

MPEG-4 Video and its Potential for Future Multimedia Services

T.Sikora¹ and L.Chiariglione²

¹ Heinrich-Hertz-Institute (HHI), Einsteinufer 37, D-10587 Berlin, Germany, Email: sikora@hhi.de

² CSELT, Italy

Abstract

The Moving Picture Experts Group (MPEG) committee, which originated the MPEG-1 and MPEG-2 video and audio compression standards, is currently developing MPEG-4 with wide industry participation. MPEG-4 is targeted for interactive Multimedia applications and will become an international standard in 1998. It is expected that MPEG-4 will become the enabling technology for multimedia audio-visual communications as much as MPEG-2 has become the enabling technology for digital television. The purpose of the paper is to discuss the scope and potential of the MPEG-4 standardization activities for networked interactive multimedia applications with particular emphasis on the MPEG-4 video standard.

I. Introduction

Anticipating the rapid convergence of telecommunications industries, computer and TV/film industries, the MPEG group officially initiated a new MPEG-4 standardization phase in 1994 - with the mandate to standardize algorithms and tools for coding and flexible representation of audio-visual data to meet the challenges of future Multimedia applications and applications environments [1][2]. In particular MPEG-4 addresses the need for

- *Universal accessibility and robustness in error prone environments* - Multimedia audio-visual data need to be transmitted and accessed in heterogeneous network environments, possibly under severe error conditions (e.g. mobile channels). Although the MPEG-4 standards will be network (physical-layer) independent in nature, the algorithms and tools for coding audio-visual data need to be designed with awareness of network peculiarities.
- *High interactive functionality* - Future Multimedia applications will call for extended interactive functionalities to assist the user's needs. In particular the flexible, highly interactive access to and manipulation of audio-visual data will be of prime importance. It is envisioned that - in addition

to conventional playback of audio and video sequences - the user need to access „content“ of audio-visual data to present and manipulate/store the data in a highly flexible way.

- *Coding of natural and synthetic data* - Next generation graphics processors will enable Multimedia terminals to present both pixel based audio and video data together with synthetic audio/speech and video in a highly flexible way. MPEG-4 will assist the efficient and flexible coding and representation of both natural (pixel based) as well as synthetic data.
- *Compression efficiency* - For the storage and transmission of audio-visual data a high coding efficiency, meaning a good quality of the reconstructed data, is required. Improved coding efficiency, in particular at very low bit rates below 64 kbits/s, continues to be an important functionality to be supported by the MPEG-4 video standard.
- *Decoder downloadability* - In particular the technological advances foreseen in the areas of general purpose DSP's and powerful high end computers will allow the software implementation and flexible downloading of tools and algorithms for decoding audio and video data, tailored to suit the needs of specific applications.

Bit rates targeted for the MPEG-4 video standard are between 5-64 kbits/s for mobile or PSTN video applications [3] and up to 2 Mbits/s for TV/film applications. Seven new (with respect to existing or emerging standards) key video coding functionalities have been defined which support the MPEG-4 focus and which provide the main requirements for the work in the MPEG video group [2]. The requirements cover the main topics related to "Content-Based Interactivity", "Compression" and "Universal Access". The release of the MPEG-4 International Standard is targeted for July 1998.

II. MPEG-4 Content-Based Interactivity

In addition to standard MPEG-1 or MPEG-2 like provisions for efficient coding of conventional image or audio sequences [4], MPEG-4 will enable an efficient coded representation of the audio and video data that can be „content based“, with the aim to use and present the data in a highly flexible way [1][2]. In particular it is envisioned to allow the access and manipulation of audio-visual objects in the compressed domain at the coded data level - to assist future Multimedia data-base access applications such as the flexible presentation of image or audio content in the World-Wide-Web, computer games, and related applications.

Video: The basic concept of the envisioned MPEG-4 „content-based“ functionality for image/video applications is illustrated in Figure 1 for a simple example of an image scene containing a number of video objects, here the background, several items and a text overlay. The attempt is to encode the sequence in a way that will allow for the user the separate decoding and reconstruction of the objects - to assist the presentation and manipulation of the original scene in a flexible way.

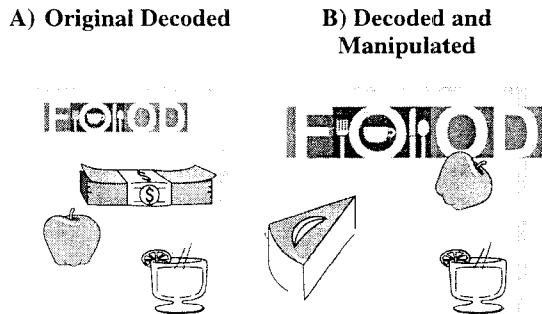


Fig. 1: An example for the flexible content-based access and manipulation of objects in MPEG-4 image sequences.

The MPEG-4 video coding standard will provide an „object layered“ bitstream to assist this functionality. Each object is coded into a separate object bitstream layer. The shape and transparency of the object - as well as the spatial coordinates and additional parameters describing scales and location, such as object zoom, rotation, translation or related - are included in the bitstream. The user can either reconstruct the original sequence in its entity, by decoding all „object layers“ and by displaying the objects at original sizes and scales and at the original location as indicated in Figure 1A. Alternatively it is possible to manipulate the image sequence by simple operations. For example in Figure 1B some objects were not decoded and used for reconstruction, while others were decoded and displayed using subsequent scaling, rotation or translation. The scaling, rotation and translation parameters employed for manipulation

of the image sequence can be altered in the bit stream by means of simple bitstream editing operations - without the need for further transcoding. In addition new objects can be included which did not belong to the original scene - or original objects may be neglected. Since the bit stream of the sequence is organized in an „object layered“ form the inclusion or deletion of additional objects in the image sequence is performed on the bitstream level by adding/deleting the appropriate object bitstreams - again, without the need for further transcoding.

It is targeted to provide to the user the different video objects also with various scales of quality, size or frame rates to assist the flexible presentation of the data.

Audio and Speech: Similar to the content-based functionalities outlined for video applications above it is targeted to provide object-layered audio bitstreams to assist the access to and manipulation of content in audio and speech sequences. An example of an application is the content-based audio coding of a violin concerto played by an orchestra - where, on demand, it is possible to extract and enhance the sound of single instruments. Alternatively the concerto may be replayed with or without the solo violin.

SNHC: It is envisioned to provide the above capabilities for both synthetic (S) and natural (N) audio-visual objects as well as for hybrid (H) coding (C) and representation of natural and synthetic objects. As an example it is targeted to allow the coding and generation of text overlays based on graphics primitives. This would greatly reduce the bits needed to store and transmit text and allow a high degree of flexibility for representing or altering text - e.g. it will be possible to select various types and fonts for display in Figure 1. Other functionalities envisioned are the efficient coding of computer animated texture mapped wire-grid faces and human bodies.

Systems: The MPEG-4 architecture will allow the separate coding of audio or video objects, natural or synthetic - and the appropriate multiplexing of the separate object elementary streams into a single bitstream. Similar to the MPEG-1 and MPEG-2 standards a MPEG-4 „Systems“ standard will be developed to assist multiplexing of elementary streams, synchronization and packetization. Additionally, as described above, the MPEG-4 systems multiplex will provide in the header of the bitstream layer of each object the basic representation/manipulation parameters - such as translation, rotation or zoom of an object in relation to reference coordinates and scales.

III. The MPEG-4 Video "Toolbox" Approach

The overall MPEG-4 applications scenario envisions the standardization of "tools" and "algorithms" for natural audio and video as well as for synthetic 2-D or 3-D audio and video to allow the hybrid coding of these components [1][2]. The MPEG-4 group has taken further steps towards an open, flexible and extensible MPEG-4 standard by anticipating the foreseen rapid developments in the area of programmable general purpose DSP technology - and the obvious advantages with respect to software implementations/downloadability of the standard. In this respect it is targeted to provide an open MPEG-4 standard by enabling mechanisms to download missing software decoder tools at the receiver. The "glue" that will bind independent coding tools together is the foreseen MPEG-4 Systems Description Language (MSDL) which will be part of the MPEG-4 „Systems“ standard and comprise several key components. Firstly, a definition of the interfaces between the coding tools. Second, a mechanism to combine coding tools and to construct algorithms and profiles, and third a mechanism to download new tools. The MSDL will transmit with the bitstream the structure and rules for the decoder - thus the way how the tools have to be used at the decoder in order to decode and reconstruct natural and synthetic audio and video.

IV. The MPEG-4 Video Standard

The MPEG-4 video coding algorithms will eventually support all functionalities already provided by MPEG-1 and MPEG-2, including the provision to efficiently compress standard rectangular sized image sequences at varying levels of input formats, frame rates and bit rates. In addition the content-based functionalities will be assisted.

A basic classification of the bit rates and functionalities currently provided by the MPEG-4 video coding model under development is depicted in Figure 2, with the attempt to clusters bit rate levels vs sets of functionalities and basic applications profiles. At the bottom end a „VLBV Core“ (VLBV: Very Low Bit Rate Video) provides algorithms and tools for applications operating at bit rates between 5...64 kbits/s, supporting image sequences with lower spatial resolutions (from a few pixels per lines and rows up to CIF resolution) and lower frame rates (ranging from 0 Hz for still images to 15 Hz). The basic applications specific functionalities supported by the VLBV Core include a) conventional VLBV coding of rectangular size image sequences with high coding efficiency and high error robustness/resilience, low latency and low complexity for real-time Multimedia communications applications, and b) provisions for „random access“ and „fast/forward“ and „fast/reverse“ operations for Multimedia data-base storage and access applications.

The same basic functionalities outlined above are also supported by a Higher Bit Rate Video Core (HBV

Core) with a higher range of spatial and temporal input parameters up to R.601 resolutions - employing identical algorithms and tools as the VLBV Core. The reader is referred to the references [2] and [3] for a more detailed description of the techniques under development for the VLBV and HBV Cores.

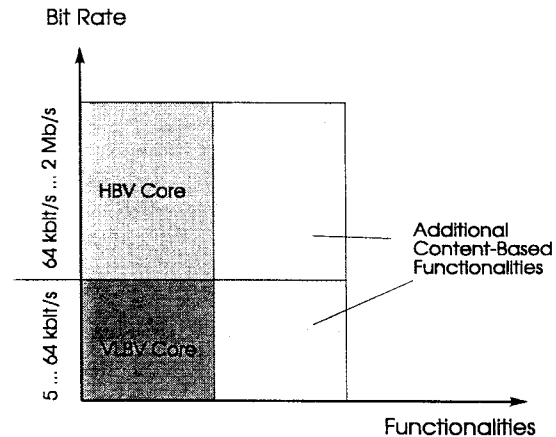


Fig. 2: Structure of the MPEG-4 Video Coding Standard

At the heart of the additional „content-based“ functionalities is the support for the separate encoding and decoding of content (i.e. physical objects in a scene) as already discussed with Figure 1. Within the context of MPEG-4 this functionality - the ability to identify and selectively decode and reconstruct video content of interest - is referred to as "Content-Based Scalability". This MPEG-4 feature provides the most elementary mechanism for interactivity and manipulation with/of content of images or video in the compressed domain without the need for further segmentation or transcoding at the receiver. The extended MPEG-4 algorithms and tools for content-based functionalities can be seen as a superset of the VLBV and HBV Cores - thus the tools provided by the VLBV and HBV Cores are complemented with additional elements [2].

To enable the content based interactive functionalities envisioned, the MPEG-4 video standard introduces the concept of Video Object Planes (VOP's). This concept is illustrated in Figure 3. It is assumed that each frame of an input video sequence is segmented into a number of arbitrarily shaped image regions (video object planes) - each of the regions may possibly cover particular image or video content of interest, i.e. describing physical objects or content within scenes. In contrast to the video source format used for the MPEG-1 and MPEG-2 standards, the video input to be coded by the MPEG-4 Verification Model is thus no longer considered a rectangular region. The input to be coded can be a VOP image region of arbitrary shape and the shape and location of the region can vary from frame to frame. Successive VOP's

belonging to the same physical object in a scene are referred to as Video Objects (VO's) - a sequence of VOP's of possibly arbitrary shape and position. The shape, motion and texture information of the VOP's belonging to the same VO is encoded and transmitted or coded into a separate VOL (Video Object Layer). In addition, relevant information needed to identify each of the VOL's - and how the various VOL's are composed at the receiver to reconstruct the entire original sequence is also included in the bitstream. This allows the separate decoding of each VOP and the required flexible manipulation of the video sequence as indicated in Figures 1 and 3. Notice that the video source input assumed for the VOL structure either already exists in terms of separate entities (i.e. is generated with chroma-key technology) or is generated by means of on-line or off-line segmentation algorithms.

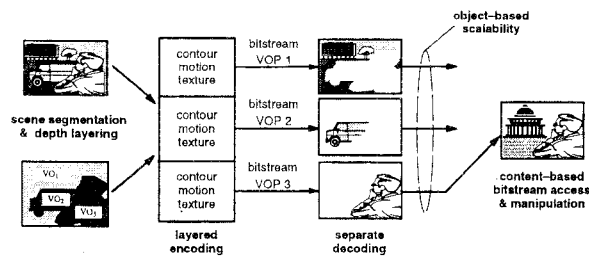


Fig. 3: The „object-layered“ coding approach taken by the MPEG-4 video coding standard.

Notice that MPEG-4 images as well as image sequences are in general considered to be arbitrarily shaped - in contrast to the standard MPEG-1 and MPEG-2 definitions which encode rectangular size image sequences. The MPEG-4 content-based approach can be seen as a logical extension of the conventional MPEG-4 VLBV and HBV Core coding approach towards image input sequences of arbitrary shape. In particular, if the original input image sequences are not decomposed into several VOL's of arbitrary shape, the coding structure simply degenerates into a single layer representation which supports coding of conventional image sequences of rectangular shape.

As illustrated in Figure 4, the MPEG-4 video standard will support the coding of rectangular size image sequences which is similar to conventional MPEG-1/2 coding approaches and involves motion prediction/compensation followed by DCT-based texture coding. For the content-based functionalities where the image sequence input is of arbitrary shape and location this approach is extended by also coding shape and transparency information.

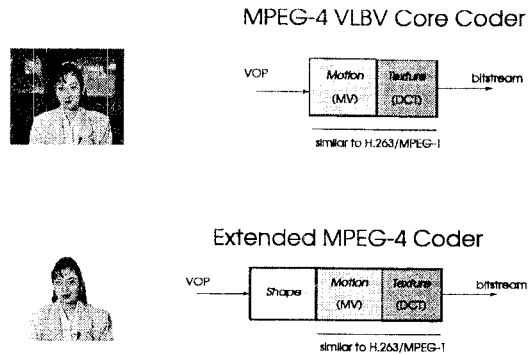


Fig. 4: VLBV Core and the Extended Core Coder

V. Summary

In this paper the scope and potential of the MPEG-4 standard was discussed in the context of future audio-visual Multimedia communications environments. The MPEG-4 standard will provide tools and algorithms for coding both natural and synthetic video, audio and speech data - and provisions to represent the data at the user terminal in a highly flexible way. The MPEG-4 video coding standard introduces the concept of object-layered video coding to assist „content-based“ functionalities for the user.

References:

- [1] *L. Chiariglione*, „MPEG and Multimedia Communications“, IEEE Trans. CSVT, Vol.7, No.1, Feb.1997.
- [2] *T. Sikora*, „The MPEG-4 Video Standard Verification Model“, IEEE Trans. CSVT, Vol.7, No.1, Feb.1997.
- [3] *T. Sikora*, "MPEG-4 Very Low Bit Rate Video ", Proc. IEEE ISCAS Conference, Hongkong, June 1997.
- [4] *R. Schäfer and T. Sikora*, „Digital Video Coding Standards and Their Role in Video Communications“, Proceedings of the IEEE, Vol.83, No.6. June 1995.