COMPARISON OF MPEG-4 BIFS AND SOME OTHER MULTIMEDIA DESCRIPTION LANGUAGES

C. Concolato, J.-C. Dufourd

Ecole Nationale Supérieure des Télécommunications, ENST
46, Rue Barrault
75013 PARIS, France

ABSTRACT

A multimedia presentation can be composed of several media elements whose types can be of natural video, natural or synthetic audio, 2D and/or 3D vector graphics or text. The spatial and temporal organisation of the different elements composing the presentation are specified using a Multimedia Description Format or Language. MPEG-4 [1] provides a binary Multimedia Description Format called BIFS (BInary Format for Scene). The purpose of this paper is to describe some technologies, like the W3C recommendations and other proprietary solutions, which could compete with the Multimedia Description Format provided by MPEG-4.

1. INTRODUCTION

A Multimedia Description is defined as a description of a multimedia presentation. It describes the spatial layout of the different media elements (video, audio, graphics, text …) as well as the temporal order in which these elements will play during the presentation. A Multimedia Description may contain some direct or indirect references to the media elements composing the presentation.

In this paper, we present different approaches of Multimedia Description. In the first part, we describe the ISO technologies for Multimedia Descriptions which are VRML and BIFS. In a second part, we present some recommendations proposed by W3C, namely SMIL and SVG. Finally, we describe some proprietary tools developed by Macromedia and Apple and that are related to the creation or the storage of multimedia presentations.

2. ISO Standards: VRML and MPEG-4 BIFS

ISO, International Standardisation Organisation, specifies two Multimedia Description formats that will be presented in this section: VRML and BIFS.

2.1 From VRML to X3D

The Virtual Reality Modelling Language (VRML) ([1],[2]) is a textual format for describing interactive 3D objects and worlds. It was designed in 1997 and its purposes were to be used on the Internet, intranets, and local client systems. VRML was also intended to be a universal interchange format for integrated 3D graphics and multimedia. To achieve authorability, VRML was designed as a textual format and with the ability to compose files together through inclusion, to relate files together through hyperlinking and to easily reuse complex contents through the use of author-defined macros (PROTO, EXTERNPROTO). Therefore, VRML is widely used in a variety of application areas such as engineering and scientific visualisation, multimedia presentations, entertainment and educational titles, web pages, and shared virtual worlds.

VRML is capable of representing static and animated 3D and multimedia objects with hyperlinks to other media such as text, sounds, movies, and images. VRML browsers (SoNG, Cosmo), as well as authoring tools for the creation of VRML files (3D Studio Max, Amapi 3D), are widely available for many different platforms. Each VRML textual file represents a 3D time-based space that contains graphic and aural objects that can be dynamically modified through mechanisms like routes and interpolators. Simple interactivity is possible with pointing-devices like mice and joysticks.

The first version of VRML has encountered a widespread success among the 3D authors community. A second version of this standard and a new standard, called X3D (eXtensible 3D) are being worked on by the Web3D Consortium. The main added features of these works are the use of XML as textual format for representing 3D worlds and the addition of 22 new nodes among them:

- **Contour2D** and **Polyline2D**, which allow for the use of some 2D primitives in a 3D world;
- 10 **Geo** nodes which bring in the Geospatial coordinate;
- **KeySensor**, **StringSensor** which enable better interactivity by the use of keyboard devices or remote control;
- And 6 **Nurbs** nodes which allow the use of nurbs primitive in a 3D world.

Despite its widespread success in the 3D community, VRML has not reached the expected success in the World Wide Web community. There are three main reasons for this:

- First, VRML *does not support dynamic composition*. Indeed, a scene is enterily known at the loading of the VRML file except for user interaction. No server-side interaction is handled.
- Second, VRML worlds often needs to contain a lot of primitives to be smooth and realistic. This is not a problem in itself. The problem lies in the fact that VRML was intended for the internet but *it does not support any streaming facility*. The only way to reduce the playback delay is to zip the VRML file prior to sending it over the internet.

- Third, VRML *does not offer any 2D primitive* nor the possibility to mix 2D and 3D worlds.
- Fourth, VRML *cannot insure fine synchronization* on a per frame basis.
- Finally, VRML *does not provide hooks to enable encryption, watermarking and in general digital rights management* of 3D content.

2.2 BIFS: ISO/IEC 14496-1:2001

MPEG-4 BIFS, BInary Format for Scene, is specified in chapter 8 of ISO/IEC 14996-1 ([3],[4]). BIFS is a binary Multimedia Description format. The development of the MPEG-4 scene description format started when the Virtual Reality Modelling Language (VRML) was getting momentum in the 3D community. VRML was brought to the MPEG Systems group as a candidate for the core of the MPEG-4 scene description tool and became the kernel of MPEG-4 BIFS. Being a 3D textual language for download-and-play, VRML did not address many of the MPEG-4 scene description requirements and thus there was need for many improvements and additions which are described hereafter.

Like VRML, BIFS describes interactive 3D objects and worlds but it can as well describe 2D or synthetic aural objects and worlds. A mechanism of layer allows for mixing 2D and 3D environments. Moreover, BIFS adds a crucial feature compared to VRML: the mechanism of update. Indeed, a BIFS scene can be updated, i.e. new objects can be added, deleted or replaced. This mechanism transforms a static binary scene into a stream that can be sent over a network and synchronized with other streams (video, audio and meta-data). This update mechanism is composed of two main tools: the BIFS-Command tool whose goal ranges from the modification of the 2D/3D position of an object to the addition, deletion and/or replacement of the entire or part of the scene. The second tool, BIFS-Anim, allows for continuously updating values of attributes of graphical objects such as their positions, sizes, scales, in order to achieve fast animation. This feature is very important because it shows how BIFS is stream-oriented. BIFS is not appropriate for every kind of transport because it is very sensitive to transmission errors, but it is the first standardized binary format for scene descriptions.

MPEG-4 BIFS works in conjunction with MPEG-4 Object Descriptor Framework (ODF). The ODF is the tool that enables uniform handling of various media types. Hence, a JPEG picture and an MPEG-4 video stream are handled in the same way in the scene by using the OD identifier either of the picture or of the video. The OD provides information for the synchronization, the encryption, the description of the media referred to in this object. This tool shares some functionalities with the content control SMIL module, namely the ability to switch from one media to another.

MPEG-4 has a textual format for representing BIFS streams called XMT, eXtended MPEG-4 Textual format. XMT is an XML language and therefore offers all the possibility that XML offers, namely human readability, content transformation through XSL (XML Style Sheet) and XSLT (XSL Transformation). XMT is specified by MPEG with two levels of abstraction. The first language, called XMT-A, is the perfect mapping of the binary syntax of BIFS with an XML language. The XMT-A language is close to the X3D language. XMT-A is a low level description language that allows for authoring a wide range of scenes. The second language specified by MPEG is called XMT-O. It is a high level description language which shares some common part with SMIL. It allows for rapidly creating scenes with a lower level of complexity.

Contrary to other standards, MPEG-4 BIFS is a standard that allows to represent a scene containing any kind of media. It may sometimes be heavy or complex to take advantage of all the tools that MPEG-4 BIFS offers. Therefore, MPEG-4 BIFS follows the MPEG approach of profiles. A profile is a set of tools which are grouped together to target a set of applications that require the same resources. Hence, companies will only implement all the tools included in a profile.

3. W3C Recommendations: SMIL and SVG

The World Wide Web Consortium (W3C), [5], develops interoperable technologies (specifications, guidelines, software, and tools) for the Web. It has namely specified (X)HTML, XML, PNG. The approach chosen by W3C to specify a Multimedia Description format is quite different from the approach of MPEG. While MPEG specifies the 2D and 3D primitives as well as the timing and the animation information in the same part of the standard (BIFS), W3C splits the recommendations in several pieces taking advantage of the possibility offered by XML and the concepts of modularization and profiling. Among all the recommendations approved or being approved by the consortium three of them are relevant for this study:

- SMIL, described in the first subsection,
- SVG, described in the next subsection,
- And XHTML, which is HTML expressed with the strict syntax of XML. It will not be further described here.

In particular, SMIL 2.0 components can be used for integrating timing into XHTML (XHTML+SMIL profile) and SVG documents.

3.1 SMIL

SMIL, Synchronized Multimedia Integration Language, [6], defines an XML-based language that authors can use to write interactive multimedia presentations. The specification of SMIL is divided into ten functional areas: animation, content control, layout, linking, media objects, meta-information, structure, timing and synchronization, time manipulations and transition effects. Each functional area is itself divided into modules. A module is a set of elements, attributes and values that form an atomic set of tools to achieve a certain functionality.

Using SMIL 2.0, an author can describe the temporal behavior of a multimedia presentation in terms of synchronization of the different media. The author can also associate hyperlinks with media objects and describe in simple cases the layout of the presentation on a screen. The author is able to create a unique presentation which will adapt to the settings of the user. As said

previously, SMIL 2.0 does not allow to describe the media objects but rather to describe the organization of the different objects. These media objects can be of type audio, video, still pictures, still text, text stream and animations.

SMIL benefits from all the technologies designed for XML technologies (scripting, style sheets, linking, compression) and is gaining a lot of success in the Web community because of its integration in Microsoft Internet Explorer and RealNetworks RealPlayer. Still, this format is not stream-oriented because no update mechanism exists.

Since SMIL does not define how the different types of media fit in its architecture, SMIL is not a complete Multimedia Description Format. For example, if one wants to have formatted text using SMIL, the SMIL+XHTML specification is needed. If one wants to have 2D graphics, a SMIL+SVG specification is needed. If one wants to have 3D content, one would need a specification that tells you how to use a 3D XML language like X3D in conjunction with SMIL. This need for several specifications at a time could lead to some interoperability problems. It could be the case if one wants rich content that mixes 2D/3D graphics, text and video. Furthermore, SMIL performs synchronization of media, say a video track and an audio track, by referencing external players. No decoding model is specified by SMIL. So, it does not provide a fine synchronization mechanism. Finally, it does not provide any means of compression nor of encryption of the scene description.

3.2 SVG

SVG, Scalable Vector Graphics [7], defines an XML-based language for describing two-dimensional vector and mixed vector/raster graphics in XML. SVG allows for three types of graphic objects: vector graphic shapes (e.g. paths consisting of straight lines and curves (bézier or elliptical arcs), images and text. In addition to that, graphical objects can be grouped, styled, transformed and composited into previously rendered objects. The feature set includes nested transformations, clipping paths, alpha masks, filter effects and template objects. SVG drawings can be interactive (simple interaction based on pointing devices) and dynamic (using deterministic animations). Animations can be defined and triggered either declaratively (i.e. by embedding SVG animation elements in SVG content) or via scripting. In both cases, the animation is entirely contained in the SVG document and known at the loading of this latter.

Like SMIL, SVG benefits from a long list of specifications and standards efforts. Among them we can name CSS, XSLT, SMIL, XHTML and so on. But, like SMIL, SVG lacks the mechanism of update and the adaptation to streaming. In particular, a W3C Proposed Recommendation annex explains how to use a gzip compression to obtain small description files. But the fact that zip is not adapted to streaming is a well-known result. In its latest development, SVG is specifying some profiles for Mobile terminals.

SVG allows to do scripting. It requires to use the DOM and another script language, which increase dramatically the memory footprint and the complexity of the implementation.

The main drawback of SVG like SMIL is that none of them are oriented towards streaming. Indeed, it is expected that vector graphics and scene description become more important in Rich Media presentation. Hence, the initial delay to get the whole structure or graphics will become to important and therefore viewing such scenes will require streaming. It is for example unlikely that a user will wait for 5 minutes to watch a cartoon that will last as long, while it could be streamed and viewed in the same time.

Due to the involvement of big companies leader in the market of vector graphics on the Internet, SVG has some very intricate features. It allows for specifying decorated text, performing nice effects through the use of filtering. As presented in the section on SMIL, SVG can be used together with SMIL and therefore benefit from the SMIL animation module to have complex animation. It is also possible to use SVG with SMIL and XHTML. This would allow for having mixed graphics and formatted text. But, this would require a quite complex XML browser architecture.

4. Proprietary Solutions

Of course, lots of proprietary Multimedia Description Formats exist. Each of them uses its own way to describe a multimedia presentation. Hence, there is little interoperability between different solutions. Nevertheless some of them need to be investigated because they gather a great amount of users of their authoring tools and because they can adapt more easily to the users' needs than standards. In this part, we will particularly look at some technologies from Macromedia, Inc. and Apple Computer Inc..

4.1 Macromedia Technologies

Macromedia Inc., [8], is a well-known software company which develops about 20 products for the Web. These products range from a text editor to full studios for creating multimedia presentations. The tools that interest us are Flash and Director Shockwave Studio.

4.1.1 Flash

Flash is a piece of software which helps people creating 2D vector-based graphics for the Web. Flash is currently the leading product for designing 2D graphics, animations and recently 3D graphics. Flash uses a proprietary binary format for publishing called Swiff (.swf) which is very efficient in terms of compression. Though proprietary, the specification of Swiff is public and therefore a lot of tools are available for editing or conversion from other graphical format to this one and vice and versa. A binary format for editing is also used (.fla) but this latter format is not public.

Swiff is a very popular format because it is close to how graphics designers work. It uses a display list which is quite similar to an exposure sheet used by cartoon designers and allows for defining and reusing of primitives. The graphical

primitives used in Flash are approximately the same as the ones used in SVG and close to the ones used in MPEG-4 BIFS. The differences between the three formats lay in how these primitives can be organised, reused or modified. The Swiff binary format uses a concept very close to QuickTime atoms, with tag, length and payload, which makes it easy to skip content you do not understand. The main graphic primitive in Flash is the path. The path is really a set of segments from any of a dozen of types, from nothing to line to spline or bezier. Most of the other graphic primitives are usually mapped onto paths: rectangles, ellipses, circles, curves, even text.

Even if Flash encounters a real success, some criticisms could be made. First, the need to maintain two versions of the same content: one for editing and one for publishing is a real problem. Second, when the amount of content increases, it is difficult for a content provider to handle this set of binary files. Specific applications need to be developed to perform some filtering or some adaptation of the content.

4.1.2 Director Shockwave Studio

Director Shockwave Studio is to Flash what SMIL is to SVG. Director is a piece of software that allows users to create multimedia presentation embedding audio, video, text, Flash content (2D and more recently 3D content). It is the reference tool to create multimedia presentations for the Web or for offline presentations. Director's file format for editing, respectively for publishing, is a binary format called a Director Movie (.dir), respectively a Director Shockwave Movie (.dcr). A third file format, called a projector (.exe), can be used to play the presentation in its own application.

Director uses the metaphor of *cast* (in the theatrical sense of a set of actors) to integrate the media in a presentation. Then, the user can specify the spatial organization on the main stage and animate the different objects in time thanks to the score. The use of a scripting language, called Lingo, to attach elaborated behaviors to the different media is at the core of Director. Among other possibilities, Lingo allows the user to handle interactivity (mouse and keyboard), to have some simple synchronization between different medias (key points) or to animate some 3D worlds.

Since Director is really close to Flash, the same criticisms for Flash apply to Director concerning the handling of the content. It is also interesting to point out that Director was developed at first with a 2D stage and that now it is possible to add 3D content to this 2D stage but it seems rather difficult to do the opposite operation, i.e. to have a 3D scene within which you have complex 2D content (video, 2D animations).

4.2 Apple's QuickTime

Apple Computer Inc., [9], is one of leading company in multimedia thanks to its player: QuickTime. QuickTime allows to play all sorts of video, flash content, text stream, VR movies. A special kind of stream allows content creators to animate the presentation using scripting. This stream is stored in the Sprite Track.

5. SUMMARY

The table on the next page summarizes some key and missing features of the previously described technologies.

6. REFERENCES

[1] ISO/IEC JTC1 14772-1:1997, "*Information technology -- Computer graphics and image processing -- The Virtual Reality Modeling Language (VRML) -- Part 1: Functional specification and UTF-8 encoding*."

[2] Web3D Consortium, http://www.web3d.org

[3] ISO/IEC JTC1 14496-1:2001, "*Information technology -- Coding of audio-visual objects -- Part 1: Systems*".

[4] MPEG official web site, http://www.tilab.com/mpeg

[5] World Wide Web Consortium, http://www.w3.org

[6] World Wide Web Consortium, "*Synchronized Media Integration Language (SMIL 2.0)*", http://www.w3.org/TR/smil20/

[7] World Wide Web Consortium Recommandation, "*Scalable Vector Graphics (SVG) 1.0*", http://www.w3.org/TR/SVG/

[8] Macromedia, Inc., http://www.macromedia.com

[9] Apple Computer, Inc., http://www.apple.com

Feature Description	Presence in VRML	Presence in BIFS Version 4	Presence in SMIL	Presence in SVG	Presence in Flash	Presence in QuickTime
Spatial and temporal composition of text, graphics, images, videos and sounds	3D composition of 3D graphics, text, video, sound and images	2D and/or 3D composition of simple and complex scenes	2D composition only	2D graphics, text and images only. No support for video or audio.	2D graphics, text and images only. No support for video or audio.	2D and 3D
Composition of media coming from several sources	Supported but without any decoding model to insure synchronisation	Fully supported	Supported but without any decoding model to insure synchronisation	Supported only for SVG remote media	Not supported	Fully supported

Feature Description	Presence in VRML	Presence in BIFS Version 4	Presence in SMIL	Presence in SVG	Presence in Flash	Presence in QuickTime
Animation	Can be performed by means of interpolators but not in a streaming fashion	Includes VRML way of doing animation and adds the mechanisms of BIFS-Anim	SMIL Animation module only performs anim. in a non-streaming way. All the animation par. are known at time 0	Uses the SMIL Animation	Fully supported	Proprietary format
Compression	Use of the Zip compression	Native binary format with efficient compression	Not supported	Use of the Zip compression	Native compression	Proprietary format for scene, many formats for other media
Streaming facility	Not supported	Designed around this central point	Not supported	Not supported	Fully supported	Fully supported
Dynamic composition	Not supported	Supported through the update mechanism	Not supported	Not supported	Fully supported	Fully supported
Digital Rights Management	Not supported	MPEG-4 Systems in general provides hooks for DRM	Not supported	Not supported	Not supported	Not supported
Integration with other environments (Broadcast, TV)	Restricted to Web environment	Supported through the mapping of MPEG-4 over MPEG-2 and IP	Restricted to Web applications	Restricted to Web applications	Restricted to Web applications	Restricted to Web applications
Authoring tools and players	Lots of software available on the market	Some tools are available	Some tools are available	Some tools are available	Many authoring tools and players exist	Many authoring tools and players exist
Fine Synchronization between the media	Not supported on a per frame basis	Enables synchronization of several medias on a per frame basis	Not supported on a per frame basis	Not supported on a per frame basis	Not supported on a per frame basis	Fully supported
Scripting facilities	Supported by means of Javascript	Supported by means of Javascript and Java	Supported through DOM interface	Supported through DOM interface	Supported via ActionScripts	Fully supported
Independence of media source description & scene descr.	Not supported, reference to the media are made directly in the text	Fully supported thanks to the ODF	Not supported, reference to the media are made directly in the text	Not supported, reference to the media are made directly in the text	Not supported	Fully supported
Broadcast carrousel	Not supported	Fully supported	Not supported	Not supported	Not supported	Not supported
Error Resilience	Not Supported	Not supported	Not supported	Not supported	Not supported	Not supported
Formatted Text	Not supported	Under development	Through the use of XHTML	Not supported	Not supported	Not supported
Decorated text	Not supported	Under development	Not supported	Fully supported	Fully supported	Not supported
Filtering Effects	Not supported	Partially supported	Not fully supported	Fully supported	Fully supported	Fully supported
2D Graphics	Not supported	Fully supported	Through the use of SVG	Fully supported	Fully supported	Fully supported
Interactivity	Supported for pointing devices	Fully supported	Supported for DOM events	Supported for DOM events	Fully supported	Fully supported
3D Graphics	Fully supported	Fully supported	Not supported	Not supported	Under development	Fully supported