# Content Description for Face Animation

**Ali Arya**
*Simon Fraser University, Canada*

## INTRODUCTION

Face animation is a challenging area of computer graphics and multimedia systems research (Parke, 1996). Realistic personalized face animation is the basis for virtual software agents that can be used in many applications, including video conferencing, online training and customer service, visual effects in movies, and interactive games. A software agent can play the role of a trainer, a corporate representative, a specific person in an interactive virtual world, and even a virtual actor. Using this technology, movie producers can create new scenes including people who are not physically available. Furthermore, communication systems can represent a caller without any need to transmit high volume multimedia data over limited bandwidth lines. Adding intelligence to these agents makes them ideal for interactive applications such as online games and customer service. In general, the ability to generate new and realistic multimedia data for a specific character is of particular importance in cases where prerecorded footage is unavailable, difficult, or expensive to generate, or simply too limited due to the interactive nature of the application.

Personalized face animation includes all of the algorithms, data, modules, and activities required to create a multimedia presentation resembling a specific person. The input to such a system can be a combination of audiovisual data and textual commands and descriptions. A successful face animation system needs to have efficient yet powerful solutions for providing and displaying content, for example, a content description format, decoding algorithms, creating required content, and finally, an architecture to put different components together in a flexible way. Multimedia modeling and specification languages play a major role in content description for face animation systems both as general tools and in customized forms. Such descriptions can be used for authoring and also as run-time input to the animation system.

## BACKGROUND

The diverse set of works in multimedia content description involves methods for describing the components of a multimedia presentation and their spatial and temporal relations. Historically, some of the first technical achievements in this area were related to video editing where temporal positioning of video elements is necessary. The SMPTE (Society of Motion Picture and Television Engineers) time coding that precisely specifies the location of audio/video events down to the frame level is the base for EDL (Edit Decision List) which relates pieces of recorded audio/video for editing (Ankeney, 1995; Little, 1994). Electronic program guide (EPG) is another example of content description for movies in the form of textual information added to the multimedia stream. More recent efforts by SMPTE are focused on a metadata dictionary that targets the definition of metadata description of content (see *http://www.smpte-ra.org/mdd*).

Motion Picture Expert Group (MPEG) is another major player in the field of standards for multimedia content description and delivery. MPEG-4 standard (Battista et al., 1999), introduced after MPEG-1 and MPEG-2, is one of the first comprehensive attempts to define the multimedia stream in terms of its forming components (objects such as audio, foreground figure, and background image). Users of MPEG-4 systems can use object content information (OCI) to send textual information about these objects. MPEG-7 standard, mainly motivated by the need for a better and more powerful search mechanism for multimedia content over the Internet, can also be used in a variety of other applications including multimedia authoring. The standard extends OCI and consists of a set of descriptors for multimedia features (similar to metadata in other works), schemes that show the structure of the descriptors, and a description/schema definition language based on eXtensible Markup Language (XML).

Most of these methods are not aimed at, and customized for, a certain type of multimedia stream or object. This may result in a wider range of applications but limits the capabilities for some frequently used subjects such as the human face. The study of facial movements and expressions started from a biological point of view by scientists such as John Bulwer and Charles Darwin (see Ekman & Friesen, 1978). More recently, one of the most important attempts to describe facial activities (movements) was the Facial Action Coding System (FACS). Introduced by Ekman and Friesen (1978), FACS defines 64 basic facial action units (AUs), such as raising brows, talking, and turning left and right. It should be noted though, that FACS does not provide any higher-level construct to describe the sequence of actions and spatial and temporal relations between facial activities. In other words, it is not designed to be a face animation description language.

MPEG-4 standard uses an approach similar to FACS to integrate face animation into multimedia communication, by introducing face definition parameters (FDPs) and face animation parameters (FAPs). FDPs define a face by giving coordinates and other information for its major feature points such as eyes' and lips' corners. They allow personalization of a generic face model to a particular face, and are more suitable for synthetic faces. FAPs, on the other hand, encode the movements of these facial features. There are more than 70 FAPs defined similar to FACS AUs.

Although MPEG-4 defines two sets of higher-level codes, that is, visemes and expressions, compared to low-level FACS AUs, but it still has only a set of animation commands and not an animation language. Synchronized Multimedia Integration Language (SMIL), an XML-based language, is designed to specify temporal relation of the components of a multimedia presentation, especially in Web applications (Bulterman, 2001).

There have also been different languages in the fields of virtual reality and computer graphics for modeling computer-generated scenes. Examples are Virtual Reality Modeling Language (VRML, *http://www.vrml.org*), its XML-based version known as X3D, and programming libraries like OpenGL (*http://www.opengl.org*). MPEG-4 standard includes Extensible MPEG-4 Textual format (XMT) framework to represent scene description in a textual format providing interoperability with languages such as SMIL and VRML.

None of these languages are customized for face animation, and they do not provide any explicit support for it, either. Recent advances in developing and using embodied conversational agents (ECAs), especially their Web-based applications, and growing acceptance of XML as a data representation language have drawn attention to markup languages for virtual characters. The basic idea is to define specific XML constructs related to agents' actions such as moving and talking. Virtual Human Markup Language (VHML, *http://www.vhml.org*) is an example in this regard (Marriott & Stallo, 2002). It comprises a number of special purpose languages, such as Emotion Markup Language (EML), Facial Animation Markup Language (FAML), and Body Animation Markup Language (BAML). A simple VHML document looks like this:

```
• <vhml>
  <person disposition="angry">
  <p>
  First I speak with an angry voice and look very
  angry,
  <surprised intensity="50">
  but suddenly I change to look more surprised.
  </surprised>
  </p>
  </person>
  </vhml>
```

Multimodal Presentation Markup Language (MPML, *http://www.miv.t.u-tokyo.ac.jp/MPML/en*) is another XML-based markup language developed to enable the description of multimodal presentation on the WWW, based on animated characters (Prendinger et al., 2002). MPML addresses the interactivity and decision-making not directly covered by VHML, but both suffer from a lack of explicit compatibility with MPEG-4 (XMT, FAPs, etc.).

Another important group of related works are behavioural modeling languages and tools for virtual agents. Behaviour Expression Animation Toolkit (BEAT) is an XML-based system, specifically designed for human animation purposes (Cassell et al., 2001). It is a toolkit for automatically suggesting expressions and gestures, based on a given text to be spoken. BEAT uses a knowledge base and a rule set, and provides synchronization data for facial activities, all in XML format. Although BEAT is not a general content description tool, it demonstrates some of the advantages of XML-based approaches together with the power of behavioural modeling.

Another group of researchers have proposed the concept of cognitive modeling for character animation (Funge et al., 1999). Their system is based on a set of geometric, behavioural, and cognitive models for the characters to be animated. In this approach not only the physical capabilities but also the behavioural and cognitive processes are defined and modeled. A special Cognitive Modeling Language (CML) is also developed to support this system. CML does not provide any explicit support for face animation and, unlike BEAT, is not XML-based. Neither is Parameterized Action Representation (PAR), another language proposed to describe and model actions of an agent, based on interaction with environment and the personality and emotions (Allbeck & Badler, 2002). In PAR, the agent personality is defined in terms of parameters such as openness, agreeableness, and extraversion. Similar parameters are defined for other aspects affecting the behaviour.

## FACE MODELING LANGUAGE (FML)

Face Modeling Language (FML) is part of the ShowFace Personalized Face Animation Framework (Arya & Hamidzadeh, 2002). FML (*http://www.raminsoftworx.com/Research/fml_1.html*) is an XML-based language to provide structure content description, that is, a hierarchical