MPEG-4 for Networked Collaborative Virtual Environments

Igor Pandzic[1], Tolga Capin[2],
Nadia Magnenat-Thalmann[1], Daniel Thalmann[2]

[1] MIRALab - CUI
University of Geneva
24 rue du Général-Dufour
CH1211 Geneva 4, Switzerland
{Igor.Pandzic,Nadia.Thalmann}@cui.unige.ch
http://miralabwww.unige.ch/

[2] Computer Graphics Laboratory
Swiss Federal Institute of Technology (EPFL)
CH1015 Lausanne, Switzerland
{capin, thalmann}@lig.di.epfl.ch
http://ligwww.epfl.ch/

Abstract

MPEG is traditionally committed to coding and compression of audio-visual data from natural sources. However, the emerging MPEG-4 standard aims not only at multiple natural audio-visual objects composing the scene, but also synthetic audio and video to be integrated with the natural. It will also allow more interaction with both synthetic and natural objects.

Networked Collaborative Virtual Environments (NCVE) have a wide range of different requirements on networking. This paper analyzes how these requirements could potentially be met by MPEG-4 and shows that MPEG-4 tools could be extremely useful for development of such environments.

Keywords: Networked Collaborative Virtual Environments, MPEG-4

Introduction

Networked Collaborative Virtual Environments (NCVEs) are systems that allow multiple geographically distant users to evolve in a common virtual environment. The users themselves are represented within the environment using a graphical embodiment. An example of such an environment is shown in figure 1.

The user can evolve within the environment and interact with it. All events that have an impact on the environment are transmitted to other sites so that all environments can be updated and kept consistent, giving the impression for the users of being in the same, unique environment. The users become a part of the environment, embodied by a graphical representation that should ideally be human-like.

All users' actions, movements, facial expressions, changes in appearance etc. are transmitted to all

participating sites in the simulation, enabling the interaction between the users.

The need to maintain environment/user state coherent on multiple networked sites generates an extensive set of requirements on the networking part - various data types to be transmitted, each with different requirements and priorities.

Moving Pictures Expert Group (MPEG) is currently working on the new MPEG-4 standard, scheduled to become International Standard in November 1996. In a world where audio-visual data is increasingly stored, transferred and manipulated digitally, MPEG-4 sets its objectives beyond "plain" compression. Instead of regarding video as a sequence of frames with fixed shape and size and with attached audio information, the video scene is regarded as a set of dynamic objects. Thus the background of the scene might be one object, a moving car another, the sound of the engine the third etc. The objects are spatially and temporally independent and therefore can be stored, transferred and manipulated independently. The composition of the final scene is done at the decoder, potentially allowing great manipulation freedom to the consumer of the data.

Figure 1: An example session of a Networked Collaborative Virtual Environment

Video and audio acquired by recording from the real world is called natural. In addition to the natural objects, MPEG-4 aims to enable integration of synthetic objects within the scene. Synthetic, computer generated graphics and sounds are being produced and used in ever increasing quantities and it is the role of the Synthetic/Natural Hybrid Coding (SNHC) group of MPEG to integrate the coding of these objects with the natural data. Current work of SNHC concentrates on Face and Body Animation, Generic Object Coding, Media Integration of Text and Graphics and Synthetic Audio.

Currently there are four groups that work on producing MPEG-4 standards: Systems, Audio, Video and SNHC. The standard will finally consist of Systems, Audio and Video parts, and the specifications produced by SNHC will be integrated in either Audio or Video.

The Systems layer supports demultiplexing of multiple bitstreams, buffer management, time identification, scene composition and terminal configuration.

MPEG-4 video provides technologies for efficient storage, transmission and manipulation of video data in multimedia environments. The key areas addressed are efficient representation, error resilience over broad range of media, coding of arbitrarily shaped video objects, alpha map coding.

MPEG-4 Audio standardizes the coding of natural audio at bitrates ranging from 2 Kbit/sec to 64 bits/sec, addressing different bitrate ranges with appropriate coding technologies.

Synthetic/Natural Hybrid Coding (SNHC) deals with coding of synthetic audio and visual data. SNHC is described in detail in following sections.

In this paper we analyze how MPEG-4 tools could be applied to meet the requirements of Networked Collaborative Virtual Environment (NCVE) systems. The following section presents the networking requirements of NCVEs, focusing on the types of network traffic encountered in NCVE systems. After that we attempt to match these requirements to tools that will be available in MPEG-4. Conclusions are presented at the end.

The analysis of NCVE systems is based on several existing systems [Barrus96, Carlsson93, Macedonia94, Ohya95, Singh95] and in particular on the Virtual Life Network (VLNET) system [Capin97, Pandzic97]. The analysis of MPEG-4 tools is based on various MPEG documents [MPEG-N1361, MPEG-N1365, MPEG-N1545] and related papers [Koenen97, Doenges97], and in particular on the MPEG-4 Requirements definition [MPEG-N1595].

Networking requirements of NCVE systems

Networking for Collaborative Virtual Environments can be thought of in a simplified manner as represented in figure 2.

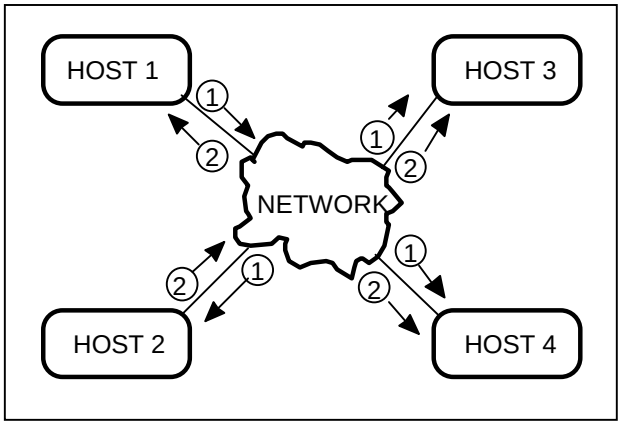

Figure 2: Simplified View of Networking for Collaborative Virtual Environments

All updates from any host are passed through the network (represented by the cloud in the figure) to all other hosts in form of bitstreams (bullets in the figure). The requirements on networking can now be divided in two categories: network topology (what is in the cloud?) and bitstream content (what is in the bullets?).

We treat these two issues separately in the following two subsections.

Network topologies for NCVE systems

The basic requirement for the network topology for an NCVE system is simple: when one host sends a message to the network, all other hosts should receive it. This is easy enough to do with a client-server or peer-to-peer approach if the number of participants is small. Unfortunately, if a system must support more then a few users, scalability becomes an important and potentially difficult issue. More sophisticated solutions have to be sought with multicasting, networks of multiple servers or hybrid approaches [Funkhouser96]. Usually, not all data needs to be transmitted to all participants, e.g. if two users are standing back to back or if they are in different (virtual) rooms, there is no need to communicate facial expressions between them since they don't see each other anyway. These "as-needed" data transfers can be achieved using an intelligent server to do data filtering or using dynamic multicast groups, to name two common examples.

Bitstream contents

Regardless of the network topology we analyze various data types that are transmitted through the network in NCVE systems. Figure 3 presents an

overview of all data types usually encountered. Current systems usually support only a subset of the data types presented here.

Download. The main need for download arises when a new user joins a NCVE session. At this moment the complete description of the Virtual Environment has to be downloaded to the new user. This includes 3D objects structured in a scene hierarchy, textures and possibly behaviors in form of scripts or programs. The new user also has to download the embodiment descriptions of all users and send his own to everyone. The embodiment might be a simple geometrical object, but in a more sophisticated system it should be a body and face description in a form that allows later animation of both body and face.

Downloads are not restricted to the session establishment phase, they can also occur anytime during the session if new objects are introduced in the scene - they have to be distributed.

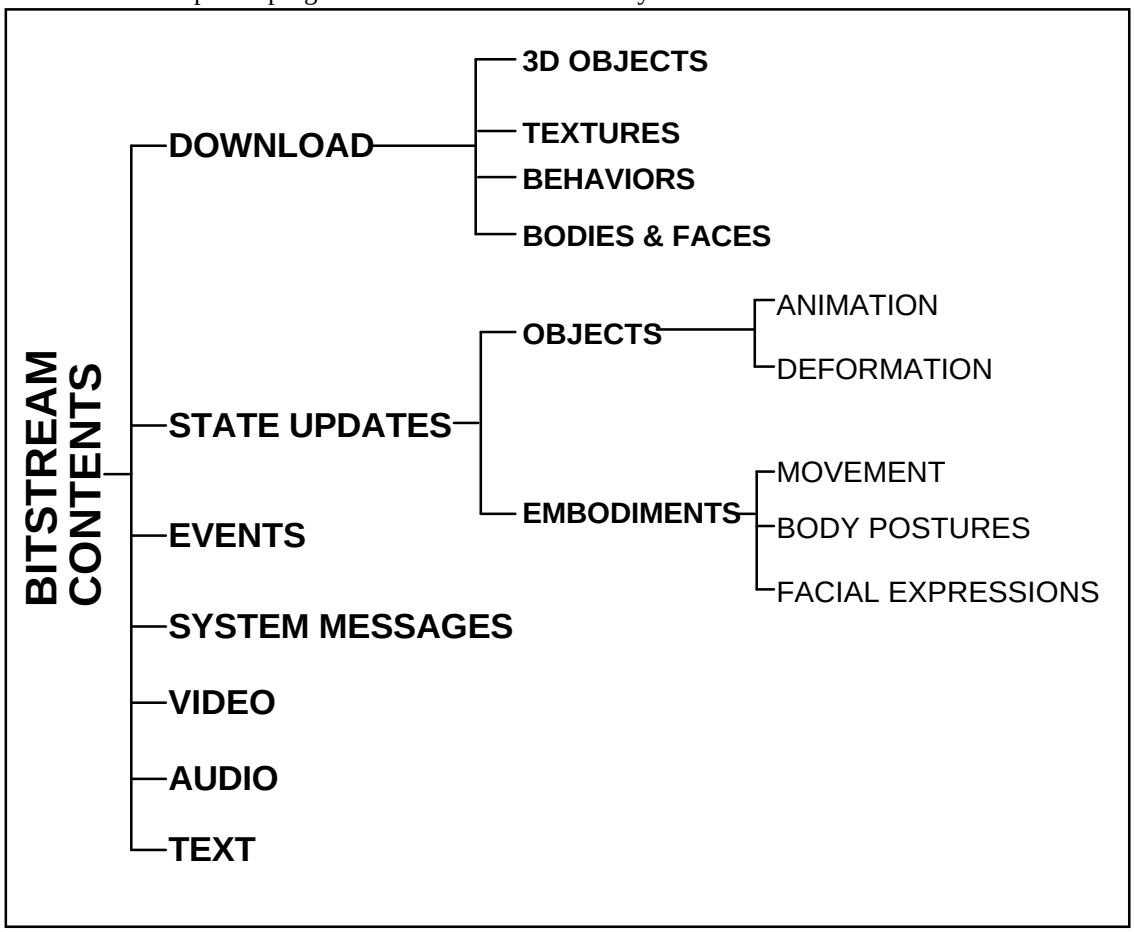

Figure 3: Bitstream contents breakdown for NCVE systems

State updates. All state changes for both the environment itself and the users' embodiments as special part of the environment have to be propagated through the network.

For the objects in the environment, this commonly involves the change of position/orientation, i.e. animation of rigid objects. In non-networked VEs objects are often deformed and not only animated rigidly. Free deformation of objects is not commonly supported in NCVE systems because of increased bandwidth needs - all displaced vertices have to be updated. It is however a very desirable feature to be included in NCVE systems.

For user embodiments, state updates also involve basic movement, and in the case of simple, rigid embodiments this is enough. For articulated, human-like embodiments means must be provided to communicate body postures and facial expressions to allow the simulation of natural body movements and actions.

Events. These are typically short messages about events happening in the environment. The basic difference from

state updates is that a state message makes all previous state messages for the same object obsolete (e.g. a new position of the object makes all previous positions obsolete) [Kessler96], which is not the case with event messages. An event is sent only once and can influence the environment state potentially for a long time. Therefore the security of event messages must be higher then for state updates.

System Messages. These messages are used typically during session establishment and log-off. Their security is essential because errors can cause serious malfunction of the system.

Video. Video can be streamed and texture-mapped on any object in the environment producing real-time video textures. This can be used in different ways for various applications. Examples include virtual video presentations, facial texture mapping for facial communication [Pandzic96] and mixing of real and virtual worlds for augmented reality applications or virtual studios.

The requirements on the video quality may vary from application to application.

Audio. The most common use of audio in NCVEs is for speech communication. However, synthetic 3D sound can also be an important part of a Virtual Environment.

Text. The most common use of text is for text-based chat between users. There are however more exciting ways to use it. Autonomous Virtual Humans with built-in AI decision-making algorithms can also be participants in NCVEs [Capin 97, Noser 96]. The best way to communicate orders/questions/dialogues to them is through text, and that is also the easiest way for them to respond. Speech recognition and Text-to-Speech (TTS) systems (possibly coupled with facial animation for lip sync) can be used to interface naturally with the users.

How MPEG-4 tools fit NCVE requirements

Based on the requirements laid out in the previous section and MPEG-related documentation [MPEG-N1361, MPEG-N1365, MPEG-N1545, MPEG-N1595, Koenen97, Doenges97] we study how MPEG-4 tools can be deployed to solve the problems of NCVE.

As for the network topologies, the solutions are out of scope of MPEG which mostly concentrates on bitstream contents. It is however worthwhile noting that an MPEG-4 system will be capable of receiving objects from up to 8 different sources, a fact to be considered when planning network topologies for particular applications [MPEG-N1595].

In the following subsections we study various components of bitstream contents with respect to MPEG-4.

Download

The MPEG-4 standard shall provide the means to download and store audio, video and synthetic objects [MPEG-N1595]. Furthermore, progressive transmission based on scalable coding techniques will be supported.

MPEG-4 will support a VRML-like 3D geometry hierarchy with all attributes, as well as behavior data. It will provide means for efficient compression of 3D meshes. Efficient still texture coding will be supported, as well as spatial- and quality- scalable coding to fit available bandwidth and rendering capabilities.

MPEG-4 will support body and face objects. Using Body Definition Parameters (BDPs) and Face Definition Parameters (FDPs) [MPEG-N1365, MPEG-N1545] it is possible to define body and face representation. BAPs and FAPs are scalable, offering a wide choice of tradeoffs between definition quality and bandwidth required. In case of absence of FDPs and/or BDPs generic bodies and faces can be used.

State updates

MPEG-4 supports scaling, rotation and translation of any video object (natural or synthetic, i.e. 3D objects included) about any axis in 2D or 3D space [MPEG-N1595]. Changes in audio objects localization are supported.

As for deformable 3D objects, at the time of writing this paper the need was recognized in MPEG to support efficient coding of object deformations, though this specification was not yet included in the documents.

MPEG-4 will support efficient coding of face and body animation. The parameters are defined to express body postures and facial expressions in an efficient manner and independently of a particular face/body model. These parameters are compressed to obtain very small bitrates (e.g. approx. 2 Kbit/sec for facial expressions).

Events and system messages

These messages are specific to a particular NCVE system and as such are not explicitly covered by MPEG-4.

Video

Video is a traditional part of MPEG and the video tools are mature and extensive. MPEG-4 Video will support all types of pixel-based video with high compression efficiency. Tools will be provided to achieve error resilient video streams over a variety of networks with possibly severe error conditions, including low-bitrate networks [MPEG-N1595]. Scalability in terms of content and spatial and temporal quality will be supported. Various delay modes, including low delay modes for real-time communication will be supported. MPEG-4 tools are optimized for the following bitrate ranges: < 64 Kbit/sec (low), 64 - 384 Kbit/sec (intermediate) and 384 Kbit/sec - 1.8 Mbit/sec (high). Various video formats will be supported.

Audio

MPEG-4 will support following types of audio content: high quality audio (> 15 KHz), intermediate quality audio (<15 KHz), wideband speech (50 Hz - 7 KHz), narrowband speech (50 Hz - 3.6 KHz) and intelligible speech (300 Hz - 3.4 KHz) [MPEG-N1595]. Tools will be provided to achieve error resilient audio streams, including support for low bitrate applications. In particular, speech coding compression will support intelligible speech at 2 Kbit/sec. A number of audio formats, as defined by sampling frequency, amplitude resolution, quantizer characteristics and number of channels will be supported.

Text

Simple text is not explicitly supported by MPEG-4. However, there will be tools for Text-to-Speech functionality [MPEG-N1595] which requires at least simple text, and possibly auxiliary information such as phoneme duration, amplitude of each phoneme etc. Capability to synchronize TTS output with facial animation system visualizing the pronunciation will be supported.

Integration

For NCVE systems it is not only important to support all the data types described in previous sections, but also to achieve an orderly integration of all data types with respect to relative priorities and synchronization.

MPEG-4 shall support dynamic multiplexing of all objects [MPEG-N1595]. Means will be provided to identify relative importance of parts of coded audio-visual information with at least 32 levels of priority. Synchronization between all objects is supported, with specified maximal differential delays (e.g. between two video objects or between an audio and a video object).

Conclusions

We have analyzed the networking requirements of Networked Collaborative Virtual Environments, and how MPEG-4 tools can be used to fulfill these requirements.

The building of network topology, session establishment/destruction and system-particular message passing are out of MPEG-4 scope and should be dealt with on another level. However, most of the data types that are important for NCVE systems will be very well supported by MPEG-4 tools (video, audio, 3D objects, textures, bodies, faces). On top of this, MPEG-4 will offer reliable multiplexing, mechanisms for establishing priorities among data, as well as synchronization. We believe that MPEG-4 tools should play an important role in building future Networked Collaborative Virtual Environment systems.

Acknowledgments

This research is partly financed by "Le Programme Prioritaire en Telecommunications de Fonds National Suisse de la Recherche Scientifique" and the ACTS project VIDAS.

References

[Barrus96] Barrus J. W., Waters R. C., Anderson D. B., "Locales and Beacons: Efficient and Precise Support For Large Multi-User Virtual Environments", Proceedings of IEEE VRAIS, 1996.

[Capin97] Capin T.K., Pandzic I.S., Noser H., Magnenat Thalmann N., Thalmann D., "Virtual Human Representation and Communication in VLNET Networked Virtual Environments", *IEEE Computer Graphics and Applications, Special Issue on Multimedia Highways*, March-April 1997.

[Carlsson93] Carlsson C., Hagsand O., "DIVE - a Multi-User Virtual Reality System", *Proceedings of IEEE VRAIS '93*, Seattle, Washington, 1993.

[Doenges97] Doenges, P.K., Capin, T.K., Lavagetto, F., Ostermann, J., Pandzic, I.S., Petajan, E.D., "MPEG-4: Audio/Video & Synthetic Graphics/Audio for Mixed Media", *Image Communication Journal, Special Issue on MPEG-4*, Vol.9, No. 4, May 1997. (to appear)

[Funkhouser96] Funkhouser T.A., "Network Topologies for Scalable Multi-User Virtual Environments, *Proceedings of IEEE VRAIS '96*, 1996

[Kessler96] Kessler G.D., Hodges E.F., "A Networked Communication Protocol for Distributed Virtual Environment Systems", *Proceedings of IEEE VRAIS '96*, 1996

[Koenen97] Koenen R., Pereira F., Chiariglione L., "MPEG-4: Context and Objectives", *Image Communication Journal, Special Issue on MPEG-4*, Vol. 9, No. 4, May 1997 (to appear)

[Macedonia94] Macedonia M.R., Zyda M.J., Pratt D.R., Barham P.T., Zestwitz, "NPSNET: A Network Software Architecture for Large-Scale Virtual Environments", *Presence: Teleoperators and Virtual Environments*, Vol. 3, No. 4, 1994.

[Noser96] Noser H., Pandzic I.S., Capin T.K., Magnenat Thalmann N., Thalmann D., "Playing Games through the Virtual Life Network", *Proceedings of Artificial Life V*, Nara, Japan, 1996.

[MPEG-N1316] Doenges P., Jordan F., Pandzic I.S., "SNHC Application Objectives and Requirements", ISO/IEC JTC1/SC29/WG11 N1316, MPEG96/July 1996.

[MPEG-N1365] Petajan E., Pandzic I.S., Capin T.K., Ho P.H., Pockaj R., Tao H., Chen H., Shen J., Chaut P.E., Osterman J., "Face and body definition and animation parameters", ISO/IEC JTC1/SC29/WG11 N1365, MPEG96/October 1996.

[MPEG-N1545] "SNHC Verification Model Version 3.0", SNHC, ISO/IEC JTC1/SC29/WG11 N1545, MPEG97/February 1997.

[MPEG-N1595] "MPEG-4 Requirements version 2", ISO/IEC JTC1/SC29/WG11 N1545, MPEG97/February 1997.

[Ohya95] Ohya J., Kitamura Y., Kishino F., Terashima N., "Virtual Space Teleconferencing: Real-Time Reproduction of 3D Human Images", *Journal of Visual Communication and Image Representation*, Vol. 6, No. 1, pp. 1-25, 1995.

[Pandzic96] Pandzic I.S., Capin T.K., Magnenat Thalmann N., Thalmann D., "Towards Natural Communication in Networked Collaborative Virtual Environments", *Proceedings FIVE 96* , Pisa, Italy, 1996.

[Pandzic97] Pandzic I.S., Capin T.K., Lee E., Magnenat Thalmann N., Thalmann D., "A flexible architecture for Virtual Humans in Networked Collaborative Virtual Environments", *Proceedings Eurographics 97* (to appear)

[Singh95] Singh G., Serra L., Png W., Wong A., Ng H., "BrickNet: Sharing Object Behaviors on the Net", *Proceedings of IEEE VRAIS '95*, 1995.