MPEG-4: ONE MULTIMEDIA STANDARD TO UNITE ALL

K. Karpouzis, A. Raouzaiou, P. Tzouveli, S. Ioannou and S. Kollias

Image, Video and Multimedia Systems Laboratory
National Technical University of Athens
157 80 Zographou, Athens, Greece

ABSTRACT

A wide variety of multimedia coding and representation standards have emerged during the past years, in the quest to establish a common denominator for both the research community and the related industry to incorporate results and algorithms in commercial products. MPEG-4 aims to improve where past standards proved ineffective or isolated and integrate different modalities to supply different forms of material, suitable for diverse applications and devices. In this paper, we describe the general concepts that MPEG-4 builds upon and how they are related to multi-user interactive environments and gaming.

1. INTRODUCTION

MPEG-4 is an ISO/IEC standard which was developed by MPEG (Moving Picture Experts Group), the committee also responsible for MPEG-1 and MPEG-2, the standards that made interactive video on CD-ROM and DVD possible. MPEG-4 was finalized in late 1998, while the first extensions (MPEG-4 Version 2) acquired the formal International Standard Status early in 2000. While this is a fully functional standard, several extensions were added since and work on specific notions work is still in progress.

MPEG-4 builds on the proven success of the previous MPEG standards in the fields of digital television, interactive graphics applications and interactive multimedia. In home video applications, MPEG-1 video and audio (i.e. MP3) are still the most widely accepted formats, providing one of the best compression/quality ratios. MPEG-2 is presently an established A/V standard for entertainment video applications, providing better compression ratio on the cost of computational complexity, both on the encoder and the decoder sides. It is utilized in digital television and DVDs, with millions of MPEG-2 decoders already deployed in PCs, set-top boxes and DVD players; besides this, most TV programs broadcast today are coded in MPEG-2.

MPEG-7 and MPEG-21 are additional toolsets which extend the functionality of MPEG-4 and interface with MPEG-4 to handle content management requirements. The integration of MPEG-4 with MPEG-7 and MPEG-21 is accomplished by multiplexing MPEG-7 metadata or MPEG-21 specifications into discrete content streams and representations. MPEG-7 is a recently finalized standard used to describe multimedia content, for indexing, program selection and intelligent content description. The standard comprises of syntax and semantics of multimedia descriptors and descriptor schemes and caters for the management; search and retrieval of local or on-line content; MPEG-21 is an emerging standard aiming to describe how different objects interact to build an infrastructure for the delivery and utilization of multimedia content.

2. MPEG-4: THE WORLD OF OBJECTS

In the case of an MPEG-2 program, content is created from various resources such as video, graphics, text, which are “composited” into a plane of pixels. The encoding process results in a “flat” medium, that does not retain any information on the initial inputs. This enables MPEG-2 content to be easily decoded by the client, but makes it static, with no editing possible at the object level, thus making content reuse almost unfeasible.

On the other hand, the MPEG-4 approach is inherently dynamic, since different objects can be encoded and transmitted separately to the decoder in respective elementary streams (ESs). In order to compose the final output, the individual multimedia objects are described in BIFS (Binary Format for Scenes) [1]. The BIFS language caters for timing and spatial placement information, as well as deterministic and event-driven object behavior. Different coding schemes are supported for each medium, thus retaining their individual properties and making content scalability more efficient. As a result, while they are not standardized, MPEG-4 does cater for efficient coders for audio, speech, video and even synthetic content such as animated faces and bodies.

Objects in MPEG-4 audiovisual scenes are organized in a hierarchical fashion (see Figure 1), where the leaves of the hierarchy consist of primitive multimedia objects, e.g. still images or video objects. MPEG-4 standardizes a number of such primitive media objects, capable of representing both natural and synthetic content types, which can be either 2- or 3-dimensional [2]. In addition to that, MPEG-4 defines the coded representation of objects such as text and graphics, and even talking synthetic heads and bodies, driven by text and producing synthesized speech [3], [4].

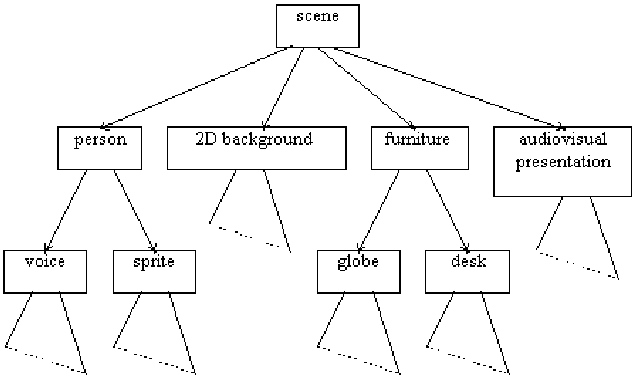


Figure 1: Logical structure of a scene [5]

3. THE ANIMATION FRAMEWORK EXTENSION (AFX)

The Animation Framework extension framework defines a collection of interoperable tools for interactive animated contents. AFX (pronounced "effects") enhances existing MPEG-4 tools by offering higher-level descriptions of animation (e.g. inverse kinematics, AU-based [6] facial animation [2]), advanced rendering (e.g. multi-texturing, procedural texturing), and compact and scaleable representations (e.g. subdivision surfaces).

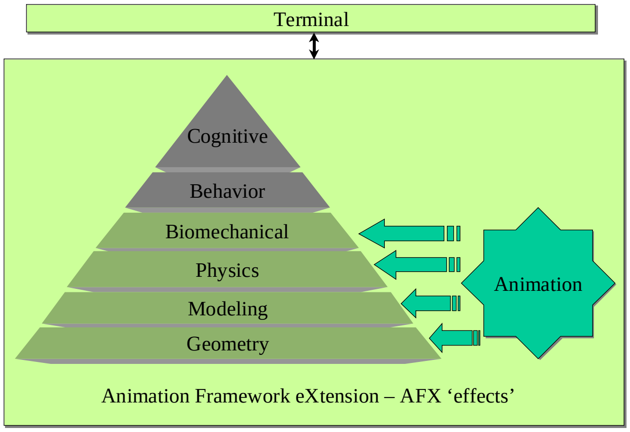


Figure 2: Models stack in computer games [7].

AFX treats objects and scene hierarchies from a bottom-up point of view, i.e. lower-level objects are combined to define the higher hierarchical categories. In this framework, six categories of tools are defined (see Figure 2):

- Geometric models defining object appearance
- Modeling tools that provide easier data control
- Physical models that add engineering notions to the animation
- Biomechanical models that perform semantic grouping and coordination
- Behavior models handling communicative movements, and
- Cognitive models that add learning capabilities.

Starting from the bottom part, geometric models describe the form and appearance of a synthetic object. Most characters in games can be efficiently controlled at this low-level; however, higher-level models for motion control generally make character animation simpler, since the form of motion is known or predictable by the designer. Geometric models are extended with linear and non-linear deformations to illustrate non-rigid transformation, such as muscle movement. Animation can then be designed by changing the deformation parameters independently of the geometric models.

Physical models are one step higher by recreating aspects of the environment, such as inertia or gravity. The use of physical models allows for the automation of realistic animation despite being computationally complex. Applications such as collision detection and compensation, deformable bodies, and rigid articulated bodies use these models intensively, usually in accordance with biomechanical models.

On the top of this pyramid lie behavioral and cognitive models. Virtual characters may expose a reactive expression when they possess no memory of previous situations and merely react to stimuli and events. Finite-States Machines are often used to encode behaviors based on multiple states, while goal-directed behaviors can be used to define a cognitive character's goals. One step further, the ability to learn defines a cognitive model, able to adapt the character's behavior.

In this context, each model utilizes all underlying concepts. For example, an autonomous agent (behavioral model) may respond to stimuli from the environment it is in and may decide to adapt its way of walking (biomechanical), which in turn can modify physics equation (physical properties) or have influence on some underlying deformable models (modeling - geometry). If cognition is supported, the agent may also learn from the stimuli and adapt or modify its behavioral models.

4. FACIAL ANIMATION

Faces make good interface elements since they are accepted as the most expressive means for communicating, as well as recognizing emotions. Thus, a lifelike human face can enhance interactive applications by providing straightforward feedback to and from the users and stimulating emotional responses from them. Besides this, the gaming and entertainment industries can benefit from employing believable, expressive characters since such features significantly enhance the atmosphere of a virtual world and communicate messages far more vividly than any textual or speech information [8].

Talking head applications are be grouped into two types:

§ Assistants, personalized secretaries or salespersons, guide for virtual tours of museums, aids to non-hearing persons, etc. Such characters may react to a specific demand: users want intuitive interfaces, personalized and humanized for the more and more numerous services offered to them.

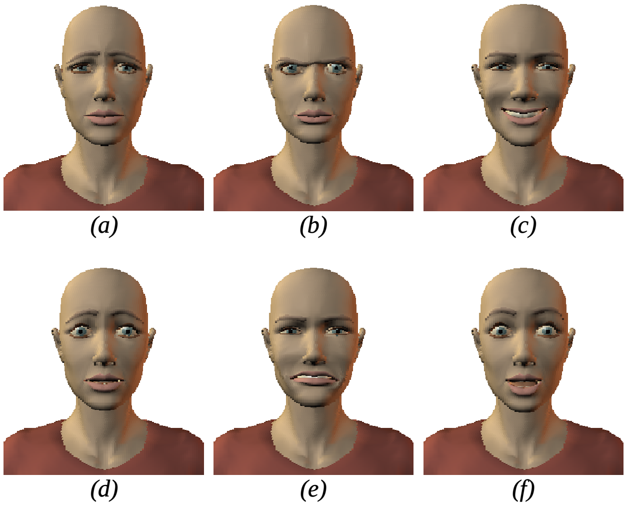
§ Clones/avatars in communication: tele-conferencing, games, virtual meetings, chat rooms, etc., are all possible fields of application.

4.1. Facial Expression Synthesis-Experimental Results

Animated profiles were created using the 3D model of the Curious Labs Poser software. This model has separate parts for each moving face part. The Poser model interacts with the controls in poser and has joints that move realistically, as in a real person. This allows us to manipulate the figure based on those parameters. To achieve this, a mapping from MPEG-4 FAPs to Poser parameters is necessary. We did this mapping mainly experimentally. The relationship between FAPs and Poser parameters is more or less straightforward.

The first set of experiments shows synthesized archetypal expressions (see Figure 3) created by using the Poser software package. The 3D nature of the face model renders the underlying emotions in a natural way.

The second set of experiments (see Figure 4) shows particular examples of non-archetypal expressions [9].


(a) *(b)* *(c)*

(d) *(e)* *(f)*

Figure 3: Synthesized archetypal expressions created using the 3D model of the POSER software package: (a) sadness, (b) anger, (c) joy, (d) fear, (e) disgust and (f) surprise

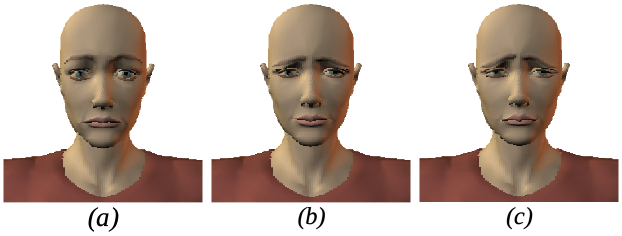

(a) *(b)* *(c)*

Figure 4: Poser face model: Animated profiles for emotion terms (a) **afraid**, (b) **guilty** and (c) **sad**

5. BODY ANIMATION

In the framework of MPEG-4 standard, parameters have been specified not only for Face but also for Body Animation (FBA) by defining specific Face and Body nodes in the scene graph.

In general, an MPEG body is a collection of nodes. The Body Definition Parameter (BDP) set provides information about body surface, body dimensions and texture, while Body Animation Parameters (BAPs) transform the posture of the body. BAPs describe the topology of the human skeleton, taking into consideration joints limitations and independent degrees of freedom in the skeleton model of the different body parts. MPEG-4 provides a high-level language for the description of a motion, since every BAP is described by its name.

5.1. BBA (Bone Based Animation)

The MPEG-4 BBA offers a standardized interchange format extending the MPEG-4 FBA [10]. In BBA the skeleton is a hierarchical structure made of bones. In this hierarchy every bone has one parent and can have as children other bones, muscles or 3D objects. For the movement of every bone we have to define the influence of this movement to the skin of our model, the movement of its children and the related inverse kinematics.

In the BBA stream, rotation is represented as Euler angles. It contains all the animation frames or data at the temporal key frames, where decoder will compute the intermediate frames by temporal interpolation (*linear* interpolation for *translation* and *scale* and *spherical linear quaternion* interpolation for *rotation* and *scaleOrientation*).

Bone based representations benefit both the synthesis and the analysis of hand gestures and other body movements, since they are closer to human conceptions of body posture than mere motion information.

6. ONLINE GAMING

The MPEG Group identified the unique requirements and characteristics of multi-user, online games and initiated the On-Line Gaming (OLGA) ad-hoc group (AhG). Since

MPEG-4 aims to be a transparent protocol, thorough standardization is in order, given the fact that a wide range of terminals for games and on-line games are available (consoles, PCs, set-top boxes, PDAs, mobile phones, etc.).

The OLGA AhG mission statement is to handle aspects related to client-server infrastructure and interoperability, by providing a common scene-graph for a variety of terminals which is transparent to the network profile. Inherent concepts of MPEG-4, such as network scalability and multimedia integration, are indispensable, especially in the case of different modules working together, e.g. in the case of a network of different consoles or software modules from different vendors.

In the case of multi-user worlds, the standard specifies the low level protocol and infrastructure for sharing sub trees of scene graphs between multiple users. It also addresses how individual nodes are shared and how shared sub-trees are protected against unwanted modifications from external sources. The architecture for handling multiple users in an MPEG-4 environment is based largely on the *Drone/Pilot* mechanism proposed in the Living Worlds specification [11]. In this context, Drones are local instances of nodes that contact their corresponding Pilots (master copies of nodes) when changes are to be distributed.

Another interesting field of on-line gaming is the support in mobile devices and especially telephones. The mobile telephone is a multi-faceted gadget with extremely high penetration rates and ever increasing uses: a personal, secured and reassuring communicating object for its user that is constantly enriched with applications and empowered with multimedia capacities. The calculation power and the high-end quality interfaces (color screen, mini-joystick, vibration, polyphonic sound, etc.) of modern telephones, make it possible to play games of similar quality as those found on hand-held consoles. To see 3D games from the late 90s, such as Doom, running on mid-range mobile phones proves that headsets are catching up with hand-held consoles. While there are still a few years left before telephones catch up with PCs, 3D games of unbelievable quality are now being released. With the wide adoption of Bluetooth technology, it will soon be possible to have a Bluetooth joystick to control games, while supporting multi-user communication over high-speed wireless networks, such as GPRS and UMTS.

7. COMMERCIAL SUPPORT

MPEG-4 creates business opportunities by reducing costs related to deploying the standard. This is invaluable in today's economic environment that seeks to introduce flexible, scalable and multi-purpose solutions, something that none of the proprietary technology offerings will ever guarantee.

The MPEG-4 Industry Forum (M4IF) is a not-for-profit organization with the goal to "further the adoption of the MPEG-4 Standard, by establishing MPEG-4 as an accepted and widely used standard among application developers, service providers, content creators and end users" [12]. The forum represents more than 100 companies from diverse industries evenly distributed across North America, Europe and Asia, addressing MPEG-4 adoption issues that go beyond the charter of ISO/IEC MPEG. Order to ensure wide adoption, M4IF picks up where MPEG stops, by providing marketing work and handling licensing issues. In addition, M4IF has an advanced program of cross-vendor product interoperability testing.

8. REFERENCES

[1] M. Bourges-Sévenier, A. Walsh. MPEG-4 Jumpstart, Prentice Hall, December 2001.

[2] M. Tekalp, J. Ostermann, "Face and 2-D mesh animation in MPEG-4", *Image Communication Journal,* vol.15, Nos. 4-5, January 2000, pp.387-421.

[3] G. Breton, C. Bouville, D. Pelé, FaceEngine: A 3D Facial Animation Engine for Real Time Applications

[4] I. Pandzic, "Talking Virtual Characters for the Internet", in Proc. of ConTel 2001, Zagreb, Croatia

[5] R. Koenen, Overview of the MPEG-4 Standard, ISO/IEC JTC1/SC29/WG11, March 2002.

[6] P. Ekman, W. Friesen, Facial Action Coding System, Consulting Psychologists Press, Palo Alto, 1978.

[7] M. Abrash, "Michael Abrash's graphics programming black book, special edition. The Coriolis Group, Inc., 1997.

[8] J. Bates, The role of emotion in believable agents, Communications of the ACM, 37(7): 122-125, 1992.

[9] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie and E. Douglas-Cowie, "Emotion Recognition and Synthesis based on MPEG-4 FAPs," in *MPEG-4 Facial Animation,* Igor Pandzic, R. Forchheimer (eds), John Wiley & Sons, UK, 2002.

[10] M. Preda, F. Prêteux, Advanced animation framework for virtual characters within the MPEG-4 standard, in *Proc. of the International Conference on Image Processing,* Rochester, NY, September 2002.

[11] Living Worlds Proposal Draft 2,
http://www.vrml.org/WorkingGroups/living-worlds/

[12] MPEG-4 Industry Forum, http://www.m4if.org/