VIDEO COMPRESSION: The MPEG Standards

Luís Miguel L. Teixeira and M. Isabel Martins

INESC Porto
Largo Mompilher 22, 4000 PORTO,
PORTUGAL
Tel.: +351 - 2 - 208.78.30 - Fax.: +351 - 2 - 208.78.29
E-mail: lmt, mmartins@inescn.pt

ABSTRACT

In order to ensure compatibility among video codecs from different manufacturers and applications and to simplify the development of new applications, intensive efforts have been undertaken in recent years to define digital video standards. These standards were the result of joint development efforts of video and audio compression as well as other system aspects required to support all the applications. Thus they often represent an optimal compromise between performance and complexity. This paper describes the main features of MPEG 1 and MPEG2 video standards, discusses the emerging standard MPEG4 and presents some of its main characteristics.

1. INTRODUCTION

Video is going to touch every area of information technology over the next few years. Video will be incorporated into applications, captured off the TV or the VCR for use as attachments to multimedia e-mail items. Video conference and video-on-demand services will be used by everyone to beam favourite programmes and films down the line to users.

Within a few years the television, the video camera, the VCR, the telephone and the PC will work together. Although there is still a long way to go, intensive efforts have been undertaken in recent years to set digital video standards. Thus standardisation activities in video coding developed from the beginning of the 1980's within CCITT, followed by CCIR (actually ITU-R) and ISO, later on. As a result, in 1989 and in 1990 the CCITT Rec. H.120 and H.261, for video conference services [2] [3], in 1989 and 1992 the CCIR Rec. 721 and 723 for digital television services [4] [5], in 1993 the ISO/IEC 11172 (MPEG1) for coding applications for digital storage media [6], in 1994 the ISO/IEC 10918 standard (JPEG) for still pictures compression was published [1] and in 1994 the ISO/IEC 13818 (MPEG2) for generic coding of multimedia signals up to HDTV quality for telecommunications and storage applications [7]. Aiming higher compression ratios, the ITU-T (former CCITT) started activities in 1993 with the objective of issuing a recommendation for video coding for narrow telecommunication channels. The video coding work was divided into two main areas: near term work directed towards Rec. H.263 [18] and long term work towards Rec. H.263/L. To achieve the schedule requirement the H.263 video coding algorithm is an extension of H.261. The objective of Rec. H.263/L is to achieve a video coding algorithm which significantly outperforms the H.263 technique. ISO/MPEG4 started activities in 1993, aiming not only higher compression ratios but also the incorporation of multimedia functionalities. An international standard is expected in November 1998. This emerging standard should support new ways (notably content-based) for communication, access and manipulation of digital audio-visual data [20]. The ITU work to develop Rec. H.263/L is being accomplished in close collaboration with the ISO/MPEG4 activity.

These generic standards are extremely important as they allow the development of VLSI and several of the basic blocks required for a large number of applications. They are the result of joint development efforts of audio and video compression experts taking into consideration requirements of all applications considered. In the first phase of standardisation called, "divergence phase", the requirements for specific applications or fields of applications are identified, and several algorithms, developed by independent laboratories, in competition, are presented and compared [8] [9]. In the second phase, "convergence phase", based on the selection of a basic(s) coding technique made in the first phase, joint efforts are co-ordinated in order to refine and optimise the selected coding techniques. In the last phase, "verification/validation", the results obtained in previous phases are validated using hardware tests/software simulations. Using this approach standards have been pulling together a enormous worldwide research effort and often achieved optimal compromises between performance and complexity.

The main goal of this paper is to provide an overview of the most recent video coding algorithms and standards: the MPEG1 and MPEG2 standards. The paper is organised as follows: in section II MPEG video compression techniques are presented, section III refers to MPEG1 and section IV to MPEG2 standard. Section V discusses the main characteristics of the emerging MPEG4 standard. Finally, section VI, discusses the future of MPEG1, MPEG2 and MPEG4.

2. MPEG VIDEO COMPRESSION TECHNIQUES

The major goal of video compression is to represent a video source with as few bits as possible while preserving the level of quality required for the given application. The bit-rate reduction is only possible by removing redundant information from the signal during the coding process and reinserting it during the decoding process. In video signals, there is a significative amount of redundancy between frames that can be classified as statistical and psychovisual redundancy [10]. The statistical redundancy results from the fact that pixel values are correlated with their neighbours in spatial and temporal directions. The psychovisual redundancy is a consequence of the human visual system (HVS) sensitivity. The human vision has a limited response to fine spatial or temporal detail and bit-rate reduction is possible by allowing distortions that should not be visible to human eyes.

This section will mention the video compression techniques used by MPEG video algorithm: subsampling of the chrominance information to match the sensitivity of the HVS, quantisation, motion compensation to exploit temporal redundancy, discrete cosine transform (DCT) to exploit spatial redundancy, variable length coding (VLC), and picture interpolation.

2.1 Subsampling

Usually video sequences are digitalised in RGB format. However RGB colour components are highly correlated, resulting in the presence of psychovisual redundancy. To reduce the correlation, the redundancy and thus the bitrate the RGB components are converted to YCbCr colour space through a linear transformation. The HVS is more sensible to the luminance component (Y) than to the chrominance components (Cb and Cr). Therefore the chrominance components are generally sub-sampled.

2.2 Predictive Coding

Predictive coding is used to reduce the statistical redundancy in the data. A prediction of the pel to be encoded is made from previously encoded data that has been transmitted. The prediction error is computed by calculating the difference between the pel to be encoded and the prediction. The prediction error is then usually quantised and entropy coded.

2.3 Motion Compensation

A video sequence is just a succession of still images shown at a fixed frame rate to give the perception of continuous motion. The motion, in most natural scenes, is organised, and can, in most cases, be approximately represented as a translation represented by a limited number of motion parameters (i.e. estimated motion vectors and a prediction error). Thus instead of encoding the original video data, the motion parameters are transmitted and only a small error component needs to be transmitted. Due to spatial correlation the video data is associated in blocks and a motion vector is determined for blocks of pels. In MPEG standard, motion vectors are defined for each 16-col by 16-line region of the picture (macroblocks).

2.4 Picture Interpolation

A picture, in MPEG, may be reconstructed from a picture in the past and a picture from the future by the technique of interpolation, or bidirectional prediction.

2.5 Transform Domain Coding

One of the key ideas in video compression techniques is to decorrelate the video information. Transform coding maps correlated video data, by applying an orthogonal linear transformation, to non correlated video information [11]. It is therefore an advantage to encode the transform coefficients instead of the original video data.

Karhunen-Loeve transformation (KLT) is the only transformation that produces always non correlated coefficients for finite images (Rosenfeld and Kak [12]) presenting thus the best compression results. Nevertheless it is computationally complex. Among the several feasible alternatives the discrete cosine transform (DCT), proposed by Ahmed, Natarajan and Rao [13], has been widely used in image and video coding standards such as JPEG, MPEG1 and MPEG2.

The forward DCT is defined as

$$F(u,v)=\frac{2}{N}C(u)C(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\left[\frac{(2x+1)v\pi}{2N}\right] \quad (1)$$

$$C(u)=\begin{cases}\dfrac{1}{\sqrt{2}} & u=0\\ 1 & u\neq 0\end{cases} \quad (2)$$

with *u*, *v*, *x*, *y* = 0, 1, 2, ..., N - 1, where *x*, *y* are spatial coordinates in the sample domain and *u,v* are coordinates in the transform domain [6,7]. The inverse DCT (IDCT) is defined as:

$$f(x,y)=\frac{2}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1}C(u)C(v)F(u,v)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (3)$$

The DCT converts an 8 by 8 block of pel values to an 8 by 8 matrix of horizontal and vertical spatial frequency coefficients. The DCT coefficient in location (0,0) (upper left) of the block represents zero horizontal and zero vertical frequency and is called DC coefficient. The others DCT coefficients are called AC coefficients. Human eyes are more sensitive to low order DCT coefficients. Thus exploiting the spatial frequency properties of the human visual system, the DCT coefficients can be encoded to match the HVS so that only the perceptually important DCT coefficients are encoded and transmitted.

2.6 Quantisation

By exploiting of the perceptual irrelevancy (DCT) and statistical redundancy (entropy coding) within the DCT domain representation, a suitable bit allocation can yield significant improvements in performance. In this context, quantisation is used to reduce the number of possible values to be transmitted, reducing the required number of bits. The HVS is not uniformly sensitive to coefficient quantisation error. The visibility of the quantisation noise, for a given coefficient, depends on the coefficient number, or frequency, the local brightness in the original image, and the duration or temporal characteristic of the error. Experience shows that high-frequency coefficients are more coarsely quantised than the low-frequency coefficients. The quantisation is a lossy process as information is lost and cannot be recovered in later stages.

2.7 Variable-Length Coding

Variable-length coding (VLC) is a statistical coding technique that assigns codewords to values to be transmitted. The length of the code word is chosen depending on the frequency of occurrence of each value. Values with high frequency of occurrence are assigned short codewords and values with sparse frequency of occurrence are assigned long codewords.

3. MPEG 1 VIDEO STANDARD

ISO and IEC in collaboration with ITU, have created joint working groups with the purpose of specifying international standards in several communications domains, including audiovisual digital communications. In 1988, ISO and IEC created a Technical Committee on Information Technology to coordinate the development of standards for coded representation of moving pictures, associated audio and their combination when used for storage and retrieval on digital storage media (DSM) at bit rates not exceeding 1.5 Mbit/s. The group was named ISO/IEC JCT1/SC29/WG11 and universally known as the Moving Pictures Experts Group (MPEG). The original work items of the group were three: coding the audiovisual information at bit rates up to 1.5, 10 and 40 Mbit/s, respectively known as MPEG1, MPEG2 and MPEG3. The MPEG3 work item was dropped in July 1992, when it became apparent that the functionality supported by MPEG2 made MPEG3 redundant. A new work item, MPEG4, is currently being addressed targeted to audiovisual coding at high compression levels with increased functionalities.

The MPEG activity is basically divided into three fields: MPEG-video, defining the video compression algorithm; MPEG-audio specifying audio coding algorithms; MPEG-systems, responsible for the multiplexing and inter-media synchronisation aspects. The target quality for the audio component was stereophonic sound with CD quality at rates lower than 512 Kbit/s. The MPEG1 video coding algorithm although very flexible, was optimised to give its best performance at bit rates around 1.2 Mbit/s working with picture spatial resolutions of 350 pixels per 250 lines and frame repetition rates between 24 and 30 images per second.

3.1 MPEG 1 Video Requirements

The MPEG standard is a generic standard. Generic means that the standard is independent of a particular application and of the delivery media; however, it could not ignore the requirements of the applications [14]. It was developed in response to the growing need for a common format for representing compressed video on various DSM such as CDs, DATs, Winchester disks and optical drives. Applications using compressed video on DSM need to be able to perform a number of operations in addition to normal forward playback of the video sequence. The compression algorithm must have features that make it possible to fulfil all the requirements. The following features have been identified as important in order to meet the need of the applications of MPEG [6,14]:

3.1.1 Random Access

Random access requires that any picture can be decoded in a limited amount of time. It implies the existence of access points.

3.1.2 Fast Search (Forward/Reverse)

It should be possible for an application, depending of the storage media, to scan a compressed bitstream and, using selected suitable access points, display images to obtain a fast forward or a fast reverse effect.

3.1.3 Reverse Playback

Some applications may require the video signal to be played in reverse order. This effect may be achieved by storing in memory groups of images after they have been previously decoded.

3.1.4 Error Robustness

Most communication channels and DSM are not error-free. Although appropriate channel coding schemes are beyond the scope of MPEG1 the video compression scheme is robust to residual errors. The decoder, after erroneous data, may resynchronise through the slice structure.

3.1.5 Coding/decoding delay

There are applications, such as videotelephony and video conference, to which the total system delay is critical and should be under 150 ms. There are also applications that can support fairly long encoding delays. There is a trade-off between delay and picture quality but nevertheless the algorithm should perform well over a range of acceptable delays.

3.2 Bitstream Hierarchy

The encoded video data is organised into a layered structure which allows integration of the different coding modes and provides the means to implement the required facilities for DSM applications. This structure comprises 6 hierarchical layers. The **sequence** is the top of the hierarchy in which the video information is structured. It comprises one or more Group of Pictures (GOP). A **GOP** comprises a certain number of encoded images. The length of the GOP may be dictated by the random access requirements. It is used as a random access unit. A **picture** is a frame classified as intra, predicted or interpolated (bi-directional) according to the mode which was used to encode it as described bellow. A picture is divided into **slices**. A slice is a collection of an integer number of macro blocks, in raster-scan order. Usually, a slice is a horizontal stripe within a frame. The first slice of a picture must start with the upper-left macro block of that picture and the last slice must end with the lower-right macro block. A macroblock (**MB**) associates 4 blocks of luminance with the spatially corresponding block of each chrominance component. It is used as a motion compensation unit. Finally, a **block** is a picture section of 8 pixels by 8 lines either of the luminance or the chrominance components. It is used as a DCT unit.

Due to the introduction of the motion compensated temporal interpolation technique, the MPEG1 algorithm is able to produce three different types of encoded images, by using the following three different coding modes:

- *Intra mode*; images encoded individually without using temporal prediction (without reference no any other picture);
- *Predicted mode* (P-pictures); inter frame coded pictures using unidirectional MC prediction;
- *Bi-directional mode*, which generates inter frame coded pictures using bi-directional motion compensated prediction. B-frames may be encoded using either forward prediction where reference is made to an image in the past, backward prediction where reference is made to a future image, and finally to an image in the past and one in the future.

The use of temporal interpolation, generating coded images with possible reference to images in the future, forces a rearrangement of the order of the coded pictures before transmission. I and P pictures must be transmitted before interpolated (B) frames.

A typical GOP in display order might be as in (4) whereas the bitstream order would be as in (5).

$$I_0 B_1 B_2 P_3 B_4 B_5 P_6 B_7 B_8 P_9 B_{10} B_{11} I_{12} \tag{4}$$

$$I_0 P_3 B_1 B_2 P_6 B_4 B_5 P_9 B_7 B_8 I_{12} B_{10} B_{11} \tag{5}$$

3.3 Encoding

Encoding process is divided into the following steps: motion estimation, prediction calculation, DCT type estimation, subtraction of prediction from picture, DCT calculation, DCT coefficients quantisation and generation of VLC data, inverse DCT coefficients quantisation and finally IDCT calculation and addition of predicted image. These steps are clearly indicated in block diagrams described bellow.

The algorithm uses block-based motion estimation. The motion estimation step can be subdivided in:

- calculation of optimum motion vectors for each of the possible motion compensation types
- selection of the best motion compensation type by calculating and minimising a prediction error based on a cost function
- selection of either motion compensation or any other possible encoding type (intra coding, No MC coding)

The decoder and encoder block diagrams are shown in Fig. 1. In Figure 1, "Q" and "IQ" denote the quantisation and dequantisation operations; the operation "Clip" performs truncation to the nearest integer within the range [-255;255]; "MC" represents motion compensation; "VLC" represents variable length coding of the quantised DCT coefficients; "VLD" represents variable length decoding; and "Store" represents a frame-store that holds up to two pictures.

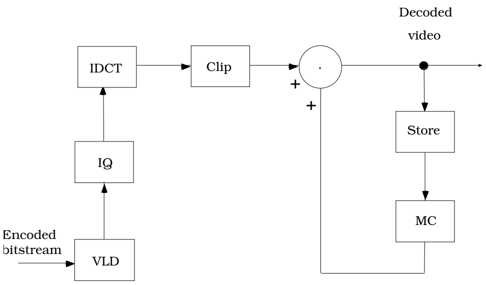

Fig. 1.a) MPEG decoder block diagram

In the decoder, the quantised DCT coefficients are reconstructed and inverse transformed to produce the prediction error. This is added to the motion-compensated prediction generated from previously decoded pictures to produce the decoded output.

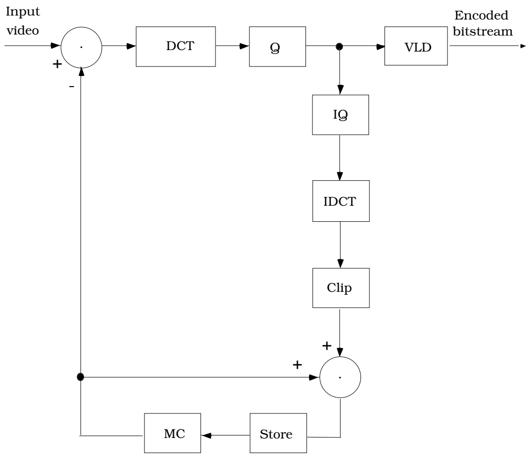

Fig. 1.b) MPEG encoder block diagram

The encoder subtracts the motion-compensated prediction from the source picture to form a 'prediction error' picture. The prediction error is transformed with the DCT, the coefficients are quantised and these quantised values coded using a VLC. The coded prediction error is combined with 'side information' required by the decoder, such as motion vectors and synchronising information, and formed into a bitstream for transmission.

3.4 Constrained Parameter Set

Because of the large range of the characteristics of bitstreams that is supported by the standard, a special subset of the coding parameter, known as "Constrained Parameters Set" (CPS), has been defined. A flag in the bitstream indicates whether or not it is a CPS.

Horizontal picture ≤ 720 pels
Vertical Size ≤ 576 pels
Picture area ≤ 396 macroblocks (mb)
Pel rate ≤ $396 \times 25 = 330 \times 30$ (mb/sec)
Picture rate ≤ 30 frames/sec
Input buffer size ≤ 327 680 bits
Bitrate ≤ 1 856 000 bits/sec

Table 1) CPS MPEG Parameters

4. MPEG 2 STANDARD

In 1991, MPEG started a second phase of work (MPEG2 - ISO/IEC 13818) with the goal of developing a standard to cover a wider range of applications rather than just storage and retrieval in DSM, offering much higher picture resolutions and bitrates. Since the early stages that MPEG2 principal application was the all-digital transmission of broadcast TV quality video at coded bitrates between 4 and 9 Mbit/sec. However, the MPEG2 syntax has been made suitable to other applications such as those at higher bit rates and sample rates (e.g. HDTV). The most significant enhancement over MPEG1 is the addition of syntax for efficient coding of interlaced video (e.g. 16x8 block size motion compensation, Dual Prime, et al). Several other more subtle improvements (e.g. 10-bit DCT DC precision, non-linear quantisation, VLC tables, improved mismatch control) are included which have a clear improvement on coding efficiency, even for progressive video. Other key features of MPEG2 are the scalable extensions which permit the division of a continuous video signal into two or more coded bitstreams representing the video at different resolutions, picture quality (i.e. SNR), or picture rates.

4.1 Non-scalable syntax

The full syntax can be divided into two major categories [7]. One is the non-scalable syntax, which is structured as a super set of the MPEG1 syntax. The second is the scalable syntax.

The main difference between MPEG2 non-scalable syntax and MPEG1 syntax is the additional compression tools for interlaced video signals. The compression algorithm is similar to MPEG1. First it uses block-based motion compensation to reduce the temporal redundancy. Motion compensation is used both for causal prediction of the current picture from a previous picture, and for non-causal, interpolated prediction from past and future pictures. The prediction error is further compressed using DCT to remove spatial correlation before it is quantised in a lossy process. Finally, the motion vectors are combined with the residual DCT information, and encoded using VLC codes.

4.2 Scalable syntax

The total bitstream may be structured in layers, starting with a base layer (can be decoded by itself) and adding a number of enhancement layers. The base layer can use the non-scalable syntax, or in some situations conform to the MPEG1 syntax. Scalable techniques are being suggested for digital terrestrial broadcasting, in conjunction with layered modulation systems, to provide graceful picture degradation in the presence of channel errors [16]. The MPEG2 standard provides several different forms of scalabilities that can be used by any application with different implementation complexities. The basic scalability tools offered are: *data partitioning*, *SNR scalability*, *spatial scalability* and *temporal scalability*. Combinations of these basic scalability tools are also supported.

4.2.1 Spatial Scalability

In spatial scalability the base layer is coded at lower spatial resolution than the upper layers. The upsampled reconstructed lower (base) layer is then used as a prediction for the higher layers. Spatial scalability is the appropriate tool to be used in applications where interworking of video standards is necessary as well as in simulcasting.

4.2.2 SNR Scalability

SNR scalability is a frequency domain method where channels are coded with the same spatial resolution, but with differing picture quality (achieved through macroblock quantisation step sizes). The lower layer provides the basic video quality while the enhancement layer carries the information which, when added to the lower layer, regenerates a higher quality reproduction of the input video. SNR scalability is intended to be used in HDTV applications with embedded TV or in video services with multiple qualities.

4.2.3 Temporal Scalability

In temporal scalability, the base layer is coded at a lower frame rate, and the intermediate frames can be coded in a second bitstream using the first bitstream reconstruction as prediction. Only sophisticated systems of the future, will be able to regenerate and display such a full temporal resolution video signal. In stereoscopic vision, for example, the left video channel can be predicted from the right channel.

4.2.4 Data Partitioning

Data partitioning is a frequency domain method that breaks the block of 64 quantised transform coefficients into two bitstreams. The first, higher priority bitstream, contains the more critical lower frequency coefficients and side information (such as headers, DC values, motion vectors). The second, lower priority bitstream, carries higher frequency AC data. It is appropriate when two transmission channels are available. Unlike the other scalable tools, neither layer may be decoded by itself.

4.3 Profiles & Levels

MPEG2 is intended to be generic, as was the case with MPEG1. Different algorithmic elements or 'tools', developed for many applications, have been combined into a single syntax to meet the requirements of different applications [15]. However to avoid the implementation of the full syntax in all the decoders the concept of "Profiles" was introduced. A Profile is a defined subset of the entire bitstream syntax. MPEG2 specifications define two non-scalable profiles, Simple Profile (SP) and Main Profile (MP), and three scalable profiles: SNR Profile, Spatial Profile and High Profile (HP) [7]. The profiles are defined such that a higher profile is a superset of a lower one. Within the limits imposed by the syntax of a given profile it is still possible to require a very large variation in the performance of the encoders/decoders depending upon the values taken by parameters in the bitstream. To deal with this problem, within each Profile, quality 'levels' were defined (low level, main level, high 1440 level, and high level). A level is a defined set of constraints imposed on parameters in the bitstream. Not all the levels are defined for each profile (table 2). These constraints may be simple limits on numbers (table 3).

Level	SP	MP	SNR	Spatial	HP
High		X			X
High-1440		X		X	X
Main	X	X	X		X
Low		X	X		

Table 2) Defined levels for each profile

The SP uses no backward or interpolated prediction. Therefore no picture reordering is required. This makes SP suitable for low-delay applications such as video telephone or video conferencing.

Level	Horiz. Size	Vert. Size	Frame Rate	Bitrate (Mb/s)
High	1920	1152	60	80
High 1440	1440	1152	60	60
Main	720	576	30	15
Low	352	288	30	4

Table 3) Parameters upper bounds at each level

The MP adds support for B pictures. Thus picture quality may be increased and a higher degree of compression may be achieved. MP decoder should also decode MPEG1 video bitstreams.

The SNR profile adds support for two levels of pictures quality. The addition of an extra quantisation stage does not essentially change its nature and the codec works like a MP codec. The error introduced by the first quantisation is itself quantised, run-length and VLC coded, and transmitted as the enhancement layer.

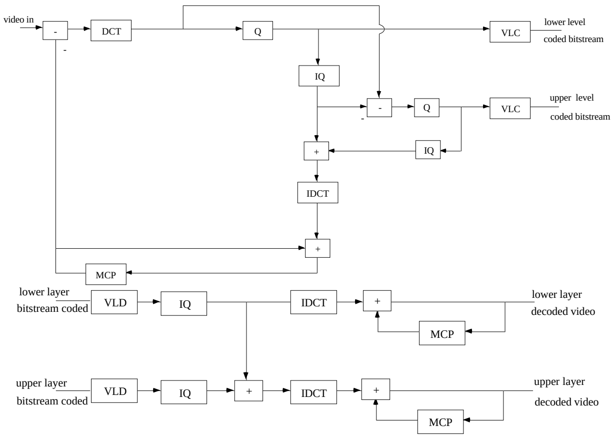

Figure 2) SNR coder/ decoder block diagrams

The spatial profile adds support for enhancement layers carrying the coded image at different resolutions. As a result, the decoded picture from the base layer must be sample rate converted to the higher resolution by means of an 'up-converter' [16]. The two coder loops are operating at different picture resolutions (Fig. 3). The adaptive weighting function, W, selects between the prediction from the upper and lower layers.

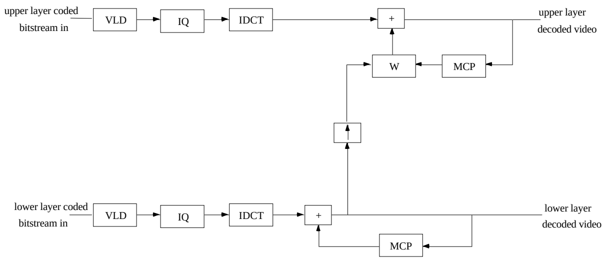

Figure 3.a) Spatial decoder

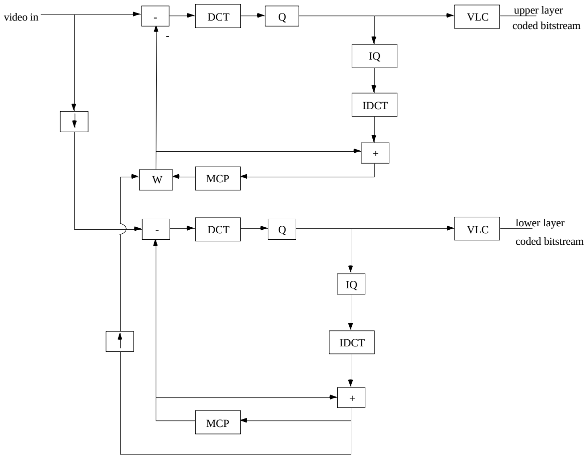

Figure 3.b) Spatial coder

The HP adds support for 4:2:2 sampled component video.

4.4 MPEG2 extensions

Since the final approval of MPEG-2 Video in November 1994, one additional profile has been developed. This uses existing coding tools of MPEG-2 Video but is capable to deal with pictures having a colour resolution of 4:2:2 and a higher bitrate (up to 50 Mbps). This allows flexibility in choosing video quality and latency that are specific to the needs of studio applications, such as multiple generation coding (repeated encoding and decoding), digital effects and digital distribution. The 4:2:2 profile has been finally approved in January 1996 and is now an integral part of MPEG-2 Video.

The Multiview Profile (MVP) is an additional profile currently being developed. By using existing MPEG-2 Video coding tools it is possible to encode in an efficient way two video sequences issued from two cameras shooting the same scene with a small angle between them. This profile will be finally approved in July 1996.

ISO/IEC are now studying new problems related to the interconnection of MPEG with different environments.

4.4.1 Digital Storage Media Command Control

The standard DSM-CC will be necessary to connect the different applications which require interaction with MPEG information stored on a disk or other storage media. The use of DSM-CC is specified by many network providers around the world, and DSM-CC has been adopted by many working forums on standardisation, such as the Digital Audiovisual Council (DAVIC). The goal of DAVIC is to promote broadband digital services using a variety of delivery media such as optical fiber, cable or satellite, by ensuring compatibility and interoperability on a world-wide basis. This will be advantageous to all parties - equipment manufactures, network operators, content producers, service providers and most importantly *consumers*.

4.4.2 Non Backwards Compatible (NBC)

NBC audio is needed, as it has been proven that backwards compatibility entails a quality cost that some applications are not expected to tolerate. The standard is expected to be approved in March 1997.

4.4.3 10 bits coding

Part 8 of MPEG-2 was originally planned to be coding of video when input samples are 10 bits. Work on this part was discontinued when it became apparent that there was insufficient interest from industry for such a standard.

4.4.4 Real Time Interface (RTI)

The goal of RTI is to provide a specification for a real-time interface to Transport Stream decoders which may be used for adaptation to every appropriate network carrying Transport Streams.

5. MPEG4

The development of this new standard reflects new trends on the standardisation of multimedia information resulting from the merging of three worlds: telecommunications, TV/film, and computers, with elements that have historically belonged to each of these areas being introduced into the others and resulting in the convergence of common applications of the three associated industries [19][20]. Audio-Visual coding and 2D/3D computer graphics are converging relative to their enabling technologies and used in real-time/interactive applications. These applications should be accessible through a wide range of storage and transmission media, including mobile networks.

The emerging MPEG4 standard will give users the possibility to achieve various forms of interactivity with the audio-visual content of a scene and to mix synthetic and natural audio and video information. The standard will provide for a high degree of flexibility and extensibility in order to take advantage of rapidly evolving technologies. These flexibility and extensibility shall be provided by a syntactic description language, called “MPEG-4 Syntactic Description Language (MSDL)”. This syntax will allow the incorporation of new coding techniques, tools and functionalities, providing the standard with the ability to adapt and evolve continuously.

5.1 The Functionalities

The new or improved functionalities described in the MPEG4 Proposal Package Description (PPD) [20], have been clustered in three main groups:

- *content-based interactivity*, addressing the ability to interact with meaningful objects in an audio-visual scene. Within this class, four key functionalities have been identified: content-based multimedia data access tools; content-based manipulation and bitstream editing; hybrid natural and synthetic data coding; improved temporal random access.
- *high compression*, important not only to enable low bit rate applications, but also to provide the ability to efficiently code multiple views of a scene. Within this class, two key

functionalities have been identified: improved coding efficiency; coding of multiple concurrent data streams.

- *universal accessibility*, meaning that access to audio-visual data should be available over a wide range of storage and transmission media. Particularly important, considering the rapid growth of mobile communications, is the access to the applications via wireless networks. Within this class, two key functionalities have been identified: robustness in error-prone environments; content-based scalability.

In addition, several other important functionalities, that may be provided by already existing standards, are also considered in MPEG4.

It is clear that some of the functionalitities referred above, namely content-based interactivity, represent not only additional facilities but also a significant evolution in the way that the audio-visual information is represented. In March 1996, a Call for Proposals has been issued, specifically concerned with coding techniques for environments containing mixed synthetic and natural data [21]. MPEG4 is soliciting technology for standardising the coding of hybrid data that combines important features of traditional Audio-Visual and 2D/3D graphical data based on natural and synthetic sources.

5.2 Representation of Audio-Visual Scenes

One fundamental goal of MPEG4 is to efficiently encode interactive 2D and 3D environments consisting of real-time audio, video and synthetic objects [22]. One of the main consequences of the need to develop a standard providing content-based interactivity is the consideration of structures to represent the visual information more complex than the pixel, like the region or the object. These structures must be easily associated with meaningful semantic units that are part of the scene [19].

In order to provide the ability to interact with the audio-visual content of a scene, it is necessary that the scene is structured in terms of Audio-Visual Objects (AVO), which become accessible Audio-Visual units. An AVO is a representation of a real or virtual object, that may have associated only a video component or only an audio component or both components, may be 2D or 3D, time-varying or static, natural or synthetic, or a combination of these [23]. This way, a scene may be understood as a composition of AVOs, according to their spatial-temporal relationships.

5.3 Structure

To achieve the desired ability in terms of flexibility and extensibility, the structure of the MPEG4 standard is foreseen to be composed of four different elements (Fig. 4) [20]:

- the *MPEG4 Syntactic Description Language* (MSDL) [23] that must allow selection, description and downloading of tools, algorithms and profiles and describe how the elementary data streams are to be parsed and processed;
- *tools*, that are techniques accessible via the MSDL or described using the MSDL;
- *algorithms*, that are organised collections of tools providing one or more functionalities;
- *profiles*, that are algorithms or combinations of algorithms that provide solutions to particular classes of applications.

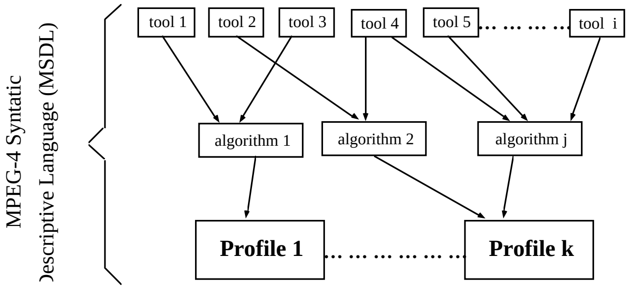

Figure 4) MPEG-4 elements.

It is not yet defined which of these elements will be standardised. The MSDL will certainly be normative. Theoretically, this would be enough. But, in practice, there will be the need to standardise some more elements. It should be noted that some profiles, corresponding to the present standards, for instance H.261, MPEG1, MPEG2, H.263, can be created, making this evolution more smooth.

5.4 Testing

The specification of the testing and evaluation methodologies for the new MPEG4 functionalities is a new challenge in the framework of standardisation, since there is no significant experience for the type of tests necessary to evaluate the new functionalities envisaged by MPEG4 [24]. A first Call For Proposals (CFP) for video algorithms and tools was issued with the deadline of September 1995. The first round of tests for video has taken place at the beginning of November 1995, and addressed only a limited set of functionalities, namely: scalability (spatial scalability, temporal scalability and object scalability), compression, and error robustness (error resilience and error recovery). The type of tests performed were conventional subjective viewing tests. Algorithms addressing other functionalities and tools have been evaluated by a panel of experts, based on the technical description. At this time, no tests were performed to evaluate combined audio-visual algorithms, but it is expected that these tests will be performed at a later stage. The sequences proposed to be tested, according to its content complexity, are summarised in table 4 [24]:

Class	Content complexity	Video test material
A	Low spatial detail & low amount of movement	Mother & daughter, Akiyo, Hall monitor, Container ship, Sean
B	Medium spatial detail & low amount of movement or vice versa	Foreman, News, Silent voice, Coast guard
C	High spatial detail & medium amount of movement or vice versa	Bus, Table tennis, Stefan, Mobile & Calendar
D	Stereoscopic	Tunnel, Fun fair
E	Hybrid natural and synthetic	Children, Bream, Weather, Destruction

Table 4) MPEG4 Video Library.

To compare the performance of the submitted coding schemes with respect to the existing standards, suitable anchor conditions were used for each range of bitrates. The standards used to generate the quality anchor sequences were ISO/MPEG1 and ITU-T H.263.

Class	Bitrates (kbit/s)	Anchor condition
A	10, 24, 48	H.263 at the same bitrate
B	24, 48, 112	H.263 at the same bitrate
C	320, 512, 1024	MPEG-1 at the same bitrate
D	512 for layer 1 1024 for layer 1	-
E	48. 112 320	H.263 at the same bitrate MPEG-1 at the same bitrate

Table 5) Test conditions proposed and anchors.

A second CFP for video algorithms and tools was issued with the deadline of January 1996. The proposals were evaluated by a panel of experts without formal subjective tests.

A wide range of algorithms were evaluated and are being discussed. Among them, it can be found the successful hybrid DCT/DPCM coding schemes and also many new algorithms like object-based and segmentation-based coding schemes.

The outcome of these two rounds of video proposals evaluation led to the definition of the first video Verification Model (VM), in January 1996.

5.5 MPEG4 Video VM

A VM is a completely defined encoding and decoding environment, composed of tools and algorithms, together with a bitstream syntax, such that an experiment performed by multiple independent parties will produce essentially identical results. A VM evolves through versions during the core experiment process [25]. In March 1996, the MPEG4 video VM was further developed into version 2.0 [26].

The representation architecture adopted in the present MPEG4 video VM is based on the concept of Video Object Planes (VOP). A VOP is an arbitrarily shaped 2D video object component. VOPs correspond to entities in the bitstream that the user can access and manipulate, and can carry semantically meaningful information. The encoder sends together with the VOP, composition information to indicate where and when each VOP is to be displayed. At the decoder side the user may be allowed to change the composition of the scene displayed by interacting on the composition information. The encoder and decoder architectures are presented in figure 4.

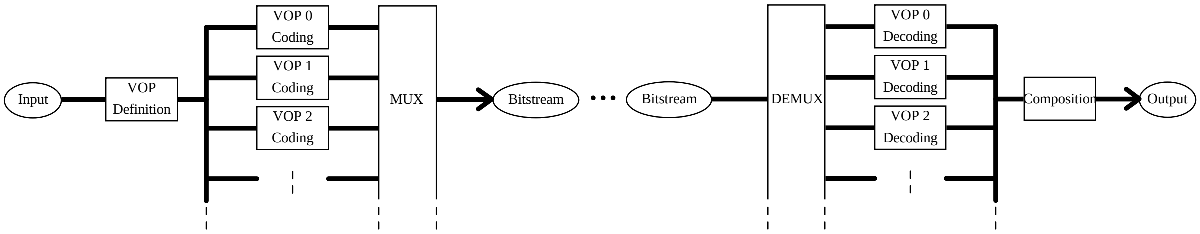

Figure 4) VM encoder and decoder structure.

A VOP is made up of Y, U, V components plus shape information, referred as alpha planes. The method to produce the VOPs is not defined in the MPEG4 video VM.

The VOP encoder is mainly composed of two parts: the shape coder and the traditional motion and texture coder. Alpha planes may be binary or grey scale (8 bits). Binary alpha planes are encoded by quadtree and grey scale alpha planes are encoded by quadtree with vector quantisation. To encode each VOP, the coding tools used in the VM are essentially those already used in existing video coding standards, with the difference that it is possible to separate, at the VOP level, the motion and texture information.

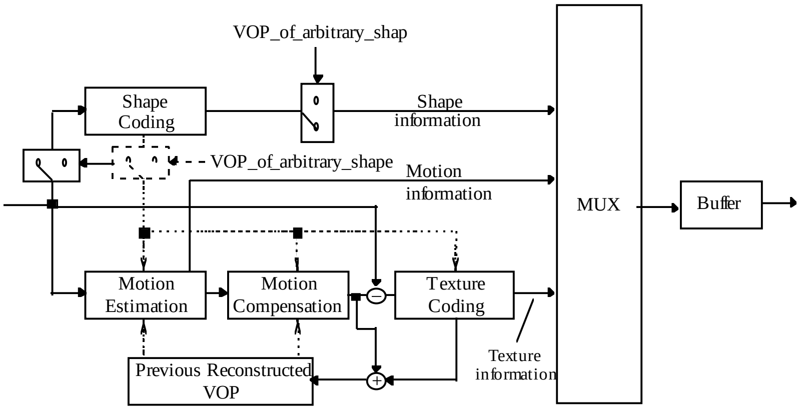

Figure 5) VOP encoder structure.

The VM bitstream syntax is structured into two layers: the session layer and the VOP layer. The session layer is the top layer. It contains the session width and height and comprises one or more VOPs. The VOP layer contains the VOP identifier, the VOP temporal reference, the VOP visibility, the VOP composition order, the VOP spatial reference, the VOP width and height, a VOP scaling factor, and the VOP coded data. Many of these values are not used for decoding but for picture composition.

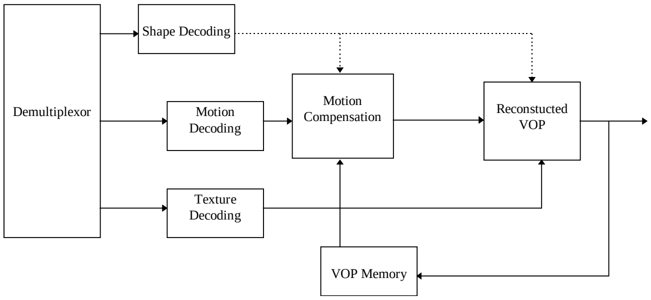

Figure 6) VOP decoder structure.

The VOP decoder is mainly composed of two parts: the shape decoder and the motion and texture decoder. The reconstructed VOP is obtained by the right combination of the shape, texture and motion information. The reconstructed VOPs are then passed to the compositor, where they are blended in the order specified by the VOP composition order.

Since this is still an early version of the VM, it is expected that many changes will be incorporated during the MPEG4 development process.

6. DISCUSSION

The success of the MPEG family of standards is based on the fact that they were designed to be generic. Thus, they can be used in a wider range of applications when compared with their predecessors that were designed to meet the requirements of one single application (e.g. H.120 and CCIR 721).

MPEG1 was designed for multimedia storage applications, providing considerable flexibly. There are currently many boards (encoders/decoders) available on the market and commercial products such as Philips CD-I. MPEG-1 bitstreams can be stored on digital media and read by MPC-1 CD-ROM compliant drives. So far the consumer market has been held up because of the amount of video that could be stored on a single disc (about 72 minutes).The combination of MPEG2 and DVD (Digital Video Disc.) format, has extended the potential running time of continuous video to 260 minutes. Nevertheless, this doesn't mean the end of MPEG1. MPEG1 is perfectly suitable for computer-based training and multimedia. By the way, 520 minutes of video is the limit that can be stored on a dual-sided, dual layer SD disc, that translates to 17Gb in digital terms. The first versions will have a single-layer capacity of 4.7 Gb (dual-layer versions will move this up to 9Gb).

MPEG2 standard is expected to cover a larger field of applications going from Digital Terrestrial Television Broadcasting (DTTB), Video-on-demand, pay TV, pay-per-view, HDTV, Direct Broadcasting Satellite (DBS), Satellite and Electronics News Gathering (SNG and ENG), distribution by cable or fibre, interactive applications, retrieval system on DSM, disc or tape, to the same storage applications but allowing considerably higher bitrates and quality levels. Thus MPEG2 is expected to be used in almost all applications involving transmission of video and sound. This success is not strange to the fact that there was a strong commitment from industries, cable and satellite operators and broadcasters in the development of this standard. However the encoding process is complex and not object of standardisation. To obtain high quality sequences considerable knowledge of the algorithm is required.

The Grand Alliance was formed by seven organisations (AT&T Corp., GI, MIT, Philips Consumer Electronics, David Sarnoff research center, Thomson, Zenith Electronics Corp.) to evaluate technologies and to decide on key elements that will be at the heart of the US HDTV system. The video compression and transport protocol were selected based on the MPEG-2 standards. In Europe, the development of standards for Digital Video Broadcasting (DVB) as well as the preparation of the introduction of services is coordinated by the European Project on Digital Video Broadcasting. Techniques for the transmission of DVB signals via satellite have been devised as well as a specification for retransmission of DVB signals via cable and (S)MATV networks. DVB group decided to use for the source coding of video signals MPEG 2 video standard. As MPEG 2 is a generic standard, a subset has been defined in the form of "Implementation Guidelines" to specify the services that are to be realised [27].

MPEG4, driven by the emerging needs of new Audio-Visual applications and the new ways that Audio-Visual information is being produced and consumed, moves to a radically new style of Audio-Visual coding, based on the concept of objects. Thus, decoding techniques are being extended to include scene composition and manipulation at the receiver side.

MPEG4 foresees providing functionalities based on the Audio-Visual content of the sequence. This standard will allow the user to access and control the content of an audio-visual sequence. It provides means to describe the content with a lot of flexibility. Decomposing a scene into VOPs does not only provide means for implementing interactivity, but it also provides means for a smart usage of the decoder capacity. Besides it is expected that the target hardware architectures will be partially or fully programmable.

In MPEG4, it is expected that the range of applications, or services, will be much broader than it was in MPEG2. Potential applications include content based Audio-Visual database access, games, virtual environment simulations, conferencing, training, education, mobile Audio-Visual terminals, improved PSTN Audio-Visual communications, tele-shopping, remote monitoring and control.

REFERENCES

[1] ISO/IEC 10918-1, "Information Technology - Digital compression and coding of continuos-tone still images: requirements and guidelines", Geneva, 1994.

[2] CCITT Recommendation H.120, "Codecs for videoconferencing using primary digital group transmission", Geneva, 1989.

[3] CCITT Recommendation H.261, "Video codec for audiovisual services at p×64kb/s", Geneva, 1990.

[4] CCIR Recommendation 721, "Transmission of component-coded digital television signals for standardization", Signal Processing: Image Commun., vol. 1, no. 1, pp. 29-43, 1989.

[5] CCIR Recommendation 723/ETS 300 174, "Digital coding of component television signals for contribution quality applications in the range 34-45 Mbits/s", Sophia Antipolis, 1992.

[6] ISO/IEC 11172-2, "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1,5 Mbit/s-video", Geneva, 1993.

[7] ISO/IEC IS 13818-2: Generic coding of moving pictures and associated audio, November 1994

[8] Sakae Okubo, "Reference Model Methodology - A Tool for the Collaborative Creation of Video Coding Standards", Proceedings of the IEEE, vol. 83, no. 2, pp. 139-150, February 1995

[9] Ralf Schafer, Thomas Sikora, "Digital Video Coding Standards and their Role in Video Communications", Proceedings of the IEEE, vol. 83, no. 6, pp. 907-924, June1995.

[10] Paul A. Wintz, "Transform Picture Coding", Proceedings of the IEEE, vol. 60, no. 7, pp. 809-820, July1972.

[11] Rama Chellappa, "Digital Image Processing - 2nd Edition", IEEE Computer Society Press, pp. 401-408, 1992.

[12] A. Rosenfeld, A. C. Kak, "Digital Picture Processing", vol. 1, chapter 5, Academic Press., Orlando, Fla., 1982

[13] N. Ahmed, T. Natarajan, K. R. Rao, "Discrete Cosine Transform", IEEE Trans. Computer, vol. C-23, no. 1, pp. 90-93, Jan 1974.

[14] Didier Le Gall, "MPEG: A Video Compression Standard for Multimedia Applications", Communications of the ACM, vol. 34, no. 4, pp. 47-58, April 1991.

[15] MPEG-2 Video Requirements Subgroup, "Agreements on Profile/Level", ISO/IEC JTC1/SC29/WG11 Doc. N0489, New York, July 1993.

[16] P. N. Tudor, "MPEG-2 Video compression tutorial", January 1995.

[17]ISO/IEC-JTC1/SC29/WG11 MPEG93/457 MPEG Video Test Model 5 (TM-5), April 1993.

[18] Draft ITU-T Recommendation H.263, "Video coding for low bitrate communication”, December 1995.

[19] Fernando Pereira, "MPEG-4: A New Challenge for the Representation of Audio-Visual Information", Keynote speech at Picture Coding Symposium’96, March 1996.

[20] ISO/IEC JTC1/SC29/WG11 N998, "MPEG-4 Proposal Package Description (PPD) - Revision 3”, July 1995.

[21] ISO/IEC JTC1/SC29/WG11 N1195, "MPEG-4 SNHC Call For Proposals”, March1996.

[22] ISO/IEC JTC1/SC29/WG11 N1199, "MPEG-4 SNHC Proposal Package Description”, March1996.

[23] ISO/IEC JTC1/SC29/WG11 N1246, "MSDL specification - Version 1.1”, March1996.

[24] ISO/IEC JTC1/SC29/WG11 N999, "MPEG-4 Test and Evaluation Procedures Document”, July1995.

[25] ISO/IEC JTC1/SC29/WG11 N1191, "MPEG-4 Development Procedures”, March1996.

[26] ISO/IEC JTC1/SC29/WG11 N1260, "MPEG-4 Video Verification Model - Version 2.0”, March1996.

[27] U. Reims, "The European Project on Digital Video Broadcasting - Achievements and Current Status ", International Broadcasting Convention, September 1994.