SMILE: A Multilayered Facial Animation System

Prem Kalra, Angelo Mangili, Nadia Magnenat-Thalmann, Daniel Thalmann

ABSTRACT

This paper describes a methodology for specifying facial animation based on a multi-layered approach. Each successive layer defines entities from a more abstract point of view, starting with phonemes, and working up through words, sentences, expressions, and emotions. Finally, the high level layer allows the manipulation of these entities, ensuring synchronization of the eye motion with emotions and word flow of a sentence.

Keywords: Facial Animation, Speech, Synchronisation

1. INTRODUCTION

1.1 The human facial motion problems

Three dimensional modeling and animation of the human face has been one of the major research fields in human animation. One of the ultimate objectives of this research is to model exactly the human facial anatomy and movements to satisfy both structural and functional aspects. This, however, involves many problems to be solved simultaneously. Some of these are: the geometric representation of the actual facial structure and shape, the modeling of interior facial details such as muscles, bones, and tissues, the incorporation of dynamics for the motion involved in making expressions, and the synchronization of speech and emotions.

This paper addresses the problems associated with facial animation and discusses some methods to solve them using a multi-layered animation system. A language for synchronizing speech, emotions and eye motions is developed to provide a way to naturally specify animation sequences.

1.2 Complexity of the human facial motion

The difficulty of the modeling of human facial motion is mainly due to the complexity of the physical structure of the human face. Not only are there a great number of specific bones, but there is also an interaction between muscles and bones and between the muscles themselves. This complex interaction results in what are commonly called facial expressions. To create a model of these expressions, we must first analyze in more detail the role of the components of the human face: bones, muscles, skin and organs.

Bones in the face may be divided into two main parts: the cranium itself, which surrounds the brain and the eyes, and the lower jaw which is articulated and plays an important role in speech. These bones force a more or less rigid shape to the skin which may only slip on the cranium. The skin covers the bony structure: it is elastic and flexible. There are 40 Muscles which are an intermediate between the skin and the bones. They force the skin to move in a certain direction and in a given way. Face muscles have various shapes: long, flat, wide, thin, etc. In addition to their action, muscles also have volume.

1.3 Basic facial animation

As stated by Ekman (1975), humans are highly sensitive to visual messages sent voluntarily or involuntary by the face. Consequently, facial animation requires specific algorithms able to render with a high degree of realism the natural characteristics of the motion. Research on basic facial animation and modeling has been extensively studied and several models have been proposed.

For example, in the Parke models (1975,1982) the set of facial parameters is based on both observation and the underlying structures that cause facial expression. The animator can create any facial image by specifying the appropriate set of parameter values. Motions are described as a pair of numeric tuples which identify the initial frame, final frame, and interpolation. Pearce et al. (1986) introduced a small set of keywords to extend the Parke model.

Platt and Badler (1981) have designed a model that is based on underlying facial structure. The skin is the outside level, represented by a set of 3D points that define a surface which can be modified. The bones represent an initial level that cannot be moved. Between both levels, muscles are groups of points with elastic arcs.

Waters (1987) represents the action of muscles using primary motivators on a non-specific deformable topology of the face. The muscle actions themselves are tested against FACS (Facial Action Coding System) which employs action units directly to one muscle or a small group of muscles. Two types of muscles are created: linear/parallel muscles that pull and sphincter muscles that squeeze.

Magnenat-Thalmann et al. (1988) defined a model where the action of a muscle is simulated by a procedure, called an Abstract Muscle Action procedure (AMA), which acts on the vertices composing the human face figure. It is possible to animate a human face by manipulating the facial parameters using AMA procedures. By combining the facial parameters obtained by the AMA procedures in different ways, we can construct more complex entities corresponding to the well-known concept of facial expression.

Nahas et al. (1987) propose a method based on the B-spline. They use a digitizing system to obtain position data on the face from which they extract a certain number of points, and organize them in a matrix. This matrix is used as a set of control points for a 5-dimensional bicubic B-spline surface. The model is animated by moving these control points.

1.4 Synchronization between speech, emotions and eye motion

Lip synchronization in computer animation was first studied by Parke (1982) using its parameterized model. Three recent papers report studies of problems in computer animated speech: Hill et al. (1988) introduce an automatic approach to animate speech using speech synthesized by rules; the extra parameters needed to control lips, jaw and facial expression are simply added into the table of parameters needed to control the speech itself. Lewis and Parke (1987) automate the lip synchronization between computer generated imagery and real speech recorded from a real actor. Magnenat-Thalmann et al. (1988) describe a lip synchronization based on AMA procedures.
Previous works on facial animation (Pearce et al. 1986) propose methods where synchronization is manual (e.g. "do action from frame x to y"). However, the parametrization of an emotion is hard to control because once defined it is always played the same way.

Magnenat-Thalmann et al. (1988) used a collection of multiple tracks. A track is a chronological sequence of keyframes for a given facial parameter. Tracks are independent, but they may be mixed in the same way as sound is mixed in a sound studio. With such an approach it is easy to define, for example, an eye movement in an expression corresponding to a phoneme. Although the approach works and was used for the film Rendez-vous à Montréal (Magnenat-Thalmann and Thalmann, 1987), the process of synchronization is manual and must be performed by the animator.

This paper describes a method that can be used to solve the synchronization problem.

2. THE MULTI-LAYERED APPROACH

2.1 The layers

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 the results somewhat unpredictable. This suggest that more **abstrac**t **entities** should be defined in order to create a system that can be easily manipulated. A multi-layered approach is convenient for this.

The system proposed in this paper is independent of the animation system. The results are specified in terms of perceptible movements (e.g. elevate the eyebrows with an intensity of 70%).

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. This approach has the following advantages:

- the system is extensible.
- the independence of each layer allows the behavior of an element of the system to be modified without impact on the others.

Five layers are defined in our approach: (Fig. 1)

layer 0: definition of the entity **muscle** or equivalent.
layer 1: definition of the entity **minimal perceptible action**.
layer 2: definition of the entities **phonemes** and **expressions**.
layer 3: definition of the entities **words** and **emotions**.
layer 4: **synchronization** mechanism between emotions, speech and eye motion.

2.2 Layer 0: Abstract muscles

This level correspond to the basic animation system. In our case, the software implementation is currently based on the Abstract Muscle Action procedures as already introduced in a previous work (Thalmann, 1988). These actions are very specific to the various muscles and give the illusion of the presence of a bony structure. More generally, basic facial animation is based on independent facial parameters simulated by specific AMA procedures. A problem with such an approach is that deformations are based on empirical models and not on physical laws.

An interactive and more general model is currently under development. The model consists on a generic representation for the facial components, namely skin, bones, and muscles. The skin is represented as a polygonal mesh. The skull is considered rigid and immobile except the mandible. The muscles are the links between the skin points and the bone. These muscles act as directional vectors for determining the deformations on the skin and their direction can be changed interactively by the designer.

Using region mapping, the designer can interactively map muscles on the model. With another interactive interface the designer can compose the actions of various muscles resulting into an expression. Each muscle has parameters like max, min and value for the amount of contraction. The action units (AUs) of FACS may be used as the guide for constructing the expressions. The deformation of the muscles is based on the simple force equation (F=kU). To improve the realism of the simulation, we are currently considering the possibility of using finite elements methods.

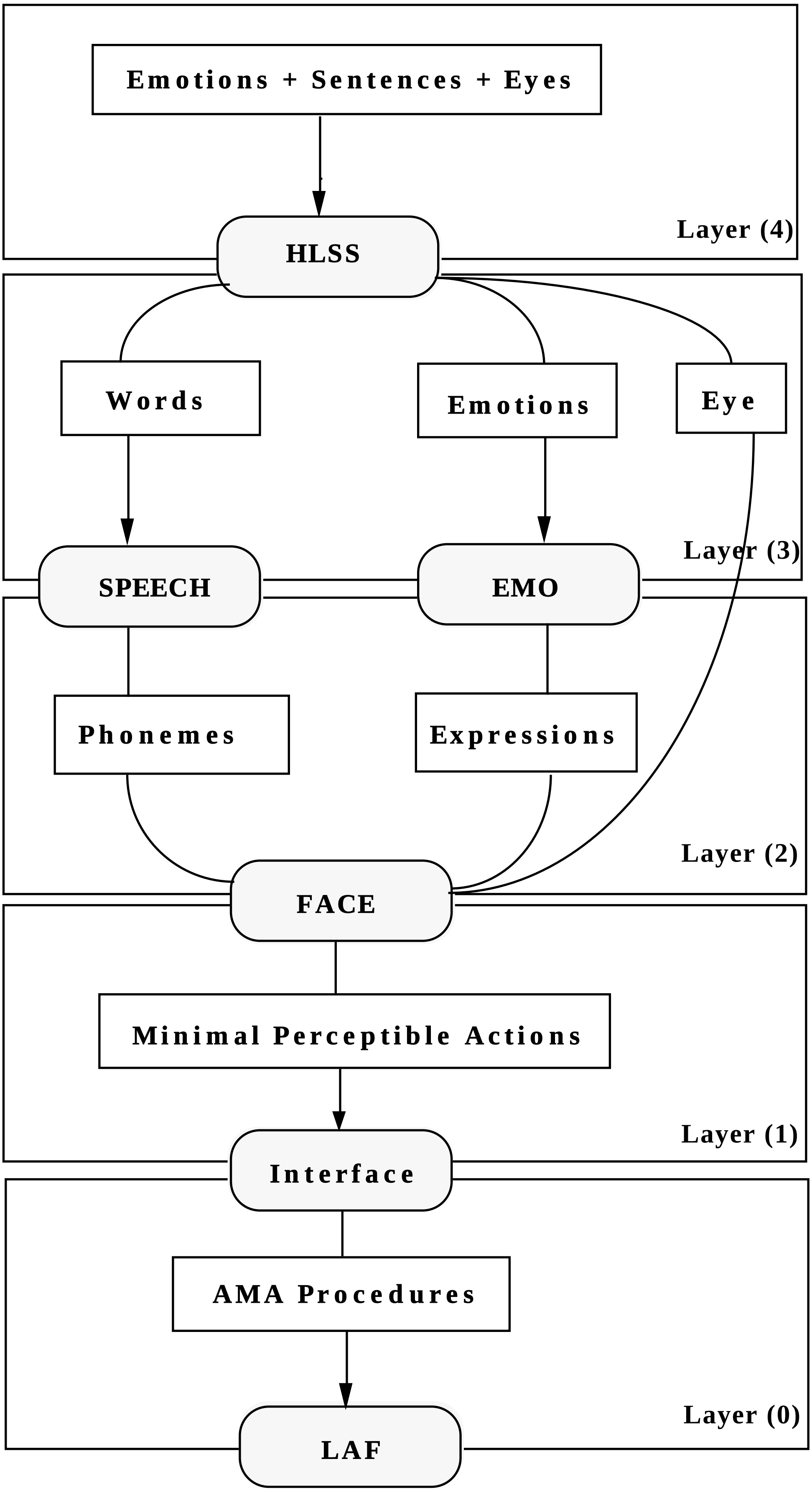


Fig. 1: Entities defoned at each level

2.3 Layer 1: Minimal Perceptible Action

A minimal perceptible action is a basic facial motion parameter. The range of this motion is normalized between 0% and 100% (e.g. raise the right eyebrow 70%). An instance of the minimal action is of the general form <frame number> <minimal action> <intensity>. The animation is carried out by traditional keyframing.

2.4 Layer 2: Facial snapshot

A facial snapshot is obtained by specifying the value of each minimal action. Once defined, a snapshot has the form that follows: <frame number> <snapshot> <intensity>. It should be noted that several snapshots may be active at the same time, this allows for example to specify a phoneme and a smile at the same time. (Fig. 2)

2.4.1 Layer 2a: Phonemes

A phoneme snapshot is a particular position of the mouth during a sound emission. It is possible to represent a phoneme instance by a set of minimal actions interacting with the mouth area.
e.g.

```
    [ snapshot    pp =>
        [ action            raise_sup_lip   30%]
        [ action            lower_inf_lip   20%]
        [ action            open_jaw        15%]
    ]
```

Normally, each word of a language has its phonetic representation according to the International Phoneme Alphabet. A representative subset of this is encoded in form of snapshots.

2.4.2 Layer 2b: Expressions

An expression snapshot is a particular position of the face at a given time. This is generated by a set of minimal actions in the same way as phonemes.

Based on Ekman's work on facial expressions (Ekman), several primary expressions may be classified: surprise, fear, disgust, anger, happiness, sadness (Fig. 3-6). Basic expressions and variants may be easy defined using snapshots.

2.5 Layer 3: Sequences of snapshots

2.5.1 Layer 3a: Words

As already mentioned, a word may be specified by the sequence of component phonemes. However there is no algorithm for automatic decomposition of a word into phonemes (Allen 1987). The solution to this problem is to use a dictionary that may be created using a learning approach: each time an unknown word is detected, the user should enter the decomposition, which is then stored in the dictionary.

Another problem is the adjustment of the duration of each phoneme relative to the average duration of the phoneme and its context in the sentence (previous and next phonemes). Several heuristic methods have been proposed to solve this problem by researchers in the area of speech synthesis from text (Allen 1987). In our case the system is able to generate the correct sequence of phonemes in the given time using a specification such as "How are you (pause 200 ms) Juliet". Optional commands may act on intensity, duration and emphasis of each word and pauses may be also added in order to control rhythm and intonation of the sentence.

2.5.2 Layer 3b: Emotions

An emotion is defined as the evolution of the human face over time: it is a sequence of expressions with various durations and intensities. The emotion model proposed here is based on the general form of an envelope: signal intensity = f(t) (Ekman 1978b).

An envelope may be defined using 4 stages:

- ATTACK:	transition between the absence of signal and the maximum signal.
- DECAY:	transition between the maximum signal and the stabilized signal.
- SUSTAIN:	duration of the active signal.
- RELEASE:	transition to the normal state.

For each stage of an emotion, the average duration of the stage and the sequence of expressions may be specified. One of the major problems is how to parameterize the emotions. To solve this, we introduce the concept of **generic emotion**.

An emotion has a specific average duration, but it is context-sensitive. For example, a smile may have a 5-6 second duration, but it may last 30 seconds in case of a laughable situation. It is also important to note that the duration of each stage of the emotion is not equally sensitive to the time expansion. If you expand the overall duration of the emotion envelope, the ATTACK and RELEASE stages will expand proportionally less than the SUSTAIN stage. To take into account this proportional expansion, we introduce a **sensitivity factor** associated to each stage.

In order to naturally render each emotion, mechanisms based on **statistical distribution** have been introduced. For example, we may define a stage duration of 5 ± 1 seconds according to a uniform distributed law, or an intensity of 0.7± 0.05 according to a Gauss distribution.

These parameterization mechanisms allow the creation of generic emotions. Once a generic emotion is introduced in the emotion dictionary, it is easy to produce an instance by specifying its duration and its magnitude.

2.6 Layer 4: Synchronization mechanism

We already mentioned the needs for synchronizing the various facial actions: emotions, word flow in a sentence and eye motion. In this layer we introduce mechanisms for specifying the **starting time**, the **ending time** and the **duration** of an action. This implies that each action can be executed independently of the current state of the environment, because the synchronization is dependant on time alone.

3. THE HLSS LANGUAGE

HLSS (High Level Script Scheduler) is a formalism for specifying the synchronization and the dependence between the various actions. An action is an entity defined by a starting time and a duration of execution using the general model:

while <duration> **do** <action>.

3.1 Action duration

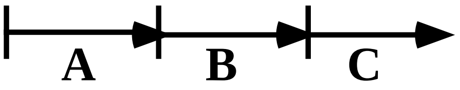
The action duration may be specified in various ways:

- no specification (corresponds to *default value*)
  e.g. [**emotion** SMILE]
- *relative* duration (% of the default value)
  e.g. [**emotion** SMILE **while** 80% **myduration**]
- *absolute* duration
  e.g. [**emotion** SMILE **while** 30 seconds]
- *relative to another action*
  e.g. [**emotion** SMILE **while** [**say** "How are you"]]

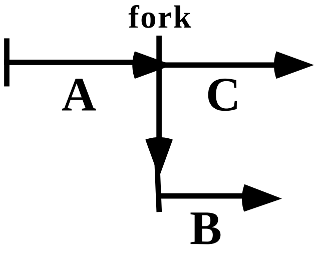
3.2 Starting time of an action

The starting time of an action may be specified as follows:

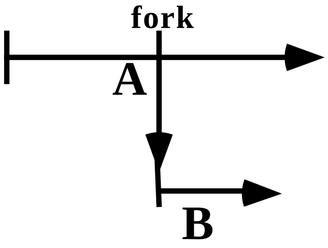
[A] [B] [C]



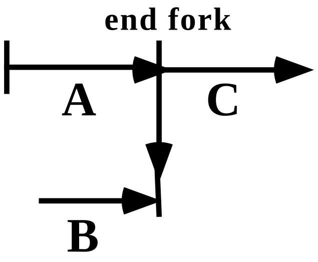
[A] [fork B] [C]



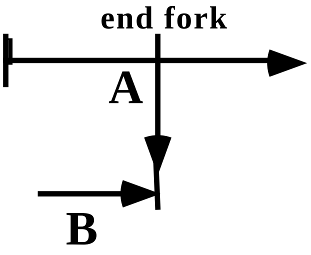
[A... [fork B] ...]



[A] [end fork B] [C]



[A... [end fork B] ...



3.3 Action types

3.3.1 Actor synchronization

A mechanism of actor names allows several actors to be synchronized at the same time.

```
e.g.    [actor JULIET while
            [       [ say  "What's the time?"]
                    [ actor ROMEO while
                            [ say  "It's midnight..."]
                    ]
                    [ say  "Oh, it's late..."]
            ]
        ]
```

3.3.2 Emotion synchronization

This action generates a facial emotion of an actor; the emotion is assumed to be in the emotion dictionary.

```
e.g.    [ emotion  FEAR]
        [ emotion  ANGER while
            [ say  “Aghh“]
        ]
```

3.3.3 Sentence synchronization

For the purpose of the synchronisation of the word flow with other expressions, each word is considered as an independent action with a starting time and a duration. Therefore, it is possible to execute any action between two words.

```
e.g.    [ say   "My name is Juliet"
            [ emotion  WINK]
            "and your's ?"
        ]
```

4. IMPLEMENTATION

This multi-layered facial animation system is part of a new system for the intelligent animation of human characters in their environment, developed in the C language for the IRIS Silicon Graphics Workstations network at the Swiss Federal Institute of Technology in Lausanne and the University of Geneva. Some of the aspects of this system include task-level animation, behavioral walk based on a vision approach, and local deformations of the body.

5. CONCLUSION

The facial animation system described in this paper is based on a multi-layered model. At each level, the degree of abstraction increases. This results in a system where the degree of complexity is relatively low and therefore the animation is simple to specify. Also, the defined entities correspond to intuitive concepts such as phonemes, expressions, words and emotions, which make them natural to manipulate.

We also have introduced a manipulation language HLSS which provides simple synchronization mechanisms. These mechanisms completely hide the time specification. Moreover the mechanisms are general and may be extended to any type of action which may be calculated independently from the actor position and the the environment state.

6. ACKNOWLEDGEMENTS

The authors are grateful to Enrico Gobbetti and Russell Turner for the revision of the english text. The research was supported by le Fonds National Suisse pour la Recherche Scientifique, the Natural Sciences and Engineering Council of Canada, and the FCAR foundation.

7. REFERENCES

Allen J. and al. (1987), From Text to Speech: The MITalk System, Cambridge University Press.
Darwin C. (1872), The Expression of Emotion in Man and Animals, Curr. ed: University of Chicago Press, 1965.
Guenter B.(1889), A System for Simulating Human Facial Expression, State-of-the-art in Computer Animation, Springer-Verlag, pp.191-202.
Ekman P. and Friesen WV. (1975), Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues, Prentice-Hall.
Ekman P and Friesen W (1978a) Facial Action Coding System, Consulting Psychologists Press, Palo Alto.
Ekman P., Friesen WV. (1978b), Facial Action Coding System, Investigator's Guide Part 2, Consulting Psychologists Press Inc.
Ekman P. (1980), L'Expression des Emotions, La Recherche, No 117, pp. 1408-1415.
Hill DR, Pearce A and Wyvill B (1988) Animating Speech: an Automated Approach Using Speech Synthesised by Rules, The Visual Computer, Vol.3, No.5
Jonsson A. (1986), A Text to Speech System Using Area Functions and a Dictionary, Göteborg.
Lewis JP, Parke FI (1987) Automated Lip-synch and Speech Synthesis for Character Animation, Proc. CHI '87 and Graphics Interface '87, Toronto, pp.143-147.
Magnenat-Thalmann N, Thalmann D (1987) The Direction of Synthetic Actors in the film Rendez-vous à Montréal, IEEE Computer Graphics and Applications, Vol.7, No.12.
Magnenat-Thalmann N, Primeau E, Thalmann D (1988), Abstract Muscle Action Procedures for Human Face Animation, The Visual Computer, Vol.3, No.5
Nahas M, Huitric H and Saintourens M (1988) Animation of a B-spline Figure, The Visual Computer, Vol.3, No.5.
Parke F.I. (1972) Animation of Faces, Proc. ACM Annual Conf., Vol.1.
Parke F.I. (1974) A Parametric Model for Human Faces, PhD dissertation, University of Utah, department of Computer Science.
Parke FI (1975) A Model for Human Faces that allows Speech Synchronized Animation, Computers and Graphics, pergamon Press, Vol.1, No.1, pp.1-4.
Parke FI (1982) Parameterized Models for Facial Animation, IEEE Computer Graphics and Applications, Vol.2, No.9, pp.61-68.
Pearce A, Wyvill B, Wyvill G and Hill D (1986) Speech and expression: a Computer Solution to Face Animation, Proc. Graphics Interface '86, pp.136-140.
Platt S, Badler N (1981) Animating Facial Expressions, Proc. SIGGRAPH '81, pp.245-252.
Waters K (1987) A Muscle Model for Animating Three-Dimensional Facial Expression, Proc. SIGGRAPH '87, Vol.21, No.4, pp.17-24.

Fig. 2: Two different snapshots mixed together

Fig 3: Smile (mouth region)

Fig 4: Disgust (mouth region)

Fig 5: Surprise (mouth region)

Fig 6: Anger (mouth region)

Prem Kalra is a researcher at MIRALab, University of Geneva. He received his MSc in Computer Science from the University of New Brunswick. Then, he was researcher at the University of Arizona. His research interests include computer animation and geometric modeling.
E-mail: `kalra@uni2a.unige.ch`

Angelo Mangili is a researcher at the Computer Graphics Laboratory of the Swiss Federal Institute of Technology in Lausanne, Switzerland. He received his diplôme d'ingénieur informaticien from the same institute. His research interests include computer animation and software engineering.
E-mail: `mangili@eldi.epfl.CH`

Nadia Magnenat Thalmann is currently full Professor of Computer Science at the University of Geneva, Switzerland and Adjunct Professor at HEC Montreal, Canada. She has served on a variety of government advisory boards and program committees in Canada. She has received several awards, including the 1985 Communications Award from the Government of Quebec. In May 1987, she was nominated woman of the year in sciences by the Montreal community. Dr. Magnenat Thalmann received a BS in psychology, an MS in biochemistry, and a Ph.D in quantum chemistry and computer graphics from the University of Geneva. She has written and edited several books and research papers in image synthesis and computer animation and was codirector of the computer-generated films *Dream Flight*, *Eglantine*, *Rendez-vous à Montréal*, *Galaxy Sweetheart*, *IAD and Flashback*. She served as chairperson of Graphics Interface '85, CGI '88, Computer Animation '89 and Computer Animation '90.
E-mail: `thalmann@uni2a.unige.ch`

Daniel Thalmann is currently full Professor and Director of the Computer Graphics Laboratory at the Swiss Federal Institute of Technology in Lausanne, Switzerland. Since 1977, he was Professor at the University of Montreal and codirector of the MIRALab research laboratory. He received his diploma in nuclear physics and Ph.D in Computer Science from the University of Geneva. He is coeditor-in-chief of the *Journal of Visualization and Computer Animation*, member of the editorial board of the *Visual Computer* and cochairs the EUROGRAPHICS Working Group on Computer Simulation and Animation. Daniel Thalmann's research interests include 3D computer animation, image synthesis, and scientific visualization. He has published more than 100 papers in these areas and is coauthor of several books including: *Computer Animation: Theory and Practice* and *Image Synthesis: Theory and Practice*. He is also codirector of several computer-generated films.
E-mail: `thalmann@eldi.epfl.CH`

The authors may be contacted at:

Computer Graphics Lab
Swiss Federal Institute of Technology
CH 1015 Lausanne
Switzerland

tel: ++41-21-693-5214
fax: ++ 41-21-693-3909

MIRALab, CUI
University of Geneva
12 rue du Lac
CH 1207 Geneva
Switzerland
tel: ++41-22-787-6581
fax: ++ 41-22-735-3905