MPEG-4: Way Beyond Video

New Standard Melds Multiple Media, Low Data Rates, Interactivity

by Peter N. Glaskowsky

Though today's MPEG-1 and MPEG-2 standards are widely used in personal computers and consumer electronics, these standards are limited to video and audio data. The MPEG-4 standard, ratified last October, integrates these basic data types, plus many more, into a sophisticated new model for multimedia communication.

In addition to ordinary video and audio data, MPEG-4 supports application-specific forms of video, such as videoconferencing, optimized audio-compression schemes for speech and music, 2D and 3D graphics, and methods to produce animated bodies and faces to create virtual actors.

An MPEG-4 bitstream contains definitions of objects as well as the rules by which those objects may be combined and presented. In addition to defining the content itself, an MPEG-4 program can define algorithms for music synthesis, using a new high-level music-description language. This algorithmic approach may be extended in the next version of MPEG-4, due by the end of the year, to cover other media.

The copyright status and ownership of MPEG-4 content can be defined in the bitstream, and the standard includes an extension mechanism to allow the use of encryption techniques to protect MPEG-4 content. The encryption algorithms are not predefined, restricting their use to applications where the content source and consumer have agreed to a specific cryptographic solution.

With so many data types and algorithms contained in the MPEG-4 standard, and others allowed by extensions, fully hardwired MPEG-4 codecs are unlikely. Though some fixed-function devices already exist to handle specific algorithms, media processors are likely to be the most successful MPEG-4 engines. Programmable devices such as Equator's MAP1000 (see MPR 12/7/98, p. 1) offer, at a reasonable cost, the performance and flexibility that MPEG-4 requires.

MPEG-4 Effort Is International in Scope

The MPEG-4 standard is being developed by the Motion Pictures Experts Group (MPEG) working group within the International Standards Organization (ISO). The MPEG effort (*www.cselt.it/mpeg*) includes representatives from more than 200 companies in 20 nations, making it a truly international effort.

Like most large committee-based standards, MPEG-4 has been slow to develop. The work began in 1993 with the goal of combining audio, video, and graphical data from both natural and synthetic sources. This was more easily said than done, as evidenced by the plethora of techniques in MPEG-4, but the goal was achieved.

The MPEG-4 standard is based on a new scheme for describing multimedia content. The standard's binary format for scenes (BIFS) allows the content creator to define multiple objects and coordinate spaces, then to define the relationships among these entities and how they are to be combined for presentation. The key goal of BIFS was to communicate all this information using the minimum bandwidth.

The presentation may include any or all of the basic MPEG-4 object types, and objects may be altered after creation. Some objects can receive input from the user, and this input can be used to alter the content stream. These features allow highly interactive presentations. Such presentations can range from simple ones—objects that move with mouse clicks—to complete applications such as 3D games. Because of its complex nature and multiple levels of abstraction, however, MPEG-4 is unlikely ever to compete with conventional high-level computer languages as a game-development environment. This flexibility could come in handy for designing platform-independent user interfaces, however.

Various Video Options Offered

MPEG-4 supports most of the features found in MPEG-1 and MPEG-2, including video compression based on the discrete-cosine-transform (DCT) and motion-compensation algorithms used in the earlier standards. MPEG-4's video capabilities surpass those of MPEG-2 in two major ways. First, MPEG-4 includes support for very low bit-rate video (VLBV), where video is compressed to achieve bit rates from 5 to 64 Kbits/s for low-resolution, low-frame-rate content.

Even at these low data rates, MPEG-4's improved compression techniques and object-oriented architecture will allow decent video quality, substantially better than that found on today's videoconferencing systems running at similar or faster data rates. The same algorithms may be used to encode broadcast-quality video at higher data rates, typically up to 4 Mbits/s. The working group is currently evaluating ways to provide high-definition video at still higher data rates.

The second major advance over MPEG-2 is support for nonrectangular video images. Where MPEG-2 works only on complete video frames, MPEG-4 can encode and decode video representations of individual objects, then define how these objects are overlaid to create the complete frame. One classic example is a TV weatherman. When encoded in MPEG-2, a weatherman standing in front of a weather map is encoded as a single video image. In MPEG-4, the image of the weatherman can be encoded in one stream and the background image encoded as a still image in another stream. The static background image must be sent only once; only the moving weatherman requires motion video. The two streams are transmitted in a single TransMux stream (described later) and combined by the MPEG-4 decoder.

The new standard's support for nonrectangular images is achieved by specifying a mask that defines the outline of the image, an outline that can change from frame to frame as the object changes shape. The mask can use one-bit values to identify transparent and opaque pixels, or it can define eight bits of transparency for each pixel. The transparency option allows even more sophisticated compositing of translucent objects such as smoke effects and overlaid logos.

To implement this nonrectangular coding option, MPEG-4 defines a new shape-adaptive DCT algorithm. Although the DCT algorithm used in MPEG-2 works only on 8 × 8-pixel blocks, MPEG-4's enhanced DCT works on blocks of any shape.

New Graphics Technology Good To See

MPEG-4's ability to represent visual information goes far beyond that of MPEG-2, with the addition of several sophisticated techniques for coding synthetic objects—those composed from 2D and 3D graphics. Ordinary 2D still images can be coded using the DCT algorithm or a new wavelet-based technique. Wavelet image compression is very efficient and scalable, making it a natural choice for MPEG-4.

The standard's support for 3D graphics is similar to that of the Virtual Reality Modeling Language (VRML) used on the Internet, but it differs in minor ways—most notably in MPEG-4's need for real-time operation. Like VRML, MPEG-4's 3D features allow content creators to define 3D objects and specify how these objects are combined to create 3D scenes. The MPEG working group and the Web3D Consortium (*www.web3d.org*), owners of the VRML specification, are working to reconcile differences and establish formal mapping functions from one standard to the other.

Only conventional triangle-based 3D models are supported by MPEG-4. The working group chose not to adopt higher-level geometric modeling techniques such as non-uniform rational B-splines (NURBS) or constructive solid geometry (CSG). These techniques would have permitted more compact representations of some 3D objects—particularly those composed of curved surfaces—but they would also have required more computing power in the decoder.

MPEG-4 also provides specific support for 3D modeling of the human body. The standard defines a model for facial animation that begins with a generic face with a neutral expression. The model includes parameters that can be altered to change the basic appearance of this default face. Other parameters allow the face to produce expressions and simulate the facial movements that accompany speech. For example, MPEG-4 defines an "open_jaw" function that takes a parameter to specify how far open the jaw should be. The next version of the MPEG-4 standard will extend these concepts to models of the complete body.

The standard also provides a method for mapping an animated video texture onto 3D objects. This technique allows a video image of a real person's face to be mapped onto a 3D model of a human head, which itself could be driven by MPEG-4's facial-animation scheme. This is a complex way to depict a talking head, but it may require less bandwidth than the conventional approach and allow more flexibility in playback, including inherent support for multiple viewpoints.

The 2D and 3D elements of a program can be combined with video streams, using a full 3D-coordinate model. In some programs, 3D objects can be presented in front of 2D or video background images. In others, a 2D or video image can be mapped onto a TV screen in a fully synthetic model of a room.

Each video and graphic element in an MPEG-4 image can be encoded or rendered separately. For example, important foreground characters can be displayed in full-resolution video, with background objects rendered at lower resolution, to save bandwidth or processing power.

New Audio Alternatives Worth a Listen

Audio coding in MPEG-4 is also well advanced beyond previous MPEG standards. MPEG-4 supports both natural and synthetic audio coding. Natural audio can be encoded at rates of 6–24 Kbits/s to achieve quality superior to that of typical AM-radio broadcasts. Lower and higher quality levels are supported using several different audio algorithms.

Two codecs optimized for speech coding are included in the standard: harmonic vector excitation coding (HVXC) at fixed rates of 2–4 Kbits/s and code excited linear predictive (CELP) for rates of 4–24 Kbits/s. Using a variable bit-rate mode, HVXC can also operate at an average rate of just 1.2 Kbits/s.

General audio coding is handled by TwinVQ, a vector-quantization algorithm, or the advanced audio coding (AAC) algorithm from MPEG-2. As with speech, the content creator selects the algorithm according to the quality and bit rate required by the program material.

Even lower data rates for speech can be achieved through a text-to-speech (TTS) algorithm. Instead of digitizing and compressing the sound of a person's voice and re-creating it at the other end, TTS sends text over the communication channel and synthesizes the spoken words in the decoder. The MPEG-4 implementation allows the synthetic voice to be defined in terms of parameters such as pitch and speed. The text can be annotated with additional parameters to make the result more intelligible and convey control functions such as synchronized facial animation. The TTS bitstream can even be used to create an animated model of a human hand making the gestures of sign language.

For high-quality audio at low data rates, MPEG-4 defines a new music-synthesis language called SAOL (Structured Audio Orchestra Language). SAOL programs define the sound and behavior of instruments, then they define a virtual orchestra made up of these instruments. A bitstream expresses the music to be played (creating a sort of digital sheet music), and the decoder generates audio from this bitstream at any desired quality level.

MPEG-7: A Standard for Multimedia Content Description

It's one thing to be able to generate and display a complex multimedia presentation such as those enabled by MPEG-4. It's entirely another matter to be able to describe the content in a presentation, especially in a way that allows a user to search a library of presentations for a specific element. The MPEG-7 standard, currently in development by the MPEG working group and planned for final adoption by July 2001, is meant to provide a comprehensive framework for describing multimedia content using multiple levels of abstraction. (MPEG-5 and -6 have never been assigned to a standards effort.)

The simplest element of an MPEG-7 content description will be a set of text keywords related to the multimedia object being described, which need not be an MPEG-4 presentation. MPEG-7 doesn't attempt to define how its descriptions are generated or used; it simply defines standard formats for those descriptions.

Some descriptive tags could be generated automatically. MPEG-4 provides a method to identify the copyright owner of each element in a program; this information could be extracted to an MPEG-7 description. It might even be possible to automatically identify actors or other recognizable elements in a presentation.

For more complex descriptions, such as credits for special-effects artists referenced to the specific scenes of a movie they worked on, automatic extraction might be impossible. In such cases, the MPEG-7 description would be entered by hand.

A photo library, once properly described, could then be searched for all pictures of 1957 Chevrolet automobiles. A collection of music could be searched for all passages performed on a Stradivarius violin. On a more mundane level, fans would be able to find every post–Star Trek television appearance of William Shatner. MPEG-7's benefits to future historians—and couch potatoes—are immeasurable.

SAOL does not describe a method of synthesis; rather, it defines a way to describe synthesis, allowing the content creator to select the synthesis method best suited to the music. Many different synthesis algorithms can be used by SAOL, including frequency modulation (FM), wavetable synthesis, and physics-based modeling.

Data Format Defines Multiple Levels of Encoding

MPEG-4 transport streams may contain several multiplexed components. Transport streams are therefore known as TransMux streams. TransMux streams are demultiplexed into one or more FlexMux streams, which in turn may contain several Elementary Streams. Each type of Elementary Stream is assigned to a specific type of decoder. The decoders render the content into a format ready for presentation.

The standard also defines how TransMux streams are transferred over a network. This portion of the standard is called the Delivery Multimedia Integration Framework (DMIF). DMIF defines the way a device requests and receives MPEG-4 bitstreams. DMIF is session oriented; that is, a device must establish a DMIF session before it can start to receive content. DMIF is designed to work with a wide variety of network interfaces. The DMIF session allows the device to identify the available protocol(s) and issue the necessary requests. DMIF is also capable of communicating quality of service (QoS) requirements between the server and client so devices will not request more than the server or network can provide.

DMIF is designed to cover three different ways to deliver MPEG-4 content: interactive networks, broadcasting, and local storage. These three alternatives can be combined in a single session. An MPEG-4 broadcast over digital cable TV, for example, could contain links to interactive content on the Internet as well as links to user-interface features preloaded on the user's hard disk.

Scalability Supports Multiple Platforms

The flexibility present in MPEG-4's multimedia and networking components allows a single presentation to be viewed on a wide range of platforms. This capability greatly exceeds that of MPEG-2. A video presentation encoded at high resolution using MPEG-2 can be decoded and displayed at lower resolutions, using various digital filtering techniques. The entire bitstream must still be transmitted to the decoder, however. With MPEG-4, the program can be encoded with a base (lower-quality) stream and one or more enhancement streams that build on the base stream to provide higher quality. Up to 3 levels of quality are supported for video images; up to 11 quality levels may be defined for still images using wavelet compression.

When multiple objects are used in a program, some of the objects may be placed in the enhancement streams. This placement means that the low-quality version of the program is different from the high-quality versions of the same program, but it enables even more dramatic scalability. Objects can be assigned to the base or enhancement streams according to a priority scheme, ensuring that critical elements of the program are seen on all playback platforms but allowing less important objects to be dropped if necessary.

A program that takes advantage of MPEG-4's scalability features can easily be decoded on low-end playback hardware using only the base stream. The playback device can also use as many of the enhancement streams as appropriate, based on the display device, available processing power, or other local resources.

If a scalable bitstream is being sent over a network to a playback device with known characteristics, the enhancement streams may be omitted by the server. Similarly, if such a program is sent over a network with limited bandwidth, the enhancement streams may be dropped to prevent network congestion.

Because enhancement streams are part of the top-level TransMux stream, only a device with knowledge of the MPEG-4 transport layer can remove enhancement streams. Current Internet routers aren't intelligent enough to handle this task, but if MPEG-4 becomes popular, future networking devices could be designed to modify TransMux streams in order to reduce or prevent network congestion.

MPEG-4 Faces Competition From Many Quarters

The breadth of the MPEG-4 standard is unique, but many alternatives exist for the individual components of the standard. Apple's QuickTime has long been able to mix audio, video, 2D, 3D, and text content in a single presentation. RealNetworks's RealPlayer supports multimedia transmission over low-bandwidth Internet connections.

Microsoft's ActiveMovie, part of the company's ActiveX multimedia API, represents another serious competitor for MPEG-4. ActiveMovie is an object-oriented multimedia presentation specification based on Microsoft's Common Object Model (COM). Virtually any multimedia algorithm can be implemented as a COM filter, and multiple algorithms can be combined to achieve the same effects as MPEG-4. Microsoft's support for ActiveMovie makes it unlikely that the company will put much effort into MPEG-4 for the PC platform.

Microsoft's Chromeffects work (see MPR 4/20/98, p. 21), which combines 2D/3D graphics with video and online content, also overlaps MPEG-4 but to a lesser extent. Chromeffects was withdrawn from the market after the release of the first software-development kit, but work continues on the technology within Microsoft.

MPEG-4's high-level 3D scene-description model is incompatible with similar high-level programming models in other 3D APIs such as OpenGL, the Java 3D API, Microsoft's Direct3D, and the forthcoming SGI/Microsoft Fahrenheit API. The 3D content-creation industry has little room for multiple incompatible APIs, a factor that is likely to reduce support for MPEG-4's 3D technology.

Version 2 of the MPEG-4 standard, which should be ratified by the end of the year, will add new tools for additional media types, multiuser interaction, and a file format based on Apple's QuickTime. Another useful feature in the new version will be support for multiple viewpoints. This feature will enable the use of stereoscopic displays and virtual-reality applications.

Version 2 will even include an MPEG-specific subset of Java. The MPEG-J language will be used to embed control and processing operations too complex to represent in the first MPEG-4 release. The MPEG working group stresses that MPEG-J will not be used to define new downloadable codecs. This suggests that MPEG-J will not be optimized for media processing. Such a use would require MPEG-J implementations to achieve some specified level of performance, a requirement that would be difficult to meet.

Healthy Patient But Uncertain Prognosis

Though the standard itself is well defined, its prospects are unclear. ISO and the MPEG working group have made no effort to determine in advance what elements of the MPEG-4 standard may be covered by patents. It is currently unclear what license fees a chip or system vendor might have to pay for the intellectual property contained in MPEG-4. Adoption of the MPEG-2 standard was slightly delayed by similar issues; the greater complexity of MPEG-4 could make for longer delays.

Authoring content for MPEG-4 will also be much more difficult than for MPEG-2. Currently, any source of video or audio can easily be converted into an MPEG-2 bitstream. Adding the extra features allowed by the DVD or digital-TV delivery options is also simple. A wide variety of free or inexpensive tools are available for generating and viewing MPEG-2 content, and these tools have helped build a broad base of MPEG-2 experience among content creators.

Although it would be possible to generate an MPEG-4 program using ordinary linear audio and video sources, this approach would not leverage MPEG-4's unique capabilities. It will be years before producers of TV shows, educational films, and computer games gain the expertise necessary to make proper use of MPEG-4. The lack of freely available tools for MPEG-4 content creation, encoding, and playback will only delay the commercial success of the standard.

The greatest threat to MPEG-4's success comes from the Internet, where Microsoft and hundreds of smaller developers are busy releasing tools and applications that accomplish many of MPEG-4's goals. None of these competing efforts is as ambitious or as comprehensive as MPEG-4, but some—particularly ActiveMovie—may be good enough to prevent MPEG-4 from getting a foothold in the PC market.

None of these alternatives is a single standard that provides all the capabilities of MPEG-4, however. Most are proprietary solutions. It seems likely that MPEG-4 will receive serious consideration from makers of consumer-electronics products, who place a high value on international standards efforts. The benefits to be gained from MPEG-4 may ultimately outweigh the costs of adopting it, but considerable work remains to be done before this evaluation can be completed.

MPEG-4 is also useful as an example of how bandwidth can be traded for computing resources. MPEG-4 allows programs to be created to suit a very wide range of communication channels. The bandwidth needed to transmit an MPEG-4 program can be reduced by doing more work when creating the program, or requiring more work to play it back, or both. As computers get faster and low-bandwidth communication channels continue to be widely used (especially wireless solutions), MPEG-4 should become a more attractive alternative to existing multimedia standards. M