MODELING HUMAN FACIAL EXPRESSIONS AT MULTIPLE RESOLUTIONS

Antai Peng and Monson H. Hayes

School of Electrical and Computer Engineering
Georgia Tech, Atlanta, GA 30332-0250

ABSTRACT

Human facial expression modeling has been an active research area recently. Most of the existing systems do not provide an easy way to adjust the model such that different levels of detail of expressions can be modeled. In this paper, we propose a method for modeling the human facial expressions at different resolutions. Our method is based on the FACS developed by Ekman and Freisen [8]. Although the facial expressions we are mainly interested in are those related to speech either directly or indirectly, the modeling method can be extended to apply to facial expressions that are not related to speech.

1. INTRODUCTION

Modeling human facial expressions, especially those expressions that are generated from speech, has been an interesting area of research for quite some time. Practical application such as teleconferencing, where two parties exchange not only auditory signals (speech) but also visual signals (images) , has demonstrated a need to perform intelligent image coding so that a lower bit-rate can be achieved without significant distortion of the images. Another research area is computer user interface design where researchers are concerned with how to improve the computer interface so that the interaction between the computer and the human becomes more natural and spontaneous. Since facial expressions are used in human-human communication, they have been targeted as a possible enhancement to computer interface design. Still another interesting research topic is the so-called Text-To-Speech-To-Facial-Animation (TSFA). A TSFA system is one that, given a text string and a speaker's facial image, generates a sequence of images with the changes of the facial expressions corresponding to the input text string. Although these research applications differ significantly and their goals vary, modeling facial expressions is one of the common important fundamental problems that must be solved. Our research work has been concentrating on modeling the human facial expressions and using them in a TSFA system.

Research in the area of facial expression modeling and animation include pioneering work by F. Parke on parameterized facial animation modeling[1, 2], facial animation based on the Abstract Muscle Action (AMA) model [4], facial image analysis and synthesis using physical and anatomical models [5], knowledge-based facial expressions analysis and synthesis [7] and many others. Moreover, in [6] a facial image synthesis system driven by speech and text is described, and in [3] the problems related to facial expression by drawing parallels between speech synthesis and facial expression synthesis are discussed. Although encouraging results have been obtained thus far from various research groups, most of the systems that have been developed do not provide a simple way to adjust the model such that the facial expressions are modeled differently for different individuals. In these systems it is also very difficult, if not impossible, to re-use one modeling system, which is developed for one application, in another application that has a different set of system requirements.

The method that we are proposing will allow the human facial expressions to be modeled at different levels of details. For applications where bit-rate is a main concern, such as teleconferencing, the facial expressions may be modeled at a low resolution. When higher resolution is desired, such as that required by the movie industry, the proposed method will allow for the production of more realistic and real-life like expressions.

In Section 2 we describe the facial expression representation that we have developed. Using this representation, facial expression at different resolutions can be modeled as discussed in Section 3. Finally, in Section 4 we present some illustrative examples.

2. REPRESENTATION OF FACIAL EXPRESSION

The facial expression representation system that we are developing is based on the "Facial Action Coding System" (FACS) developed by P. Ekman and W. V. Friesen [8]. Using this system, our goal is to map speech phonemes into realistic facial expressions and, eventually, to generate realistic video sequences for spoken languages. In [8], there are mainly four groups of lower face Facial Actions that are related to speech. They are: Vertical Action Units, Horizontal Action Units, Orbital Action Units and Oblique Action Units. The Vertical Action Units we use are as follows:

0-7803-2431-5/95 $4.00

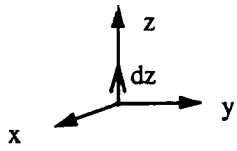

Figure 1 Vertical Action Unit

- AU10 = Upper Lip Raise
- AU15 = Depress Lip Corner
- AU 25 = Mouth Apart
- AU 26 = Mouth Open
- AU 27 = Mouth Wide Open

The Horizontal Action Unit we use is:

- AU20 = Lip Corner Stretch

The Orbital Action Unit that we use is:

- AU 18 = Lip Pucker

and the Oblique Action Units that we are using include:

- AU 12 = Lip Horizontal Stretch

Besides these four categories, there are several Action Units grouped as miscellaneous that are related to the speech. We have considered the following two Action Units:

- AU 29 = Jaw Thrust
- AU 32 = Bite

The Action Units are additive, i.e. AUi can be simply added on top of AUj. Because of this property, the AUs can be viewed as basic components for facial expression. For those expressions correspond to phonemes, the AUs we have listed above form an AU space:

[AU10, AU12, AU15, ..., AU32]

If we let vector v_1 represent AU10, v_2 represent AU12 and so on, then we may form a vector of Facial Action Units as follows:

$$V = [v_1, v_2, v_3, ..., v_{10}]'$$

Since each Facial Action Unit can be specified either as a translation on x, y or z axis or as a rotation with respect to a certain point, it is possible to represent each AU as follows:

$$AU = [dx, dy, dz]'$$

or

$$v = [dx, dy, dz]'$$

For example, a Vertical Action Unit is basically a translation in the z-direction as shown in Figure 1 whereas an Oblique Action Unit consists of two translations in the y-axis and z-axis as shown in Figure 2.

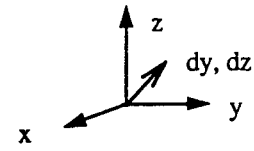

Figure 2 Oblique Action Unit

With this representation for each Action Unit, our AU-space may now be represented as:

$$V = \begin{bmatrix} dx_1 & dy_1 & dz_1 \\ dx_2 & dy_2 & dz_2 \\ \cdots & \cdots & \cdots \\ dx_{10} & dy_{10} & dz_{10} \end{bmatrix}' \quad (1)$$

It is assumed that each phonemic facial expression of interest may be constructed from an appropriate linear combination of AUs. We represent this construction as follows,

$$f = w_1 * v_1 + w_2 * v_2 + ... + w_{10} * v_{10} = VW$$

Here V is defined as in (1). W is a vector of weights that must be determined. Once we have this representation, constructing the facial expression is narrowed down to adjusting the weight vector. For instance, to model the upper lip raising, simply set $w_1 = 1$ and $w_i = 0$, i = 2, 3, ..., 10. To model the mouth opening, we can set $w_4 = 1$ and the rest of the weights to zero.

3. FACIAL EXPRESSION MODELING

As can be seen from our discussion above, this representation of facial expression provides several advantages. First, it is simple and straight forward. A weight of 1 means the corresponding AU is one that is used in the synthesis of a particular facial expression. Secondly, fine tuning the facial expression becomes a problem of simply adjusting the weights. By varying the weights, the same facial expression at different scales can be easily modeled. For instance, lip puckering at a strong and weak scale can be implemented by giving a bigger and smaller weights respectively to the Action Unit 18. Figure 3 shows a strong lip pucker and a weak one.

In addition to the ability to create strong or weak version of the same expression, adjusting the weights can also be used to model different styles of spoken language. This is important for any facial expression system since each person speaks very differently with different facial expressions and lip movements.

Figure 3. A weak and strong Lip pucker

The third advantage of the proposed representation lies in the ease by which the system can be expanded or changed to satisfy different goals. Facial expression modeling can be used in various applications. Video-conference, for instance, requires a high bit-rate while concerns less about the fine details. Weights of less importance can be set to zeros in this case. The same implementation can be used for other applications without modifications but with more accurate weights. Such applications include human computer interaction where distortions of facial image may be less acceptable. Since weights are digitally stored in the system, their accuracy is mainly impacted by the number of bits in which they are coded. For crude facial image, a binary representation using 1 bit per weight is possible. The more bits that are used, the more detailed the model may become and the higher the resolution then can be achieved. Thus, the method we are proposing allows us to model human expression at different levels of details.

It can be seen that this method can be easily extended to model facial expressions non-related to speech. To do that, we will add Action Units belonging to the upper face, i.e., Action Units that cause eyes to open/close , eye brows to raise/lower and so on. We can either add these Action Units to the existing AU space or form a separate one for the upper face.

4. DISCUSSION

Currently, the weights are manually selected. We first give some initial values to the weight vector and construct the expressions. The images generated are then viewed by the human. Weights are then adjusted afterwards. This process continues until the output images look satisfactory to the viewers. This is a tedious and subjective process. Table 1 shows some of the weights that we have used for the modeling of speech and the synthesized lip images.

We are currently investigating methods to automate the weight assigning process by incorporating facial image analysis and image understanding techniques.

Phonemes Weight Vector	Synthesized Lip Images
/e/ or EY (0,0,0,0,1,0,1,0,0,0)	
/f/ (0,0,0,0,0,0,0,0,1,1)	
/ ʌ / or AH (0,0,0,0.5,0,0,1,0,0,0)	

Table 1. Sample Weights for phonemes and synthesized lips. The weight vectors include Action Units AU10, AU12, AU15, AU18, AU20, AU5, AU26, AU27, AU29 and AU32.

5. REFERENCE

1. F.I. Parke, "A model for human faces that allows speech synchronized animation", Computer Graphics, Vol. 1, pp. 3-4 (1975)

2. F.I. Parke, Parameterized models for facial animation", IEEE Computer Graphics Applications, Vol. 2, no. 9, pp. 61-68 (1982)

3. D. Hill, A. Pearce, and B. Wyvill, "Animating speech: A automated approach using speech synthesized by rules", Visual Computer, Vol. 3, pp. 277-287 (1988)

4. N.Magnenat-Thalmann, E. Primeau, and D. Thalmann, "Abstract muscle action procedures for face animation", Visual Computer, Vol. 3, pp. 290-297(1988)

5. D.Terzopoulous, K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models", IEEE Trans-PAMI, Vol. 15, No. 6, pp. 69-579 (1993)

6. S. Morishima, K. Aizawa, and H. Harashima, "A real-time facial action image synthesis system driven by speech and text", SPIE Vol. 1360 Visual Comm. Image Proc., pp. 1151-1158(1990)

7. S. Morishima, K. Aizawa, and H. Harashima, "An intelligent facial image coding driven by speech and phoneme", IEEE ICASSP89, Pres. No.M8.7, pp. 1795-1798 (1989)

8. P. Ekman and W.V. Friesen, "Facial Action Coding System", Consulting Psychologists Press, 1977