A 3D Agent with Synthetic Face and Semiautonomous Behavior for Multimodal Presentations

Istvan Barakonyi, Mitsuru Ishizuka
Department of Information and Communication Engineering,
School of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Email: {bara,ishizuka}@miv.t.u-tokyo.ac.jp

Abstract

Numerous researchers in the human-computer interaction field develop applications for and test theories with lifelike animated agents. In this paper we present a three-dimensional agent with synthesized face and speech, which is capable of automatic motion generation with a little user support of text annotation. Users without artistic skills are also capable of easy creation and extension of their own agents with tailor-made appearance and motion. The agent can be reused in a wide range of applications as a stand-alone component in addition to an editor program, which enables authors to immediately try the parameter settings for facial expressions, head and eye movements, facial coloration and motion generation. Its implementation and compact size allows for easy use on the Internet, which is demonstrated by an already existing application called the MPML project.

1. Introduction

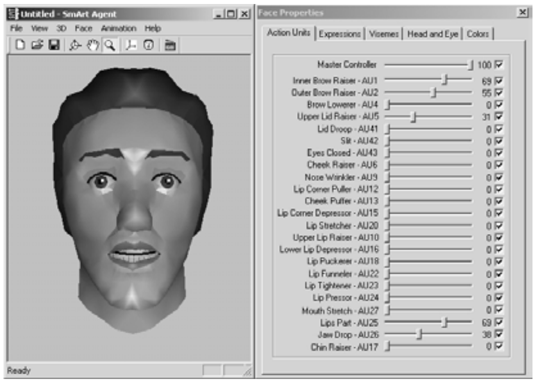
This paper describes a three-dimensional facial agent with synthetic speech, face and behavior. The background field of our research includes facial animation, presentational and conversational agents. Lifelike, animated agents with realistic behavior have gained much interest in the recent years as they seem to make dreams about having an intelligent “computer friend” come true [2]. Their function and application domain is versatile; so far they have been utilized as virtual actors [16], personal interactive tutors [17] and presentation agents [1], [7]. We have focused on the latter category, as distributed, interactive presentations over the Internet without a human presenter are becoming a promising way for teaching or demonstrating ideas.

We visualize our agent as a three-dimensional human face (see the screenshot of Figure 1). In [13] a direct parameterized model was introduced that provides control parameters for both facial conformation and expression. The action units of the Facial Action Coding System described in [8] and the abstract muscle action procedures of [10] both define atomic facial movements to build complex expressions as well as head and eye motion, following a pseudomuscle-based animation concept [14]. The work described in [18] presents a dynamic face model with physics-based synthetic muscles and facial tissue.

In our project we didn’t aim to achieve photorealism but rather easy modification and animation, in addition to reasonable speed and quality at today’s average PC platform. The compact size and the implementation method (see Section 6) allow for the agent’s easy use on the Internet. Our facial model’s approach is pseudomuscle-based, mainly because of limited computation time (there are speed concerns with the physics-based model) and its close relation to the natural face structure. The face is represented by a polygon mesh, which is colored conforming to the facial parts. We obtained the geometric data from Harashima Laboratory’s FaceTool project [9].

Multimodal user input and output [3] have been a key design issue to utilize numerous communication channels. Currently besides traditional GUI input and output methods we utilize speech synthesis and recognition, and consider extensions during the future development.

2. System overview



Master Controller 100 ☑
Inner Brow Raiser - AU1 69 ☑
Outer Brow Raiser - AU2 55 ☑
Brow Lowerer - AU4 0 ☑
Upper Lid Raiser - AU5 31 ☑
Lid Droop - AU41 0 ☑
Slit - AU42 0 ☑
Eyes Closed - AU43 0 ☑
Cheek Raiser - AU6 0 ☑
Nose Wrinkler - AU9 0 ☑
Lip Corner Puller - AU12 0 ☑
Cheek Puffer - AU13 0 ☑
Lip Corner Depressor - AU15 0 ☑
Lip Stretcher - AU20 0 ☑
Upper Lip Raiser - AU10 0 ☑
Lower Lip Depressor - AU16 0 ☑
Lip Puckerer - AU18 0 ☑
Lip Funneler - AU22 0 ☑
Lip Tightener - AU23 0 ☑
Lip Pressor - AU24 0 ☑
Mouth Stretch - AU27 0 ☑
Lips Part - AU25 69 ☑
Jaw Drop - AU26 38 ☑
Chin Raiser - AU17 0 ☑

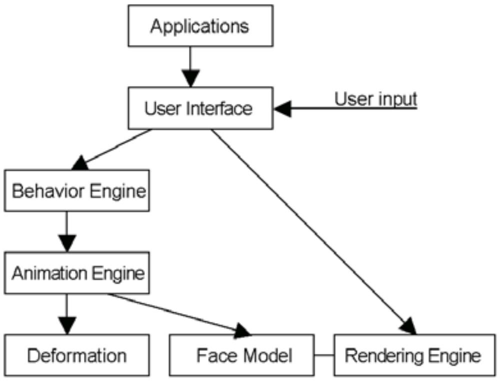


Figure 1. Screenshot of the agent editor

Figure 2. Architecture of the agent

The synthesis and animation of human faces is a complex procedure with several abstraction levels. We use the multi-layered structure of Figure 2, where these levels are treated independently rather than as a monolithic structure. The object at the lowest level represents the rendering engine. Besides rendering the vertices and facets of the facial mesh, it is also responsible for calculating smooth surface for the skin, while preserving creases and wrinkles with a shading algorithm. The face model and the deformation component contain the geometric data of the face model and the pseudomuscles for facial expression composition. The animation engine, which will be discussed more in details in Section 3, is responsible for animating the face at an atomic level like contracting muscles or moving the eye to a certain position, and provides smooth, natural transition between these movements. The behavior engine discussed in Section 4 is built on the top of these primitives and treats higher-level units like facial expressions at a certain intensity, lip shapes for the speech or maintaining/breaking eye contact. Moreover, its task is to "bring life" into the synthetic head by automatically generating motion to support speech and give a lifelike impression. The top-level object in the architecture is the user interface level providing a standardized communication interface towards the outside world. This is the service access point for other applications, e.g. commands that can be called from author scripts are defined here. It also handles user input like mouse events, speech input for recognition and in the future possibly facial recognition. The facial agent component can either be reused inside another application's code or scripted as an individual component.

3. The presenter

The user's attitude to a computer agent is influenced by its appearance and behavior. The appearance in our case is mainly determined by the visual features of the face. In cartoons and movies positive characters most often have a charming look conforming to what most people consider attractive, while negative characters usually seem rather unappealing or have some kind of asymmetry. By changing the agent's facial mesh a wide range of facial shapes can be created (see Figure 3). In our agent editor users can modify facial parameters from an extendable parameter library, which is similar to the method of drawing an identikit for the police. Different parts of the face (nose, lips, eyes, head etc.) can be changed separately. The result of this higher-level modification can be then "fine tuned" at the vertex level.

We often say expressions like "he is red with anger" showing the importance of another significant visual feature: facial colors. We put special emphasis on it since besides the facial shape it enables the creation of various faces with different skin, hair, iris etc. color, and it can simulate secondary emotion effects like blushing (cheek color) or pallor (skin color). Other effects like the stubble of an unshaved face (cheek and moustache color) or make-up can be also realized. Facial colors related to the skin (e.g. cheek, nose, eyelid etc.) are always blended with the current skin tone as its changes affect all the other facial shades. For instance, when the skin gets red, the additional redness of the cheek also gets a darker shade, and if unshaved, the stubble's bluish color becomes a bit red as well because the skin shows through.

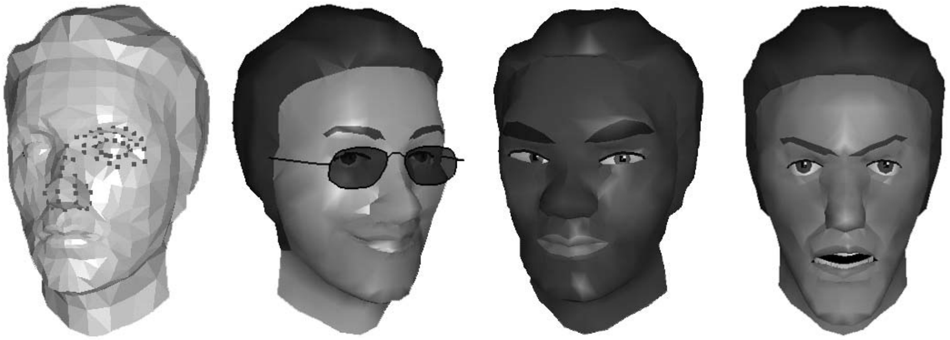


Figure 3. Creating different faces by changing facial parameters

As mentioned in the introduction, to animate the face we try to imitate the function of facial muscles or muscle groups, which are implemented as Ekman's action units. By assigning intensity values (0-100%) to these action units, we can simulate the contraction of the facial muscles at a desired degree and thus compose facial expressions. The creation of visemes – visually indistinguishable phoneme groups - for lip synchronization to the synthesized speech follows the same approach as facial expressions, because mouth shapes are also formed by facial muscles. Other controllable low-level animation parameters are eye and head position.

4. Synthesizing behavior

The other important factor how an agent appeals to us is its behavior. We build it up from atomic components that we call behavioral primitives. These are the following: making a facial expression or a mouth shape with a certain intensity, moving the head or the eyes to a certain position and changing the color of a facial part. We have to take special care for making smooth and natural transition between these facial gesture elements. Real muscles cannot contract or release immediately, they need a sort of ease-in and ease-out period. In Figure 4 you can see the transition function called Hermite function of third order (see [12] for details), which can produce realistic motion with the desired features.

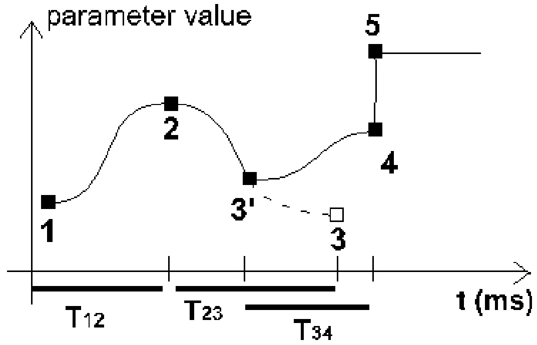


Figure 4. Smooth transition function

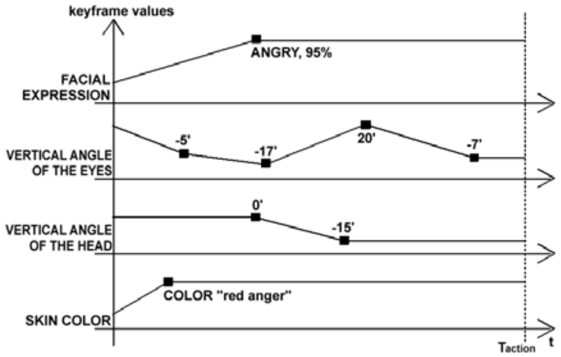


Figure 5. Action composition with keyframing

By adding precise timing and duration information to the previously defined behavioral primitives we can compose higher level, complex actions with the keyframing animation technique (see Figure 5), for example making a half smile, blushing and turning the head down at the same time can be stored and reused as an action called "*ashamed*". Commands like "make/break eye contact" or "turn to/away from user" are also introduced. The agent talks by using a text-to-speech (TTS) engine for synthesis. The lip synchronization is done automatically, our algorithm uses the look-ahead coarticulation model [5]. Lip shapes coming from the speech and the facial gestures are blended naturally. There are some sounds that cannot be synthesized by the TTS engine, therefore we use sound files, e.g. to simulate crying or hiccup.

An important feature of our agent is its ability to execute all the previously defined complex actions while speaking. It is implemented by inserting tags into the speech text. For example passing the following speech string to the agent makes it act out an uneasy restaurant situation: *"\action='satisfied',120,90\ Oh, this was a great dinner! \action='shocked',90,100\ What? Is it really so expensive?"* The values after the action name scale the intensity and the time of the originally defined action. To provide the dynamic, realistic feature of the human face, natural movements are generated by the behavior engine. The occurrence and type of these events are depending on whether the face is idle like gazing around to avoid a static, staring face, or an action is being executed like stressing a spoken word with a head nod or raising the eyebrows at a meaningful break. We referred to [6] and [15] for background work on this topic and we are extending it with our own studies. Additional speech tags can control the voice parameters like pitch, speed or volume, which allows for a more natural intonation and emotional content in the voice.

5. Application

One of our laboratory's main projects is to develop a system including a language called Multimodal Presentation Markup Language (MPML) and related multimodal presentation authoring software [7]. It enables authors to easily enhance their already existing presentation content with interface agents having believable behavior. It integrates a powerful, XML-based language to author scenarios for the agents, i.e. giving instructions when, what and how to perform a certain part of the presentation. As the language is standardized and independent from the agent system, it can control several different agents including ours, the popular Microsoft Character Agent and other scriptable, animated interface agents. A converter software generates commands and tagged speech strings for our agent containing facial gestures information and voice intonation parameters to perform the presentation on the top of the content pages.

6. Implementation

Currently the standalone agent is realized as an ActiveX component coded in C++, and the editor software with GUI uses C++ as well with MFC classes. The rendering engine is implemented in OpenGL specification 1.1. The current facial model uses 825 vertices and 1386 triangles. Our application produced a 50 FPS frame rate in full screen mode in the following test environment: *Pentium III 750 MHz, 256Mb memory, NVIDIA GeForce 256 video card, Win2000 Professional, 1280x1024 screen resolution, 16-bit color depth.* Speech synthesis and recognition are realized with the Microsoft Speech API 5.0, thus all text-to-speech and speech recognition engines conforming to this interface can be used with the agent.

7. Conclusion and future developments

We designed a scriptable three-dimensional facial character agent with realistic behavior. The intention of our work is to provide an easy-to-use and extensible tool for both application developers and users. Our agent's capabilities are especially suitable for being a presentation agent. To enhance the multimodality of the agent's interface, facial gesture recognition can be added. Experiments are considered with more detailed and sophisticated 3D face models. The behavior also needs more thorough investigation, for instance by capturing personal behavior to make the agent fake the gestures of observed persons. Additional secondary emotion effects may be supported like tears or sweating. Psychologists claim [4] that it is rather hard to find precise description on the relation between the spoken text and the facial expressions supporting it. Instead of trying to get the computer to guess the content and the stressing of a text, we are thinking about a system that tracks the user's face while reading aloud his/her presentation content and tagging the speech automatically with facial action tags conforming to the captured information.

8. References

[1] E. Andre, T. Rist and J. Muller: WebPersona: A Life-Like Presentation Agent for the World Wide Web, *Knowledge-Based System*s, Vol. II, pp. 25-36, 1998.

[2] J. Bates: The Role of Emotion in Believable Agents, *Communications of the ACM,* 37(7):122-125, 1994.

[3] C. Benoit, J. C. Martin, C. Pelachaud, L. Schomaker and B. Suhm: Audio-Visual and Multimodal Speech Systems, *In D. Gibbon (Ed.) Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume*, to appear.

[4] G. Collier: Emotional Expression, Lawrance Erlbaum Associates Inc., 1985.

[5] M. M. Cohen, D. W. Massaro: Modeling Coarticulation in Synthetic Visual Speech, *Models and Techniques in Computer Animation, N. M. Thalmann, and D. Thalmann (Eds.), Tokyo: Springer-Verla*g, pp. 139-156. 1993.

[6] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost and M. Stone: Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents, *SIGGRAPH*, pp. 413-420, 1994.

[7] S. Descamps, H. Predinger, M. Ishizuka: A Multimodal Presentation Mark-up Language for Enhanced Affective Presentation, *Proc. Int'l Conf. on Intelligent Multimedia and Distant Learning (ICIMADE-01)*, pp.9-16

[8] P. Ekman, W. Friesen: Facial Action Coding System, *Consulting Psychologists Press, Inc.,* 1978.

[9] Harashima Laboratory homepage: http://www.hc.t.u-tokyo.ac.jp

[10] P. Kalra, A. Mangili, N.M. Thalmann and D. Thalmann: SMILE: A Multilayered Facial Animation System, *Proc IFIP WG 5.10, Tokyo, Japan (Ed Kunii Tosiyasu L)* pp. 189-198, 1991.

[11] T. Noma, L. Zhao, N. Badler: Design of a Virtual Human Presenter, *IEEE Computer Graphics and Appli., Vol.20., No.4*, pp.79-85, 2000.

[12] J. Ostermann: Animation of Synthetic Faces in MPEG-4, *Computer Animation'98, Philadelphia, Pennsylvania, USA, IEEE Computer Society,* 1998.

[13] F. Parke: A Parameterized Model for Facial Animation, *IEEE Computer Graphics & Applications*, 2(9):61-68., 1982.

[14] F. Parke, K. Waters: Computer Facial Animation, *A K Peters, Wellesley, Massachusetts*, 1993.

[15] C. Pelachaud, N. I. Badler, and M. Steedman: Generating facial expressions for speech, *Cognitive Science*, 20(1):1-46, 1996.

[16] K. Perlin, A. Goldberg: Improv. A System for Scripting Interactive Actors in Virtual Worlds, *Computer Graphics. 29(3), 1996.*

[17] J. Rickel, W. L. Johnson: STEVE: A Pedagogical Agent for Virtual Reality. *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, 1998.

[18] K. Waters: A Muscle Model for Animating Three-Dimensional Facial Expression, *SIGGRAPH 87, Computer Graphics Annual Conference series*, 21(4):17-24. Addison Wesley, July 1987.