such fields as interactive video, CD-ROM, and digital TV, it soon became apparent that multimedia applications required more than the established achievements. Thus, in 1993 MPEG started working to provide the standardized technological elements enabling the integration of the production, distribution, and content access paradigms of digital TV, interactive graphics applications (synthetic content), and interactive multimedia (distribution of and access to enhanced content on the Web). MPEG-4 version 1, formally called ISO/IEC 14496, has been available as an international standard since December 1998. The second version will be finished in December 1999.

MPEG-4 aims to provide a set of technologies to satisfy the needs of authors, service providers, and end users, by avoiding the emergence of a multitude of proprietary, incompatible formats and players.

The standard should allow the development of systems that can be configured for a vast number of applications (among others, real-time communications, surveillance, and mobile multimedia). To achieve this requires providing standardized ways to

- Interact with the material, based on encoding units of aural, visual, or audio-visual content, called media objects. These media objects can be natural or synthetic, which means they could be recorded with a camera or microphone, or generated with a computer.
- Interact with the content, based on the description of these objects' composition, to create compound media objects that form audio-visual scenes. The composition of these audio-visual MPEG-4 objects mirrors the real world, where spatial and temporal relations between objects let users interact with these objects in a way similar to everyday use.
- Integrate different data types, allowing the harmonization of natural and synthetic objects such as 2D and 3D; mono and stereo video or multiview video; mono, stereo, and multichannel audio; and so on.
- Multiplex and synchronize the data associated with media objects so that they can be transported over network channels providing a quality of service (QoS).
- Interact with the audio-visual scene generat-

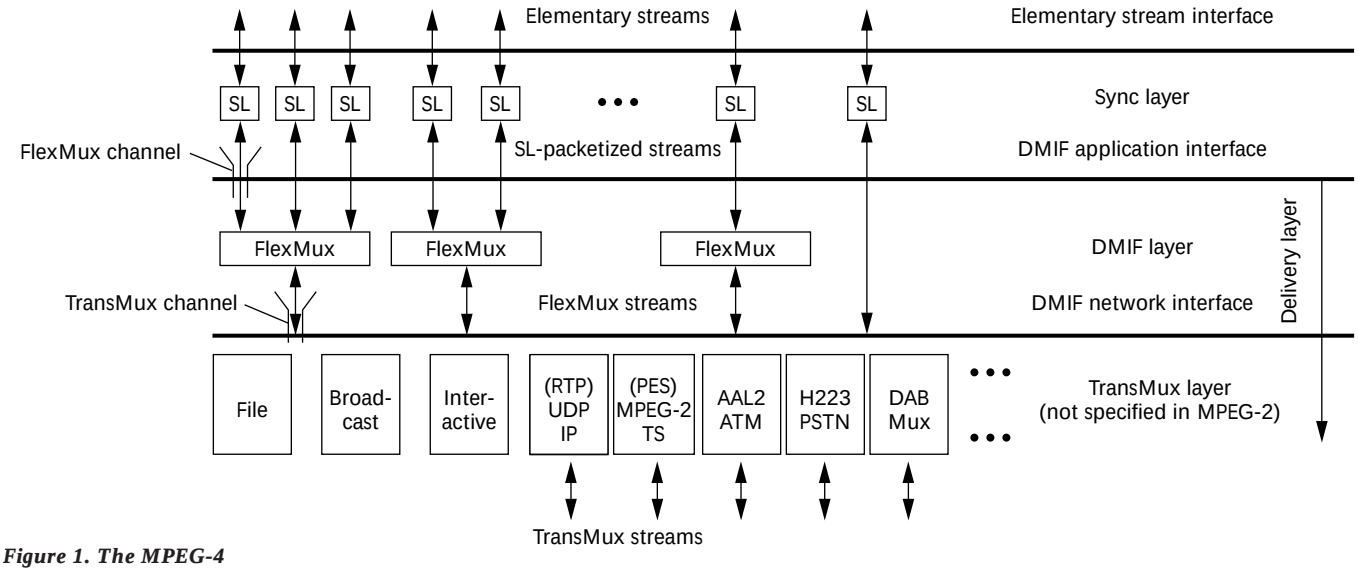

Figure 1. The MPEG-4 system layer model.

ed at the receiver's end. For example, manipulate some characteristics of an object, access selected parts of a scene, remove an object from one scene and integrate it into another, and so on.

To meet the swift technological progress within multimedia, MPEG-4 developed the Syntactic Description Language (MSDL). This approach not only aimed to integrate new algorithms through defined tools, but also to adopt at any time new tools, techniques, and concepts that should have improved or new functionality. In other words, MSDL was the way to guarantee the flexibility of the standard, preventing eventual obsolescence and narrowness of scope. However, MSDL didn't become part of the standard and was replaced by the Binary Format for Scene Description (BIFS), which served a limited but ultimately more robust role.

The major extensions of MPEG-4 over MPEG-2, with respect to the three main goals (content/interaction, flexibility/extensibility, and integration) follow:

- A standard functionality, such as synchronization of audio and video, modes for short delay, and usage via networks.
- Scalability, where content-based scalability is important, since this mechanism prioritizes objects within a scene.
- Content-based manipulation of a bit stream without transcoding.
- Content-based access (indication, hyperlink, request, up- and downloading, deleting, previewing, and so on).
- Content combination such as text and graphics overlay, mixing synthetic and video and/or audio data, and so on.
- Efficient coding of several streams simultaneously, such as stereo video or several views of the same event.
- Improved efficiency in coding (improvement of data quality with low bit rates compared to existing standards such as H.263).
- Robustness in error-susceptible environments due to an elasticity toward remaining errors (a selective look-ahead error correction, error control, error masking, and so on).
- Improved random access on parts of an audio-visual sequence.

Like its predecessors, MPEG-4 deals with streams. Since MPEG-4 subdivides audio-visual content into objects, a stream's standardized characteristics concerned multiplexing, demultiplexing, and synchronizing multiple steams. Figure 1 describes the relationship between different streams based on the MPEG-4 System Layer Model.

MPEG-4 allows you to attach metadata about content to objects. Users of the standard can use this Object Content Information (OCI) data stream to send textual information along with

MPEG-4 content. It's also possible to classify content according to predefined tables, which must be defined outside of MPEG. However, because there's no standardized structure and format defined for the metadata, we see this as having limited long-term usefulness.

The abandonment of MPEG-3 caused much speculation about the next number. Should it be 5 (the next) or 8 (creating an obvious binary pattern)? MPEG, however, decided not to follow either logical expansion of the sequence, but chose the number 7 instead. So, just like MPEG-3, MPEG-5 and MPEG-6 aren't defined.

MPEG-7: Describing multimedia material

The preceding MPEG standards have mainly addressed coded representation of audio-visual information. MPEG-7, on the other hand, focuses on the standardization of a common interface for describing multimedia materials (representing information about the content, but not the content itself—"the bits about the bits"). In this context, MPEG-7 addresses aspects such as facilitating interoperability and globalization of data resources and flexibility of data management.

Thus, the commonalties between previous MPEG standards and MPEG-7 rely on the fact that previous standards can use MPEG-7 descriptions to improve their facilities of content description. On the other hand, important differences between the standards involve technical tools and applications (as we'll show later). This means that we can use MPEG-7 independently of the other MPEG standards—for example, conceptually a description might even be attached to a celluloid film.

However, the major difference between the previous standards and MPEG-7 actually concerns human nature. MPEG-7 must reconcile the approaches that the different communities—such as database and signal processing—favor.

The database world and others who need high-level descriptions typically believe that MPEG-7 only needs to provide standardized structure and linking mechanisms. The signal processing community, which has primarily focused on image analysis, sees success in only standardizing the representation of the content (standardizing features). Thus the different technical insights, and the different ways of formulating the challenge presented by MPEG-7 have caused the most difficulty within MPEG-7.

Next, we provide a brief overview on the kind of applications MPEG-7 is addressing.

MPEG-7 applications

The increased volume of audio-visual data available in our everyday lives requires effective multimedia systems that make it possible to access, interact, and display complex and inhomogeneous information. Such needs relate to important social and economic issues. Plus they're imperative in various cases of professional and consumer applications such as

- education
- journalism (for example, searching for speeches of a certain politician using his name, voice, or face)
- tourist information
- cultural services (history museums, art galleries, and so on)
- entertainment (searching for games, karaoke)
- investigation services (human characteristics recognition, forensics)
- geographical information systems
- remote sensing (cartography, ecology, natural resources management)
- surveillance (traffic control, surface transportation, nondestructive testing in hostile environment)
- biomedical applications
- shopping (searching for clothes you like)
- architecture, real estate, and interior design
- social applications (such as dating services)
- film, video, and radio archives

Describing all these applications, the application-specific requirements for content description, the requirements that the application places on MPEG-7, and pointing to some relevant work and references for an application exceeds the scope of this article. However, we refer you to the MPEG-7 Applications Document,[4] which provides this information.

Here we'll outline only a few applications to

provide a better understanding of what MPEG-7 should be and what functionality it should deliver. This is not to imply an ordering or priority of applications, it simply reflects our interest in video or audio. Therefore, we'll first investigate aspects of storage and retrieval in audio-visual databases (indexing and retrieval). Then we'll investigate applications that follow a paradigm more akin to broadcasting and Webcasting (selecting and filtering). Finally, we'll present some aspects for specialized professional applications.

Making audio-visual material as searchable as text

MPEG-7 began as a scheme for making audio-visual material as searchable as text is today. Indeed, it's conceivable that the structure and discipline to even minimally describe multimedia may exceed the current state of textual information retrieval. Although the proposed multimedia content descriptions now serve as much more than search applications, they remain the primary applications for MPEG-7. These retrieval applications involve databases, audio-visual archives, and the Web-based Internet paradigm (a client requests material from a server).

TV and film archives represent a typical application in this domain. They store vast amounts of multimedia material in several different formats (digital or analog tapes, film, CD-ROM, and so on) along with precise descriptive information (metadata) that may or may not be precisely time-coded. This metadata is stored in databases with proprietary formats. An enormous potential interest exists in an international standard format for the storage and exchange of descriptions that could ensure

- interoperability between video archive operators,
- perennial relevance of the metadata, and
- a wider diffusion of the data to the professional and general public.

To support these goals, MPEG-7 must accommodate visual and other searches of such existing multimedia databases. In addition, a vast amount of the older, analog audio-visual material will be digitized in years to come. This creates a tremendous opportunity to include content-based indexing features (extractable during the digitization/compression process[5]) into those existing databases.

For new audio-visual material, the ability to associate descriptive information within video streams at various stages of video production can dramatically improve the quality and productivity of manual, controlled vocabulary annotation of video data in a video archive. For example, preproduction and postproduction scripts, information captured or annotated during shooting, and postproduction edit lists would be very useful in the retrieval and reuse of archival material.[6]

MPEG-7's specific requirements for such applications include

- Full-text descriptions as well as structured fields (database descriptions).
- A mechanism by which different MPEG-7 descriptions can support the ability to interoperate between different content-description semantics (such as different database schemas, different thesauri, and so on).
- A robust linking mechanism that allows referencing audio-visual objects or object instances and time references (including descriptions with incomplete or missing time references) even in an analog format.
- A structure to handle multiple versions of the same document at several stages in the production process and descriptions that apply to multiple copies of the same material.

For audio databases we face a similar situation. The consumer music industry is currently struggling with how to reach consumers with increasingly fragmented tastes. Music, as with all broadcast media artifacts, is undergoing the same Internet-flavored transformation as cable TV. An ideal way of presenting consumers with available music is to let them search effortlessly for what they want. Searchers may hum approximate renditions of the song they seek from a kiosk or from the comfort of their own home.[7] Alternately, they may seek out music with features (musicians, style, tempo, year of creation) similar to those they already know. From there, they can listen to an appropriate sample (and perhaps view associated information such as lyrics or a video) and buy the music on the spot. The requirements for such types of audio-oriented applications on MPEG-7 include

- A mechanism that supports melody and other musical features that allow for reasonable

errors by the indexer to accommodate query-by-humming.

- A mechanism that supports descriptors based on information associated with the data (such as textual data).

- Support description schemes that contain descriptors of visual, audio, and/or other features, and support links between the different media (cross-modal).

Other interesting applications related to audio include sound effects libraries, historical speech databases, and movie scene retrieval by memorable auditory events.

Supporting push and pull information acquisition methods

Filtering is essentially the converse of search. Search involves the "pull" of information, while filtering implies information "push." Search requests the inclusion of information, while filtering excludes data. Both pursuits benefit strongly from the same sort of meta-information. Typical domains for such applications include broadcasting and the emerging Webcasting. These domains have very distinct requirements, generally dealing with streamed descriptions rather than static descriptions stored on databases.

User-agent-driven media selection and filtering in a broadcasting environment has particular interest for MPEG-7. This approach lets users select information more appropriate to their uses and desires from a broadcast stream of 500 channels, using the same meta-information as that used in search. Moreover, this application gives rise to several subtypes, primarily divided among types of users. A consumer-oriented selection leads to personalized audio-visual programs, for example. This can go much farther than typical video-on-demand in collecting personally relevant news programs, for example. A content-producer-oriented selection made on the segment or shot level is a way of collecting raw material from archives. The requirements for such types of applications on MPEG-7 include

- Support for descriptors and description schemes that allow multiple languages.

- A mechanism by which a media object may be represented by a set of concepts that may depend on locality or language.

- Support efficient interactive response times.

However, new ways of automating and streamlining the presentation of that data also requires selecting and filtering. A system that combines knowledge about the context, user, application, and design principles with knowledge about the information to be displayed can accomplish this.[8] Through clever application of that knowledge, you can have an intelligent multimedia presentation system. For MPEG, this requires a mechanism by which to

- encode contextual information and

- represent temporal relationships.

Finally, selecting and filtering facilitates accessibility to information for all users, especially those who suffer from one or several disabilities such as visual, auditory, motor, or cognitive disabilities. (For more information visit http://www.yuri.org/webable/library.html#guidelinesandstandards/, http://www.cogsci.ed.ac.uk/, or http://www.hcrc.ed.ac.uk/.) Providing active information representations might help overcome such problems. The key issue is to allow multimodal communication to present information optimized for individual users' abilities. Consider, for example, a search agent that doesn't exclude images as an information resource for the blind, but rather makes the MPEG-7 metadata available. Aided by that metadata, sonification (auditory display) or haptic display becomes possible. Similarity of metadata helps provide a set of information in different modalities, in case the user can't access the particular information. Thus, MPEG-7 must support descriptions that contain descriptors of visual, audio, and/or other features.

Enabling nontraditional control of information

The following potential MPEG-7 applications don't limit themselves to traditional, media-oriented, multimedia content, but are functional within the metacontent representation in development under MPEG-7. Interestingly, they're neither push nor pull, but reflect a certain amount of control over information through metadata. These applications reach into such diverse, but data-intensive domains as medicine and remote sensing. Such applications can only increase the usefulness and reach of this proposed international standard.

One of the specific applications is semi-

automated video editing.[9] Assuming that sufficient information exists about the content and structure of a multimedia object (see the previous section), a smart multimedia clip could start to edit itself in a manner appropriate to its neighboring multimedia. For example, a piece of music and a video clip from different sources could combine in such a way that the music stretches and contracts to synchronize with specific hit points in the video, creating an appropriate customized soundtrack.

This could be a new paradigm for multimedia, adding a method layer on top of MPEG-7's representation layer. (We by no means suggest that such methods for interaction be standardized in MPEG-7. As with many other advanced capabilities building on the standard, it's an issue for implementers to address.) Making multimedia aware to an extent opens access to novice users and increases productivity for experts. Such hidden intelligence on the part of the data itself shifts multimedia editing from direct manipulation to loose management of data.

Semi-automated multimedia editing encompasses a broad category of applications. It can aid home users as well as experts in studios through varying amounts of guidance or assistance through the process. In its simpler version, assisted editing can consist of an MPEG-7-enabled browser for selecting video shots, using a suitable shot description language. In an intermediate version, assisted editing could include planning—proposing shot selections and edit points—thereby satisfying a scenario expressed in a sequence description language.

The education domain relates closely to semi-automated editing. The challenge of using multimedia in educational software lies in exploiting the intrinsic information as much as possible to support different pedagogical approaches such as summarization, question answering, or detection of and reaction to misunderstanding or nonunderstanding.[10] By providing direct access to short video sequences within a large database, MPEG-7 can promote the use of audio, video, and film archive material in higher education in many areas, including

- *History*: Radio, TV, and film provide detailed accounts of many contemporary events and prove useful for classroom presentations, provided that a sufficiently precise (MPEG-7) description can be queried based on dates, places, personalities, and so on.
- *Performing arts* (music, theater): Fine-grained, standardized descriptions can bring a selection of relevant documents into the classroom for special classes, using online video archives as opposed to costly local tape libraries. For instance, users can compare several productions of a theatrical scene or musical work.[11] Because classic and contemporary theater are widely available in translation, this application can target worldwide audiences.
- *Film music*: The right tool can improve the knowledge and skills of users in the domain of film theory/practice and film music (music for film genres).[12] Depending on the user's background, the system should provide enough material to improve the user's ability to understand the complexity of each medium and also handle the complex relationships between the two media. To achieve this, the system should offer an environment in which the student can perform guided or supported experiments such as editing film, mixing sound, or combining both, which requires that the system analyze and criticize the user's results. Thus, this system must automatically generate film/sound sequences and their synchronization.

The resulting requirements for MPEG-7 for these scenarios include

- A mechanism by which descriptions may link to external information such as Hypertext Markup Language (HTML), Standardized Markup Language (SGML), World Wide Web services, and so on.
- Support for interoperation with descriptions.
- The ability to allow real-time operation in conjunction with a database.
- Pointers as handles that refer to the data directly, to let users manipulate the multimedia.

We'd like to end this brief overview on MPEG-7 applications with surveillance applications.[13] Here, a camera monitors sensitive areas, and the system must trigger an action if some event occurs. The system may build its database from no information or limited information, and accumulate a video database and metadata as time elapses. Metacontent extraction (at an encoder site)

and metadata exploitation (at a decoder site) should employ the same database. When the database becomes sufficiently large, the system at both sides should have the ability to support operations on the database such as

- Search on the audio/video database for a specific event (synthetic or current data), which represents a sequence of audio/video data.
- Find similar events in the past.
- Make decisions on the current data related to the accumulated database and/or to a priori known data.

A related application comes from security and forensics—matching faces or fingerprints. The requirements that this type of application puts on MPEG-7 include

- Real-time operation in conjunction with a database.
- Support for descriptors for unique data types.

Having reviewed the applicability of MPEG-7, we'll next discuss its goals.

MPEG-7 goals

MPEG-7 aims to

- describe multimedia content,
- manage data flexibly, and
- globalize data resources.

We'll discuss the issues these goals represent in further detail.

Multimedia content description

MPEG-7's most important goal is to provide a set of methods and tools for the different classes of multimedia content description. When we discuss description classes,[14] we actually mean different possible aspects that a description of audio-visual content might cover. A key concept to remember is that many different ways exist to describe any entity, depending on how it will be used. Thus, MPEG-7 must accommodate these several methods and make them complementary rather than mutually exclusive.

Four fundamental description classes relate to the data—that is, the material to be described—and not so much to each other. Transcriptive, physical, perceptual, and medium-based descriptions represent largely independent views of the data. On top of these schemes lies an architectural description that draws relationships between large sections of the data and relationships between and within the description(s) below it. The annotative description, a home for human annotation and other sorts of commentary on the data itself, sits on top of all the layers and touches each of them.

Most likely, any real-life description for use in MPEG-7 applications would employ only one or two of these classes. We now discuss in more detail the different possible types of description that may exist.

- *Medium-based*: We need to describe the medium in which the data is expressed. What occurred in the translation from scene to video? What's the sampling rate of the digital file? Is it an analog source? Where are the shot cuts? What's the camera's focal length? A medium-based description can address these sorts of low-level, surface features that describe the recording/playback medium itself.[15]

  Many techniques exist for obtaining some of the descriptions suggested above through image or sound analysis. Other features are simple forms of metadata such as frame rate. Encoding these descriptions would be easiest during the content-creation process.

- *Physical*: Potentially grouped with the perceptual class, the physical approach may cover all the computational features that don't correspond to human perception. Practically speaking, we can derive these features easily from the raw multimedia data. They have unambiguous values and well-established algorithms for derivation. Examples include "level" or "power" (as opposed to a perceptual "loudness") and "frequency" (as opposed to a perceptual "pitch").

- *Perceptual*: The perceptual view segments the media into objects, that is, descriptions of such characteristics as color, texture, and timbre. Recognizing patterns or classes/clusters are mainstays of today's technology. The objects describe the image or sound itself and don't generally address the structure of the scene recorded (or created) within the data.

- *Transcription*: This description class typically represents a reconstruction (or transcoding) of the world's structure as captured by the data. In music, this class was intended to serve as the actual transcription of the music—that is, the notes played (or more likely, intended by the composer). This class is the natural home for dialog transcripts and their links into the matching audio-visual material. For visual material, it may be akin to the 3D positions and characteristics of visible objects or a meticulous storyboard.

  Naturally, capturing that richness of structure from simple video or audio (without any other information) exceeds today's technology. Still, that doesn't diminish the usefulness of such a description class. There's always more than one way of obtaining a description. Smart cameras may use camera parameters (such as motion or lens) to help solve the structure-from-image problem. Musical listening systems can't transcribe automatically (other than in the most simple cases), but the majority of western classical music is scored, and many of these scores exist electronically.

- *Architectural*: Above the three previously mentioned classes lies the architectural description class. It describes the structural elements of the other classes and therefore the data they describe. This class represents the domain of document structuralists, in which a user only wants to know the relationship between segments of a document and doesn't need to know what these segments contain.[16] This class is also the domain of so-called syntactic structures, which must necessarily build on lower level semantics.

  This kind of description can be obtained in many ways. It may be hand-generated, a byproduct of an automatic analysis, or implicit in a literal transcript and merely serve to make those relationships explicit.

- *Annotative*: This description class sits atop all other classes, as well as the data itself. It's the domain of human annotation and other (typically) human analysis of the existing metadata. In its original incarnation, in music, it was the place for musicologists to comment on various other features of a piece of music or recording, such as general aspects of musicological analysis. Common subjects for musicology include musical forms (referring to the architecture), the notes themselves (linking to the transcript), commenting on the emotional content of the music (linking to the data itself), and relationships to other pieces (linking to other descriptions).

MPEG-7 tries to provide the formalisms to support the requirements for the different description classes. We should note that although the apparent complexity may seem daunting, the diversity of description levels allows flexible and expressive ways to represent adequately the content by some formal structure.

To reuse MPEG-7 descriptions efficiently, users will need to adapt them to their specific needs. This leads to modifying and manipulating existing structures. Manipulating document structure will benefit operations that make the traversal and manipulation of trees, linked lists, and webs natural—either to prune or reorganize the structural framework or to transform the values stored in some nodes to a more user-friendly representation. To avoid multiplying ad hoc solutions requires a generic way of defining structure transformation and manipulation.

We anticipate, though, that the underlying data structures and their composition will remain independent from the applied extraction mechanisms. In other words, MPEG-7 structures provide an application-independent description framework that extraction mechanisms can instantiate.

Whichever features describe an audio-visual document will either be extracted automatically by an algorithm running on a computer or annotated by a human expert. To perform such a task automatically requires a formal specification of the extracted entity or feature. This specification might be atomic or might represent the weighted sum or some other derivation of a number of features. Examples for such features from music are timbre or density; in the visual domains it might be the composition of an image. Finally, since multimedia content builds on temporal and spatial constraints—that is, presentational constraints—it's obvious that spatial and temporal requirements influence a description's semantic and syntactic structure.

Flexibility in data management

The aspect of flexibility described for combining features and creating documents leads directly to the second important concern of the MPEG-7 group, flexibility in data management. MPEG-7 aims to provide a framework that allows

references to parts of a document, to a whole document, and to a series of documents. The question of multimodality relates to this flexibility. This means it should be possible to describe multimedia content in such a way as to allow queries based on visual descriptions to retrieve audio data and vice versa. Figure 2 provides an example of different descriptions such as the score, storyboard, and compositional structure, and their relationship to the film "Alexander Nevsky" from Sergei Eisenstein. Note that the synchronization is achieved here by using time codes (represented in Figure 2 using spatial ordering) as pointers from the different types of descriptions to the actual data (the video).

Flexibility also helps ensure the longevity of the standard, which means that description schemes designed for a given task should be easily modifiable for different but related applications. MPEG-7 will address applications that can be stored (online or offline) or streamed (for example, broadcast, push models on the Internet) and can operate in both real-time and non-real-time environments.

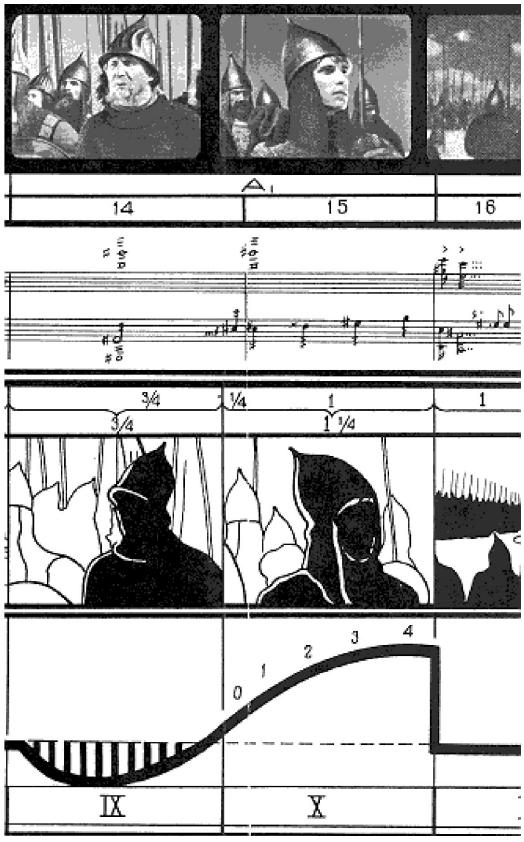

Figure 2. Relationship among different content descriptions.[17]

Globalization of data resources

MPEG-7's third goal aims to support the globalization of data resources. MPEG-7 descriptions may be physically located with the associated audio-visual material, in the same data stream, or on the same storage system, but the descriptions could also live somewhere else. When the content and its descriptions aren't co-located, mechanisms that link audio-visual material and their MPEG-7 descriptions prove useful. These links should work in both directions. The combination of flexibility and globalization of data resources allows humans as well as machines—in the form of agents—to exchange, retrieve, and reuse relevant material. An agent serves as an autonomous computational system acting in computer networks or in a computer based on a set of goals it tries to achieve.[18]

This final issue leads to the foremost goal of MPEG-7: to provide a means to allow the interoperability of content descriptions. It's absolutely essential that MPEG-7 exist to serve this goal. Standardization chiefly seeks to reach beyond any single, proprietary solution and provide not only a framework, but the concrete means by which industry solutions may work together. If MPEG-7 becomes too generic or focused on one application, it will fail in this respect.

MPEG-7 doesn't extract descriptions/features automatically. Nor does it specify the search engine (or any other programs such as audio-visual content recognition tools or tools to generate the description) that can use the description. Those are outside the scope of the planned standard. Rather, MPEG-7 will concentrate on standardizing a representation that can be used for description. In developing the standard, however, MPEG might build some coding tools, just as it did with the predecessors of MPEG-7, namely MPEG-1, -2, and -4. Also, for these standards, coding tools were built for research purposes, but they didn't become part of the standard itself.

Note that while MPEG-7 aims to standardize a Multimedia Content Description Interface, MPEG emphasizes audio-visual content. That is, MPEG-7 doesn't aim to create description schemes or descriptors for text. However, MPEG-7 will consider existing solutions for describing text documents such as SGML and it's derivations like the Extensible Markup Language (XML), Resource Description Framework (RDF), and so on, and support them with suitable, necessary interfaces between audio-visual content descriptions and the textual-content descriptions.

As a compatible extension for content description to MPEG-4, MPEG-7 will consider MPEG-4 Object Content Identification (OCI), an MPEG-4-

specific solution for providing limited amounts of information about MPEG-4 content as a subset of MPEG-7.

The MPEG-7 working group recognizes the fact that other standards for the description of multimedia content are under development. Thus, they'll consider other standardization activities such as the Society for Motion Picture and Television Engineers/European Broadcasting Union (SMPTE/EBU) task force, Digital Video Broadcasting-Service Information (DVB-SI), European Committee for Standardization/ Information Society Standardization System (CEN/ISSS), and so on. For more details regarding goals of MPEG-7 please see "MPEG-7: Context and Objectives Document."[19]

Having introduced the basic direction MPEG-7 intends to take here, we'll introduce the terminology and provide a more in-depth discussion of MPEG-7 in the next issue of *IEEE MultiMedia*. MM

Acknowledgments

Some of the achievements described in this article were carried out within the Advanced Communication Technologies and Services (ACTS) project Distributed Internet Content Exchange with MPEG-7 Agent Negotiations (Diceman). The work was partly funded by the European Commission. The views expressed are those of the authors and should not be taken to represent, in any way, the views of the European Commission or its services.

References

1. D.L. Gall, "MPEG: A Video Compression Standard for Multimedia Applications," *Comm. of the ACM*, Vol. 34, No. 4, 1991, pp. 46-58.
2. R. Steinmetz, *Multimedia Technologie: Grundlagen, Komponenten und Systeme (Multimedia Technology: Computing, Communications, and Applications)*, Springer-Verlag, Berlin, 1998 (in German).
3. U. Reimers, ed. *Digitale Fernsehtechnik: Datenkompression und Übertragung für DVB (Digital TV Technology: Data Compression and Transmission for Digital Video Broadcasting)*, Springer-Verlag, Berlin, 1997 (in German).
4. *ISO/MPEG N2728, Applications for MPEG-7*, MPEG Requirements Group, Int'l Organization for Standardization, March 1999, http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/N2728.html.
5. M. Davis, "Media Streams: An Ionic Visual Language for Video Annotation," *Telektronikk*, Vol. 89, No. 4, 1993, pp. 59-71.
6. F. Nack, and A. Steinmetz, "Approaches on Intelligent Video Production," *Proc. of European Conf. on Artificial Intelligence (ECAI 98) Workshop on AI/ALife and Entertainment*, E. André et al., eds. European Coordinating Council for Articial Intelligence, Brighton, UK, Aug. 1998, pp. 48-56.
7. A. Lindsay, *Using Contour as a Mid-Level Representation of Melody*, master's thesis, Massachusetts Institute of Technology, MIT Media Lab, Cambridge, Mass., 1996, http://sound.media.mit.edu/~alindsay/thesis.html.
8. E. Andre and T. Rist, "Generating Coherent Presentations Employing Textual and Visual Material," *Artificial Intelligence Review* (Special Volume on the Integration of Natural Language and Vision Processing), Vol. 9, No. 2-3, 1995, pp. 147-165.
9. F. Nack and A. Parkes, "The Application of Video Semantics and Theme Representation in Automated Video Editing," *MultiMedia Tools and Applications*, Vol. 4, No. 1, 1997, pp. 57-83.
10. R.C. Schank, "Active Learning through Multimedia," *IEEE MultiMedia*, Vol. 1, No. 1, Spring 1994, pp. 69-78.
11. P. Tagg, "Musicology and the Semiotics of Popular Music," *Semiotica*, Vol. 66, No. 1/3, 1987, pp. 279-298.
12. P. Tagg, ed., *Film Music, Mood Music, and Popular Music Research: Interviews, Conversations, Entretiens*, SPGUMD 8002, Göteborg Stencilled Papers, Musicology Dept., Göteborg University, Göteborg, Sweden, 1980.
13. J.D. Courtney, "Automatic Video Indexing by Object Motion Analysis," *Pattern Recognition*, Vol. 30, No. 4, 1997, pp. 607-626.
14. A. Lindsay and W. Kriechbaum, "MPEG-7 for Music and Music for MPEG-7," presented at ACM Multimedia workshop on Content Processing for Music, Sept. 1998, http://www.acm.org/sigmm/MM98/aigrain.html.
15. E. Hartley, F. Nack, and A. Parkes, "Proposed Revision of w1734 Terminology," Doc. ISO/MPEG M2808, MPEG Fribourg Meeting, Int'l Organization for Standardization, October 1997.
16. W. Kriechbaum, "Some Remarks on Document Structure and Description Schemes," Doc. ISO/MPEG M3649, MPEG Dublin Meeting, ISO, July 1998.
17. S.M. Eisenstein, *Selected Works: Towards a Theory of Montage*, BFI Publishing, London, UK, 1991.
18. P. Maes, "Modeling Adaptive Autonomous Agents," *J. of Artificial Life*, Vol. 1, No. 1/2, 1994, http://pattie.www.media.mit.edu/people/pattie/cv.html#Refereedjournals.
19. *ISO/MPEG N2729, MPEG-7 Context & Objectives*, MPEG Requirements Group, ISO, Geneva, March 1999, http://www.darmstadt.gmd.de/mobile/MPEG7/Documents/N2729.html.

Frank Nack is a member of the Mobile Interactive Media (Mobile) division at the German National Research Center for Information Technology-Integrated Publication and Information Systems Institute (GMD-IPSI). His research interests include video representation, digital video production, interactive storytelling, and media-networked-oriented agent technology. He earned his PhD in computing at Lancaster University, UK. He is an active member of the MPEG-7 standardization group, where he served as editor of the *Context and Objectives Document* and the *Requirements Document*. He now chairs the MPEG-7 Description Definition Language (DDL) development group.

Adam Lindsay is the principal investigator of the multimedia research division at Starlab, a research firm based in Brussels, Belgium. He joined Starlab after earning his MS in mid-level representation of musical melody at the Massachusetts Institute of Technology Media Lab, Cambridge, Massachusetts. His current research is on applying MPEG-7-style metadata to multimedia to make it more intelligent about itself. He is an active member of the MPEG-7 standardization group, where he served as the editor of the MPEG-7 *Applications Document* and as the leader in MPEG-7 Audio activities.

Readers may contact Nack at GMD-IPSI Dolivostr. 15, 64293 Darmstadt, Germany, e-mail Frank.Nack@darmstadt.gmd.de. Contact Lindsay at Starlab, Escelsiorlaan 40-42, B-1930, Zaventem, Belgium, e-mail adam@starlab.net.

Get Wireless

Cell phone for voice communication have become so commonplace that everyone from grandmothers concerned for their safety to high school students with busy social lives are linked in to local satellite networks. But while competing data devices—laptop computers, PDAs—abound, the push for wireless data transfer has lagged years behind. The July-August issue of *IT Professional*, the Computer Society's bimonthly resource offering technology solutions for the enterprise, spotlights an article on mobile technologies. "Get Wireless: A Mobile Technology Spectrum" by Prathima Agrawal and Cormac J. Sreenan (pp.18-23) describes the trade-offs between device complexity and network dependence, examining the state of technology and networks today and projecting what might be possible in the near and distant future.

A related article ("New Protection from Bad Mobile Code," p. 10) in the same issue describes new software that can scan mobile code and make sure it's bug-free before it reaches its destination. With the expected e-commerce boom, using filtering technology at the network level to detect threats such as the Melissa virus and ExploreZip worm will become increasingly important to help IT managers check content and neutralize bad code at the network level before it can execute.

To read these and other selected articles from *IT Pro* online, visit http://computer.org/itpro/. To subscribe to *IT Pro* or any of the Computer Society's magazines and transactions, visit http://computer.org/subscribe/index.htm.

IT Professional