CHAPTER 5

Multimedia Communication Standards

Chapter Overview

Multimedia communication standards have to rely on compromises between what is theoretically possible and what is technologically feasible. Standards can only be successful in the marketplace if the cost performance ratio is well balanced. This is specifically true in the field of audio-video coding where a large variety of innovative coding algorithms exist, but may be too complex for implementation.

In this chapter, we discuss MPEG-1, MPEG-2, MPEG-4, MPEG-4 VTC, JPEG2000, MPEG-7, MPEG-21, ITU-T and Internet standards. MPEG-1 is targeted at CD-ROM with applications at a bit rate of about 1.5 Mb/s. It has also proved useful for computer-generated multimedia where transmission bandwidth and storage capacity are limited or expensive. MPEG-2 addresses high-quality coding for all digital multimedia transmissions at data rates of 2 to 50 Mb/s. It can produce the video quality needed for multimedia entertainment piped to the home and for more demanding business and scientific applications. The scope and potential of the MPEG-4 standard is discussed in the context of audio-visual multimedia communication environments. We show that this standard provides tools and algorithms for coding both natural and synthetic audio and video, as well as provisions to represent the audio-visual data at the user terminal in a highly flexible manner. JPEG-2000

not only provides rate distortion and subjective image quality performance superior to existing standards, but also provides functionalities that current standards can either not address efficiently or not address at all. The objective of the MPEG-7 standardization process is to facilitate the browsing and retrieval of multimedia. We discuss audiovisual content presentation issues from the MPEG-21 multimedia framework. We discuss the ITU-T standardization process in multimedia communications from the video and speech coding, as well from the multimedia multiplex and synchronization points of view (H.32x, H.26x, H.22x). The Internet standardization process concludes the chapter.

5.1 Introduction

In a broad sense, multimedia is assumed to be a general framework for interaction with information available from different sources. With the digital revolution, it became possible to exploit a well-known concept further: the more that is known about the content means the better can be its representation, processing, and so forth, in terms of efficiency and allowed functionalities. After we break down the boundary between speech research and image research, a large number of new techniques and applications can be developed [5.1].

A multimedia standard is expected to provide support for a large number of applications. These applications translate into a specific set of requirements that may be very different from one another. One theme common to most applications is the need for supporting interactivity with different kinds of data. Communications mean standards, but the production of standards for multimedia communications is beset by the problem that the many industries having a stake in multimedia communications have radically different approaches to standardization. Standards play a major role in the multimedia revolution because they provide interoperability between hardware and software provided by multiple vendors.

Example 5.1 Although, in practice, one could use a nonstandard coder, this would lead to a closed architecture system, which would discourage the widespread use of the system in which the coder was embedded. As a result, all of the major vendors of both terminals and software for multimedia communications have embraced the concept of standardization so that their various products will operate at a basic level.

The success of the MPEG [2.63] is based on a number of concurrent elements. MPEG appeared at a time when the coding algorithms of audio and video were reaching asymptotic performance. By relying on the support in terms of technical expertise, of all industries interested in digital audio and video applications, MPEG contributed to the practical acceptance of the audio-visual representation layer, independent of the delivery system. A last element of success has been the focus on the decoder instead of the traditional encoder-decoder approach.

Therefore, MPEG could provide the standard solution to the major players who were considering the use of digital coding of audio and video for innovative mass-market products and could allow a faster achievement of a critical mass than would have been possible without it. The different industries have been diverging, but multimedia communications necessarily need some convergence zone that can only be achieved by standardization in key areas. Putting every stakeholder together and producing communication standards accepted by all is a big task. After the great success of the MPEG-1 and MPEG-2 standards, which opened the digital frontiers to audiovisual information and allowed the deployment of high performance services, MPEG is striking again with the emerging MPEG-4 standard. The MPEG-4 standard is the acknowledgement by MPEG, the leading standardization body in audiovisual representation technology, that the data models underpinning MPEG-1 and MPEG-2 were limited and could not fulfill new needs of emerging multimedia applications, such as hyperlinking, interaction and natural and synthetic data integration. MPEG-4 is the answer to the requirements coming from the new ways in which audio-visual information is nowadays produced, delivered and consumed. To reach this target, MPEG-4 follows an object-based representation approach where an audiovisual scene is coded as a composition of objects, natural as well as synthetic, which provides the first powerful hybrid playground. The objective of MPEG-4 is thus to provide an audiovisual representation standard supporting new ways of communication access and interaction with digital audiovisual data, and it offers a common technical solution to various services. It also extends to layered coding (scalabilities), multiview (stereoscopic video), shape/texture/motion coding of objects and animation. Its role extends to the Internet, Web TV, large databases (storage, retrieval and transmission) and mobile networks [5.2]. MPEG-4 Version 1 became an international standard in February 1999, and Version 2 became a standard in November 1999. Version 2 with extended functionalities is backward compatible with Version 1.

Multimedia databases on the market today allow searching for pictures using characteristics, such as color, texture and information about the shape of objects in an image. MPEG started a new work item to provide a solution to the problem of facilitating multimedia search engines. One of the members of the MPEG family (called Multimedia Content Description Interface) is MPEG-7 [5.3, 5.4, 5.5]. It extends the limited current search capabilities to include more information types, such as video, image, audio, graphics and animation. In other words, MPEG-7 specifies a standardized description of various types of multimedia information. This description is associated with the content itself and allows fast and efficient searching for multimedia that is of interest to users. The description can be attached to any kind of multimedia material, no matter what the format of the description is. Stored material that has this information attached to it can be indexed, searched and retrieved.

When the scope of new work has been sufficiently clarified, MPEG usually makes open requests for proposals. So far proposals have been requested for the following:

- MPEG-1 Audio and Video (July 1989)
- MPEG-2 Audio and Video (July 1991)

- MPEG-4 Audio and Video (July 1995)
- MPEG-7 and MPEG-21

In the original ITU-T work plan, the goal was to define a near-term recommendation in 1996, followed by a long-term recommendation several years later. The near-term recommendation is referred to as H.263. The long-term recommendation H.26L (previously called H.263L) is scheduled for standardization in the year 2002 and may adopt a completely new compression algorithm. After H.263 was completed, it became apparent that incremental changes could be made to H.263 that could visibly improve its compression performance. Thus, ITU-T decided in 1996 that a revision to H.263 would be created that incorporated these incremental improvements. This is H.263 plus with several new features. Hence, the name H.263+ (now called H.263 Version 2). H.263+ contains approximately 12 new features that do not exist in H.263. These include new coding modes that improve compression efficiency, support for scalable bit streams, several new features to support packet networks [5.6] and error-prone environments, added functionality and support for a variety of video formats.

5.2 MPEG Approach to Multimedia Standardization

MPEG was established in January 1988 with the mandate to develop standards for the coded representation of moving pictures, audio and their combination. It operates in the framework of the Joint ISO/IEC Technical Committee (JTC 1) on Information Technology under WG11 of SC29.

Starting from its first meeting in May 1988 when 25 experts participated, MPEG has grown to an unusually large committee. Usually some 350 experts from some 200 companies and organizations from about 20 countries take part in MPEG meetings. As a rule, MPEG meets three times a year (in March, July and November), but meets more frequently when the workload so demands.

Depending on the nature of the standard, documents of different nature may be produced. For audio and video coding standards, the first document is called a *Verification Model* (VM). In MPEG-1 and MPEG-2, this was called Simulation and Test Model, respectively. The VM describes, in some sort of programming language, the operation of the encoder and the decoder. The VM is used to carry out simulations to optimize the performance of the coding scheme. When MPEG has reached sufficient confidence in the stability of the standard under development, a *Working Draft* (WD) is produced. This is already in the form of a standard, but is kept internal to MPEG for revision. At the planned time, the WD has become sufficiently solid and becomes *Committee Draft* (CD).

A WD usually undergoes several revisions before moving to the CD stage. A key role is played by core experiments where different technical options are studied by at least two different partners. Each revision involves a large number of experts who draw the committee's attention to possible errors contained in the document.

A list of work items in MPEG multimedia standardization is as follows:

- *ISO/IEC IS 11172*—Coding of moving pictures and associated audio at up to about 1.5 Mb/s (MPEG-1)
  - Part 1 Systems
  - Part 2 Video
  - Part 3 Audio
  - Part 4 Conformance testing
  - Part 5 Software simulation
- *ISO/IEC IS 13818*—Generic coding of moving pictures and associated audio (MPEG-2)
  - Part 1 Systems
  - Part 2 Video
  - Part 3 Audio
  - Part 4 Conformance testing
  - Part 5 Software simulation
  - Part 6 System extensions—Digital Storage Media-Command and Control (DSM-CC)
  - Part 7 Audio extension—Advanced Audio Coding (AAC)
  - Part 8 VOID—(withdrawn)
  - Part 9 System extension—Real Time Interface (RTI) for system decoders
  - Part 10 Conformance extension for DSM-CC
- *ISO/IEC IS 14496*—Coding of audiovisual objects (MPEG-4)
  - Part 1 Systems
  - Part 2 Visual
  - Part 3 Audio
  - Part 4 Conformance testing
  - Part 5 Reference software
  - Part 6 Delivery Multimedia Integration Framework (DMIF)
  - Part 7 Optimized reference software
  - Part 8 Carriage of MPEG-4 content across IP networks
  - Part 9 Reference hardware description
- *ISO/IEC IS 15938*—Multimedia content description interface (MPEG-7)
  - Part 1 Systems
  - Part 2 Description Definition Language (DDL)
  - Part 3 Visual
  - Part 4 Audio
  - Part 5 Multimedia Description Schemes (MDS)
  - Part 6 Reference software
  - Part 7 Conformance
  - Part 8 Extraction and use of MPEG-7 descriptors

- *ISO/IEC IS 18034*—Multimedia framework (MPEG-21)
  - Part 1 Vision, technologies and strategy
  - Part 2 Digital item declaration
  - Part 3 Digital item identification and description
  - Part 4 Intellectual property management and protection
  - Part 5 Rights expression language
  - Part 6 Rights data dictionary

Although the television paradigm dominated audiovisual communications for many years, the situation now is evolving very quickly in terms of the ways audiovisual content is produced, delivered and consumed [5.7]. Moreover, hardware and software are getting more and more powerful, opening new frontiers to the technologies used and to the functionalities provided.

Producing content today is made very easy. Digital still cameras directly storing images in JPEG format have hit the mass market. Together with the first digital video cameras recording directly in MPEG-1 format, this represents a major step for the acceptance in the consumer market of digital audiovisual acquisition technology. This step transforms every one of us into a potential content producer, capable of creating content that can be easily distributed and published using the Internet. Moreover, more content is being synthetically produced, computer generated and integrated with natural material as a truly hybrid audiovisual content. The various pieces of content, digitally encoded, can be successively reused without the quality losses typical of the previous analog processes.

Although audiovisual information was, until recently, only carried across very few networks, the trend is now toward the generalization of visual information in every single network. Moreover, the increasing mobility in telecommunications is a major trend. Mobile connections will not be limited to voices, and other types of data, including real-time media, will be the next. Because mobile telephones are replaced every two to three years, new mobile devices can finally make the decade-long promise of audiovisual communications turn into reality [5.8]. The explosion of the Web and the acceptance of its interactive mode of operation have clearly shown in the last few years that the traditional television paradigm would no longer suffice for audiovisual services. Users will want to have access to audio and video like they now have access to text and graphics. This requires moving pictures and audio of acceptable quality at low bit-rates on the Web and Web-type interactivity with live content.

Standardization items have been identified well in advance and no MPEG standard has endorsed an industry standard. MPEG standards do not specify complete systems. Therefore, it is possible that industry standards are needed with MPEG standards to make full-fledged products.

Example 5.2 Industries by definition need to make vertically integrated specifications in order to make products that satisfy some needs. Audiovisual decoding may well be a piece of technology that can be shared with other communities, but, in the event industries need to sell a satellite receiver or a video CD player, these require an integrated standard. However, if different industries need the same standard, they quite likely will have different systems in mind. Therefore, only the components of a standard, the tools as they are called in MPEG, can be specified

in a joint effort. The implementation of this principle requires the change of the nature of standards from system standards to component standards. Industries will assemble the tool specification from the standards body and build their own product specification.

If tools are the object of standardization, a new process must be devised to produce meaningful standards. The following sequence of steps has been found to be practically implementable and to produce the desired result:

1. Select a number of target applications for which the generic technology is intended to be specified.
2. List the functionalities needed by each application.
3. Break down the functionalities into components of sufficiently reduced complexity so that they can be identified in different applications.
4. Identify the functionality components that are common across the systems of interest.
5. Specify the tools that support the identified functionality components, particularly those common to different applications.
6. Verify that the tools specified can actually be used to assemble the target systems and provide the desired functionalities.

These standardized sets of tools have been called profiles in MPEG-2 Video [5.9]. It is advisable that certain major contributions of tools be specified as normative, making sure that these are not application specific, but functionally specific.

In some environments, it is proper to add those nice little things to a standard that bring a standard nearer to a product specification. This is, for example, the case of industry standards or when standards are used to enforce the concept of guaranteed quality so important to broadcasters and telecommunication operators because of their public service nature. However, in the case when a standard is to be used by multiple industries, only the minimum that is necessary for interoperability can be specified. The profile-level philosophy successfully implemented by MPEG provides a solution: within a single tool one may define different grades called levels in MPEG [5.10].

Example 5.3 When a standard is defined by a single type of industry generally, an agreement exists on where a certain functionality resides in the system. In a multi-industry environment, this is not possible. Take the case of encryption. Depending on our role in the audiovisual distribution chain, you would like to have the encryption function located where it serves your place in the chain best, because encryption is an important value-added function. If the standard endorses our business model, we will adopt the standard. If it does not, we will antagonize it.

After the work is nearing completion, it is important to make sure that it does indeed satisfy the requirements (product specification) originally set. MPEG does that through a process called verification tests, with the scope of ascertaining how well the standard produced meets the specification. We give now a brief account of MPEG multimedia communications standards—some established and some under development—by giving a description of their functionalities and usage.

5.3 MPEG-1 (Coding of Moving Pictures and Associated Audio)

The first standard developed by the group, nicknamed MPEG-1, was the coding of the combined audiovisual signal at a bit rate around 1.5 Mb/s. This was motivated by the prospect, becoming apparent in 1988, of storing video signals on a CD with a quality comparable to VHS cassettes. In 1988, coding of video at such low bit rates had become possible thanks to decades of research in video-coding algorithms. These algorithms, however, had to be applied to subsampled pictures—a single field from a frame and only half of the samples in a line—to show their effectiveness. Also, coding of audio, as separate from speech, allowed reduction by one-sixth of the PCM bit rate, typically 256 Kb/s for a stereo source, with virtual transparency. Encoded audio and video streams, with the constraint of having a common time base, were combined into a single stream by the MPEG systems layer. As previously indicated, MPEG-1, formally known as ISO/IEC 11172, is standardized in five parts [5.11]. The first three parts are Systems, Video and Audio. Two more parts complete the suite of MPEG-1 standards; Conformance Testing, which specifies the methodology for verifying claims of conformance to the standard by manufacturers of equipment and producers of bitstreams, and Software Simulation, a full C-language implementation of the MPEG-1 standard (encoder and decoder).

Part 1 addresses the problem of combining one or more datastreams from the video and audio parts of the MPEG-1 standard with timing information to form a single stream (Figure 5.1). This is an important fuction because, after being combined into a single stream, the data is in a form well suited to digital storage or transmission.

Part 2 specifies a coded representation that can be used for compressing video sequences—both 625-line and 525-line—to bit rates around 1.5 Mb/s. Part 2 was developed to operate from storage media offering a continuous transfer rate of about 1.5 Mb/s. Nevertheless, it can be used more widely than this because the approach taken is generic [5.12].

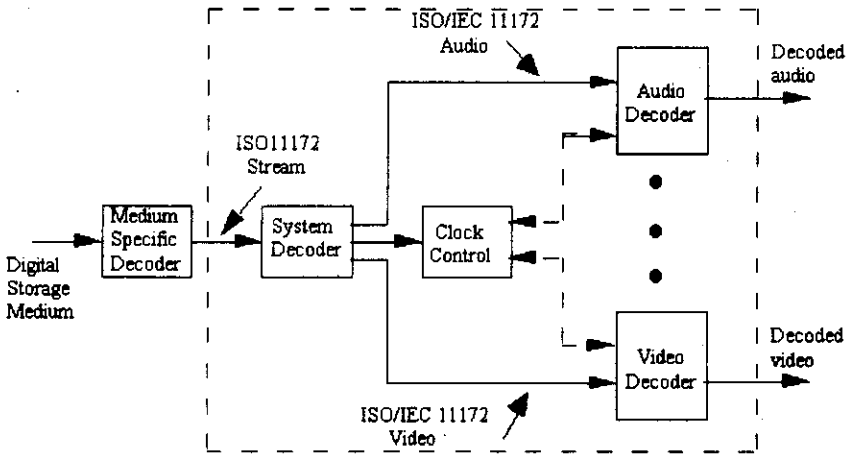

Figure 5.1 ISO/IEC 11172 decoder [5.11]. ©1993 ISO/IEC.

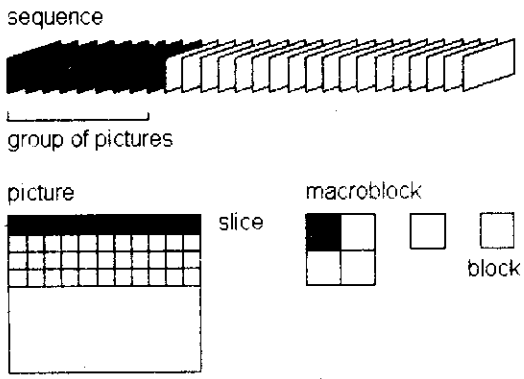

Figure 5.2 Temporal picture structure [5.12]. ©1993 ISO/IEC.

A number of techniques are used to achieve a high compression ratio. The first is to select an appropriate spatial resolution for the signal. The algorithm then uses block-based motion compensation to reduce the temporal redundancy. Motion compensation is used for causal prediction of the current picture from a previous picture, or noncausal prediction of the current picture from a future picture or for interpolative prediction from past and future pictures. The difference signal, the prediction error, is further compressed using DCT to remove spatial correlation and is then quantized. Finally, the motion vectors are combined with the DCT information and coded using variable length codes. Figure 5.2 illustrates a possible combination of the three main types of pictures that are used in the standard.

Part 3 specifies a coded representation that can be used for compressing audio sequences, both mono and stereo as shown in Figure 5.3. Input audio samples are fed into the encoder. The mapping creates a filtered and subsampled representation of the input audio stream. A psychoacoustic model creates a set of data to control the quantizer and coding. The quantizer and coding block create a set of coding symbols from the mapped input samples. The block frame packing assembles the actual bit stream from the output data of the other blocks and adds other information (for example error correction) if necessary [5.13].

Part 4 specifies how tests can be designed to verify whether bit streams and decoders meet the requirements as specified in Parts 1, 2 and 3 of the MPEG-1 standard. These tests can be used by the following:

- Manufacturers of encoders and their customers to verify whether the encoder produces valid bit streams
- Manufacturers of decoders and their customers to verify whether the decoder meets the requirements specified in Parts 1, 2 and 3 of the standard for the claimed decoder capabilities
- Applications to verify whether the characteristics of a given bit stream meet the application requirements, for example, whether the size of the coded picture does not exceed the maximum value allowed for the application [5.14].

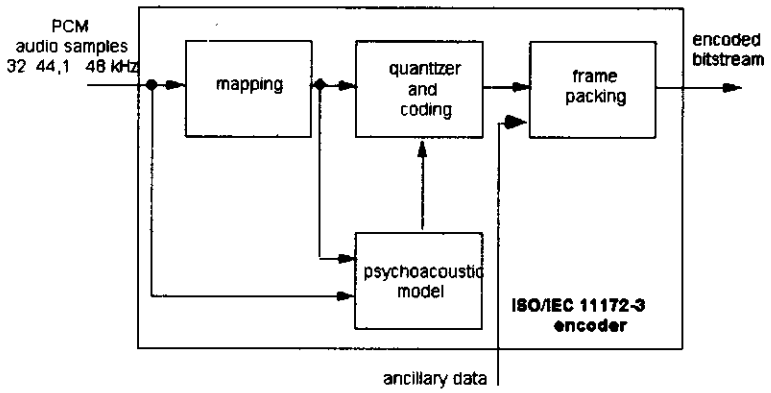

Figure 5.3 Basic structure of the MPEG-1 audio encoder [5.13]. ©1993 ISO/IEC.

Part 5, technically not a standard, but a technical report, gives a full software implementation of the first three parts of the MPEG-1 standard [5.15].

Example 5.4 The different layers have been defined because they all have their merits. Basically, the complexity of the encoder and decoder, the encoder delay and the coding efficiency increase when going from Layer I through Layer II to Layer III. Layer I has the lowest complexity and is specifically suitable for applications where the encoder complexity also plays an important role. Layer II requires a more complex encoder and a slightly more complex decoder and is directed toward one-to-many applications, that is, one encoder serves many decoders. Compared to Layer I, Layer II is able to remove more of the signal redundancy and applies the psychoacoustic threshold more efficiently. Layer III is again more complex and is directed toward lower bit-rate applications due to the additional redundancy and irrelevancy extraction from enhanced frequency resolution in its filterbank.

MPEG-1 was also the first signal-processing standard developed and eventually documented using the C-programming language. This facilitated the recognition that what matters in a coding standard is just the syntax used to represent the operations carried out on the signal. Bit rate, frame rate, number of lines, number of pixels per line, and so forth are just parameters with their overriding importance in the analog domain reduced to size in the digital domain.

MPEG-1 not only was a technological achievement, but also contributed to define the highly politicized issues of television standards. MPEG recognized that what matters in a television signal is not the number of lines or the number of fields per second, but the bandwidth of the signal in the analog domain and the number of pixels per second in the digital domain. The result has been the normative definition of the Constrained Parameter Set (CPS) for MPEG-1, where there is no reference to television standards. In terms of memory requirement, what matters is the total number of pixels in a frame, and, in terms of processing requirements, what matters is the number of macroblocks, that is, number of coded 16x16 pixels. Elements of CPS are given in Table 5.1.

Table 5.1 MPEG-1 Video CPS [5.10].

Parameter	Value
Horizontal size	≤ 768
Vertical size	≤ 576
Number of macroblocks / picture	≤ 396
Number of macroblocks / second	≤ 9900
Picture rate	≤ 30 Hz
Interpolated pictures	≤ 2
Bit rate	≤ 1856 Kb/s

The audio part of the MPEG-1 standard has become the key component for radio broadcasting at CD quality offered by digital audio broadcasting, which is actively being deployed in several countries. MPEG-1 Audio Layer II and, more recently, Layer III have become the standard form for music distribution on the Web. The full MPEG-1 standard (Audio-Video-Systems) is the standard format for distribution of video material across the Web.

MPEG-1 provided the first concrete opportunity for the microelectronics industry to invest in digital video technology. MPEG-1 decoder chips are produced by multiple sources, some of which incorporate the electronics needed to read bits from a CD. Several suppliers exist for MPEG-1 encoder chips. One consumer-electronics manufacturer has already put its digital video camera on the market. This is made up of an optical part, audio and video sensors, a single chip for audio-video systems encoding and a hard disk for 20 minutes of recording. These devices and the growing number of personal computers are creating the conditions for the popularization of multimedia contents production. Detailed information on MPEG-1 Audio can be found in ISO/IEC IS13818-3 (MPEG-1) and Brandenburg et al. [5.16, 5.17].

The MPEG-1 video algorithm was primarily targeted for multimedia CD-ROM applications, requiring additional functionality supported by both encoder and decoder. Important features provided by MPEG-1 include frame-based random access of video, fast forward/fast reverse (FF/FR) searches through compressed bit streams, reverse playback of video and editability of the compressed bit stream.

5.3.1 The Basic MPEG-1 Interframe Coding Scheme

The basic MPEG-1 video-compression technique is based on a macroblock structure, motion compensation and the conditional replenishment of macroblocks. As outlined in Figure 5.4a, the

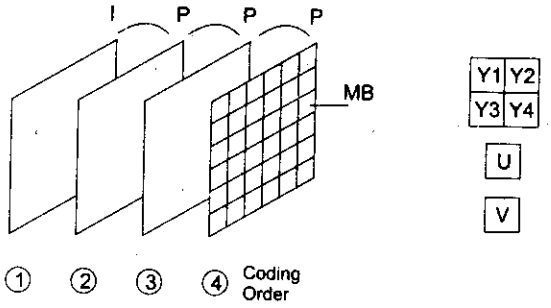

Figure 5.4 Illustration of I-pictures (I) and P-pictures (P) in a video sequence [5.12]. ©1993 ISO/IEC.

MPEG-1 coding algorithm encodes the first frame in a video sequence in intraframe coding mode (I-picture). Each subsequent frame is coded using interframe prediction (P-pictures), which means that only data from the nearest previously coded I- or P-frame is used for prediction. The MPEG-1 algorithm processes the frames of a video sequence as block based. Each color input frame in a video sequence is partitioned into nonoverlapping macroblocks as depicted in Figure 5.4b. Each macroblock contains blocks of data from both luminance and cosited chrominance bands: four luminance blocks (Y1, Y2, Y3 and Y4) and two chrominance blocks (U and V), each with size 8x8 pels. Thus, the sampling ratio between Y:U:V luminance and chrominance pels is 4:1:1. P-pictures are coded using motion-compensated prediction based on the nearest previous frame (I or P). Each frame is divided into disjoint macroblocks. With each macroblock, information related to four luminance blocks (Y1, Y2, Y3 and Y4) and two chrominance blocks (U and V) is coded. Each block contains 8x8 pels.

The block diagram of the basic hybrid DPCM/DCT MPEG-1 encoder and decoder structure is depicted in Figure 5.5. The first frame in a video sequence (I-picture) is encoded in INTRA mode without reference to any past or future frames. At the encoder, the DCT is applied to each 8x8 luminance and chrominance block, and, after output of the DCT, each of the 64 DCT coefficients is uniformly quantized (Q). The quantizer Step Size (SZ) used to quantize the DCT-coefficients within a macroblock is transmitted to the receiver. After quantization, the lowest DCT coefficient (DC coefficient) is treated differently from the remaining coefficients (AC coefficients). The DC coefficient corresponds to the average intensity of the component block and is encoded using a differential DC prediction method. The nonzero quantized values of the remaining DCT coefficients and their locations are then zig-zag scanned and run-length entropy coded using VLC tables.

The concept of zig-zag scanning of the coefficients is outlined in Figure 5.6. The scanning of the quantized DCT-domain 2D signal followed by variable-length code-word assignment for the coefficients serves as a mapping of the 2D image signal into a 1D bit stream. The nonzero AC coefficient quantized values (length) are detected along the scan line as well as the distance (run) between two consecutive nonzero coefficients. Each consecutive (run, length) pair is

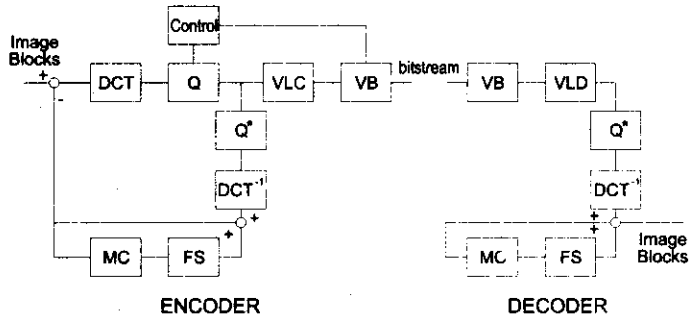

Figure 5.5 Block diagram of a basic DPCM/DCT encoder and decoder structure [5.12]. ©1993 ISO/IEC.

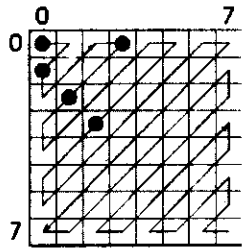

Figure 5.6 Zig-zag scanning of the quantized DCT coefficients [5.12]. ©1993 ISO/IEC.

encoded by transmitting only one VLC codeword. The purpose of zig-zag scanning is to trace the low-frequency DCT coefficients (containing most energy) before tracing the high-frequency coefficients.

Example 5.5 Zig-zag scanning of the quantized DCT coefficients in an 8x8 block is presented in Figure 5.6. Only the nonzero quantized DCT coefficients are indicated. The possible locations of nonzero DCT coefficients are indicated. The zig-zag scan attempts to trace the DCT-coefficients according to their significance. The lowest DCT coefficients (0,0) contain most of the energies within the blocks. The energy is concentrated around the low-frequency DCT coefficients.

The decoder performs the reverse operations, first extracting and decoding (VLD) the variable-length coded words from the bit stream to obtain locations and quantized values of the non-zero DCT coefficients for each block. With the reconstruction (Q*) of all nonzero DCT coefficients belonging to one block and subsequent inverse DCT (IDCT), the quantized block pixel values are obtained. By processing the entire bit stream, all image blocks are decoded and reconstructed.

For coding P-pictures, the previously I- or P-picture frame *N-1* is stored in a Frame Store (FS) in both encoder and decoder. Motion compensation (MC) is performed on a macroblock basis. Only one motion vector is estimated between frame *N* and frame *N-1* for a particular Macroblock to be encoded. These motion vectors are coded and transmitted to the receiver. The motion-compensated prediction error is calculated by subtracting each pel in a Macroblock with

its motion-shifted counterpart in the previous frame. An *8x8* DCT is then applied to each of the *8x8* blocks contained in the macroblock followed by Q of the DCT coefficients with subsequent run-length coding and entropy coding (VLC). A Video Buffer (VB) is needed to ensure that a constant target bit rate output is produced by the encoder. The quantization SZ can be adjusted for each macroblock in a frame to achieve a given target bit rate and to avoid buffer overflow and underflow.

The decoder uses the reverse process to reproduce a macroblock of frame *N* at the receiver. After decoding the variable length words contained in the video decoder buffer, the pixel values of the prediction error are reconstructed (Q* and DCT^{-1}-operations). The motion-compensated pixels from the previous frame *N-1* contained in the FS are added to the prediction errors to recover the particular macroblock of frame *N*.

5.3.2 Conditional Replenishment

An essential feature supported by the MPEG-1 coding algorithm is the possibility of updating macroblock information at the decoder only if needed, for example, if the content of the macroblock has changed in comparison to the content of the same macroblock in the previous frame (conditional macroblock replenishment). The key for efficient coding of video sequences at lower bit rates is the selection of appropriate prediction modes to achieve conditional replenishment. The MPEG standard allows three different macroblock coding types (MB types):

- *Skipped MB*—Prediction from previous frame with a zero motion vector is used. No information about the macroblock is coded or transmitted to the receiver.
- *Inter MB*—Motion-compensated prediction from the previous frame is used. The macroblock type, the MB address and, if required, the motion vector, the DCT coefficients and quantization SZ are transmitted.
- *Intra MB*—No prediction is used from the previous frame (intraframe coding only). Only the macroblock type, the MB address and the DCT coefficients and quantization SZ are transmitted to the receiver.

5.3.3 Specific Storage Media Functionalities

For accessing video from storage media, the MPEG-1 video compression algorithm was designed to support important functionalities, such as random access and FF/FR playback functionalities. To incorporate the requirement for storage media and to explore the significant advantages of motion compensation and motion interpolation further, the concept of B-pictures (bidirectional predicted/bidirectional interpolated pictures) was introduced in MPEG-1. This concept is depicted in Figure 5.7 for a group of consecutive pictures in a video sequence. Three types of pictures are considered. Intrapictures (I-pictures) are coded without reference to other pictures contained in the video sequence. I-pictures allow access points for random access and FF/FR functionality in the bit stream, but achieve only low compression. Interframe-predicted pictures (P-pictures) are coded with reference to the nearest previously coded

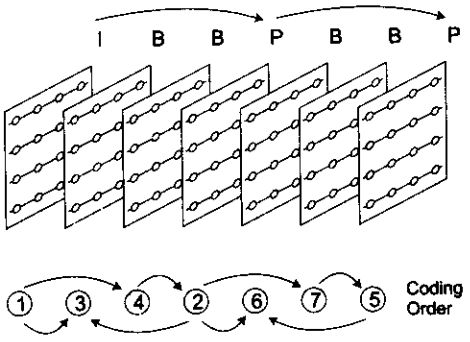

Figure 5.7 I-pictures (I), P-pictures (P) and B-pictures (B) used in a MPEG-1 video sequence [5.12]. ©1993 ISO/IEC.

I-picture or P-picture, usually incorporating motion compensation to increase coding efficiency. Because P-pictures are usually used as reference for prediction for future or past frames, they provide no suitable access points for random access functionality or editability. Bidirectional predicted/interpolated pictures (B-pictures) require both past and future frames as references. To achieve high compression, motion-compensation can be employed based on the nearest past and future P-pictures or I-pictures. B-pictures themselves are never used as references. B-pictures can be coded using motion-compensated prediction based on the two nearest already coded frames (either I-picture or P-picture). The arrangement of the picture-coding types within the video sequence is flexible to suit the needs of diverse applications. The direction for prediction is indicated in Figure 5.7.

The user can arrange the picture types in a video sequence with a high degree of flexibility to suit diverse application requirements. As a general rule, a video sequence coded using I-pictures only (I I I I I . . .) allows the highest degree of random access, FF/FR and editability, but achieves only low compression. A sequence coded with a regular I-picture update and no B-pictures (that is I P P P P P P I P P P P . . .) achieves moderate compression and a certain degree of random access and FF/FR functionality. Incorporation of all three pictures types, as depicted in Figure 5.7 (I B B P B B P B B I B B P . . .), may achieve high compression and reasonable random access and FF/FR functionality but also increases the coding delay significantly. This delay may not be tolerable for videotelephony or videoconferencing applications.

5.3.4 Rate Control

An important feature supported by the MPEG-1 encoding algorithm is the possibility of tailoring the bit rate (and thus the quality of the reconstructed video) to specific applications requirements by adjusting the quantizer SZ in Figure 5.5 for quantizing the DCT coefficients.

Coarse quantization of the DCT coefficients enables the storage or transmission of video with high compression ratios, but, depending on the level of quantization, may result in significant coding artifacts. The MPEG-1 standard allows the encoder to select different quantizer values for each coded macroblock. This enables a high degree of flexibility to allocate bits in

images where needed and to improve image quality. Furthermore, it allows the generation of both constant and variable bit rates for storage or real-time transmission of the compressed video.

Compressed video information is inherently variable in nature. This is caused in general by the variable content of successive video frames. To store or transmit video at a constant bit rate, it is therefore necessary to buffer the variable bit stream generated in the encoder in a video buffer (VB) as depicted in Figure 5.5. The input into the encoder VB is variable over time, and the output is a constant bit stream. At the decoder, the VB input bit stream is constant, and the output used for decoding is variable. MPEG encoders and decoders implement buffers of the same size to avoid reconstruction errors.

A rate-control algorithm at the encoder adjusts the quantizer SZ depending on the video content and activity to ensure that the VB will never overflow. At the same time, it targets to keep the buffers as full as possible to maximize image quality. In theory, overflow of buffers can always be avoided by using a large enough VB. However, besides the possibly undesirable costs for the implementation of large buffers, there may be additional disadvantages for applications requiring low-delay between encoder and decoder, such as for the real-time transmission of conversational video. If the encoder bitstream is smoothed using a VB to generate a constant bit rate output, a delay is introduced between the encoding process and the time the video can be reconstructed at the decoder. Usually the larger the buffer means the larger the delay introduced.

MPEG has defined a minimum VB size that needs to be supported by all decoder implementations. This value is identical to the maximum value of the VB size that an encoder can use to generate a bit stream. However, to reduce delay or encoder complexity, it is possible to choose a virtual buffer size value at the encoder smaller than the minimum VB size that needs to be supported by the decoder. This virtual buffer size value is transmitted to the decoder before sending the video bit stream.

The rate-control algorithm used to compress video is not part of the MPEG-1 standard, and it is thus left to the implementers to develop efficient strategies. It is worth emphasizing that the efficiency of the rate-control algorithms selected by manufacturers to compress video at a given bit rate heavily impacts the visible quality of the video reconstructed at the decoder.

5.4 MPEG-2 (Generic Coding of Moving Pictures and Associated Audio)

The MPEG-2 family of standards outlines the compression technologies and bit-stream syntax that enable transmission of audio and video in broadband networks. These standards also describe the aspects needed to multiplex programs, enable clock synchronization and set up logical network links carrying video and audio content. MPEG-2 is, in many cases, associated only with video compression, which is certainly one of the most important parts of its functionality [5.18, 5.19, 5.20, 5.21]. However, the MPEG-2 standards include more than just pure video. In total, MPEG-2 has different parts, which cover the different aspects of digital video and audio delivery and representation [5.22]. Table 5.2 lists the different MPEG-2 parts.

Table 5.2 Parts of the MPEG-2 standards.

ISO/IEC 13818 MPEG-2	Description
13818-1	Systems
13818-2	Video
13818-3	Audio
13818-4	Compliance
13818-5	Software simulation
13818-6	DSM-CC
13818-9	RTI for system decoders
13818-10	DSM reference script format

Basically, MPEG-2 can be seen as a superset of the MPEG-1 coding standard and was designed to be backward compatible to MPEG-1. Every MPEG-2 compatible decoder can decode a valid MPEG-1 bit stream. Many video-coding algorithms were integrated into a single syntax to meet the diverse application requirements. New coding features were added by MPEG-2 to achieve sufficient functionality and quality, so prediction modes were developed to support efficient coding of interlaced video. In addition, scalable video-coding extensions were introduced to provide additional functionalities, such as embedded coding of digital TV and HDTV and graceful quality degradation in the presence of transmission errors.

For comparison, typical MPEG-1 and MPEG-2 coding parameters are shown in Table 5.3. However, implementation of the full syntax may not be practical for most applications. MPEG-2 has introduced the concept of profiles and levels to stipulate conformance for equipment not supporting the full implementation. Profiles and levels provide means for defining subsets of the syntax and thus the decoder capabilities required to decode a particular bit stream. As a general rule, each profile defines a new set of algorithms added as a superset to the algorithms in the profile below. A level specifies the range of the parameters that are supported by the implementation (that is, image size, frame rate and bit rates). The MPEG-2 core algorithm at the Main profile features nonscalable coding of both progressive and interlaced video sources. It is expected that most MPEG-2 implementations will at least conform to the Main Profile at the Main level, which supports nonscalable coding of digital video with approximately digital TV parameters: a maximum sample density of 720 pixels per line and 576 lines per frame, a maximum frame rate of 30 frames per second and a maximum bit rate of 15 Mb/s.

The upper bound of parameters at each level of a profile is given in Table 5.4. The MPEG-2 algorithm defined in the Main Profile is a straightforward extension of the MPEG-1 coding

Table 5.3 MPEG-1 and MPEG-2 coding parameters.

Parameter	MPEG-1	MPEG-2
Standardized	1992	1994
Main application	Digital video on CD-ROM	Digital TV (and HDTV)
Spatial resolution	SIF format (1/4 TV) 288x360pixels	TV (4xTV) 576x720 (1152x1440)
Temporal resolution	25/30 frames/s	50/60 fields/s (100/120 fields/s)
Bit rate	1.5 Mb/s	4 Mb/s (20 Mb/s)
Quality	Comparable to VHS	Comparable to NTSC/PAL for TV
Compression ratio over PCM	20-30	30-40

scheme to accommodate coding of interlaced video while retaining the full range of functionality provided by MPEG-1. Because it is identical to the MPEG-1 standard, the MPEG-2 coding algorithm is based on the general hybrid DPCM/DCT coding scheme, incorporating a macroblock structure, motion compensation and coding needs for conditional replenishment of macroblocks. The concept of I-picture, P-picture and B-picture is fully retained in MPEG-2 to achieve motion prediction and to assist random access functionality. In what follows we focus on the most essential parts of MPEG-2.

Table 5.4 Upper bound of parameters at each level of profile.

Level	Parameters
High	1,920 pixels/line
	1,152 lines/frame
	60 frames/s
	80 Mb/s
High 1440	1,440 pixels/line
	1,152 lines/frame
	60 frames/s
	60 Mb/s

Table 5.4 Upper bound of parameters at each level of profile. (Continued)

Level	Parameters
Main	720 pixels/line
	576 lines/frame
	30 frames/s
	15 Mb/s
Low	352 pixels/line
	288 lines/frame
	30 frames/s
	4 Mb/s

5.4.1 MPEG-2 Video

The main goal of the MPEG-2 Video standard is to define a format that can be used to describe a coded video bit stream. This video bit stream is the output of an encoding process, which significantly compresses the video information. MPEG-2 does not specify the encoding process. It only defines the resulting bit stream. When MPEG-2 was developed, one of the requirements was to make it flexible enough to handle a range of video applications, like broadcast (satellite) services, cable TV distribution and interactive television services, subject to flexible equipment capabilities, network bandwidth constraints and picture qualities. The MPEG-2 group managed to make the standard very generic by providing a set of tools that can be combined in different ways. The MPEG-2 Video standard consists of the following parts [5.23]:

- *Basic definitions*—Basic objects such as pictures and frames are defined.
- *MPEG-2 Video syntax*—Different syntax elements are derived.
- *Semantic description for the video stream syntax*—Semantic description is given for all syntax elements.
- *Video-decoding process*—Video decoding processes are described, including decoding in interlaced and progressive modes.
- *Scalability extensions*—Different variations of scalability of MPEG-2 Video are described, and the decoding for each mode is explained.
- *Profiles and levels*—Different profiles and levels, which are used to define subsets of MPEG-2 Video, are described.
- *Annexes*—Annexes provide variable length coding tables, tables that define profile and level constraints, and the DCT function. They also contain some information sections.

MPEG-2 Video—The Basics

MPEG-2 deals with a number of basic objects that are used to structure video information. Basic objects in MPEG-2 are shown in Figure 5.8.

The video sequence represents a number of video pictures or group of video pictures. A video sequence contains only a few pictures and not a whole movie.

A frame contains all the color and brightness information that is needed to display a picture. The color and brightness information is organized into three matrices, which contain the luminance and chrominance values. Figure 5.9 shows these matrixes for a 4:4:4 and 4:2:2 sampled frame.

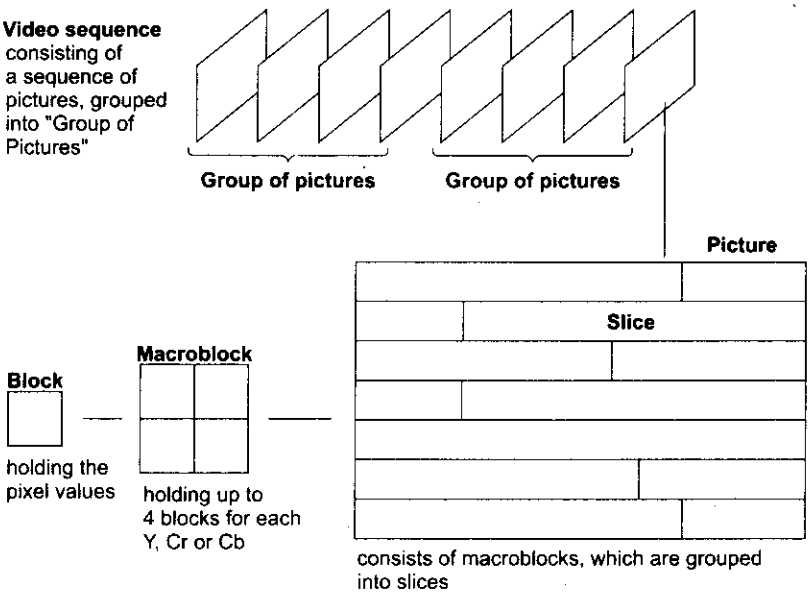

Figure 5.8 Basic objects in MPEG-2.

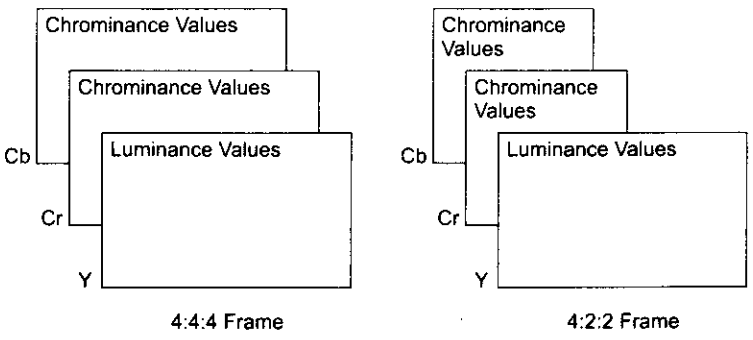

Figure 5.9 Matrixes forming a 4:4:4 and a 4:2:2 frame.

Each picture is divided into a number of blocks, which are grouped into macroblocks. Each block contains eight lines, with each line holding eight samples of luminance or chrominance pixel values from a frame. This gives 64 chrominance or luminance pixel values defining a block. Four blocks with luminance values, plus a number of blocks with chrominance values, form the luminance and chrominance information of a macroblock. The number of chrominance blocks in a macroblock depends on the sampling format used to digitize the video material. A 4:2:0 macroblock holds four blocks of luminance and two blocks of chrominance information. A 4:4:4 macroblock holds four blocks of luminance and eight blocks of chrominance information.

There are three picture types defined in MPEG-2 Video [5.24, 5.25]. Intracoded pictures (I-pictures) are pictures that are coded in such a way that they can be decoded without knowing anything about other pictures in the video sequence. In a video sequence or group of pictures, the first picture is always an I-picture and provides bootstrap information for the following pictures. Predictive-coded pictures (P-pictures) are decoded by using information from another picture, which was decoded earlier. The information that can be used from the previous picture is determined by motion estimation and is coded in what are called intermacroblocks. A P-picture consists of intracoded macroblocks and predictive-coded macroblocks. The latter are always combined with a motion vector indicating which macroblock to use from a previous picture. A P-picture requires 30 to 50% of the number of bits needed for an I-picture. Bidirectionally coded pictures (B-pictures) also use information from other pictures. Like P-pictures, they can use information provided by a picture that occurred previously. A B-picture can also use information from a picture coming in the future. As in P-pictures, picture information that cannot be found in previous or future pictures is intracoded. B-pictures require approximately 50% of the number of bits needed for a P-picture.

Example 5.6 Figure 5.10 represents an exampie to encode a picture as a B-picture. The plane that is hidden by the cloud in picture #1 starts to appear in picture #2. If this picture would be coded as a B-picture, the clouds could be borrowed from picture #1, and the front part of the plane could be taken from picture #4. As for picture #4, it would be coded as a P-picture, using the clouds from picture #1. Only the plane would actually be coded in the P-picture.

Besides the picture reordering, B-pictures also require more memory in the decoder because an additional frame needs to be stored for later reference. This makes B-pictures quite a complex feature to implement in MPEG-2 Video. Because of the complexity of B-pictures, MPEG-2 defines subsets (profiles/levels) where B-pictures are not allowed. Sequences of pictures are grouped together to form GOPs. This can be done to support random access or editing functions. A typical, widely used GOP is the sequence IBBPBBPBBPBB [5.24]. All the B- and P-pictures of this GOP can be decoded by accessing only the I-picture or P-pictures, all belonging to this GOP. To support editing, the GOP structure contains a time-stamp. The time stamp format is actually defined by the Society of Motion Picture and Television Engineers (SMPTE) and corresponds to the time code that is also used in video studio equipment. There are two types of time stamps. The first type is usually called a reference time stamp. Reference time stamps are to be found in the Packetized Elementary Stream (PES) syntax, in the program syn-

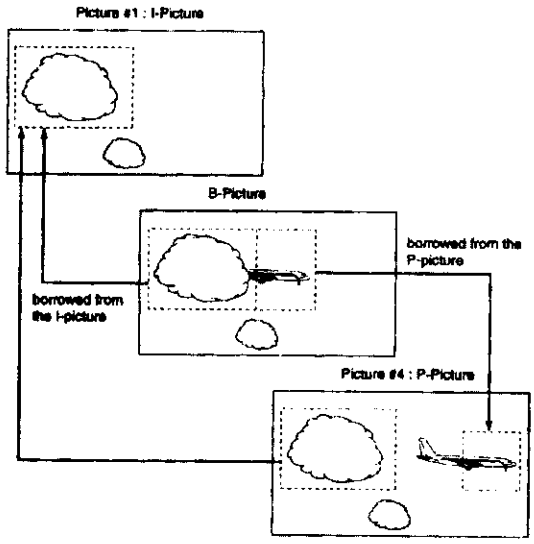

Figure 5.10 Use of a B-picture.

tax and in the transport syntax. The second type of time stamp is called Decoding Time Stamp (DTS). They indicate the exact moment where a video frame or an audio frame has to be decoded or presented to the user, respectively.

Slices are elements to support random access within a picture. A slice is a series of macroblocks. The slice contains information about where to display the contained macroblocks. In case of transmission errors and loss of picture information, the information in a slice can be used to continue the display process within a picture. Not all macroblocks of a picture must be included in slices. From a data compression point of view, slices are not really necessary. They are not coded to have resynchronization points within the picture.

MPEG-2 has introduced the concept of frame pictures and field pictures, along with particular frame prediction and field prediction modes to accommodate coding of progressive and interlaced video. For interlaced sequences, it is assumed that the coder input consists of a series of odd (top) and even (bottom) fields that are separated in time by a field period. Two fields of a frame may be coded separately (field pictures). In this case, each field is separated into adjacent nonoverlapping macroblocks, and the DCT is applied on a field basis. Alternatively, two fields may be coded together as a frame (frame pictures) similar to conventional coding of progressive video sequences. Here, consecutive lines of top and bottom fields are simply merged to form a frame. Notice that both frame pictures and field pictures can be used in a single video sequence.

New motion-compensated field-prediction modes were introduced by MPEG-2 to encode field pictures, and frame pictures efficiently. The concept of field pictures and possible field prediction are illustrated in Figure 5.11 for an interlaced video sequence, which in this figure is assumed to contain only three field pictures and no B-pictures. The top fields and the bottom fields are coded separately [5.24]. Each bottom field is coded using motion-compensated interfield prediction based on the previously coded top field. The top fields are coded using motion-

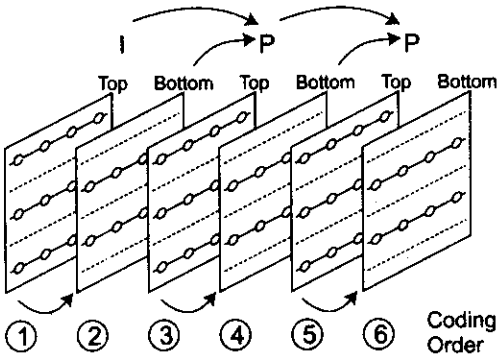

Figure 5.11 The concept of field pictures and possible field prediction.

compensated interfield prediction based on the previously coded bottom field. This concept can be extended to incorporate B-pictures. Generally, the interfield prediction from the decoded field in the same picture is preferred if no motion occurs between fields. An indication about which reference fields are used for prediction is transmitted with the bit stream. Within a field picture, all predictions are field predictions.

Frame prediction forms a prediction for a frame picture based on one or more previously decoded frames. In a frame picture, either field or frame predictions may be used and the particular prediction mode preferred can be selected on a macroblock-by-macroblock basis.
As for data compression, it is achieved by combining three techniques:

- Removing picture information that is invisible to the human eye
- Using variable length coding tables
- Using motion estimation

Because of its internal structure, the human eye is quite insensitive to high frequencies in color changes. The idea is, therefore, to represent the picture information in such a way that this characteristic of the eye is used. MPEG-2 uses a method, which is based on DCT, to approximate the original chrominance and luminance information in each block. Instead of using the real color values for each block, a set of frequency coefficients is calculated. This set describes the color transitions in the block. By dividing the resulting coefficients by a certain value, some of them can become zero after rounding. This is the step where picture information is lost. This process is called quantization, and the factors are provided by a quantization matrix. MPEG-2 defines default quantization matrixes, but also allows user-defined quantization matrixes. Quantization is also controlled by a scale factor, which allows the user to adjust the quantization level and the compression ratio. The scale factor is provided for each slice and can optionally be redefined for each macroblock. By having the scale factor in the quantization process, it becomes possible to generate constant bit rate videostreams, which fit into the constraints that might be given by a certain network architecture.

MPEG-2 defines a number of tables with codes to be used for specific patterns in a coefficient data sequence. The trick is to use very short codes with only a few bits for patterns that occur very often in the sequence. The quantization process results in a number of coefficients where certain coefficients equal zero (for example, 2, 0, 0, 1, 0, 0, 1). MPEG-2 Video is coding this sequence of coefficient data by an assigned code for a specific coefficient data pattern. The interpolation of this code returns two values. One value specifies the number of leading zeros in front of a nonzero coefficient. MPEG-2 Video uses the term "run" for this value. The other value is the actual coefficient, which is called "level" in MPEG-2 Video.

Example 5.7 The variable length codes and the corresponding run and level values can be seen in Figure 5.12. Based on this table, the sequence could be represented by the code sequence 0100, 0101 and 0101. The variable length coding tables use between 2 and 13 bits to encode run-level combinations. For the uncovered combinations, MPEG-2 Video defines an escape-coding mechanism. The level is then coded with an actual value.

Coefficient data sequence	Variable length coding table			Coded sequence
	Variable length code	Run	Level	
2, 0, 0, 1, 0, 0, 1	011	1	1	0100, 0101, 0101
	0100	0	2	
	0101	2	1	
	00101	0	3	
	00111	3	1	
	00110	4	1	
	000110	1	2	

Figure 5.12 Variable length coding.

The motion-estimation process uses the macroblocks as basic units for comparison. For each macroblock, the encoder is searching the previous picture (in the case of a P-picture) or the previous and the future pictures (in the case of a B-picture) for a macroblock that matches or closely matches the current macroblock. If such a macroblock is found, the difference between this macroblock and the current macroblock is calculated. The resulting difference is first DCT coded on an 8x8 block basis and then variable length coded, together with the motion vector of the macroblock. At decoding time, the motion vector is used to identify the macroblock in the previous or future picture. The identified macroblock will then be combined with the decoded

difference and written into the display or picture buffer. In the optimal case, the current macroblock is found at the same place in the previous picture. This would result in a zero motion vector together with a null difference. MPEG-2 would skip the coding of this macroblock. At decoding time, the previously displayed macroblock would stay on the screen.

MPEG-2 Video Syntax

Generally speaking, a syntax specifies the structure of a bitstream, such as how different parameters, tags, and so forth, are mapped and laid on the bitstream. For multiplexing purposes, it is important for the syntax to provide patterns that can be recognized with an extremely high degree of confidence. These patterns are called synchronization patterns. In addition, an indication of time and of the bit rate of the bit stream may also be provided.

The application requirements that MPEG-2 has addressed made it necessary to develop a formal syntax that supported all requirements. Some syntax elements control the appearance of other syntax elements. Many syntax elements are optional and are only present in the bitstream if a flag indicates it. A flag is mostly located in the syntax structure header [5.22]. In that way, the amount of data that has to be transmitted is further reduced. Instead of transmitting void values, like zeros or special codes, some elements are simply not present in the bit stream. The MPEG-2 standard uses a C-like pseudocode to describe the syntax.

MPEG-2 Video Scalability

The scalability is achieved by the MPEG-2 Video syntax. Video information can be separated into different information streams, which are complementary. Different applications can be realized by combining different streams of information. MPEG-2 uses the term "layer" for the different information streams. The intention of scalable coding is to provide interoperability between different services and to support receivers flexibly with different display capabilities. Receivers either not capable or willing to reconstruct the full resolution video can decode subsets of the layered bit stream to display video at lower spatial or temporal resolution or with lower quality. Another important purpose of scalable coding is to provide a layered video bit stream that is amenable for prioritized transmission. The main challenge here is to deliver video signals reliably in the presence of channel errors, such as cell loss in ATM-based transmission networks or cochannel interference in terrestrial digital broadcasting.

Example 5.8 Flexible supporting multiple resolutions is of particular interest for interworking between HDTV and Standard Definition Television (SDTV). In this case, it is important for the HDTV receiver to be compatible with the SDTV product. Compatibility can be achieved by means of scalable coding of the HDTV source, and the wasteful transmission of two independent bit streams to the HDTV and SDTV receivers can be avoided. Other important applications for scalable coding include video database browsing and multiresolution playback of video in a multimedia environment.

Scalability can be applied for different aspects of video presentation. Although some applications are constricted to low implementation complexity, others call for very high coding efficiency. As a consequence, MPEG-2 has standardized several scalable coding schemes.

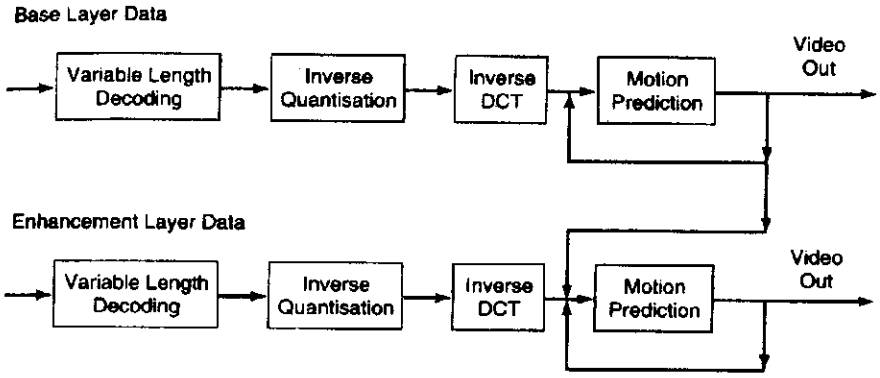

Figure 5.13 The decoding process in the case of spatial scalability.

Spatial scalability has been developed to support displays with different spatial resolutions at the receiver. Lower spatial resolution video can be reconstructed from the base layer. With spatial scalability, the enhancement layer and the base layer data are combined after the IDCT step. The process mainly affected by spatial scalability is motion compensation, which can now use motion vectors from the enhancement layer data or from the base layer data. Decoding flow in the case of spatial scalability is illustrated in Figure 5.13. This functionality is useful for many applications, including embedded coding for HDTV/TV systems, which allows a migration from a digital TV service to higher spatial resolution HDTV services [5.26, 5.27]. The algorithm is based on a classical pyramidal approach for progressive image coding [5.20, 5.28]. Spatial scalability can flexibly support a wide range of spatial resolutions, but adds considerable implementation complexity to the coding scheme.

Temporal scalability defines the possibility to handle different picture rates in a single video stream. This tool was developed with an aim simpler than spatial scalability [5.24]. Namely, stereoscopic video can be supported with a layered bit stream suitable for receivers with stereoscopic display capabilities. The base layer, which provides the basic video picture, can be combined with the enhancement layer to achieve higher frame rates. The enhancement layer uses the base layer to generate final video pictures. Layering is achieved by providing a prediction of one of the images of the stereoscopic video in the enhancement layer based on coded images from the opposite view transmitted in the base layer. Possible usage is in the support of different generations of decoder equipment with different transmission qualities. As in the case of spatial scalability, the enhancement happens after the IDCT step and mainly affects the motion compensation process.

SNR scalability allows for the handling of at least two different video qualities. The video information provided by the base layer can be improved by one or more enhancement layers carrying additional information. However, the base and enhancement layers have the same spatial video resolution [5.22]. If the base layer can be protected from the transmission errors, a version of the video with gracefully reduced quality can be obtained by decoding the base layer bitstream. The algorithm used to achieve graceful degradation is based on a frequency (DCT

domain) scalability technique. At the base layer, the DCT coefficients are coarsely quantized and transmitted to achieve moderate image quality at a reduced bit rate. The enhancement layer encodes and transmits the difference between the nonquantized DCT coefficients and the quantized coefficients from the base layer with finer quantization SZ. At the decoder, the highest quality video signal is reconstructed by decoding both the base layer and the higher layer bit streams. The main application of the SNR scalability is error concealment. In this case, the base layer would carry the most critical information while using a quite robust transport channel in a network. The enhancement layer, bearing the less critical information, could be transported across a transport channel with a lower quality of service. The enhancement process in the case of the SNR scalability happens after the inverse quantization process. The enhancement layer contains mainly DCT coefficients, which are added to the one provided by the base layer. By doing this, the picture quality is refined. Decoding flow in the case of SNR scalability is shown in Figure 5.14. It is also possible to use this method in order to obtain video with lower spatial resolution at the receiver. If the decoder selects the lowest *N*x*N* DCT coefficients from the base layer bit stream, nonstandard IDCTs of size *N*x*N* can be used to reconstruct the video at a reduced spatial resolution [5.29, 5.30]. However, depending on the encoder and decoder implementation, the lowest layer downscaled video may be subject to drift [5.31].

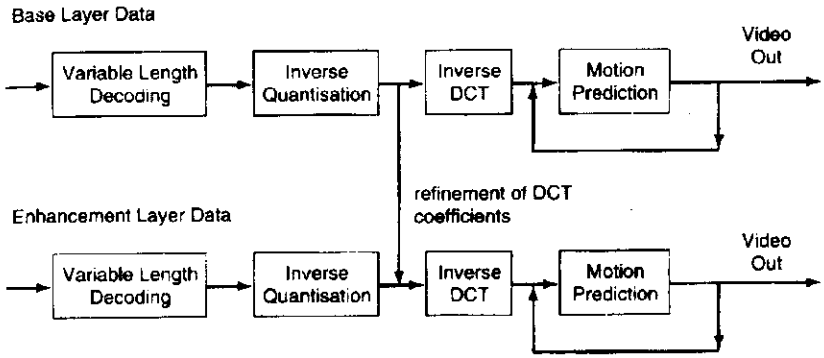

Figure 5.14 The data flow in the decoder for SNR scalability.

Data partitioning is intended to assist with error concealment in the presence of transmission or channel errors in ATM, terrestrial broadcast or magnetic recording environments. Because the tool can be entirely used as a postprocessing and preprocessing tool to any single layer coding scheme, it has not been formally standardized with MPEG-2, but is referenced in the informative Annex of the MPEG-2 Draft International Standard (DIS) document [5.27]. The algorithm is similar to the SNR scalability tool, based on the separation of DCT coefficients, and is implemented with very low complexity compared to the other scalable coding schemes. To provide error protection, the coded DCT coefficients in the bit stream are simply separated and transmitted in two layers with different error likelihoods. In data partitioning, the complete video bit stream is split into parts with relatively higher or lower importance. The most impor-

tant syntax elements are transmitted on a higher performance channel, and the less important elements are transmitted on a channel with lower performance. A special syntax element, called priority breakpoint, is used to define which parts of the video bit stream syntax are put into which partition.

MPEG-2 Video: Profiles and Levels

Because of the great range of applications, the MPEG-2 standard became quite complex. However, an application might not need the full range of the MPEG-2 Video feature set. If it had to support the complete specification, MPEG-2 equipment would be expensive. Therefore, this standard defines profiles and levels to define subsets of MPEG-2 Video. Profiles and levels defined by MPEG-2 Video are shown in Table 5.5.

Table 5.5 Profiles and levels for MPEG-2 Video.

Profiles	Algorithms	Levels
Simple Profile (SP)	Includes all functionalities provided by the Main profile, but does not support B-picture prediction modes and 4:2:0 YUV-representation.	Low Level (LL)
Main Profile (MP)	Supports functionality for coding-interlaced video, random access, B-picture prediction modes and 4:2:0 YUV-representation with nonscaling coding algorithm.	Main Level (ML)
SNR Scalable Profile (SNR)	Supports all functionalities provided by the MP plus an algorithm for SNR scalable coding (two layers allowed) and 4:2:0 YUV-representation.	High 1440 Level (H14)
Spatial Scalable Profile (Spatial)	Supports all functionalities provided by the SNR Scalable Profile plus an algorithm for spatial scalable coding (two layers allowed) and 4:2:0 YUV-representation.	High Level (HL)
High Profile (HP)	Supports all functionalities provided by the Spatial Scalable Profile plus the provision to support three layers with the SNR and Spatial scalable coding modes and 4:2:2 YUV-representation for improved quality requirements.	

A profile is described as a well-defined subset of the video syntax. Certain syntaxes defined by MPEG-2 Video are not valid and cannot be decoded if the decoder only supports a low profile. Neither the Simple nor the Main profile supports any kind of scalability. However, the lower profiles are always a subset of the higher profiles. A decoder supporting the spatial profile is required to support spatial and SNR scalabilities.

Table 5.6 MPEG-2 Main level at Main profile values.

Parameter	Main level at Main profile value
Samples/line	720
Lines/frame	576
Frames/sec	30
Luminance samples/sec	1,036,8000
Max. video data rate in Mb/s	15
Max. size of decoder buffer (bits)	1,835,008

A level defines values for certain parameters in the video bit stream. The levels describe the number of samples per line, the number of lines per frame and the number of frames per second. Profiles and levels are combined to define exactly which selection or subset from the MPEG-2 Video toolkit is used. The combination Main level at Main profile defines a sufficient subset of the MPEG-2 Video functionality. Table 5.6 shows some of the Main level–Main profile values. Profiles and levels are organized hierarchically, and MPEG-2 Video defines a forward compatibility between different profiles and levels.

5.4.2 MPEG-2 Audio

MPEG-2 Audio is a backward-compatible multichannel extension of the MPEG-1 Audio standard. Namely, the audio part of the MPEG-2 standard is to a great extent based on the MPEG-1 audio part, and a great deal of compatibility exists. The compatibility aspect is valid in two senses [5.16, 5.22]:

- Existing MPEG-1 equipment can make a partial decoding of MPEG-2 signals by extracting the MPEG-1 compatible part (backward compatibility).
- MPEG-2 equipment can decode MPEG-1 signals (forward compatibility).

The fact that the core of the MPEG-2 bit stream is an MPEG-1 bit stream enables fully compatible decoding with an MPEG-1 decoder. In addition, the need to transfer two separate bit streams, called simulcast (one for two-channel stereo and another one for the multichannel audio program), is avoided. This incurs some cost in coding efficiency for the multichannel audio signal compared to AAC, which is a Nonbackward Compatible (NBC) coding algorithm [5.16].

Any combinations of MPEG-1 and MPEG-2 audio and video can be handled by the system as specified by the MPEG Systems standards [5.11, 5.16]. For example, MPEG-2 Audio can be used with MPEG-1 Video. Also, MPEG-1 Audio can be used with MPEG-2 Video without any restrictions.

MPEG-2, as well as MPEG-1, audio compression describes three degrees of compression: layers 1, 2 and 3. These layers represent a family of coding algorithms. Basically, the complexity of the encoder and decoder, the encoder/decoder delay and the coding efficiency increases when going from layer 1 through layer 2 to layer 3. Layer 1 has the lowest complexity and is specifically suitable for applications where the encoder complexity also plays an important role. Layer 2 requires a more complex encoder and a slightly more complex decoder and is directed toward one-to-many applications, that is, one encoder serves many decoders. Compared to layer 1, layer 2 is able to remove more of the signal redundancy and to apply the psychoacoustic threshold more efficiently. Layer 3 is more complex and is directed toward lower bitrate applications due to the additional redundancy and irrelevancy extraction from enhanced frequency resolution in its filterbank. The main functional modules of the lower layers are also used by the higher layers. The subband filter of layer 1 is also used by layer 2 and layer 3. Layer 2 adds a more efficient coding of side information. Layer 3 adds a frequency transform in all the subbands. The three layers have been defined to be compatible in a hierarchical way. The level of compression, the demands for processing power and the sound quality all increase proportionally with the layer number. The required transmission bandwidth decreases with the layer number. Layer 1 has the lowest compression rate, about four times. It demands the smallest amount of processing power and has the lowest delay, realistically below 50 ms. Layer 1 also has the highest requirements for transmission bandwidth, with the highest possible level of compression, going to 448 Kbps in stereo, with the lowest possible level of compression. The sound quality of the layer 1 signal is furthermore inferior to what can be obtained by layers 2 and 3. Layer 3 is intended to yield the best sound quality of the three layers, but it achieves a compression ratio, 10:1. On the other hand, the processing time is more than three times longer. Central characteristics of the three layers of MPEG-2 audio coding are presented in Table 5.7. These three layers are compatible in the sense that a layer N decoder can decode layer N, as well as all the layers below. For example, a layer 3 decoder can decode layers 1, 2 and 3 bit streams, but a layer 2 decoder can only decode bit streams from layers 1 and 2.

The overall audio compression and encoding processes for all three layers are shown in Figure 5.15. In the case of signals with more than one channel, each channel is treated separately. The filter bank used in MPEG-2 audio coding can be one of the following types: polyphase or a hybrid polyphase and MDCT. Regardless of the type, the time domain samples are here con-

Table 5.7 Main characteristics of the three layers of MPEG-2 audio coding.

Layer	Approximate compression ratio	Target bit rate	Realistic delay	Theoretical minimum delay
1	4:1	192 Kb/s	<50 ms	19 ms
2	6:1	128 Kb/s	100 ms	35 ms
3	10:1	64 Kb/s	150 ms	58 ms

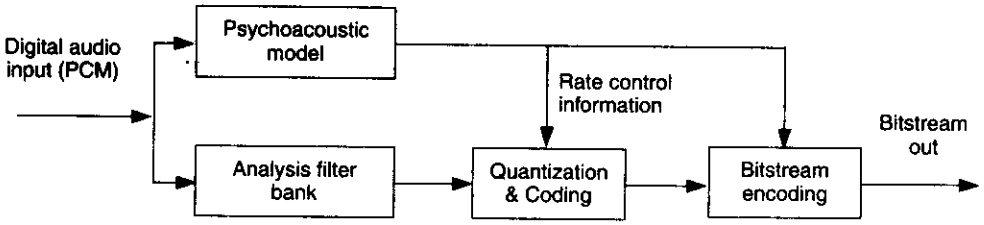

Figure 5.15 The overall audio compression and encoding processes for three layers.

verted into the same number of frequency domain samples. The output of the filter bank is a number of subbands of equal bandwidth. In layers 1 and 2, a filter bank yielding 32 subbands, each containing 12 or 36 frequency domain samples, respectively, is used. In layer 3, the number of subbands can be either 192 or 576. In parallel with the filter bank, the psychoacoustic model process calculates the Signal-to-Mask Ratio (SMR) for each subband. The primary psychoacoustic effect that a perceptual audio coder uses is called auditory masking, where parts of a signal are not audible due to the function of the human auditory system. The parts of the signal that are masked are commonly called irrelevant, as opposed to parts of the signal that are removed by a source coder (lossless or lossy), which are termed redundant.

In order to remove this irrelevance, the encoder has a psychoacoustic model. This psychoacoustic model analyzes the input signals within consecutive time blocks. It then determines for each block the spectral components of the input signals within consecutive time blocks and determines for each block the spectral components of the input audio signal by applying a frequency transform. Then it models the masking properties of the human auditory system and estimates the just-noticeable noise level, sometimes called the threshold of masking.

In parallel, the input signal is fed through a time-to-frequency mapping, resulting in spectrum components for subsequent coding. In its quantization and coding stages, the encoder tries to allocate the available number of data bits in a way that meets both the bit rate and masking requirements. The information on how the bits are distributed across the spectrum is contained in the bit stream as side information.

The principal function of the psychoacoustic model is to calculate a new bit allocation for the frequency samples in the subbands. The new bit allocation is aimed at efficiently allocating the available bits to each of the subbands. If a given subband does not have power, no bits are allocated. The principle applied is the fact that frequencies with higher power make nearby frequencies of lower power inaudible to human hearing. The new bit allocation is calculated separately for each subband.

To obtain the SMR, it is necessary to first make a time-to-frequency domain compression of the original audio signal. This works in parallel with the filter bank process. This conversion is done by the FFT technique, which allows a time-to-frequency domain transformation with a better spectral resolution than that of the polyphase filter bank. On the basis of frequency domain data, the maximum power in each subband is found, the tonal and nontonal (noiselike) parts of the audio signal are determined, the absolute masking threshold is identified, and,

finally, the masking thresholds of all the individual subbands are calculated. A global masking threshold is calculated by adding all the individual masking thresholds with the absolute threshold. Now it is possible to calculate for each subband the difference between the actual signal and the masking threshold and to obtain the SMR.

The bit or noise allocation process uses the output samples from the filter bank and the SMR information from the psychoacoustic model to determine the amount of quantization noise that is tolerable to each subband. The higher the allowable quantization noise means the lower the number of necessary bits to represent each sample. In layer 1, a scale factor is applied for each subband (containing 12 audio samples). In layer 2, each subband contains 36 audio samples, which are split into 3 groups of 12 samples. Each of the three groups can have separate scale factors. The scale factors are calculated separately for each subband. The scale factor, which will be transmitted along with the audio samples to the decoder, expresses a certain factor that the resolution steps of the audio samples will have to be multiplied by the decoder. It is possible to express small, as well as large, amplitude steps with a relatively small number of bits. For layers 1 and 2, the number of bits used to represent the audio samples in each subband is variable when the requirements of the bit rate and those of the psychoacoustic model are to be combined. In layer 3, this is done somewhat differently because of noise injected in the subbands is a variable. In both cases, the encoder starts an iterative process of increasing the accuracy of the subband quantization, to the limit possible within the specified bit rate.

In the bit stream formatting process, the subband frequency samples are joined together with the scale factor information in the audio data field of the audio frame. The audio frame contains header, error check and ancillary data fields. For layer 3, Huffman coding of the quantized frequency samples is applied at this step. It means that, instead of fixed-length PCM for each frequency sample, as used in layers 1 and 2, a variable-length code is used. Huffman coding represents the most common bit combinations in the data flow with the shortest codes, and it represents the most uncommon with the longest codes. Thus, further reduction in the bit rate can be obtained.

The decoder is much less complex. It does not require a psychoacoustic model and bit allocation procedure. Its only task is to reconstruct an audio signal from the coded spectral components and associated side information.

With MPEG-2, it is possible to use only half the sampling rate in MPEG-1 and still obtain a very good sound quality. This is especially interesting for applications such as commentary channels, multilingual channels and multimedia. The sampling rate available in MPEG-2 allows the sampling of the time domain signal to be done with 16, 22.05 or 24 KHz samples per second for all three layers. This gives an upper frequency limit of 7.5, 10.3 and 11.25 KHz, respectively. To allow transmission of more realistic stereoscopic representation, MPEG-2 supports audio channels, which together can convey a surround stereo image. The five channels are left channel (L), right channel (R), center channel (C), left rear surround channel (LS) and right rear surround channel (RS). This is also called 3/2 stereo because it makes use of three front loudspeakers and two loudspeakers in the rear. Surround setup in 3/2 stereo is shown Figure 5.16. In addition, a Low Frequency Enhancement (LFE) channel is available for a subwoofer signal in

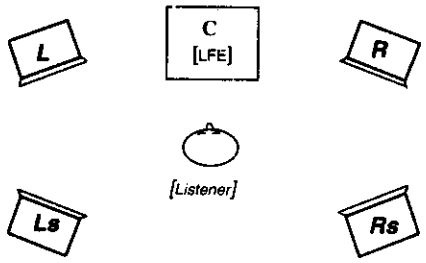

Figure 5.16 Surround sound setup in 3/2 stereo.

the range from 15 to 120 Hz. This channel is mainly used for special effects. An alternative configuration of the five-channel system of MPEG-2 Audio is the application of multilingual/commentary channels, accompanying a specific program with, for example, bilingual comments or sound tracks. The MPEG-2 specifications allow for up to seven multilingual/commentary channels per program. As the surround sound signal consists of five channels, a great deal of redundancy often exists among these five channels, so, in many cases, the same piece of audio information may appear in two or more of the five surround channels with different delays. The MPEG-2 Audio compression can use this redundancy to achieve a higher compression ratio.

The audio part of the MPEG-2 standard has focused on the compatibility with MPEG-1 Audio. However, the compatibility has some drawbacks. For example, the used techniques have not been able to eliminate completely the fact that there is a trade-off between maintaining compatibility and achieving optimal sound quality at a given bit rate. Listening tests have proved the backward compatible coding techniques to be somewhat inferior to other nonbackward compatible techniques under the condition that the same number of bits per second were available.

In the framework of MPEG-2, work was going on in the field of AAC, also known as NBC coding. The MPEG-2 AAC standard is the state-of-the art audio standard that provides very high audio quality at a rate of 64 Kb/s/channel for multichannel generation. It provides a capability of up to 48 main audio channels, 16 low frequency effects channels, 16 overdub/multilingual channels and 16 data streams. Up to 16 programs can be described, each consisting of any number of the audio and data elements. The AAC standard has three profiles called main profile (AAC), Low Complexity (LC) profile and Scalable Sampling Rate (SSR) profile. The main profile is intended for use when processing and especially memory are not a premium. The LC profile is intended for use when cycles and memory use are constrained, and SSR profile is intended for use when a scalable decoder is required. The main and LC profiles have been tested at 320 Kb/s for five-channel audio programs, and both have demonstrated better quality than competing audio-coding algorithms running at 640 Kb/s for the five-channel program.

5.4.3 MPEG-2 Systems

The MPEG-2 Systems standard is an ISO/IEC standard that defines the syntax and semantics of bit streams in which digital audio and visual data are multiplexed. Such bit streams are said to be MPEG-2 Systems compliant. However, this specification does not mandate how equipment that produces, transmits or decodes such bit streams should be designed. As a result, the specification

can be used in a diverse array of environments, including local storage, broadcast (terrestrial and satellite) and interactive environments [5.23]. This standard was industry driven and complemented the MPEG-2 activities in audio and video coding. The consumer TV industry actively participated in the definition of MPEG-2 Systems, to ensure that a low-complexity receiver could be built at a reasonable cost.

The MPEG-2 Systems standard enables the widest interoperability in digital video and audio applications and services. The video and audio part of the MPEG-2 standard defines the format with which audio or video information is presented. However, to use this data in a complete video delivery chain, some additional requirements have to be addressed. These requirements result from the applications in which the audio and video data is used. They are also related to the technology that is used to deliver the data.

Example 5.9 Let us take the standard TV broadcasting application. In TV broadcasting, there is a need to transport different programs to the consumer, who can freely choose among them. In other words, at some point, different audio-video streams have to be multiplexed together and have to be delivered together to the consumer. This multiplexing is usually done somewhere in TV broadcasting networks, like satellite or cable distribution systems. In the case of a satellite distribution system, different programs delivered by different broadcasting stations are multiplexed together at some satellite uplink station. The collection of programs, sometimes also called bouquet, is transmitted to the satellite, which then sends it down to earth in a Direct To Home (DTH) broadcasting system.

MPEG-2 Systems are using data structures that are commonly referred to as packets in the data communication world. Packets always consist of a packet header and the packet payload and can be of fixed or variable size. The basic idea behind a packet concept is to create a flexible mechanism to transport any kind of data. Usually the packet header contains the information that is needed to process the data in the packet payload. Depending on the application scenario where the packets are used, it makes sense to use variable or fixed-sized packets. For example, in a network environment, it is useful to have fixed-sized packets that are relatively short. This helps to optimize the network equipment that is processing the packets because the length of the packet is always the same. If a part of the packet is corrupted for some reason (loss of data in the network), only that information is lost. Figure 5.17 shows the scope of the MPEG-2 Systems specifications.

MPEG-2 Systems provide a two-layer multiplexing approach. The first layer is dedicated to ensuring tight synchronization between video and audio. It is a common way for presenting all the different materials that require synchronization (video, audio and private data). This layer is called PES. The second layer is dependent on the intended communication medium. The specification for error-free environments, such as local storage, is called MPEG-2 Program Stream, and the specification addressing error-prone environment is called MPEG-2 Transport Stream.

In MPEG-2, the output bit stream of an audio-video encoder or the private data bit stream is called the elementary stream. In the case of audio or video, this elementary stream can be organized into access units. An access unit is a picture in the case of a video elementary stream or an audio frame, in the case of an audio elementary stream.

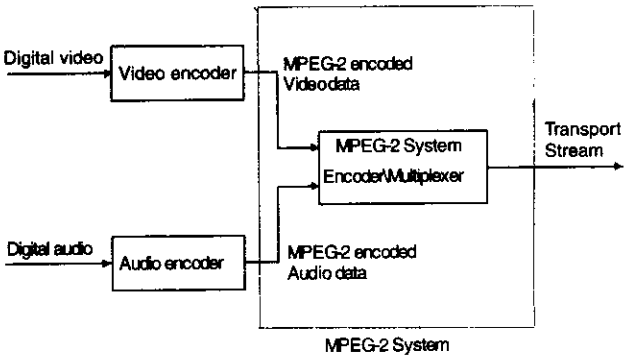

Figure 5.17 Scope of the MPEG-2 Systems standard in relation to the video and audio parts and the broadband equipment.

An elementary stream is then converted into a PES, which consists of PES packets. Each PES packet consists of the PES packet payload (which is a variable-sized part of the elementary stream) and a PES packet header. By having the size of the payload variable, the payload of the PES packets can be an exact access unit of the elementary stream.

A PES packet is a way to packetize the elementary streams uniformly. Embedded in a PES packet, elementary streams may be synchronized with time stamps. They are not protected. The PES packets may be of variable length, which allows them also to be of fixed length. PES packets may be rather long packets. However because elementary streams are continuous streams, it is also possible to know that a PES packet is finished when the next PES packet arrives. Sometimes the length is not relevant (for video PES packets).

The PES packet is mapped into the MPEG-2 Transport Stream Packet (TSP), also consisting of a header and a payload part. Consecutive transport packets form the MPEG-2 transport stream. The functionality of the MPEG-2 Systems layer processor is summarized in Figure 5.18. An MPEG-2 Systems layer processor could extract the PES packets out of the transport stream and put them into program stream packets. MPEG-2 transport streams carry transport packets. These packets carry two types of information: the compressed material and the associated signaling tables. A transport packet is identified by its Packet Identifier (PID). Each PID is assigned to carry either one particular compressed material (and only this material) or one particular sig-

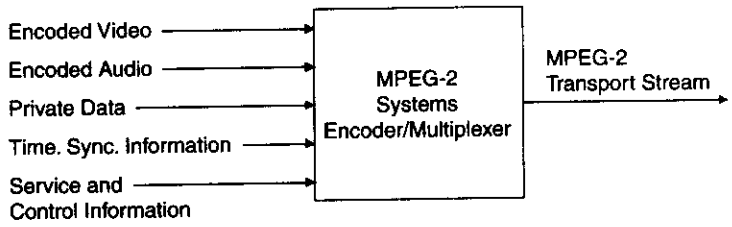

Figure 5.18 Layer processor for MPEG-2 System.

naling table. The compressed material consists of elementary streams, which may be built from video, audio or data material. These elementary streams may be tightly synchronized (because it is usually necessary for digital TV programs or for digital radio programs) or not synchronized (in the case of programs offering downloading of software or games, as an example).

The associated signaling tables consist of the description of the elementary streams, which are combined to build programs, and in the description of those programs. Tables are carried in sections. The signaling tables are called Program Specific Information (PSI).

Transport packets are 188 bytes long because MPEG-2 wanted these packets to be carried across ATM. At that time, according to the AAL, ATM cells are supposed to have a payload of 47 bytes (4x47 bytes = 188 bytes). A transport packet of 188 bytes maps exactly into the payload of four ATM cells. An ATM cell has 48 bytes of payload, but one byte of the payload is used for the overhead information of the AAL.

MPEG-2 Systems distinguish between two kinds of transport streams: Single Program Transport Streams (SPTS) and Multiprogram Transport Streams (MPTS). The SPTS contains different PES, which all share a common time base. The different PES could carry video, different audio and perhaps data information. All would be used with the same time base. An application for this is a movie transmitted in different languages. MPTS is multiples of a number of SPTS. The MPEG-2 System hierarchy based on different transport stream variations is presented in Figure 5.19.

The MPEG-2 transport packet consists of 4 bytes of header information, a variable length adaptation field and the payload, containing the PES packets. An MPEG-2 transport packet header is shown in Figure 5.20. One of the most important fields in the header is the PID. The PID is used to identify transport packets that carry PES data from the same elementary stream. It also defines

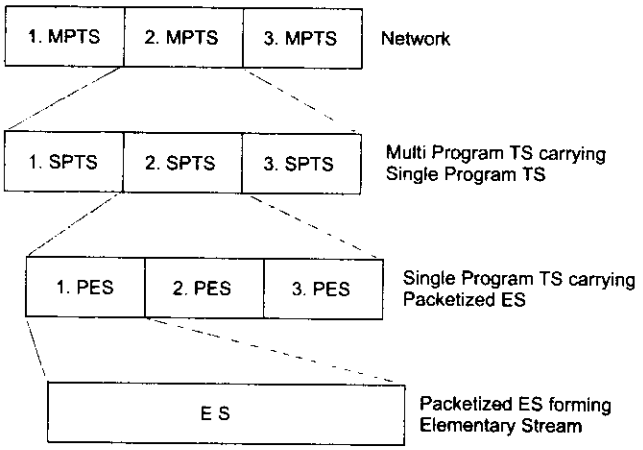

Figure 5.19 MPEG-2 Systems hierarchy based on the different transport streams.

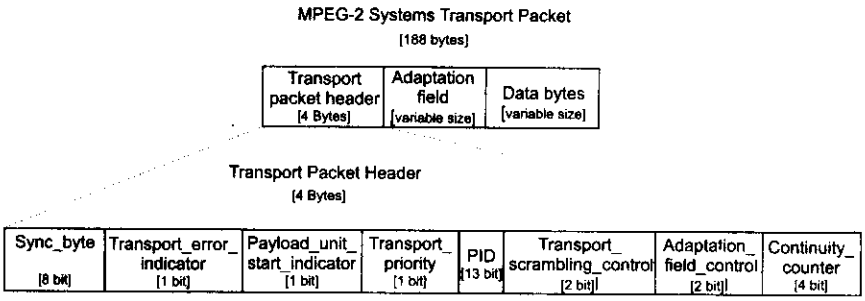

Figure 5.20 MPEG-2 transport packet header.

the type of data that is transported by the packet payload. Besides the PID, the transport packet header contains several control fields, which are used to identify the appearance of other fields in the header and also to provide the information about the payload of the transport packet. A very important field is the adaptation field. It is an optional field in the transport stream packet header, which contains additional information that is used for clock recovery and splicing functions. One of the most important fields in the adaptation field is the Program Clock Reference (PCR) field. This field contains time stamp information that is used by the decoder to synchronize its clock to the encoder clock. The adaptation field also has a section to transport private data, which is not defined by the MPEG-2 standard.

5.4.4 MPEG-2 DSM-CC

MPEG-2 DSM-CC is the specification of a set of protocols that provides the control functions and operations specific to managing MPEG-1 and MPEG-2 bit streams. These protocols may be used to support applications in both standalone and heterogeneous network environments [5.32]. In the DSM-CC model, a stream is sourced by a server and delivered to a client. Both the server and the client are considered to be users of the DSM-CC network. DSM-CC defines a logical entity called the Session and Resource Manager (SRM), which provides a logically centralized management of the DSM-CC sessions and resources. The DSM-CC reference model is shown in Figure 5.21.

To a user, DSM-CC allows the delivery of multimedia across a guaranteed end-to-end QoS irrespective of the transport technology that the user is using. Thus, it allows the end-users the choice of the transport technology and media within the locality that the service is provided and that best suits their budget.

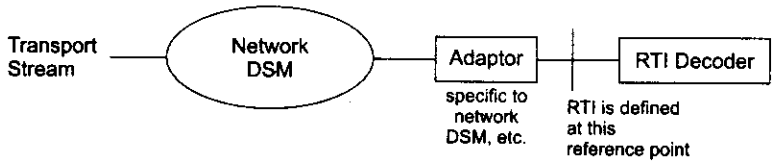

Figure 5.21 DSM-CC reference model [5.25]. ©1995 ISO/IEC.

The DSM-CC protocols supplement other networking protocols, like BISDN signaling or transport protocols, in order to cover all requirements of video networks. DSM-CC signaling assumes that the network links between the different entities are already established.

After an initial link has been set up between the two entities in the video delivery network, DSM-CC provides the functionality to continue the setup of an application session. Because this session setup happens at the interface between network and user equipment, DSM-CC defines a DSM-CC user to network protocol. After the application session has been set up, further logical links are established between a video server and a set-top box. One logical link might be used for user data (like MPEG-2-coded video), and another logical link might be used to control what is happening on the user data link.

DSM-CC defines a set of services to manipulate a videostream in the server, which can be used by the client. Because these services are only relevant between two user entities, that is, the server and the client, the DSM-CC standard refers to them as the DSM-CC user-to-user interface [5.33]. Links between server and client are shown in Figure 5.22, and Figure 5.23 illustrates the two different interfaces that DSM-CC addresses.

The user-network interface has much in common with OSI layer 3 signaling protocols. The user-user part of DSM-CC is application layer oriented and uses an object-oriented approach.

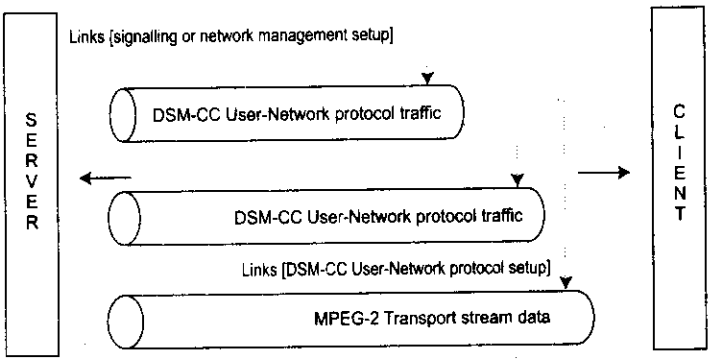

Figure 5.22 Links between server and client.

5.5 MPEG-4—Coding of Audiovisual Objects

Multimedia communication is the possibility to communicate audiovisual information with the following characteristics:

- Is natural, synthetic or both
- Is real time and non-real time
- Supports different functionalities responding to user's needs
- Flows to and from different sources simultaneously
- Does not require the user to bother with the specifics of the communications channel, but uses a technology that is aware of it
- Gives users the possibility to interact with the different information elements

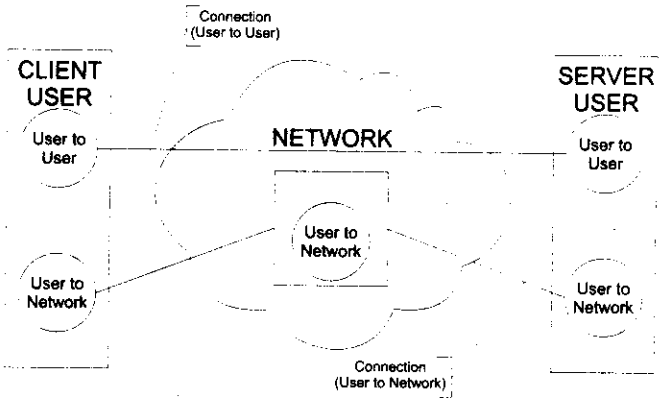

Figure 5.23 DSM-CC user-to-network and user-to-user interfaces.

- Lets the users present the results of their interaction with content in a way that suits their needs [5.9]

MPEG-4 is the MPEG project that started in July 1993 and was developed to provide enabling technology for the previous seven items. It has reached Working Draft level in November 1996, Committee Draft level in November 1997, and International Standard level in the beginning of 1999. To reach its own target, MPEG-4 follows an object-based representation approach where an audiovisual scene is coded as a composition of objects, natural as well as synthetic, providing the first powerful hybrid playground. Thus, the objective of MPEG-4 is to provide an audiovisual representation standard supporting new ways of communication, access and interaction with digital audiovisual data and offering a common technical solution to various service paradigms—telecommunications, broadcast and interactive—among which the borders are disappearing. MPEG-4 will supply an answer to the emerging needs of application fields such as video on the Internet multimedia broadcasting; content-based audiovisual database access; games; audiovisual home editing; advanced audiovisual communications, notably across mobile networks; teleshopping; and remote monitoring and control [5.34].

The fully backward compatible extensions under the title of MPEG-4 Version 2 were frozen at the end of 1999 to acquire the formal International Standard Status in early 2000. Some work on extensions in specific domains is still in progress. MPEG-4 builds on the proven success of three fields [5.35, 5.36]:

- Digital television
- Interactive graphics applications (synthetic content)
- Interactive multimedia (distribution of and access to content on the Web)

5.5.1 Overview of MPEG-4: Motivations, Achievement, Process and Requirements

MPEG-1 and MPEG-2 standards have given rise to widely adopted commercial products and services, such as Video-CD, DVD, digital television and digital audio broadcasting. The aim of the MPEG-4 standard is to define an audiovisual coding standard to address the emerging needs of the communications, interactive and broadcasting service models, as well as the needs of the mixed service models resulting from their technological convergence. The convergence of the three separate application areas, communications, computing and TV/film/entertainment, was evident in the mutual cross-fertilization with functionalities characteristic of each one of these application areas such as user interactivity, sythentic and natural hybrid coding, and Web-based services emerging more and more.

Although audiovisual information, notably the visual part, was until recently only carried across a few networks, the trend is now toward the generalization of visual information on every single network. New mobile devices can finally make the decade-long promise of audiovisual communications turn into reality. The explosion of the Web and the acceptance of its interactive mode of operation have clearly shown that the traditional television paradigm would no longer suffice for audiovisual services. Users want to have access to audio and video like they now have access to text and graphics. This requires moving pictures and audio of acceptable quality at low bit rates on the Web and providing Web-type interactivity with live content.

The MPEG-4 standard provides a set of technologies to satisfy the needs of authors, service providers and end-users. For authors, MPEG-4 enables the production of content that has greater reusability and that has greater flexibility than is possible today with individual technologies, such as digital television, animated graphics, Web pages and their extensions. For network services providers, MPEG-4 offers transparent information, which can be interpreted and translated into the appropriate native signaling messages of each network with the help of relevant standards bodies. The foregoing excludes QoS considerations, for which MPEG-4 provides generic QoS descriptors for different MPEG-4 media. For end-users, MPEG-4 brings higher levels of interaction with content within the limits set by the author. It also brings multimedia to new networks, including those employing a relatively low bit rate, and mobile networks. An MPEG-4 application document exists on the MPEG home page and describes many end-user applications, including interactive multimedia broadcast and mobile communications. MPEG-4 achieves these goals by providing standardized ways to do the following:

- Represent units of aural, visual or audiovisual content called media objects. These media objects can be of natural or synthetic origin; this means they could be recorded with a camera or microphone or generated with a computer.
- Describe the composition of these objects to create compound media objects that form audiovisual scenes.
- Multiplex and synchronize the data associated with media objects so that it can be transported across network channels providing a QoS appropriate for the nature of the specific media objects.

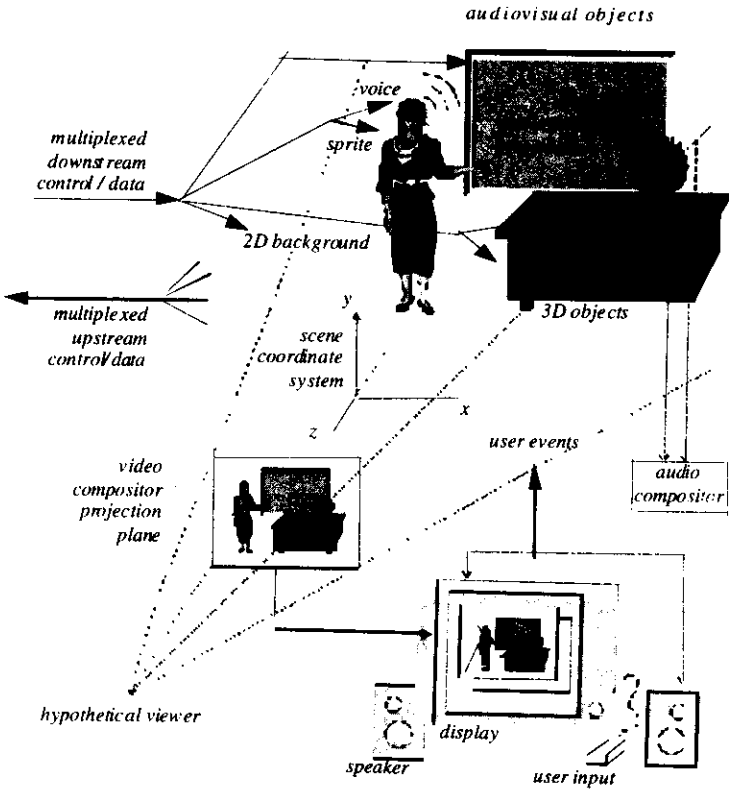

Figure 5.24 An example of an MPEG-4 audiovisual scene [5.36]. ©2001 ISO/IEC.

- Interact with the audiovisual scene generated at the receiver's end.

An audiovisual scene is depicted in Figure 5.24. The figure contains compound media objects that group primitive media objects together. Primitive media objects correspond to leaves in the descriptive tree, and compound media objects encompass entire subtrees.

A major difference from previous audiovisual standards on the basis of new functionalities, is the object-based audiovisual representation model that underpins MPEG-4. The MPEG-4 object-based architecture is shown in Figure 5.25. An object-based scene is built using individual objects that have relationships in space and time and that offer a number of advantages. First, different object types may have different suitable coded representations. A synthetic moving head is clearly best represented using animation parameters, but video benefits from a smart representation of pixel values. Second, it allows harmonious integration of data into one scene. Third, interacting with the objects and hyperlinking from them is now feasible. There are more advantages, such as selective spending of bits, easy sense of content without transcoding, providing sophisticated schemes for scalable content on the Internet, and so forth.

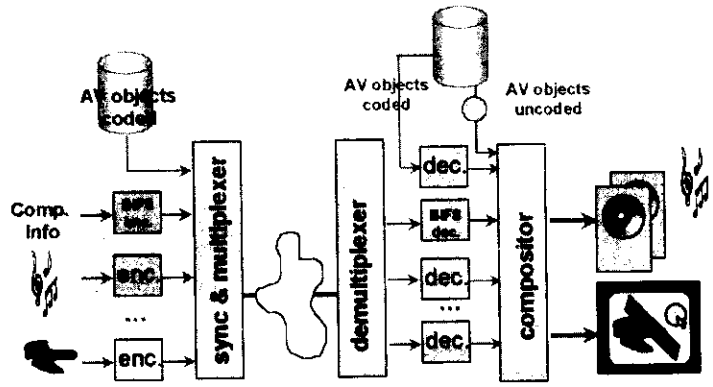

Figure 5.25 The MPEG-4 object-based architecture [5.8].
©2000 Elsevier.

MPEG-4 audiovisual scenes are composed of several media objects organized in a hierarchical fashion. At the leaves of the hierarchy, we can find primitive media objects like still images (a fixed background), video objects (a talking person without the background), audio objects (the voice associated with that person), and so forth.

MPEG-4 standardizes a number of primitive media objects and is capable of representing both natural and synthetic content types, which can be 2D or 3D. MPEG-4 also defines the coded representation of an object, such as text and graphics, talking synthetic heads and associated text used to synthesize the speech and to animate the head and synthetic sound. A media object in its coded form consists of descriptive elements that allow handling the object in an audiovisual scene. Each media object can be represented in its coded form, independent of its surroundings or background.

The applications that benefit from what MPEG-4 brings are found in different environments [5.37]. Although MPEG-4 is a rather big standard, it is structured in a way that solutions are available at the measure of the needs. The task of implementers is to extract from the MPEG-4 standard the technological solutions adequate to their needs.

Media Objects

Media objects may need streaming data, which is conveyed in one or more elementary streams. An object descriptor identifies all streams associated with one media object. Each stream itself is characterized by a set of descriptors for configuration information, for example, to determine the required decoder resources and the precision of encoded timing information. Furthermore, the descriptors may carry hints to the QoS that it requests for transmission (maximum bit rate, bit error rate, priority and so forth.).

Synchronization of elementary streams is achieved through time stamping of individual access units within elementary streams. The synchronization layer manages the identification of such access units and the time stamping. Independent of the media type, this layer allows identification of the type of access units (video or audio frames or scene-description commands) in