The MPEG STANDARDS

PART 2

In this second part, Reg Miles looks at the latest MPEG standards - MPEG-4, MPEG-7 and the new MHEG standard

Sony Digital Mavia. Digital Movie Camera

In the first part I described the work of the Motion Picture Experts Group in developing the MPEG-1 standard and the broader and more versatile, but directly descended, MPEG-2. In this part, I chart the movement of MPEG into new areas, with MPEG-4 and MPEG-7, and also introduce the Multimedia and Hypermedia Experts Group and their MHEG standard.

MPEG-3 was absorbed into an expanded MPEG-2, thus leaving MPEG-4 as the next standard. This has the potential for very high compression ratios - up to 1000:1 in extreme cases, using object-based coding to represent audiovisual objects, and the relationship of those to other objects in a scene, as well as taking into account how a user can interact with those objects. In addition to broadcasting applications, MPEG-4 uses cover internet multimedia, interactive video games, multimedia mailing, optical discs, videoconferencing, and videophones... In fact, just about everything you can think of. And the first product has already been launched - the Sharp Internet ViewCam. This can store up to one hour of video on a 32MB SmartMedia card; which can then transferred to a computer for sending as an e-mail attachment or to be posted on a web page. MPEG-4 became a standard this year; and an MPEG-4 Version 2 with backwards compatible extensions is under development.

MPEG-4

MPEG-4 defines the coded representations of two basic types of media objects: natural - recorded with a camera or a microphone, and synthetic - graphics, text, synthesised speech and sounds. These are arranged in heirarchical fashion. At the basic level are so-called primitive media objects: static objects - anything that is not moving, whether it be an object in the scene or a background, or any part of a still digital image; video objects - anything that is moving, separated from its background; and audio objects - sounds associated with the video objects. Those are the natural, primitive media objects. The synthetic ones are text and graphics; talking synthetic heads and associated text used to synthesise the speech and animate the head; and synthetic sound. Both the natural and synthetic objects can be 2-D or 3-D, plus both natural and synthetic objects can be combined.

Natural video objects are identified by algorithms that hold the parameters to describe them in terms of their shape (rectangular or arbitrary - the latter being first determined by blocks of pixels, and then the 'boundary blocks' receiving special treatment, giving a pixel-based shape), plus colour, texture, etc., and their motion.
It takes a lot of processing but, once done, very little information is actually needed to describe those aspects from frame to frame (Figure 10). However, coding is similar to that used for MPEG-1/2, with motion prediction and compensation, followed by DCT-based texture coding (using either 8 x 8 DCT or shape adaptive DCT). In addition to block-based motion prediction and

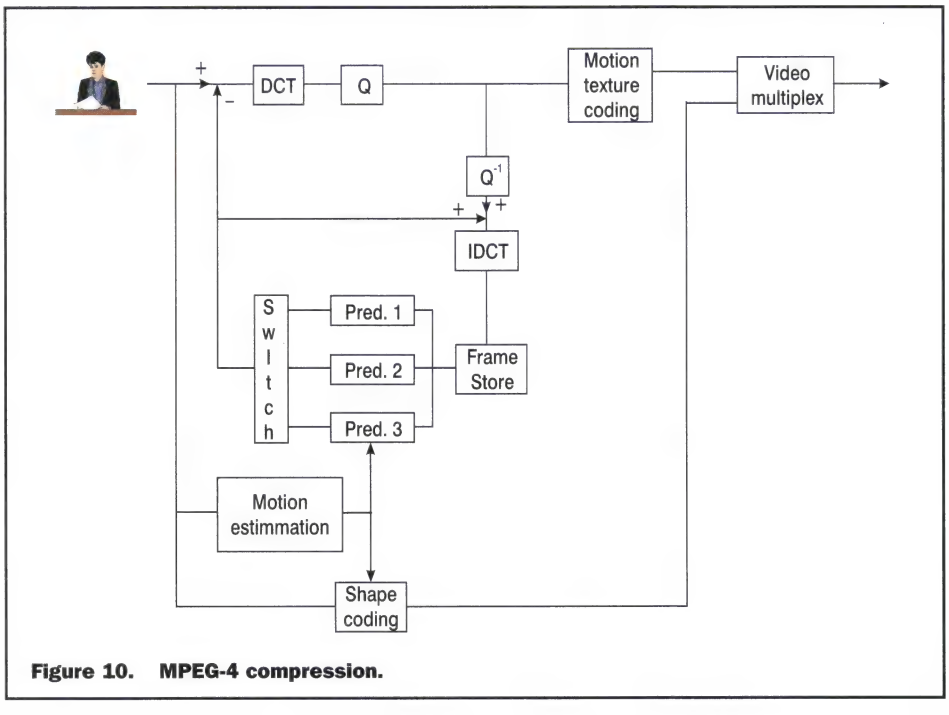


Figure 10. MPEG-4 compression.

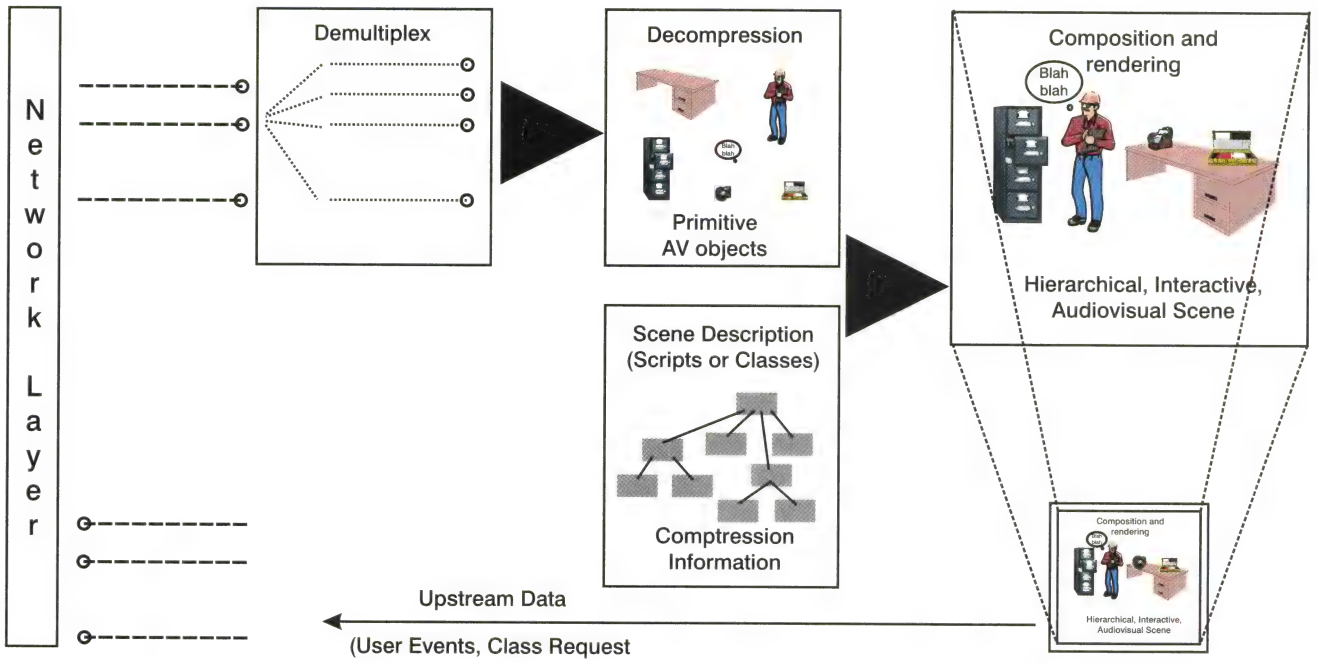


Figure 11. MPEG-4 scene description.

compensation, further compression can be achieved by using global motion compensation. This involves transmitting a static 'sprite' - a still image describing a panoramic background - to the receiver. This is then held in a sprite buffer, and only the camera motion relative to the background need be transmitted. Any moving foreground images will be coded as rectangular or arbitrary video objects and sent separately. The decoder then reconstructs the image from the sprite, the camera motions and the video object(s).

For resolution up to MPEG-1 level, combined with frame rates of up to 15Hz, very low bitrate video (VLBV) can be used - a range of about 5-64kb/s. But MPEG-4 can also be used at up to 10 Mb/s; enabling uses such as multimedia and interactive TV, with qualityequal to digital TV. MPEG-4 supports both progressive and interlaced scanning, up to CCIR 601, with sampling structures of 4:2:0 and 4:2:2 - and B&W, with pixel depths of up to 8-bits per component. And constant or variable bitrates.

Sharp Internet MPEG-4 Camera.

HVXC

The coding of natural audio objects is done by a variety of means, at bitrates ranging over 2-64kb/s - or less than 2kb/s when variable bitrate is used. Speech coding is handled by Harmonic Vector eXcitation Coding (HVXC) at 2-4kb/s; and by Code Excited Linear Predictive (CELP) for 4-24kb/s. HVXC can operate down to 1.2kb/s when using variable bitrate. CELP has two operating modes, for narrowband and wideband speech: the formeroperates up to 12kb/s, sampling at 8kHz; the latter up to 24kb/s at 16kHz. While the MPEG-2 AAC standard caters for general audio at 6-64kb/s; or TwinVQ can be used (Transform-domain Weighted Interleave Vector Quantisation), which is said to be technically superior to MP3.

Synthetic visual objects are built up using the latest techniques. A face, for example, can be constructed using shape and textures controlled by Facial Definition Parameters (FDP); and then animated by the Facial Animation Parameters (FAP) to create expressions and give the lip movements accompanying speech. And only the parameters need be transmitted. The same applies to synthesised bodies; only here it is BDP and BAP. Other objects are created by similar static and dynamic mesh coding and texture mapping.

Synthesised sounds are handled by different means depending on their type. Speech is generated from text input by a text-to-speech (TTS) coder, at bitrates from 200b/s to 1.2kb/s. Together with the means of synchronising it to facial animation, if that

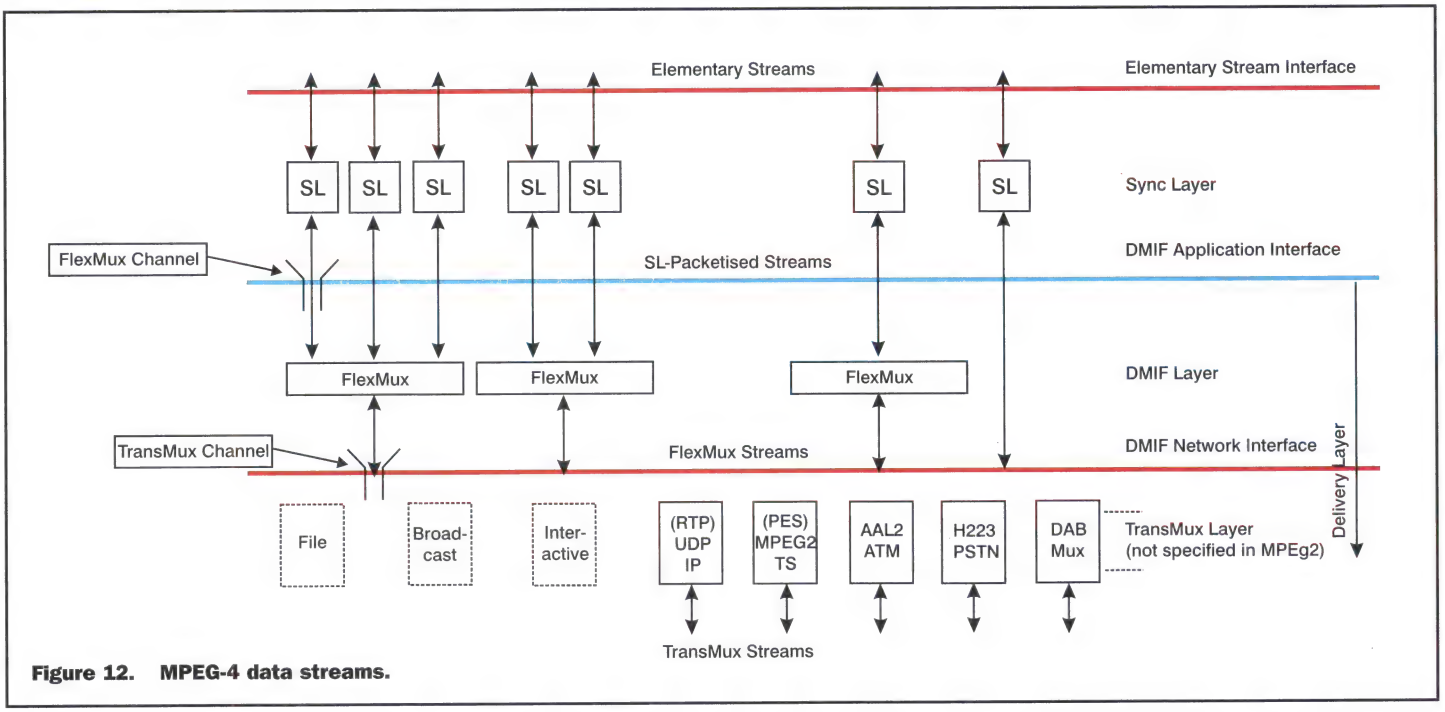


Figure 12. MPEG-4 data streams.

is required. Also, the particular language and dialect can be signalled in the bitstream. General sounds and music are produced in a synthesis language called Structured Audio Orchestra Language (SAOL - pronounced 'sail'). This defines an 'orchestra' of 'instruments' that produce the desired sounds under the control of downloaded scores in the bitstream, using the Structured Audio Score Language (SASL). It is also possible to use the MIDI protocol when such fine control is unnecesary. Or a wavetable bank format, in which sound samples are downloaded for wavetable synthesis, together with filters, reverbs, etc, for further processing.

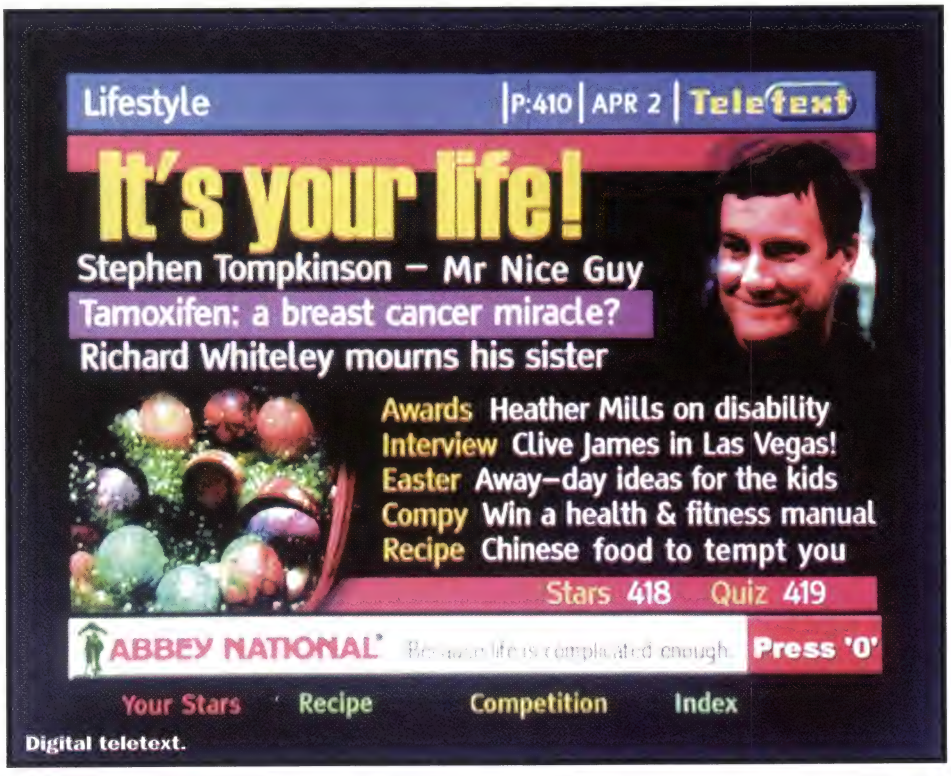
Once coded each primitive media object is an individual entity, and can be worked on separately. These primitive media objects are then collected together to form compound media objects - a dog visually barking and the sound of the barks, for example. These groupings are then used to reconstruct the scenes. MPEG-4 provides a standardised way of describing such scenes: where everything goes and what those things relate to; whether they have some interactive quality; the required changes necessary if a scene is to be navigated through; and the starting and stopping of video and/or audio streams (Figure 11). The scene description information is coded and transmitted separately from the media objects (enabling easy changes to the composition), using a programming language called Binary Format for Scenes (BIFS). This was developed from the Virtual Reality Modelling Language (VRML) increasingly used to create virtual reality environments.

Of course, there could be more than one scene description: for example, a viewer could choose between presenters or news readers speaking different languages, rather than just a different soundtrack or subtitles.

Streamed data is carried in one or more elementary streams; with an object descriptor (OD) to identify all the streams that are associated with one object (Figure 12). Each stream also has its own descriptors for configuration information, such as determining decoder resources and timing. A synchronisation layer identifies the different parts of elementary streams (video or audio frames, etc) which will have been time stamped, and synchronises all the various elements. There is also a delivery layer, which contains two multiplexed layers - FlexMux and TransMux. The Flexible (content) Multiplex can carry groups of interleaved elementary streams; while the Transport Multiplex models transport protocols to enable MPEG-4 to be used in a wide range of applications. The underlying TransMux layer can be used without FlexMux, if it provides all the necessary functions - but the sync layer must accompany it.

MPEG-4 has built on the MPEG-2 Digital Storage Media - Command and Control (DSM-CC), used to provide open application protocols to enable a variety of disparate services to be received by the user. This more advanced version has been renamed DSM-CC Multimedia Integration Framework (DMIF). As with MPEG-2 there are different Profiles to suit different applications.

The visual profiles are: Simple Visual Profile - error resilient coding of rectangular objects, for mobile networks; Simple Scalable Visual Profile - adding coding for temporal and spatial scalable objects, for



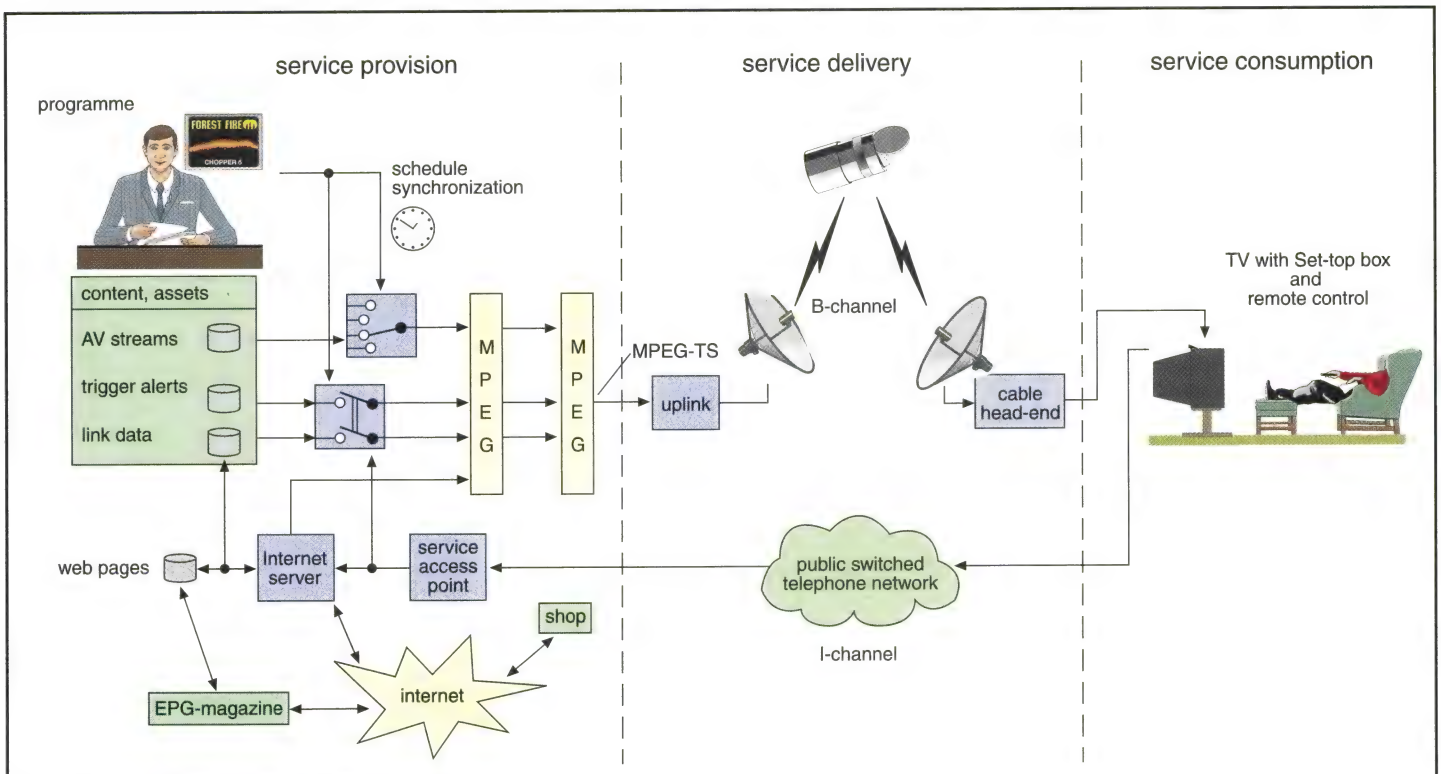


Figure 13. MHEG for interactive viewing.

Internet use - used by the Sharp camera; Core Visual Profile - adding further coding for arbitrary-shaped objects, for interactive multimedia Internet use; Main Visual Profile - adding coding for interlaced, semi-transparent, and sprite objects, for interactive and entertainment broadcast uses plus DVD; and N-Bit Visual Profile - adding coding for video objects with 4-12-bit depths, for surveillance.

The synthetic profiles are: Simple Facial Animation Visual Profile - which does as the name implies, and is suitable for such applications as A/V presentations for the hearing impaired; Scalable Texture Visual Profile: - giving spatial scalability of still images, which can be used for games and digital still cameras; Basic Animated 2-D Texture Visual Profile - gives spatial scalability, SNR scalability, mesh-based animation for still images, and simple face animation; and Hybrid Visual Profile - combines natural video objects with synthetic and/or combined synthetic and natural objects (hybrids), ideal for multimedia.

There are also four audio profiles: Speech Profile - providing HVXC, CELP and TTS; Synthesis Profile - using SAOL, wavetables and TTS for very low bitrate applications; Scalable Profile - a superset of the Speech Profile, suitable for Internet and Narrowband Audio Digital Broadcasting (NADIB) with bitrates of 6-24 kb/s and bandwidths of 3.5-9 kHz; and Main Profile - a combination of the other three.

Then three graphics profiles: Simple 2-D Graphics, Complete 2-D Graphics, and Complete Graphics; and four scene description profiles: audio (for radio, etc), Simple 2-D, Complete 2-D, and Complete - for dynamic virtual 3-D.

Version 2 of MPEG-4 will add new profiles to the list, and is completely backwards compatible with Version 1. It will also add multi-user interactions; Advanced Audio BIFS for more natural sounds, particularly in relationship to the visual environment; improved natural video, including studio quality and stereoscopic images; improved 3-D models; and improved face and body animation.

Apple's QuickTime Version

There will also be an MP4 file format for convenient handling of MPEG-4 information, based on Apple's QuickTime. It is a streamable format, which supports streaming without actually being streamed itself. Metadata, known as 'hint tracks', will tell the server how to deliver the media over a particular TransMux. Another addition is MPEG-J, an Application programming Interface that combines MPEG-4 media with Java code. With the Java application delivered by a separate elementary stream. Work is also progressing to allow MPEG-4 programme elements to be added to MPEG-2 transport streams, and for complete MPEG-4 programmes to be carried in MPEG-2 broadcast multiplexes.

As with an increasing number of things these days, the hardware can be upgraded as needed. In this case the decoder can be

Sanyo 42in. multimedia plasma display.

programmed to one of three levels using the MPEG-4 Syntactic Description Language (MSDL). In Level 0 the decoder has only standardised algorithms; in Level 1 it also has standardised tools which can be configured into an algorithm by the encoder; while in Level 2 the decoder can download new tools and algorithms from the encoder.

Having got all this audio/visual/graphics stuff for computers (MPEG-4PC), TVs, whatever, it needs something else to be able to find what you want. Which is where MPEG-7 comes in (the break in the numbers is the result of Members deciding that the new project was such a break with what had gone before that 'lucky 7' was appropriate).

MPEG-7

MPEG-7 (or Multimedia Content Description Interface) was inaugurated in 1996 to facilitate searching for audio, visual and graphics material. Whether that is for the day to day running of an archive, research, directory services, or TV viewers wanting to find programmes to suit their moods and preferences. The MPEG-7 committee will do this by specifying standard descriptions for all the different types of multimedia information and those will be used to label the contents. It is not the means of searching, but a guide for programs that will be doing the searching.

These descriptors can be held with the associated material, or called up by links from remote locations. MPEG-7 will also standardise Description Schemes (DS) for the descriptors and their relationships; and a Description Definition Language (DDL) to specify the DS. The descriptors do not rely on how the programme material is coded; they will work with both digital and analogue material, and can even be applied to printed images - it is 'What', not 'How'. And, by using MPEG-4 encoding, it will be possible to associate the descriptors with objects within scenes - natural and synthetic, vision and sound.

Different levels of discrimination will also be allowed, enabling searches to made by people with different intentions. At one level it could be abstract: shape, colour, size, movement, etc; at another level it could be a description of the scene: girl in red coat crossing road. The same would apply for audio: key, tempo, location, etc; and at another level: a brass band playing in a park. While the same street scene could be searched for in terms of architecture, shops, the types of vehicles, the fashions worn by the pedestrians, etc. The abstract descriptions obviously lend themselves to automatic searching, whereas specifics would require a lot of intervention on the part of an operator.

Descriptions of the data itself may also be necessary, to determine whether it can be made use of. How it is coded, and how large it is; whether it is copyright material, and, if so, what the conditions are for its use, and the price to be paid; and such like. It is anticipated that MPEG-7 will become an international standard in 2001.

MHEG

In the meantime, MHEG is continuing to be developed. This stands for Multimedia and Hypermedia information coding Experts Group. As the name implies, this combines multimedia with hypermedia and the use of links to call up and navigate through additional information, as is done on the Web. But, in this case, with the emphasis on AV rather than text, thus suiting it to television (the MHEG group was formed in 1989 before the advent of multimedia computers). A viewer watching a programme could therefore call up additional information on something that aroused their curiosity or interest - such as background information on a film star or the local availability of an advertised car.

The group set to work under the title of MHEG-1 to produce a coding system that would represent multimedia/hypermedia interactive applications - what they are, what they do, and what can be done to them - within the structure of an interactive programme, complete with menus, buttons, etc. (similar to the Hyper-Text Markup Language - HTML - used for Web-based material). The applications are held by the programme provider and are downloaded to the user when required; an MHEG interpreter at the user end interprets the various parts of the application, displays the results, and responds to any interactions (Figure 13).

MHEG can encompass different video and audio formats, not just those developed by MPEG, and will group them into a single presentation with everything synchronised and operating as one. It is also intended that it should remain independent of any delivery or file interchange system, just so long as that is MHEG-compliant.

MHEG-1 became a standard in 1995. An MHEG-2 variant on it, using different encoding, was cancelled before it was finished. In 1991 work began on MHEG-3: this extended MHEG-1 by the use of a scripting language, Script Interchange Representation (SIR), and it became a standard in 1996. While, in 1993, MHEG-4 was begun to define procedures for registering MHEG-1 format identifiers - and it became a standard in 1995.

MHEG-5

The year before that work began on MHEG-5. This is a cut-down version of MHEG-1, intended to be used where memory and processing power are limited such as in integrated digital TVs and set-top boxes, and for specific applications such as digital teletext and the various interactive possibilities that are now becoming available. This became a standard in 1996. It is the one used in the UK for digital terrestrial TV - the Digital Television Group having developed a very specific profile for use here. There is also a European version being developed, EuroMHEG, which will have a more advanced specification to take advantagesof new or upgraded services as they become available.

Philips 32in. MHEG-5 Compatible integrated digital TV.

MHEG-6

MHEG-6 (begun in 1995, became a standard in 1997) is a further development of MHEG-5, adding the capability for data processing and enabling it to communicate with external devices using Java code. 1997 was also the year that MHEG-7 was announced, but this is merely to specify the tests to ensure that an MHEG-5 interpreter conforms to the requirements necessary for any particular application. Incidentally, an MHEG-5 Maintenance Task Force was created in 1998 to sort out problems in the standard that have become apparent now that it is being put to use. Finally, this year has seen the creation of MHEG-8, to specify an Extensible Markup Language (XML) notation for MHEG-5.

It's just a pity that none of these technical developments will actually make the programmes themselves any better to watch.