DOCUMENT RESUME

ED 388 216 IR 017 365

AUTHOR	Thalmann, Nadia Magnenat
TITLE	Communicating with Virtual Humans.
PUB DATE	94
NOTE	7p.; In: Educational Multimedia and Hypermedia, 1994. Proceedings of ED-MEDIA 94--World Conference on Educational Multimedia and Hypermedia (Vancouver, British Columbia, Canada, June 25-30, 1994); see IR 017 359.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Animation; *Computer Graphics; Computer Mediated Communication: Computer System Design; *Facial Expressions; Foreign Countries; Input Output Devices
IDENTIFIERS	Virtual Reality

ABSTRACT

The face is a small part of a human, but it plays an essential role in communication. An open hybrid system for facial animation is presented. It encapsulates a considerable amount of information regarding facial models, movements, expressions, emotions, and speech. The complex description of facial animation can be handled better by assigning multiple input accessories. These input accessories may be a simple script, a multi-input musical keyboard, a gesture dialogue from the DataGlove, or some other type of interactive physical or virtual device. Integration of all means of control offers flexibility and freedom to the animator. The scope of such an open system is tremendous. The system described is written in C with the interface built on top of the Fifth Dimension Toolkit. The various input components are independent processes running on UNIX workstations. Communication between the processes is done through sockets in stream mode using Internet protocol. A figure illustrates the system processes. (Contains 17 references.) (MAS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 388 216

Communicating with Virtual Humans

NADIA MAGNENAT THALMANN
MIRALab, University of Geneva
24, rue du Général-Dufour,
CH 1204 Geneva, Switzerland
E-Mail: thalmann@cui.unige.ch

Abstract: In this paper, we present an open hybrid system for facial animation. It encapsulates a considerable amount of information regarding facial models, movements, expressions, emotions and speech. The complex description of facial animation can be handled better by assigning multiple input accessories. These input accessories may be a simple script or a multi-input musical keyboard or a gesture dialogue from the DataGlove or some other type of interactive physical or virtual device. Integration of all means of control offers flexibility and freedom to the animator. The scope of such an open system is tremendous. Virtual worlds would be desolate indeed without things like synthetic faces that we can relate to and understand.

Introduction

In the context of real-time multimedia animation systems, the relationship between a real person (the user) and virtual humans should be emphasized. With the existence of graphics workstations able to display complex scenes containing several thousands polygons at interactive speed, and with the advent of such new interactive devices as the Spaceball, EyePhone, and DataGlove, it is possible to create computer-generated characters based on a full 3D interaction metaphor in which the specifications of deformations or motion are given in real-time. True interaction between the animator and the actor requires a two-way communication: not only may the animator interact to give commands to the actor but the actor is also able to answer him. Finally, we may aspire to a virtual reality where synthetic actors participate fully: real dialog between the animator and the actor. The animator may now enter in the synthetic world that he/she has created, admire it, modify it and truly perceive it. Finally, computer-generated human beings should be present and active in the synthetic world. They should be the synthetic actors playing their unique role in the theater representing the scene to be simulated.

The face is a small part of a human, but it plays an essential role in communication. People look at faces for clues to emotions or even to read lips. It is a particular challenge to imitate these few details. In this paper, we present an open hybrid system for facial animation. It encapsulates a considerable amount of information regarding facial models, movements, expressions, emotions and speech. The complex description of facial animation can be handled better by assigning multiple input accessories. These input accessories may be a simple script or a multi-input musical keyboard or a gesture dialogue from the DataGlove or some other type of interactive physical or virtual device. Integration of all means of control offers flexibility and freedom to the animator.

Facial Animation

Because the human face plays the most important role for identification and communication, realistic construction and animation of the face is of immense interest in the research of human animation. The ultimate goal of this research would be to model exactly the human facial anatomy and movements to satisfy both structural and functional aspects. However, this involves the concurrent solution of many problems. The human face is a very irregular structure, which varies from person to person. The problem is further compounded with its interior details such as muscles, bones and tissues, and the motion which involves complex interactions and deformations of different facial features. Although all movements may be rendered by muscles, the direct use of a muscle-based model is very difficult. The complexity of the model and our poor knowledge of anatomy makes

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it
- Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Gary H. Marks
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

BEST COPY AVAILABLE

RC017345

the results somewhat unpredictable. This suggests that more abstract entities should be defined in order to create a system that can be easily manipulated. A multi-layered approach (Kalra et al. 1991) is convenient for this. In order to manipulate abstract entities like our representation of the human face (phonemes, words, expressions, emotions), we propose to decompose the problem into several layers. The high level layers are the most abstract and specify "what to do", the low level layers describe "how to do". Each level is seen as an independent layer with its own input and output.

There are presently three main types of facial animation systems in terms of driving mechanism or animation control. One type of systems uses a script or command language for specifying the animation (Kalra et al. 1991; Kaneko et al. 1992; Magnenat Thalmann et al. 1988; Pelachaud et al. 1991). These systems are simple but non-interactive and thus are not very appropriate for real time animation. In addition, fine-tuning an animation is difficult when merely editing the script, as there exists a non-trivial relation between textual description and animation results. Recently several authors have proposed new facial animation techniques which are based on the information derived from human performances. The information extracted is used for controlling the facial animation. These performance driven techniques provide a very realistic rendering and motion of the face. Williams (1990) used a texture map based technicue with points on the surface of the real face. Mase and Pentland (1990) apply optical flow and principal direction analysis for lip reading. Terzopoulos and Waters (1991) reported on techniques for estimating face muscle contraction parameters from video sequences. Kurihara and Arai (1991) introduced a new transformation method for modeling and animating the face using photographs of an individual face. Waters and Terzopoulos (1991) modeled and animated faces using scanned data obtained from a radial laser scanner. Saji et al. (1992) introduced a new method called "Lighting Switch Photometry" to extract 3D shapes from the moving human face. Kato et al (1992) use isodensity maps for the description and the synthesis of facial expressions. However, these techniques do not process the information extraction in real-time. Real-time facial animation driven by an interactive input device was reported by DeGraf (1989), but the external control on animation is very limited when used in isolation. Though it provides high accuracy for timings, it is extremely difficult to edit. Systems driven by speech (Lewis 1992; Hill et al. 1988) are focused in lip-synchronization and speech decomposition into phonemes. These are adequate when animation involves only speech. What would in fact be more desirable, is a system which can encapsulate different kinds of animation specifications and control mechanisms. Such a system would meet the needs of the animator for almost every situation by giving access to the different means of control. The tracking of a live video sequence may provide the basic sequence of a synthetic animation, textual data may produce speech with audio feedback, a hand gesture may govern the gesture motion of the head and eyes, and so on. Here, our attempt is to present how the information from different sources can be related and controlled to give a sequence of animation. As there does not exist what one can refer to as the 'best' framework for motion control for facial animation, this suggests having an open system where one can try several possibilities and chose the one which is subjectively the 'best.'

In order to gain flexibility and modularity in the execution of the system we need a high degree of interaction. We present some of the advanced input accessories which provide natural interaction and thus intuitive control. 3D interaction is already quite popular for many applications, and here we integrate some of the novel paradigms to experiment in the context of controlling facial animation. Possibilities for control with different interactive situations are examined; e.g. gesture dialogue using a DataGlove and musical streams from a MIDI-keyboard. We believe that it is more important to provide a wide range of interaction components than to enforce a particular style of interface. One of the interactive systems for facial expressions presented by deGraf (1989) contains the philosophy of using various puppet interfaces to drive facial animation, however, it seems to have hard wiring of devices for manipulations, which restricts flexibility and interchangeability of different device components. Figure 1 shows an example of facial expression.

A Multimedia Architecture For Facial Expressions

Facial movements rely on perception-driven behaviors. Cognitively, these can be understood as externalization or manifestation of verbal or non-verbal communication agents on a face. These agents activate certain channels of a face associatively which in turn triggers the relevant muscles and which eventually deforms the face. In a computational model, such a behavior can be interpreted as translating behavioral or cerebral activity into a set of functional units which embody the necessary activity-information. The resulting actions are then combined in a sequence of discrete actions which when applied cause the necessary movements on the face. In our system we model such a behavior by separating facial animation into three major components, namely face model, animation controls and composer. The face model primarily describes the geometric structure of the face, deformation controller and muscle actions. The model receives streams of actions to perform. These actions are decomposed into the required muscle actions and a new instance of the face is derived for each frame.

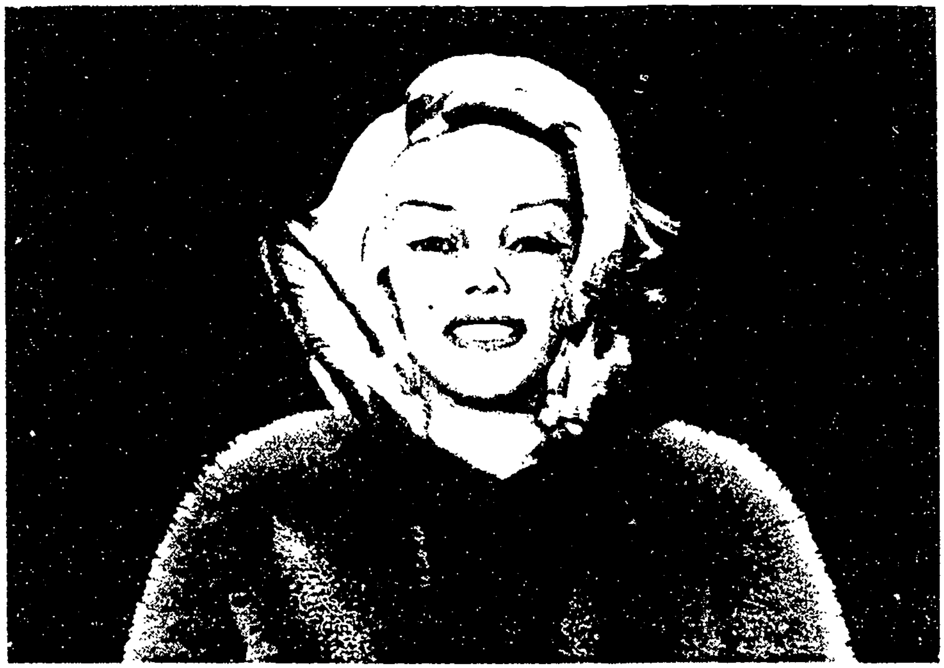

Fig.1. Facial expression

The animation controls specify animation characteristics (Magnenat Thalmann and Thalmann 1991). A facial animation system needs to incorporate adequate knowledge about its static and dynamic environments to enable animators to control its execution with (maybe) predefined, yet flexible set of commands. The system's structure therefore should embed such a know-how in a natural way. In order to satisfy this need, our system employs hierarchical structure and modular design.

Animation Control and Input Accessories

As the system considers the specification of different input components separately from the animation, we can try various ways of controlling the animation. Also, various input components can be used at the same time. Such an approach provides a platform where we can experiment with new methods of control. This analysis may allow us to identify the kind of access we may require for defining and controlling varying levels of abstractions for computational animation models in virtual environments. These accessing components may demand different kinds of interaction which may establish the need for experimentation with several types of devices. As no single mode of control can give completeness, such a test-bed environment can evaluate what device can be used for which means of control. Possibility of composing and mixing different types of controls enhances the reconfigurability of the system. This also allows cooperative group tasks for animation, where more than one person can control the animation in real time. We present here four types of input accessories we have tried. All have some advantages and disadvantages. As there could be simultaneous use of different types at the same type we can overcome some of the disadvantages.

Script-based animation

Script is a standard method of specifying animation consisting of different types of entities. It is like a special language using a few key words for specific operations. Most automated facial animation systems employ this approach.

In our system a language HLSS (High Level Script Scheduler) (Kalra et al. 1991) is used to specify the synchronization in terms of an action and its duration. From action dependence, the starting and the ending times of an action can be deduced. The general format of specifying an action is as follows: while <duration> do <action>. The duration of an action can be a default duration, a relative percentage of the default duration, an absolute duration in seconds, or deduced from other actions preceding or succeeding the present action. The starting time of each action can be specified in different ways, for example, sequentially or parallel using the normal concepts of "fork" and "end" employed in scheduling problems.

One of the major advantages of such an input accessory is that it is in text form. Users can very conveniently change it by editing a text file. On the other hand, being non-interactive it is not possible to change certain parameters while the script is running. Therefore, it is not very suitable for real time animation. It is more useful for background animation.

Animation control through MIDI-Keyboard

MIDI-keyboards can be another type of source for controlling the animation. The keyboard has a number of keys enabling us to associate several parameters with the keys. Activation of each key gives two kinds of information: initial velocity with which the key is hit and the pressure variation. System G^6 a real-time, video animation system uses Korg M1 keyboard to move different parts of the mask of a face. However, the system has special purpose processing hardware to perform the animation.

In our system, this device can be used as direct manipulator for MPAs, each key may be assigned to an MPA and the initial velocity of the key may be attached to its intensity. This type of control is at a rather low level. Higher level control can also be obtained by assigning the keys to expressions and phonemes. Fig 5 shows a sequence of facial animation using the key presses of the MIDI-keyboard. Here, some of the expressions and phonemes have been associated with particular keys of the keyboard, the intensity of the expression or phoneme is governed by the velocity of the key hit. The duration of an expression can be determined by the duration for which the key is pressed. That means an expression starts with the intensity corresponding to the velocity of the key hit and continues until the key is released. At present, the pressure variation of the key hit is not being used, however, we intend to include it to modulate the intensity of an expression during its execution. Sound output from MIDI can provide the needful feedback at the execution of an action. At expression level, it may reflect its intensity. It may be interesting to use array of keys to control a single emotion, here, the mapping of keys would be with the included expression instances or channels, for example, an emotion containing eye, head and mouth motion, each may be associated with a set of keys respectively.

The advantage of such a device is that it provides a number of keys which can be assigned individually to the motion parameters. This gives simultaneous control on many parameters. But at the same time it demands hardwiring of the meaning of desirable action with a particular key. Each time a modification or an extension in the control method will require reconfiguration of the meaning of keys. Also, one-dimensional arrangement of keys in the keyboard forces a certain type of order in the manner of control. For music the ordering is with respect to frequency. However, it is not evident how to arrange facial expressions in a one-dimensional fashion, consequently this constrains the use of the device in its natural form.

Postures and Gestures Dialogue

The type of control presented in the previous section is directly dependent on the physical structure of the device used: the mapping between a user's actions and animation controls is obtained by associating a meaning to the various key presses. This device dependency is a factor that limits the animator's expressiveness.

The use of devices which simply sense user's motions, and the use of adaptive pattern recognition can overcome these problems. This gives more freedom to the animator and allows the definition of the mapping between sensor measurements and interpretations to be more complex. Hand gesture recognition is a domain where these techniques can be used. The use of hand-gestures can provide non-verbal cues for a natural human-computer interaction.

In our system, we use posture recognition on data obtained from the DataGlove to obtain categorical and parametric information to drive facial animation. In the ensuing sections we present briefly the posture recognition technique, its continuous classification and application to facial animation control.

Once a posture is recognized, parametric information can be extracted from the location of hand and how it moves. This information can then be used to drive the facial animation. The type of posture as the categorical information can be associated with a type of action performed by the face. We experimented with two types of controls: direct control at the expression level and higher level control at the emotion level.

The recognition technique here is based on multi-layer perceptrons (MLPs) (Rumelhart et al. 1986), a type of artificial neural network which is potentially able to approximate any real function (Cybenko 1989).

One of the main disadvantages using gesture dialogue as a control means for facial animation is the poor precision offered by the DataGlove. This prohibits the use of this device for fine-tuning. However, for global action manipulation, this provides a natural means of nonverbal communication and interaction method. For duration control of emotion, it is better to associate it with the duration of the gesture itself so as to establish a temporal correspondence between animator's action and the action executed.

Live Video Digitizing and Human Performances

Another approach consists of recording a real human face using a video input like the Live Video Digitizer and extracting from the image the information necessary to generate similar facial expressions on a synthetic face. The problem with this approach is that the image analysis is not easy to perform in real-time. Our recognition method is based on snakes as introduced by Terzopoulos and Waters (1991). A snake is a dynamic deformable 2D contour in the x-y plane. A discrete snake is a set of nodes with time varying positions. The nodes are coupled by internal forces making the snake acting like a series of springs resisting compression and a thin wire resisting bending. To create an interactive discrete snake, nodal masses are set to zero and the expression forces are introduced into the equations of motion for dynamic node/spring system. Terzopoulos and Waters make it responsive to a force field derived from the image. They express the force field which influences the snake's shape and motion through a time-varying potential function. To compute the potential, they apply a discrete smoothing filter consisting of 4-neighbor local averaging of the pixel intensities allowed by the application of a discrete approximation.

Our approach is different from Terzopoulos-Waters approach because we need to analyze the emotion in real-time. Instead of using a filter which globally transforms the image into a planar force field, we apply the filter in the neighborhood of the nodes of the snake. We only use a snake for the mouth; the rest of the information (jaw, eyebrows, eyes) is obtained by fast image-processing techniques.

Implementation

The System is written in C with the interface built on top of the Fifth Dimension Toolkit (Turner et al. 1990). The various input components are independent processes running on UNIX workstations. The communication between the processes is done through sockets in stream mode (sequenced two-way communication based on byte streams) using the Internet protocol. Figure 1 shows the components of a proposed system where each component is a different 3D device or a media. The input components communicate with the central process through inter process communication (IPC). A command IPC-server starts the server on the machine, the command is executed. The server basically accepts and distributes messages to all its clients. The output of each component is further processed by IPC-filter to get the desired mapping between the IPC outgoing messages of each components to the incoming action streams to the central process. The central process basically composes the IPC messages coming from different sources and produces the relevant stream of actions (MPAs) to be performed by the face model. Sensory feedback can be provided to users by different means, e.g. real time animation serves as the visual feedback, MIDI output may provide audio feedback and a text output may provide the final animation sequence script. IPC protocol allows the flexibility of developing each component individually and then hook up with the main process.

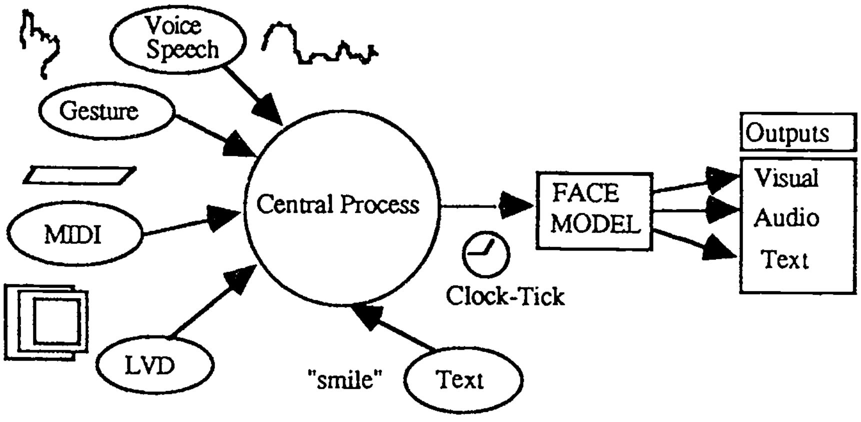

Fig. 2. Proposed system with multimedia components for facial animation control.

References

Cybenko G (1989) Approximation by Superposition of a Signoidal Function, Math, Control Signals and Systems (2), pp. 303-314.

deGraf B (1989) in State of the Art in Facial Animation, SIGGRAPH '89 Course Notes No. 26, pp. 10-20.

Hill DR, Pearce A, Wyvill B (1988), Animating Speech: An Automated Approach Using Speech Synthesized by Rules, The Visual Computer, Vol. 3, No. 5, pp. 277-289.

Kalra P, Mangili A, Magnenat-Thalmann N, Thalmann D (1991) SMILE: a Multilayered Facial Animation System, Proc. IFIP Conf. on Graphics Modeling, Tokyo, Japan, pp.189-198.

Kaneko M, Koike A, Hatori Y (1992) Automatic Synthesis of Moving Facial Images with Expression and Mouth Shape Controlled by Text, Proc CGI '92, Tokyo (Ed T L Kunii), pp. 57-75.

Kato M, So I, Hishinuma Y, Nakamura O, Minami T (1992) Description and Synthesis of Facial Expression based on Isodensity Maps, in: Tosiyasu L (ed): Visual Computing, Springer-Verlag, Tokyo, pp.39-56.

Kurihara T, Arai K (1991), A Transformation Method for Modeling and Animation of the Human Face from Photographs, Proc. Computer Animation '91 Geneva, Switzerland, Springer-Verlag, Tokyo, pp. 45-57.

Lewis JP (1992), Automated Lipsynch: Background and Techniques, The Journal of Visualization and Computer Animation, Vol. 2, No. 4, pp. 118-122,.

Magnenat Thalmann N, Primeau E, Thalmann D (1988), Abstract Muscle Action Procedures for Human Face Animation, The Visual Computer, Vol. 3, No. 5, pp. 290-297.

Magnenat Thalmann N, Thalmann D (1991) Complex Models for Animating Synthetic Actors, IEEE Computer Graphics and Applications, Vol.11, No5, pp.32-44.

Mase K, Pentland A (1990) Automatic Lipreading by Computer, Trans. Inst. Elec. Info. and Comm. Eng.,vol. J73-D-II, No. 6, pp. 796-803..

Pelachaud C, Badler NI, Steadman M (1991), Linguistic Issues in Facial Animation, Proc. Computer Animation '91, (Eds Magnenat-Thalmann N and Thalmann D), Springer, Tokyo, pp. 15-30

Rumelhart D. E., Hinton G. E., Williams R. J. (1986), Learning Internal Representations by Error Propagation. In Rumelhart D. E., McClellend J. L. (eds.), Parallel Distributed Processing, Vol. 1, pp.318-362.

Saji H, Hioki H, Shinagawa Y, Yoshida K, Kunii TL (1992) Extraction of 3D Shapes from the Moving Human face Using Lighting Switch Photometry, in Magnenat Thalmann N , Thalmann D (eds) Creating and Animating the Virtual World, Springer-Verlag Tokyo, pp. 69-86.

Terzopoulos D, Waters K (1991) Techniques for Realistic Facial Modeling and Animation, Proc. Computer Animation '91,(Eds Magnenat-Thalmann N and Thalmann D), Springer, Tokyo,, pp. 59-74.

Turner R, Gobbetti E, Balaguer F, Mangili A, Thalmann D, Magnenat-Thalmann N (1990), An Object-Oriented Methodology Using Dynamic Variables for Animation and Scientific Visualization, Proc CGI 90, Singapore, (Eds Chua TS, Kunii TL), pp. 317-328.

Williams L (1990), Performance Driven Facial Animation, Proc SIGGRAPH '90, Computer Graphics, Vol. 24, No. 3, pp. 235-242.

Acknowledgment

The research is supported by Le Fonds National pour la Recherche Scientifique.