Real-Time Facial Animation for Avatars in Collaborative Virtual Environments

Dennis Burford and Edwin Blake
Collaborative Visual Computing Laboratory
Department of Computer Science
University of Cape Town

Email: {dburford, edwin}@cs.uct.ac.za

ABSTRACT

Collaborative virtual environments (CVE's) provide opportunities for interaction, communication and cooperation between participants at different physical locations. For CVE's to be effective, participants must feel that they are present in the virtual environment. It has been postulated that avatars (virtual representations) using body-like figures increase presence. We outline a system for providing real-time facial animation for avatars in CVE's. The system shall use a performer driven approach, with fiducials providing facial feature tracking. We hypothesize that this system will increase personal and group presence, and lead to a greater emotional investment in group interaction.

1 VIRTUAL ENVIRONMENTS AND PRESENCE

The construction of virtual spaces within a computer system has been possible for some time. These *virtual environments* (VE's) have been used for a variety of applications, the most common being data visualisation, architectural walkthroughs, and simulators. The main advantage over traditional display techniques is that the user becomes an important part of the scene. Instead of delivering the data to the real world via a computer display, VE's immerse the user in the virtual world of the data. *Collaborative virtual environments* (CVE's) share a virtual world among many users. Often the users are at separate physical locations, requiring network techniques to deliver and maintain a consistent representation of the environment. Those present in a CVE are usually able to interact, communicate and thereby collaborate on various tasks.

The potential for using shared virtual spaces for collaboration and entertainment is well documented and is an active research topic for many groups, including British Telecommunications [2]. They see possibilities for a new communication paradigm facilitating interaction among people at different physical locations. These people will have a desire to "meet" because they share interests or need to cooperate on certain tasks. Whatever the reason, they believe meeting in a shared virtual environment allows multi-way communication and interaction far surpassing the potential of standard telecommunications technology.

Key to developing a usable CVE is the ability to convince the participants that they are present in the VE and that others are there with them - they must have mutual awareness. Without the sense of personal and group "presence", it is impossible for active and productive collaboration to take place. Slater et al [14] define presence as "a state of consciousness, the (psychological) sense of being in the virtual environment". Slater et al [12, 13] further classify presence into *personal presence* and *shared presence* (or co-presence). Personal presence relates to the subjective feeling of actually "being" in the VE, leading to a sense of "places visited, rather than seen" [14]. Shared presence refers jointly to the feeling that others in the VE actually exist, and to the feeling

of group membership.

This all important idea of presence relies on the user *sensing* others and feeling *immersed* in the virtual environment. To this end, objects and participants should behave as expected; they should show the same characteristic behaviour as they would in the real world. That is, they should act, react and, with respect to the participants, express themselves in a believable way.

In order to support mutual awareness, issues such as user location, availability, attitudes, and personal and group identity must be addressed. These issues can be tackled by using virtual representations of participants, or *avatars*. Avatars are crucial in a CVE, as they represent the point of view of each participant in the virtual world and thus facilitate an awareness of ongoing activities. Avatars using body-like figures are particularly useful, since animation of body movements and facial expressions can be used to enhance mutual awareness.

This project aims to provide an increased degree of expression for participants in distributed virtual environments by providing believable real-time (and hence low bit rate) facial animation for their virtual representations. The objective will be to achieve the best results possible using relatively low cost, widely available equipment. The focus will be on recognizing lip movement for vocal communication and major expressions such as smiling, frowning, surprise and so on. We hypothesize that the increased expressive ability provided by the facial animation should lead to an improved sense of presence and mutual awareness and therefore a greater emotional investment.

In the next section, we describe the various techniques used for facial animation and some of their applications. Section 3 discusses the requirements for a real-time system. We present an overview of our project in Section 4. The paper ends with a brief summary.

2 FACIAL ANIMATION

In the real world, facial expressions are the best indicator of a person's mood, emotion and general "state". On the whole the face, and specifically the eyes, are important for intimacy and trust. Mood, emotion and trust are important issues in collaboration and they must be conveyed in some way. Within a CVE, facial animation for avatars is one possible solution.

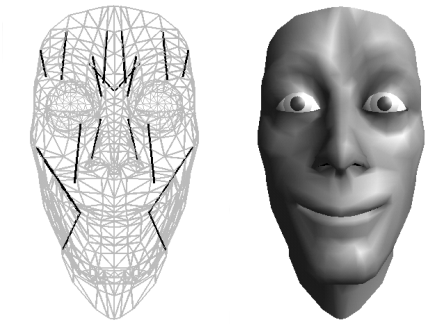

Figure 1: An implementation of Waters' muscle model showing a smiling expression (Compaq Cambridge Research Laboratory). The black line segments in the first figure indicate the directions of possible muscle contraction.

Computer assisted facial animation is a well established field that has been applied to high-compression video conferencing, synthesised actors, character animations, virtual reality, and lip-synchronisation and lip reading for the deaf. The major challenges in facial animation are:

1. *Developing realistic and flexible models.* Constructing a believable 3D facial model is perhaps the most difficult aspect to facial animation. Physically based muscle models, such as those by Lee [9], Terzopoulos [16, 17] and Waters [18], are realistic but computationally expensive, while deformations of geometry [11] and texture [8] generally trade realism for speed. Figure 1 shows an implementation of Waters' muscle model.

   Many systems try to construct a facial model that closely resembles the actor driving the animation. Guenter et al [8] use Cyberware scans with complex texture mapping while Escher and Magnenat-Thalmann [6] fit a generic mesh model to a specific face using control points obtained from two camera views.

2. *Recognition, analysis and later synthesis of expressions.* Recognition and synthesis of expressions is often achieved through performer driven animation (actor tracking) [1, 7, 19]. Alternatively, functional control can be used to replay and interpolate predefined expressions. This allows expressions to be manually

"constructed" through a control panel interface.

3. *Parameterisation of expressions and expression components.* Several techniques allow the parametric definition and deformation of facial models. Early work was done in this area by Ekman who developed the Facial Action Coding System, FACS [5]. Magnenat-Thalmann et al have since developed the Abstract Muscle Action system, AMA [10]. An important development is the more recent ISO backed protocol: the MPEG-4 Synthetic/Natural Hybrid Coding (SNHC) scheme. This protocol defines parameters for facial definition (FDP) and animation (FAP).

The various techniques discussed in this section are frequently combined - the choice dependent on the final application.

3 REQUIREMENTS FOR A REAL TIME SYSTEM

For a real time system, it is necessary to recognize, transmit and synthesize the facial expressions as they happen (or with a delay that is effectively un-noticeable to the users). In order to do this, it is important to maintain a consistent frame-rate for the animation - ideally 15-25 frames per second. This is a major aim of our system, and we will need to combine and tailor the methods discussed in the previous section to achieve this.

For virtual reality systems, performer driven animation is desirable. Functional control techniques can be used, but these require some additional type of user interface between the user and their avatar. No matter how ingenious the approach, the extra thought and effort required from the user will always detract from the spontaneity of expressions and the overall experience.

With performer driven animation, the biggest bottleneck in the system is likely to be expression recognition and tracking. For this reason, it is important to find ways to aid and simplify the recognition process. Typically, fiducials (small markers) in the form of balls or paper discs are placed onto the participant's face and tracked in time (see Figure 2). The positions of the fiducials must be determined for each frame of the video

Figure 2: An actress wearing facial fiducials - in this case, brightly coloured beads.

sequence and each fiducial should be correlated with the corresponding fiducial of the previous frame. For complex tracking, multiple cameras can be used in order to determine the positions in 3D. Fiducial correlation between each camera is then required.

The main problem is the identification of the fiducials for each frame of the sequence. This is an image segmentation problem which can be tackled using standard techniques developed in this area. One approach is to find all the connected pixels that fall within the known colour range and calculate the positions from these pixel clusters. The image analysis is computationally expensive, especially when the entire image needs to be scanned for each frame of the sequence. To ensure real-time animation, shortcuts are required:

- Additional input devices can be used to aid tracking and make predictions about the future fiducial positions. This means that only certain regions of the image need to scanned for fiducials - speeding up the recognition process. More complex motion prediction could further refine the search regions.
- Only the key facial features need be tracked - the others can be "faked" through simulated expressions.
- There is room for limited interpolation between adjacent frames. Interpolation can only be performed when the time difference between frames is sufficiently small. Buffering frames that have a large time separation will cause excessive delays.
- It is not totally necessary to use multiple cameras. If the overall head movement is

small, 2D tracking can be sufficient for real time applications.

- It is also not essential that the facial model closely resembles the user. More importantly, the avatar's expressions should be believable, smoothly animated and correctly convey the user's emotion.

A number of attempts have been made at producing a real-time animation system for applications such as virtual videoconferencing. An example is the system developed by Escher and Magnenat-Thalmann [6]. Their animation is driven by feature recognition from live video and/or phonemes extracted from the audio stream. Processing tasks are delegated amongst different machines to improve performance. The results are integrated and used to drive free form deformations to a mesh model resembling the actor.

4 PROJECT OVERVIEW

Currently, this project is work in progress. The material presented in this paper is an overview of intended work and represents our future direction. We hope to complete the project midway through 2000, and perform experiments to verify our hypotheses.

4.1 EQUIPMENT

- *CrystalEyes Sterographic* glasses and *Logitech* ultrasound head-tracker.
- Some form of facial fiducials.
- Silicon Graphics workstations with the standard SGI cameras.

4.2 APPROACH

We shall use a performer driven animation approach with 2D fiducial tracking. A single camera will be used for video input with the *CrystalEyes StereoGraphics* glasses providing head tracking (position and orientation). The changes in fiducial position will drive model deformation.

Our planned animation system is outlined in Figure 3. The two input devices, a video camera and the *CrystalEyes* glasses (for headtracking), are shown at the top. The various processing modules appear below. These modules are described in detail in Sections 4.2.1-4.2.3.

Work has been done on the controlling routines for the headtracker (module ***A***) and the image segmentation routines (module ***B***). Preliminary image segmentation is performed by thresholding the images in HSV colour space.

4.2.1 INITIALISATION

0: The initialisation module (greyed) calibrates the head tracker and camera and correlates the images from the camera with the position and orientation data from the head tracker. The routine also scans the initial image of the face and determines the positions and colours of the facial fiducials. This is the only time that the entire image is scanned for fiducials.

The position and colour information is entered into a database, with colouring information linked to the position of the corresponding fiducial. In subsequent frames, the information in the database is used to find and isolate the fiducials, as described below.

4.2.2 FIDUCIAL TRACKING

A: This module uses the current position and orientation of the headtracker to transform the fiducial positions of the previous frame. This gives a crude prediction for their positions in the new frame.

B: A recognition routine uses the predicted positions along with the colours of the previous frame to find the fiducial positions for the current frame. The new colours and positions are entered into the database.

C: It is now a simple task to compute the change in fiducial positions which are then used to deform the facial model. Changes is fiducial positions are used rather than absolute coordinates.

4.2.3 MODEL DEFORMATION

D: Once the positions of the fiducials are determined for each frame, they are used to deform a facial model. A simple polygonal mesh model for the face will probably be used, with free form deformations to the control points of the mesh driven by the positions of the fiducials. More so-

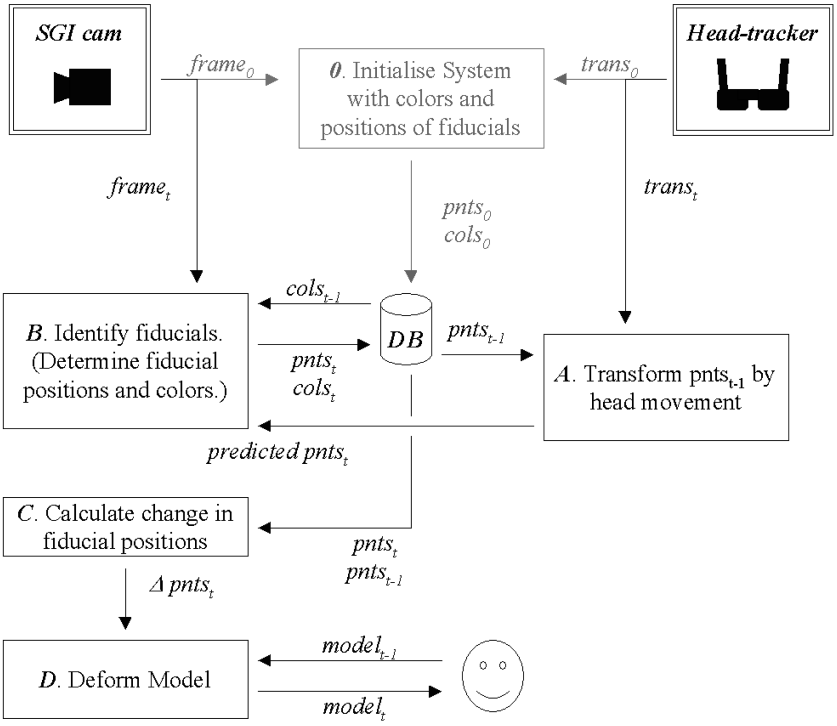

Figure 3: Flow diagram for recognition, analysis and synthesis of facial expressions. The greyed section is performed only once to initialise the system.

phisticated muscle models (see Figure 1) may be investigated for added realism.

The final system will be implemented using DIVE [4, 3], a toolkit for the development of multi-user distributed virtual environments. Only the deformations to the facial model will be transmitted to each participant, with the rendering performed on their workstations. This will limit the amount of data transmitted across the network.

4.3 EXPERIMENTS

In order to test the hypothesis mentioned in Section 1, we will perform various experiments aimed at assessing any increase in personal and group presence. Since facial expressions are so tightly linked to a person's emotion and mood, we will need to carefully design experiments with appropriate scenarios. Participants will need to communicate their feelings through their expressions and be attentive to the others in the environment.

Several scenarios involving careful listening, discussion and small group collaboration are being considered. A more ambitious aspect of the project is the possible application to distance counseling and therapy. Slater et al [15] have performed such an experiment involving public speaking anxiety in virtual environments. Their goal is to investigate the effectiveness of virtual environments in psychotherapy for social phobias.

A major issue in the design of these experiments is finding reliable and objective measures of presence, which are, at least for now, rather elusive. One may argue that if we need presence in order to work effectively, then we should use task performance as a metric. Indeed, it is commonly used for testing virtual environments, but since performance is influenced by many factors, it is not totally reliable. Instead, subjective measures such as post-experiment questionnaires are often used.

5 SUMMARY

A real time facial animation system will be developed for avatars within a collaborative virtual environment. The aim of the system will be to determine the effect of facial animation on personal

and group presence.

We will be using a performer driven approach, with fiducials providing facial feature tracking. A single camera will be used to track the fiducial motion in 2D. *CrystalEyes StereoGraphics* glasses and a *Logitech* ultra-sound head tracker will provide orientation and position information that will aid the tracking. An emphasis is on using relatively low cost and widely available equipment to provide good real time results.

Several experiments measuring the effects of facial animation on presence need to be designed. These experiments will focus on the expressive ability of the avatars, and the effect this has on the overall experience within the CVE.

REFERENCES

[1] Philippe Bergeron and Pierre Lachapelle. Controlling Facial Expressions and Body Movements in the Computer-Generated Animated Short 'Tony De Peltrie'. In *SIGGRAPH 85*, Computer Graphics Annual Conference series. Addison Wesley, July 1985.

[2] Laurence Bradley, Graham Walker, and Andrew McGrath. Shared spaces. *British Telecommunications Engineering Journal*, 15, July 1996.

[3] Carlsson and Hagsand. DIVE - A Multi User Virtual Reality System. In *IEEE Virtual Reality Annual International Symposium*, pages 394–400, September 18-22 1993.

[4] C. Carlsson and O. Hagsand. DIVE - A Platform for Multi-User Virtual Environments. *Computers and Graphics*, 17(6), 1993.

[5] Paul Ekman and Wallace V. Friesen. *Manual for the Facial Action Coding System.* Consulting Psychologists Press, Inc., Palo Alto, California, 1978.

[6] Marc Escher and Nadia Magnenat-Thalmann. Automatic 3D Cloning and Real-Time Animation of a Human Face. *MIRALab, University of Geneva*, 1997.

[7] Irfan Essa, Sumit Basu, Trevor Darrell, , and Alex Pentland. Modeling, Tracking and Interactive Animation of Faces and Heads Using Input from Video. In *Computer Animation Conference*, Computer Graphics Annual Conference series, pages 68–79, June 1996.

[8] Brian Guenter, Cindy Grimm, Daniel Wood, Henrique Malvar, and Fredrick Pighin. Making Faces. In *SIGGRAPH 98*, Computer Graphics Annual Conference series, pages 55–66. Addison Wesley, July 1998.

[9] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Realistic Modeling for Facial Animation. In *SIGGRAPH 95*, Computer Graphics Annual Conference series, pages 55–62. Addison Wesley, August 1995.

[10] N. Magnenat-Thalmann, H. Minh, M. de Angelis, and D. Thalmann. Design, Transformation and Animation of Human Faces. *The Visual Computer*, 5(1/2):32–39, March 1989.

[11] Frederic I. Parke. Computer Generated Animation of Faces. In *Proceedings ACM annual conference*, August 1972.

[12] M. Slater, A. Steed, J. McCarthy, and F. Maringelli. The Influence of Body Movement on Presence in Virtual Environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(3), September 1998.

[13] M. Slater, M. Usoh, S. Benford, D. Snowdon, C. Brown, T. Rodden, G. Smith, and S. Wilbur. Distributed Extensible Virtual Reality Laboratory (DEVRL). In *Virtual Environments and Scientific Visualisation '96*, pages 137–148. Springer Computer Science Goebel, M., Slavik, P. and van Wijk, J.J. (eds). ISSN0946-2767, 1996.

[14] M. Slater, M. Usoh, and Y. Chrysanthou. The Influence of Dynamic Shadows on Presence in Immersive Virtual Environments. In M. Goebel (ed.) Springer Computer Science, editor, *Virtual Environments '95*, pages 8–21, 1995. ISSN 0946-2767.

[15] Mel Slater, David-Paul Pertaub, and Anthony Steed. Public Speaking in Virtual Reality: Facing an Audience of Avatars. *IEEE Computer Graphics and Applications*, 19(2):6–9, April 1999.

[16] D. Terzopoulos and K. Waters. Analysis and Synthesis of Facial Image Sequences using Physical and Anatomical Models. *IEEE*

Trans. Pattern Analysis and Machine Intelligence, 15(6):569–579, June 1993.

[17] Demetri Terzopoulos and Keith Waters. Physically-Based Facial Modeling, Analysis, and Animation. *Journal of Visualisation and Computer Animation*, 1(4):73–80, March 1990.

[18] Keith Waters. A Muscle Model for Animating Three-Dimensional Facial Expression. In *SIGGRAPH 87*, volume 21 of *Computer Graphics Annual Conference series*, pages 17–24. Addison Wesley, July 1987.

[19] Lance Williams. Performance-Driven Facial Animation. In *SIGGRAPH 90*, volume 24 of *Computer Graphics Annual Conference series*, pages 235–242. Addison Wesley, August 1990.