Contact Details of Presenting Authors

Stefan Rauthenberg (rauthenberg@hhi.de), Peter Kauff (kauff@hhi.de)
Tel: +49-30-31002 266, +49-30-31002 615
Fax: +49-30-3927200

Summation

- Brief explaination of the MPEG-4 Standard of ISO/IEC.
- Discussion of its potential benefits for shared virtual environment systems.
- Presentation of a virtual conference system using MPEG-4 technology
- Details of an exisiting prototype implementation: The Virtual Meeting Point
- Outlook on future work targeting immersive tele-presence systems.
- Live demonstration of the Virtual Meeting Point (reduced demo version).

The Virtual Meeting Room

A Realtime Implementation of a Shared Virtual Environment System Using Today's Consumer Technology in Connection with the MPEG-4 Standard

Stefan Rauthenberg Peter Kauff
Heinrich-Hertz-Institute (HHI), Image Processing Department,
Einsteinufer 37, D-10587 Berlin, Germany

Andreas Graffunder
T-Nova Deutsche Telekom Innovationsgesellschaft GmbH, Berkom,
Am Kavalleriesand 3, D-64295 Darmstadt, Germany

Introduction

Roundabout five years ago the Moving Pictures Experts Group (MPEG) of ISO/IEC initiated its third standardization phase called MPEG-4 [1][3]. Meanwhile MPEG-4 has been released successfully as International Standard (IS) in early 1999 [2]. In view of the versatile requirements of multimedia, it supports a wide range of bit rates and functionalities. Outstanding novelities of MPEG-4 are the philosophy of considering scenes as compositions of audio-visual objects (AVO's), the support of hybrid coding of natural video and 2D/3D graphics in a common context (e.g. virtual 3D worlds) and the provision of advanced system and interoperability capabilities for interactive services. Especially communication systems working with shared virtual environments will benefit from these particular advantages of MPEG-4.

Against this background the paper describes the prototype of an MPEG-4 video conference system where several participants can meet in a virtual 3D meeting room. The complete prototype system has been built up at Heinrich-Hertz-Institute (HHI) in close cooperation with T-Nova Deutsche Telekom Innovationsgesellschaft (T-Nova). The main intention of this development was to study and to demonstrate its feasibility using today's consumer technology (incl. networking, compression, scene composition, interactivity, graphical user interface, etc.). It was therefore exclusively implemented in software on state-of-the-art PC's (i.e. without any support of dedicated hardware or expensive graphic boards). A real-time demonstration of a system with one PC workstation and two laptops with USB cameras will be shown during the conference.

The MPEG-4 Standard

Although conventional compression standards like ITU-T H.261/263 and ISO/IEC MPEG-1/2 have already been specified for various fields of application reaching from very low bit rate video coding to high quality video coding, their common purpose is digital transmission or storage bandwidth reduction [5][6]. Hence, video signals of a given rectangular format are usually reproduced in exactly the same format on a rectangular screen. MPEG-4 goes beyond this restricted functionality of conventional compression standards. Envisaging multimedia applications, it provides an universal description of the scene, where the particular objects contained therein are coded separately, often in combination with routes for internal event handling and user interactions. In more detail, MPEG-4 offers the following main features [2]:

- hybrid coding of natural and synthetic data including the capability to code multiple types of audio-visual objects (AVO) like frame-based and arbitrarily shaped videos or 2D and 3D graphics in a common framework
- advanced video and audio coding features (e.g. ability to encode shaped video consisting of texture, motion, binary shapes and continuous alpha planes)
- flexible scene composition by using a special description language called BIFS (Binary Format for Scene) taking into account 2D or 3D scenes, natural and synthetic AVO's as well as the definition of internal scene interrelations and content-based user interactions
- enhanced networking and systems features like the framework connecting MPEG-4 terminals over heterogeneous networks and transport layers with other terminals or media servers - the so-called DMIF (Digital Multimedia Integration Framework) - or the definition of special combinations of tools for AVO coding and scene composition, organized in profiles and levels to classify the complexity as well as the performance features of MPEG-4 terminals

Fig. 1 gives an example of a BIFS scene graph consisting of different kinds of AVO: a static background (e.g. still image or texture map), a 2D audio-visual presentation (e.g. conventional video and sound), a set of graphical objects (e.g. geometric primitives or 3D wire-frame models) and a natural portrayal of a person (arbitrarily shaped video with voice). Often, the leaf nodes of a BIFS scene graph point at an associated object descriptor (OD). The OD gives information about the object identification and contents. To this end, an OD refers to further sources containing either all information needed for streaming and decoding the linked AVO or information on a sub-graph of the scene again encoded by BIFS.

The different AVO's of a given scene are encoded separately by using specific encoder tools depending on the particular object type. As a result, each AVO is represented by one or more individual bit-streams. These AVO elementary streams are multiplexed at systems level together with all timing information needed for synchronization purposes as well as the BIFS & OD elementary streams providing the particular rules for scene composition and internal event handling. At the MPEG-4 terminal the elementary streams are de-multiplexed and then assembled to the audiovisual scene to be reproduced. This is achieved by a compositor which interprets the BIFS syntax - optionally in combination with user interactions where a particular AVO can, for example, be selected to retrieve associated information from media-servers.

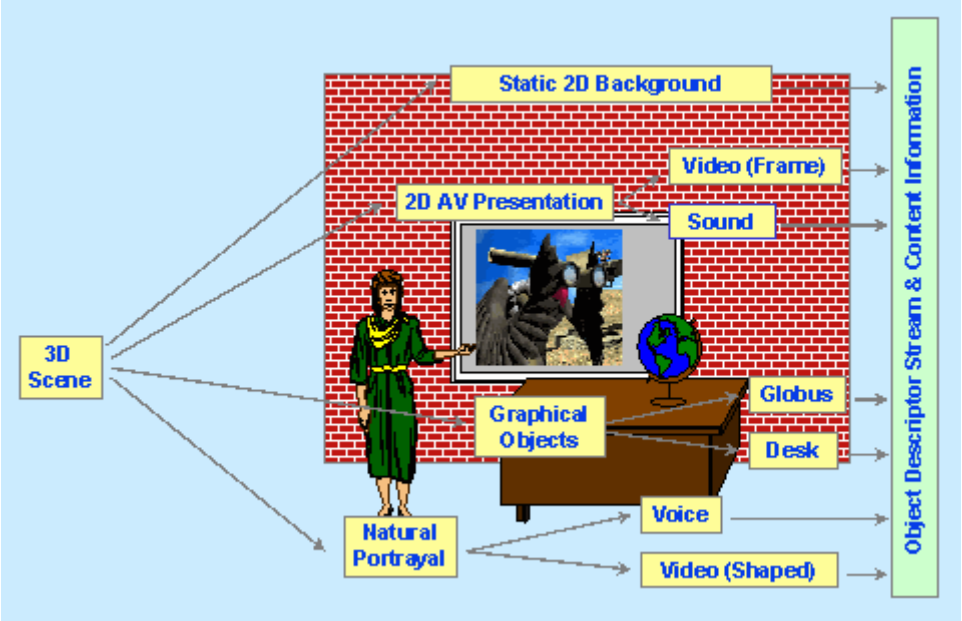

Fig. 1: Example of a BIFS scene graph in MPEG-4

The Virtual Meeting Point

The virtual meeting point belongs to the category of shared virtual environment applications which may take advantage of certain unique features of the MPEG-4 standard, such as interoperability between distributed terminals, user navigation and interaction in virtual 3D worlds, the seamless integration of natural video objects into graphical scenes and the possibility to encode arbitrarily shaped video objects [4][5]. The latter is particularly important for the integration of live video objects into virtual environments as for example AVO's representing real persons sitting behind a synthetic desk (see **Fig. 2**).

Fig.2 : Screen shot of a running conference

The outline of the prototype system which has been developed at HHI in cooperation with T-Nova and which has been demonstrated the first time at the MINT Symposium in November 1998 in Berlin is shown in **Fig. 3** [6]. It

consists of up to four MPEG-4 terminals and an additional conference server that manages control tasks. In contrast to conventional multi-point conference systems based on the ITU-T H.32X standards, which display the participants in separate rectangular windows, it integrates all participants into the 3D scene of a shared virtual meeting room as depicted in **Fig. 2** .

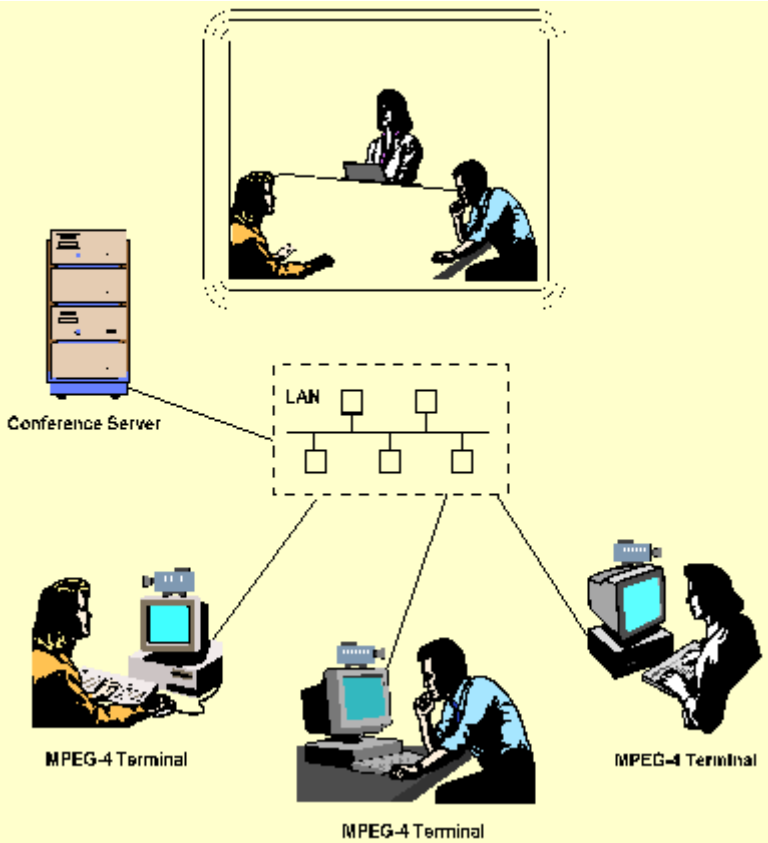

Fig. 3: System overview

Fig. 4 shows system details of one MPEG-4 conference terminal. For transmission purposes the video of a conferee is captured by a camera and automatically segmented in real-time against a stationary and pre-known but arbitrary-textured background. The segmented portrayal is then encoded as a shaped AVO by a real-time MPEG-4 video encoder. The resulting elementary stream is synchronized and multiplexed with the associated audio data by using a RTP packetizer. The data packets are transmitted via RTP/UDP/IP multicast or unicast depending on the characteristics of the network connecting the conference terminals.

In addition to real-time segmentation and encoding, each terminal has to run multiple video/audio decoders in order to process the data from other participants. The decoded AVO's are streamed as real-time data into a BIFS scene of the virtual meeting room. The composition of the 3D scene, including synchronization of all the natural AVOs, is achieved by a high-speed software rendering engine running nearly in real time video speed without relying on particular display hardware. The user can navigate through the 3D scene (e.g.: closing up to zoom into one of the conferees' portrayal or to interact with devices in the virtual world). On state-of-the-art PC's the display subsystem is able to achieve a video frame rate of 12.5 fps for scenes with approx. 16000 triangles and 8000 vertices.

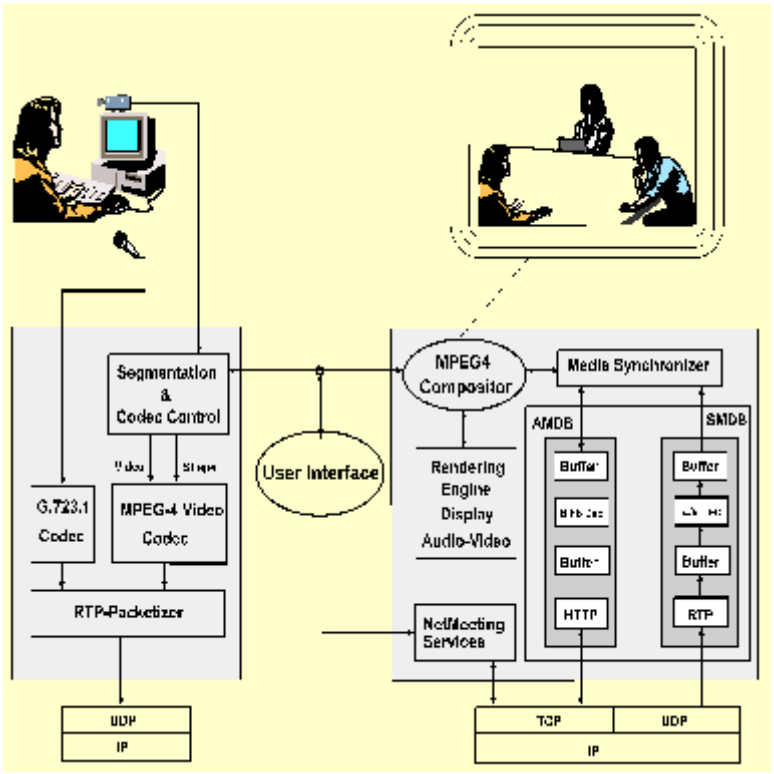

Fig. 4: Details of one MPEG-4 conference terminal

As all these processes have to run on one PC, it is quite obvious that the computational load of the virtual meeting point is quite high. Computing time is clearly dominated by the 3D-compositor followed by segmentation and MPEG-4 video encoding. A Pentium 400 MHz dual processor machine is required to achieve the maximum video frame rate of 12.5 fps. Therefore, interoperability with less well equipped terminals is a further important issue for the conference system. The terminals can operate at different complexity levels depending on their particular configuration. In the case of a full-power dual processor PC, the system supports all features of 3D composition including navigation through a virtual meeting room while up to four conferees are displayed as arbitrarily shaped video objects integrated in the common 3D scene as shown in **Fig. 2** . For terminals with less power like a single processor PC or a laptop, the system may fall back to 2D BIFS composition where up to three participants are displayed in a static layered 2D scene.

Conclusions and Future Work

By its generalized scene description language (BIFS) as well as its enlarged systems and coding concept, MPEG-4 enables a flexible processing of complex audio-visual contents, such as scene contents in interactive tele-communication. It supports streaming of various kinds of real time data from servers or between distributed terminals and it allows integration of audio-visual streams directly into 2D and 3D scenes while the user continues to interact with its content.

In order to demonstrate these potential features for shared virtual environment systems, a software prototype for interactive MPEG-4 service called *The Virtual Meeting Point* has been developed in the context of a feasibility study in close cooperation with the company T-Nova belonging to Deutsche Telekom. With this prototype system it could be demonstrated successfully that complex MPEG-4 services including networking, compression, 3D scene composition and interactive data retrieval can exclusively be implemented in software on a state-of-the-art PC without any support from dedicated hardware or graphic acceleration. Basic real-time software of MPEG-4 key components, such as video encoders and decoders or compositors, as well as the real-time segmentation of natural video objects, have been developed in this context and are now available for other system implementations in the same field or for re-designing and commercializing the existing system.

Future work will exploit these basic components to extend follow-up systems towards more telepresence. One main issue of the next steps in this direction is an increase of spatio-temporal resolution to improve quality and realizm of natural portrayals in virtual environments. Another point is the correct perspective presentation of natural video objects in the virtual environment if the user navigates through the virtual 3D world or if he moves his head in front of the display. The latter is particulary important for tele-presence applications, for example eye-contact in tele-conference systems using shared virtual tables. In principle, perspective view point adaptation becomes possible whenever multiple segmented camera views are available from the capture devices [7].In this sense, it is planned to utilize special options of the MPEG-4 gray-level shape syntax to encode disparity maps which are derived from disparity estimation between the available views, and to implement the so-called incomplete 3D representation of video objects for efficient and fast viewpoint interpolation [8][9]. The ulitmative goal of all these activities is a demonstrator of an immersive tele-conference system which runs in real-time software with the signal processor technology of the next years and most desirable without any dedicated hardware.

Acknowledgements

The authors would like to thank Mark Palkow, Ralf Tanger, Uwe Kowalik and Edward Cooke for their assistance during the implementation of the *Virtual Meeting Point* . This work was supported by the Ministry for Education, Science, Research and Technology of the Federal Republic of Germany under Grant No. 01 BN 701 and Land Berlin. The prototype was developed in cooperation with T-Nova Deutsche Telekom Innovationsgesellschaft mbH.

References

[1] T. Sikora and L. Chiariglione, **"MPEG-4 and its Potential for Future Multimedia Services"**, IEEE Proc. of ISCAS'97, Hong Kong, June 1997.

[2] ISO/IEC JTC1/SC29/WG11, **" Draft ISO/IEC FDIS 14496 - Generic Coding of Audiovisual Objects - Parts 1 to 6"**, MPEG, doc. N2501 to N2506, Oct. 1998.

[3] T. Sikora, **"MPEG digital video coding standards"**, IEEE Signal Processing Magazine, Vol. 14, No. 5, pp. 82-100, Sept. 1997

[4] P. Kauff, J. -R. Ohm, T. Sikora: **"The MPEG-4 standard and its application in virtual 3D environments"**, Proceedings of 42nd Asilomar Conference, Monterey, Nov. 1998.

[5] S. Rauthenberg, U. Kowalik, A. Graffunder, P. Kauff: **"Virtual Shop and Virtual Meeting Point - Two Prototype Applications of Interactive Services Using the New Multimedia Coding Standard MPEG-4"**, Proc. of ICCC'99, Int. Conf. on Computer Communication, Tokyo, September 1999

[6] S. Bauer, C. Herpel, A. Kaup, J.-R- Ohm and J. Spille, **"The Multimedia-Standard MPEG-4 and its Applications in the MINT Project (in German)"**, Special Issue of "Der Fernmelde Ingenieur", pp. 43-64, Vol. 52, No. 11 & 12, Symposium: The MINT Project &ndash Multimedia Communication Using Integrated Networks and Terminals, Berlin, November 1998

[7] E. Izquierdo and X. Feng, **"Image-Based 3D Modeling of Arbitrarily Natural Objects"**, VLBV Workshop, Urbana, Illinois, USA, October 1998

[8] J.-R. Ohm and K. Müller : **"Incomplete 3D - Multiview representation of video objects"**, IEEE Trans. on CSVT., Special Issue on SNHC, February 1999

[9] S. Ekmekci and J.-R. Ohm : **"Incomplete 3D Representation and View Synthesis for Video Objects Captured by Multiple Cameras"**, PCS'99, Portland, USA, April 1999