H.261 was designed for conferencing systems and video telephony.

6.7 MPEG

The MPEG standard was developed by ISO/IEC JTC1/SC 29/WG11 to cover motion video as well as audio coding according to the ISO/IEC standardization process. Considering the state of the art in CD-technology digital mass storage, MPEG strives for a data stream compression rate of about 1.2 Mbits/second, which is today's typical CD-ROM data transfer rate. MPEG can deliver a data rate of at most 1856000 bits/second, which should not be exceeded [ISO93a]. Data rates for audio are between 32 and 448 Kbits/second; this data rate enables video and audio compression of acceptable quality. In 1993, MPEG was accepted as the International Standard (IS) [ISO93a] and the first commercially available MPEG products entered the market.

The MPEG standard explicitly considers functionalities of other standards::

- *JPEG.* Since a video sequence can be regarded as a sequence of still images, and the JPEG standard development was always ahead of the MPEG standard, the MPEG standard makes use of JPEG.

- *H.261.* Since the H.261 standard was already available during the work on the MPEG standard, the working group strived for compatibility (at least in some areas) with this standard. Implementations that are capable of H.261, as well as of MPEG, may arise, however, MPEG is the more advanced technique.

MPEG is suitable for symmetric as well as asymmetric compression. Asymmetric compression requires more effort for coding than for decoding. Compression is carried out once, whereas decompression is performed many times. A typical application area is retrieval systems. Symmetric compression is known to expect equal effort for the compression and decompression processes. Interactive dialogue applications make use of this encoding technique, where a restricted end-to-end delay is required.

Besides the specification of video [Le 91, VG91] and audio coding, the MPEG standard provides a *system definition*, which specifies the combination of several individual data streams.

6.7.1 Video Encoding

In contrast to JPEG, but similar to H.261, the image preparation phase of MPEG, according to our reference scheme shown in Figure-6.1, exactly defines the format of an image. Each image consists of three components (similar to the *YUV* format); the luminance component has twice as many samples in the horizontal and vertical axes as the other two components – this is known as *color-subsampling*. The resolution of the luminance component should not exceed 768 x 576 pixels; for each component, a pixel is coded with eight bits.

The MPEG data stream includes more information than a data stream compressed according to the JPEG standard. For example, the aspect ratio of a pixel is included. MPEG provides 14 different image aspect ratios per pixel. The most important are:

- A square pixel (1:1) is suitable for most computer graphics systems.
- For an image with 702 × 575 pixels, an aspect ratio of 4:3 is defined.
- For an image of 711 × 487 pixels, an aspect ratio of 4:3 is defined.
- For an image with 625 lines, an aspect ratio of 16:9 is defined, the ratio required for European HDTV.
- For an image with 525 lines, an aspect ratio of 16:9 is defined, the ratio required for U.S. HDTV.

The image refresh frequency is also encoded in the data stream. Eight frequencies are defined: 23.976 Hz, 24 Hz, 25 Hz, 29.97 Hz, 30 Hz, 50 Hz, 59.94 Hz and 60 Hz.

A temporal prediction of still images leads to a considerable compression ratio. Moving images often contain non-translational moving patterns such as rotations or waves at the seaside. Areas in an image with these irregular patterns of strong

motion can only be reduced by a ratio similar to that of intraframe encoding. The use of temporal predictors requires the storage of a great amount of information and image data. There is a need to balance this required storage capacity and the achievable compression rate. In most cases, predictive encoding only makes sense for parts of images and not for the whole image. Therefore, each image is divided into areas called *macro blocks*. Each macro block is partitioned into 16×16 pixels for the luminance component and 8×8 pixels for each of the two chrominance components. These macro blocks turn out to be quite suitable for compression based on motion estimation. This is a compromise of costs for prediction and the resulting data reduction.

Due to the required frame rate, each image must be built up within a maximum of 41.7 milliseconds. From the user's perspective there are no advantages to progressive image display over sequential display. The user has neither the need nor possibility to define the MCUs (Minimum Coded Units) in MPEG (in contrast to JPEG).

MPEG distinguishes four types of image coding for processing, as shown Figure 6.16. The reasons behind this are the contradictory demands for an efficient coding scheme and fast random access. To achieve a high compression ratio, temporal redundancies of subsequent pictures must be exploited (interframe), whereas the demand for fast random access requires intraframe coding. The following types of images are distinguished (*image* is used as a synonym for *still image* or *frame*):

- *I-frames* (*Intra-coded images*) are self contained, i.e., coded without any reference to other images. An I-frame is treated as a still image. MPEG makes use of JPEG for I-frames. However, contrary to JPEG, compression must often be executed in real-time. The compression rate of I-frames is the lowest within MPEG. I-frames are points for random access in MPEG streams.

  I-frames use 8×8 blocks defined within a macro block, on which a DCT is performed. The DC-coefficients are then DPCM coded; differences of successive blocks of one component are computed and transformed using variable-length coding. MPEG distinguishes two types of macro blocks – the first type includes only the encoded data and the second covers a parameter used for scaling by adjustment of the quantization characteristics.

- *P-frames* (*Predictive-coded frames*) require information of the previous I-frame

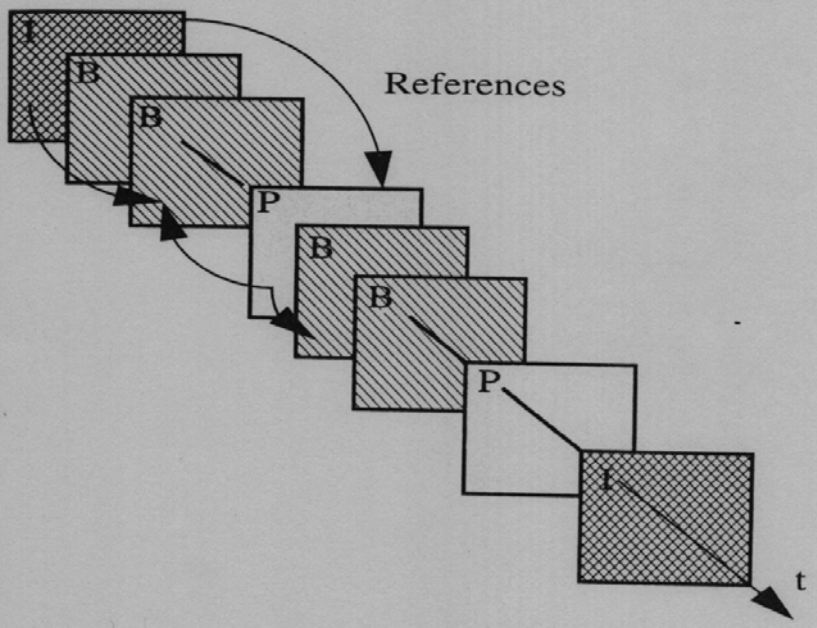

Figure 6.16: *Types of images in MPEG.*

and/or all previous P-frames for encoding and decoding.

The coding of P-frames is based on the fact that, by successive images, their areas often do not change at all but instead, the whole area is shifted. In this case of *temporal redundancy*, the block of the last P- or I-frame that is most similar to the block under consideration is determined. Several methods for motion estimation are available to the encoder. The most processing-intensive methods tend to give better results, so the following trade-offs must be made in the encoder: computational power, and hence cost, versus video quality [ISO93a]. Several matching criteria are available, e.g., the differences of all absolute values of the luminance component are computed. The minimal number of the sum of all differences indicates the best matching macro block. Thereby, MPEG does not provide a certain algorithm for motion estimation, but instead specifies the coding of the result. Only the motion vector (the difference between the spatial location of the macro blocks) and the small difference in content of these macro blocks are left to be encoded. The search range, i.e., the maximum size of the motion vector, is not defined in the standard, but it is constrained by the definable motion vector range. The larger

the search range the better the motion estimation, although the computation is slower.

Like I-frames, P-frames consist of I-frame macro blocks and six predictive macro blocks. The coder must determine if a macro block should be coded predictively or as a macro block of an I-frame, and furthermore, if there is a motion vector that must be encoded. A P-frame can contain macro blocks that are encoded using the same technique as I-frames. The coder for specific macro blocks of P-frames must consider the differences of macro blocks, as well as the motion vector. The difference of all six 8×8 pixel blocks of the best matching macro block and the macro block to be coded are transformed using a two-dimensional DCT. For further data rate reduction, blocks that only have DCT-coefficients with all values of zero are not processed further. These are stored using 6-bit values, which are added to the encoded data stream. Subsequently, the DC- and the AC-coefficients are encoded using the same technique. Note that this differs from JPEG and from the coding of macro blocks of I-frames. In the next step, a run-length encoding and the determination of a variable-length code (not according to Huffman, but similar) is applied. Since the motion vectors of adjacent macro blocks often differ only slightly, DPCM encoding is used. The result is again transformed using a table leading to a variable-length encoded word.

- *B-frames* (*Bi-directionally predictive-coded frames*) require information of the previous and following I- and/or P-frame for encoding and decoding. The highest compression ratio is attainable by using these frames. A B-frame is defined as the difference of a prediction of the past image and the following P- or I-frame. B-frames can never be directly accessed in a random fashion.

  For the prediction of B-frames, the previous as well as the following P- or I-frames are taken into account. The following example illustrates the advantages of a *bi-directional prediction*. In a video scene, a ball moves from left to right in front of a static background. In the left area of the scene, parts of the image appear that in the former image were covered by the ball. A prediction of these areas can be derived from the following but not from the previous image. A macro block may be derived from the previous or the next macro block of P- or I-frames. Apart from a motion vector from the previous

to the next image, a motion vector in the other direction can also be used. *Interpolative* motion compensation that uses both matching macro blocks is allowed. In this case, two motion vectors are encoded. The difference of the macro block to be encoded and the interpolated macro block is determined. Further quantization and entropy encoding are performed like P-frame specific macro blocks. B-frames must not be stored in the decoder as a reference for subsequent decoding of images.

- *D-frames* (*DC-coded frames*) are intraframe-encoded. They can be used for fast forward or fast rewind modes. The DC-parameters are DCT-coded; the AC-coefficients are neglected.

  D-frames consist only of the lowest frequencies of an image. They only use one type of macro block and only the DC-coefficients are encoded. D-frames are used for display in fast-forward or fast-rewind modes. This could also be realized by a suitable order of I-frames. For this purpose, I-frames must occur periodically in data stream. Slow-rewind playback requires huge storage capacity. Therefore, all images that were combined in a group must be decoded in the forward mode and stored, after which a rewind playback is possible. This is known as the *group of pictures* in MPEG.

Figure 6.16 shows a sequence of I-, P- and B-frames. For example, the prediction for the first P-frames and a bi-directional prediction for a B-frame is shown. Note that by using B-frames the order of the images in a MPEG-coded data stream often differs from the actual decoding order. A P-frame to be displayed after the related B-frame must be decoded before the B-frame because its data is required for the decompression of the B-frame. This fact introduces an additional end-to-end delay.

The regularity of a sequence of I-, P- and B-frames is determined by the MPEG application. For fast random access, the best resolution would be achieved by coding the whole data stream as I-frames. On the other hand, the highest degree of compression is attained by using as many B-frames as possible. For practical applications, the following sequence has proved to be useful, "IBBPBBPBB IBBPBBPBB ..." In this case, random access would have a resolution of nine still images (i.e., about 330 milliseconds), and it still provides a very good compression ratio.

Concerning quantization, it should be mentioned that AC-coefficients of B- and P-

frames are usually large values, whereas those of I-frames are smaller values. Thus the MPEG quantization is adjusted respectively. If the data rate increases over a certain threshold, the quantization enlarges the step size. In the opposite case, the step size is reduced and the quantization is performed with finer granularity.

6.7.2 Audio Encoding

MPEG audio coding uses the same sampling frequencies as Compact Disc Digital Audio (CD-DA) and Digital Audio Tape (DAT), i.e., 44.1 kHz and 48 kHz, and additionally, 32 kHz is available, all at 16 bits.

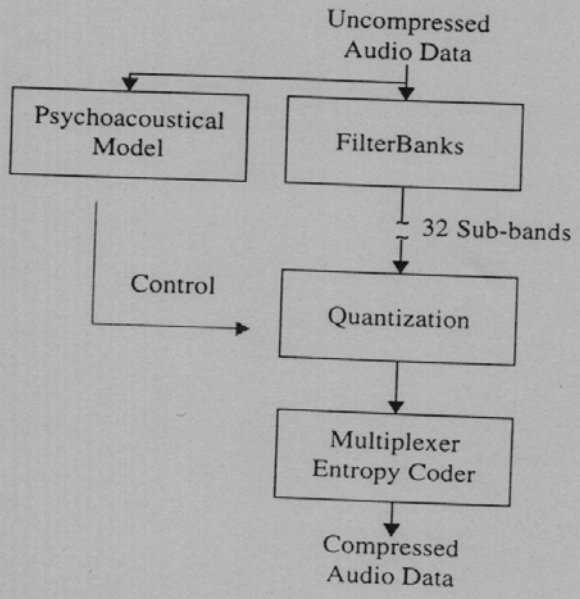

Figure 6.17: *MPEG basic steps of audio encoding.*

Three different layers (Figure 6.17) of encoder and decoder complexity and performance are defined. An implementation of a higher layer must be able to decode the MPEG audio signals of lower layers [Mus90]. Similar to two-dimensional DCT for video, a transformation into the frequency domain is applied for audio. *Fast Fourier Transformation* (FFT) is suitable for this coding, and the spectrum is split into 32 non-interleaved subbands. For each subband, the amplitude of the audio signal is

calculated. Also for each subband, the noise level is determined simultaneously to the actual FFT by using a *psychoacoustic model*. At a higher noise level, a rough quantization is performed, and at a lower noise level, a finer quantization is applied. The quantized spectral portions of layers one and two are PCM-encoded and those of layer three are Huffman-encoded. The audio coding can be performed with a single channel, two independent channels or one stereo signal. In the definition of MPEG, there are two different stereo modes: two channels that are processed either independently or as *joint stereo*. In the case of joint stereo, MPEG exploits redundancy of both channels and achieves a higher compression ratio.

Each layer defines 14 fixed bit rates for the encoded audio data stream, which in MPEG are addressed by a bit rate index. The minimal value is always 32 Kbits/second. These layers support different maximal bit rates: layer 1 allows for a maximal bit rate of 448 Kbits/second, layer 2 for 384 Kbits/second and layer 3 for 320 Kbits/s. For layers 1 and 2, a decoder is not required to support a variable bit rate. In layer 3, a variable bit rate is specified by switching the bit rate index. For layer 2, not all combinations of bit rate and mode are allowed:

- 32 Kbits/second, 48 Kbits/second, 56 Kbits/second and 80 Kbits/second are only allowed for a single channel.
- 64 Kbits/second, 96 Kbits/second, 112 Kbits/second, 128 Kbits/second, 160 Kbits/second and 192 Kbits/second are allowed for all modes.
- 224 Kbits/second, 256 Kbits/second, 320 Kbits/second, 384 Kbits/second are allowed for the modes *stereo*, *joint stereo* and *dual channel* modes.

6.7.3 Data Stream

Audio Stream

MPEG specifies a syntax for the interleaved audio and video data streams. An audio data stream consists of frames, which are divided into audio access units. Each audio access unit is composed of slots. At the lowest complexity (layer 1), a slot consists of four bytes. In any other layer, it consists of one byte. A frame

always consists of a fixed number of samples. Most important is the *audio access unit*, which is the smallest possible audio sequence of compressed data that can be completely decoded independent of all other data. The audio access units of one frame lead to a playing time of 8 milliseconds at 48 kHz, of 8.7 milliseconds at 44.1 kHz, and 12 milliseconds at 32 kHz. In the case of stereo signals, data from both channels are merged into one frame.

Video Stream

A video data stream is comprised of six layers:

1. At the highest level, the *sequence layer*, data buffering is handled. A data stream should have low requirements in terms of storage capacity. For this reason, at the beginning of the *sequence layer* there are the following two entries: the constant bit rate of the sequence and the storage capacity that is needed for decoding. In the processing scheme, a *video-buffer-verifier* is inserted after the quantizer. The resulting data rate is used to verify the delay caused by decoding. The *video-buffer-verifier* influences the quantizer and forms a kind of control loop. Several successive sequences could have a varying data rate. During decoding of several immediately following sequences there is no direct relationship between the end of one sequence and the beginning of the next one. The basic parameters of the decoder are set again and an initialization is executed at this time.

2. The *group of pictures layer* is the next layer. This layer consists of a minimum of one I-frame, which is the first frame. Random access to this image is always possible. At this layer, it is possible to distinguish the order of images in a data stream and during display. The first image of a data stream always has to be an I-frame. Therefore, the decoder decodes and stores the reference frame first. In the order of display, a B-frame can occur before an I-frame.

   Display Order:

Type of Frame	B	B	I	B	B	P	B	B	P	B	B	P
Number of Frame	0	1	2	3	4	5	6	7	8	9	10	11

Decoding order:

Type of Frame	I	B	B	P	B	B	P	B	B	P	B	B
Number of Frame	2	0	1	5	3	4	8	6	7	11	9	10

3. The *picture layer* contains a whole picture. The temporal reference is defined by an image number. Note that there are data fields defined in this layer which are not yet used in MPEG. The decoder is not allowed to use these data fields, as they are designated for future extensions.

4. The next layer is the *slice layer*. Each slice consists of a number of macro blocks that may vary from one image to the next. Additionally, the DCT quantization of each macro block of a slice is specified.

5. The fifth layer is the *macro block layer*. It contains the sum of the features of each macro block as described above.

6. The lowest layer is the *block layer* (described above).

The MPEG standard also specifies the combination of data streams into a single data stream in the system definition. The same idea was pursued in DVI to define the AVSS (Audio/Video Support System) data format. The most important task of this process is the actual multiplexing. It includes the coordination of input data streams and output data streams, the adjustment of clocks and buffer management. Therefore, the data stream defined by ISO 11172 is divided into single *packs*. The decoder gets the information necessary for its resource reservation from this multiplexed data stream. The maximal data rate is included in the first pack at the beginning of each ISO 11172 data stream. The definition of this data stream makes the following implicit assumption: for data stored on a secondary storage medium, it is possible to read such a header first (if necessary, by random access). In dialogue services like telephone or videophone applications using communication networks, the user will always get the header information first. In a conferencing application, using an MPEG stream might be inconvenient because a new user might like to join an existing conference after the data streams have already been setup. Therefore, the necessary header information would not be directly available to her/him.

For a data stream generated according to ISO 11172, MPEG provides time stamps that are necessary for synchronization. They refer to the relationship between mul-

tiplexed data streams, but not between other existing ISO 11172 data streams.

It should be mentioned that MPEG does not prescribe compression in real-time. MPEG defines the process of decoding, but not the decoder itself.

6.7.4 MPEG-2

The quality of a video sequence compressed according to the MPEG standard is near the target maximum data rate of about 1.5 Mbits/second. This optimum is in quality and not in performance. Further developments in the area of video coding techniques are based on a target rate of up to 40 Mbits/second; this is known as *MPEG-2* [ISO93b]. MPEG-2 strives for a higher resolution, similar to the digital video studio standard CCIR 601 and leading towards the video quality needed in HDTV. Note that most of the following information on MPEG-2 and MPEG-4 was gleaned by the authors from press releases and many personal communications with members of the MPEG expert group [Liu93].

To ensure that a harmonized solution to the widest range of applications is achieved, the ISO/IEC working group designated ISO/IEC JTC1/SC29/WG11, has been working jointly with ITU-TS Study Group 15 *Experts Group for ATM Video Coding*. MPEG-2 also collaborates with representatives from other parts of ITU-TS, EBU, ITU-RS, SMPTE and the North American HDTV community.

The MPEG group developed the *MPEG-2 Video Standard*, which specifies the coded bit stream for high-quality digital video. As a compatible extension, MPEG-2 Video builds upon the completed MPEG-1 standard by supporting interlaced video formats and a number of other advanced features, including those to support HDTV.

As a generic international standard, MPEG-2 Video was defined in terms of extensible profiles, each of which will support the features needed by an important class of applications. The *MPEG-2 Main Profile* was defined to support digital video transmission in the range of about 2 to 80 Mbits/s over cable, satellite and other broadcast channels, as well as to support digital storage and other communications applications. Parameters of the *Main Profile* and *High Profile* are suitable for supporting HDTV formats.