An Integrated Framework for Face Modeling, Facial Motion Analysis and Synthesis

Pengyu Hong, Zhen Wen, Thomas Huang
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana Champaign
Urbana, IL 61801, USA
{hong, zhenwen, huang} @ifp.uiuc.edu

ABSTRACT

This paper presents an integrated framework for face modeling, facial motion analysis and synthesis. This framework systematically addresses three closely related research issues: (1) selecting a quantitative visual representation for face modeling and face animation; (2) automatic facial motion analysis based on the same visual representation; and (3) speech to facial coarticulation modeling. The framework provides a guideline for methodically building a face modeling and animation system. The systematicness of the framework is reflected by the links among its components, whose details are presented. Based on this framework, we improved a face modeling and animation system, called the iFACE system [4]. The final system provides functionalities for customizing a generic face model for an individual, text driven face animation, off-line speech driven face animation, and real-time speech driven face animation.

Keywords

Face Modeling, Face Animation, Facial Motion Analysis, Speech to Facial Coarticulation Modeling, iFACE.

1. INTRODUCTION

Graphics based human face provides an effective solution for delivering and displaying multimedia information relating to human face. The applications include 3D model-based very low bit rate video coding for visual telecommunication [1], talking head representation of computer agent [15], and human audio-visual speech recognition [9].

There has been a large amount of research on face modeling and animation. One main task of face modeling is to develop a facial deformation control model for spatially deforming facial surface. The main goal of face animation research is to build a facial coarticulation model for deforming facial surface temporally. To realistically and naturally animate the face model, analysis of real facial motion is required. It is well known that facial coarticulation is highly correlated with the vocal track. Speech as an import media has been used to drive face model. Speech driven face animation not only needs to deal with face modeling and animation but also needs to learn a mapping from audio to facial coarticulation.

1.1 Face Modeling

Human faces are commonly modeled as free-form geometric mesh models [4], [6], parameterized geometric models [12], or physics-based models [7], [14]. Once the coordinates of the control points of the free-form model are decided, the remaining vertices on the model are deformed by interpolation. However, little research has been done in a systematic way to address how to chose interpolation functions, how to adjust control points, and what are the correlations among those control points. Parameterized geometric models calculate the coordinates of the vertices a set of predefined functions. Nonetheless, there is no theoretical basis for designing those functions. Physics-based models simulate facial skin, tissue, and muscles by multi-layer dense meshes. However, the physical models are sophisticated and computationally complicated. In addition, how to decide the parameters of the physics-based face models is an art.

1.2 Face Animation

Once the facial deformation control model is decided, a face model can be animated by temporally adjusting its parameters according to its facial coarticulation model. To naturally and realistically resynthesize facial motion, research turns to performance driven face animation techniques [13], [16] or speech driven face animation [10], [11]. Performance driven face animation uses real facial movements. However, it requires robust automatic facial motion analysis algorithm. Speech driven face animation takes advantage of the correlation between speech and facial coarticulation. It maps audio tracks into face animation sequences. The audio-to-visual mapping can be learned from an audio-visual database. To efficiently collect a large enough audio-visual database, robust facial motion analysis technique is required.

1.3 Facial Motion Analysis

It is well known that tracking facial features based on the low-level facial image features alone is not robust. High-level knowledge model must be used [2], [3], [8]. Those high-level models usually correspond to the facial deformation control models and encode information about possible deformations of facial features. The tracking algorithms extract control information from the tracking results using low-level image features only. The control information is used to deform the face model. The face animation results are fed back and used for facial motion tracking. The final tracking results will be greatly degraded if inaccurate control models are used. To be faithful to the real facial deformations, the high-level models should be learned from real facial deformations.

2. THE INTEGRATED FRAMEWORK

It has been shown above that face modeling, facial motion analysis and synthesis are related to each other. Their research should be carried out in a systematic way. This paper proposes an integrated framework for face modeling, facial motion analysis and synthesis (see Figure 1). Firstly, a quantitative representation of the facial deformations, called Motion Unit (MU), is introduced. MUs are learned from a set of labeled real facial deformations. It is assumed that a facial deformation can be approximated by a linear combination of MUs weighted by MU parameters (MUPs).

MM'01, Sept. 30-Oct. 5, 2001, Ottawa, Canada.

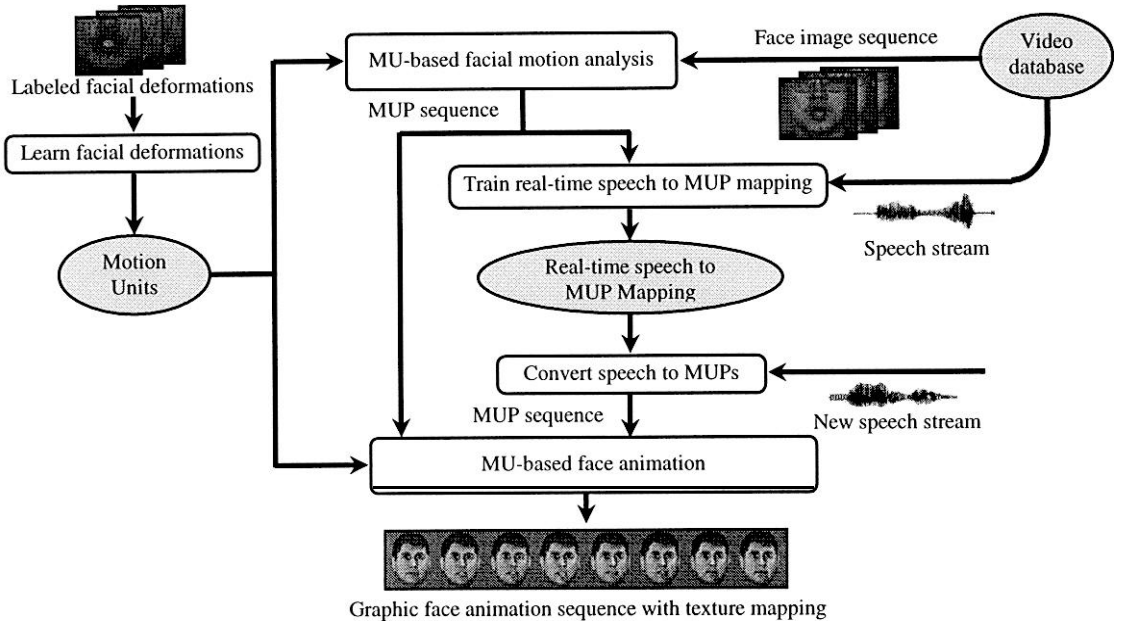


Figure 1. An integrated framework for face modeling, facial motion analysis and synthesis

A MU-based face model can be animated by adjusting the MUPs. Secondly, a robust MU-based facial motion tracking algorithm is presented. The tracking results are represented as MUP sequence. Finally, a set of facial motion tracking results and the corresponding speech are collected as the audio-visual training data. The data is used to train a real time audio to MUP mapping.

2.1 Motion Unit

MU serves as the basic information unit, which is learned from real data and links the components of the framework together. We mark 62 points in the lower face of the subject (see Figure 2). A generic mesh model is constructed to correspond those markers while the face is in its neutral position. The mesh is shown to overlap with the markers in Figure 2. In this paper, we only focus on 2D motion of the lower face because the lower face conducts the most complicated movements among face regions and is highly related to speech. Future work will apply the same methodology to the 3D deformations of the whole face. We capture the front view of a subject while he is pronouncing all English phonemes. The head of the subject is stabilized. The video is digitized at 30 fps. This results in more 1000 image frames. The markers are automatically tracked by template matching technique. A graphic interactive interface is developed to correct the positions of trackers using moose when the template matching fails due to large face or facial motions. A mesh deformation data sample vector is formed by concatenating the deformations of its vertices in the image plane.

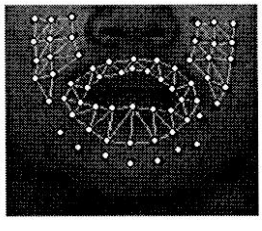


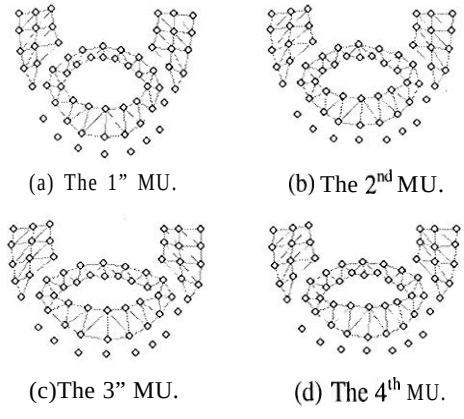
Figure 2. Markers and the generic mesh.

We assume that arbitrary facial deformation can be approximated by linear combination of MUs. Principal Component Analysis (PCA) [5] is applied to learning the significant characteristics of the mesh deformation data samples. The mean facial deformation and the first seven eigenvectors of PCA results are selected as Motion Units. Four MUs are shown in Figure 3. They respectively represent the mean deformation and local deformations around lips, mouth comers, and cheeks.

MUs have some nice properties. Firstly, MUs are learned from real data and encode the characteristics of real facial deformations. Secondly, the number of the MUs is much smaller that of the vertices on the face model. Only a few parameters need to be adjusted in order to animate the face model. This dramatically reduces the complexity of face animation.

2.2 MU-Based Facial Motion Tracking

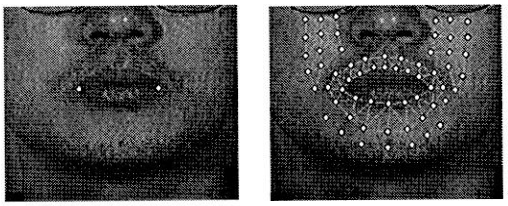
MU can be used as the high-level knowledge model to guide facial motion tracking. Currently, the tracking algorithm requires that face can only have 2D motion because MU is 2D. The algo-



(a) The 1" MU. (b) The 2nd MU.

(c)The 3" MU. (d) The 4th MU.

Figure 3. Motion Units.



(a)

Figure 4. Initialize model for tracking.

rithm requires that the face be in its neutral position in the first image frame so that the generic mesh model can be fit to the neutral face. The generic mesh model has two vertices corresponding to two mouth corners. Two mouth corners are manually selected in the facial image (see Figure 4(a)). The generic mesh model is fit to the face by scaling and rotating (see Figure 4(b)).

The tracking procedure consists of two steps, In the low-level image processing step, the locations of vertices in the next image are calculated separately by template matching. The results of template matching are usually noisy. We then constrain that the vertices can only undergo 2D global rigid motion (rotation, translation, and scaling) and local non-rigid motion within the manifold defined by MUs. Mathematically, the tracking problem can be formulated as a minimization problem:

$$(\vec{T}^*, \vec{C}^*)^{(t)} = \arg\min_{\vec{T},\vec{C}} \left\| \Psi(M\vec{C} + \vec{\zeta}) - \vec{S}^{(t)} \right\|^2 \qquad (1)$$

where

(a) t represents time or the frame number.

(b) $\Psi(\bullet)$ is the affine transformation function, whose parameter set $\vec{T}$ describes the global 2D rotation, scaling and translation transformations of the face. $\vec{T}$ is to be estimated.

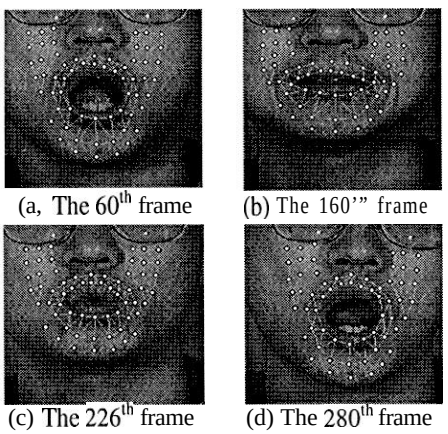
(c) $M = [\vec{m}_0\ \vec{m}_1 \cdots \vec{m}_K]$ is the MU matrix. $\vec{m}_p$ ($K \geq p \geq 0$) is a MU.

(d) $\vec{C} = [c_0\ c_1\ c_K]^T$ is the MUP vector and $c_0, c_1, \ldots, c_K$ are the MUPs. Since $\vec{m}_0$ is the mean deformation, c_0 is a constant and is always equal to 1. $c_1, .., c_K$ are unknown MUPs to be estimated.

(e) $\vec{\zeta}$ represents the concatenation of the coordinates of the vertices in their initial positions (or the neutral position) in the image plane.

(f) $\vec{S}^{(t)}$ is the facial shape estimated by using template matching technique alone at time t.

We use a least square estimator to solve eq. (1) and estimate the parameters of both global face motion and local facial motion. The local non-rigid facial motion is represented by MUPs. Figure 5 shows the tracking results of some typical facial images in an image sequence.



(a, The 60th frame (b) The 160'" frame

(c) The 226th frame (d) The 280th frame

Figure 5. Tracking examples of the MU-based facial motion tracking algorithm.

2.3 Real-time Speech Driven face animation

We videotape the front view of a speaking subject. One hundred sentences are selected from the text corpus of the DARPA TIMIT speech database. Both the audio and video are digitized at 30fps. Twelve LPC coefficients of the speech frames are calculated as the audio features. The visual features are the MUPs computed by the MU-based facial motion tracking algorithm described in Section 2.2. Overall, we have 19433 audio-visual samples. Eighty percent of the data is used for training. The remaining is used for testing.

We train a set of three layer perceptrons to estimate MUPs from audio features. Audio to visual mapping is in nature non-linear. Multilayer perceptrons (MLP) is a universal nonlinear function approximator and is suitable for modeling audio to visual mapping. Different from previous approaches using MLP for audio-to-visual mapping [10], [1 I], we divide the training data into 44 groups according to the audio feature vector of the sample. Each group corresponds to a phoneme. The audio features in each group are modeled by a Gaussian model. Each audio-visual data is classified into one of the 44 groups whose Gaussian model gives the highest score for the audio component of the audio-visual data. A MLP is trained for each group to estimate MUPs from the audio features. The input of MLP is the audio feature vector taken at seven consecutive time frames (3 backward, current and 3 for-

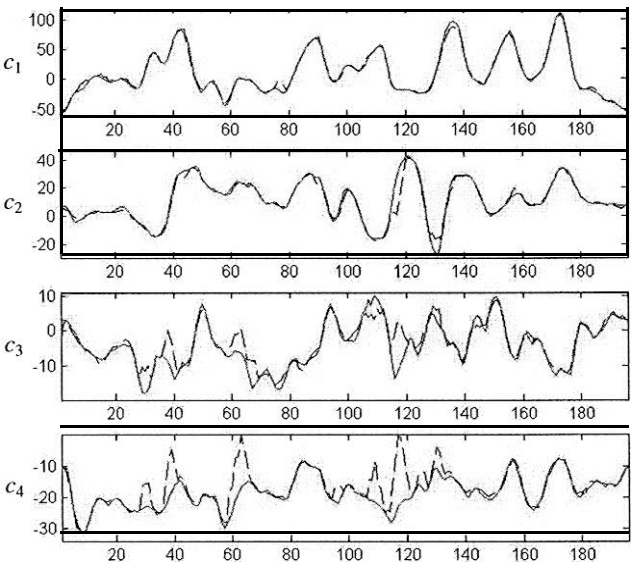


Figure 6. An example of audio-to-visual mapping.

ward time windows). In the estimation phase, an audio feature is classified into one of the groups. The corresponding MLP is selected to estimate MUPs. By dividing the data into 44 groups, lower computational complexity is achieved. In our experiments, the maximum number of the hidden units used in those MLPs is only 25. Therefore, both training and estimation have very low computational complexity. A method using triangular average window is used to smooth the jerky mapping results.

Good estimation results are obtained. Figure 6 illustrates the estimated visual results of a selected audio track. The text of the audio track is "Don't ask me to carry an oil rag like that." The figure shows the trajectories of the values of four Motion Unit Parameters (c_1, c_2, c_3, and c_4) versus the time. The horizontal axis represents time. The vertical axis represents the magnitudes of the Motion Unit Parameters. The solid red line is the ground truth. The dash blue line represents the estimated results.

The facial deformations are reconstructed using the estimated MUPs. The Pearson product-moment correlation coefficients of the original facial deformations and the reconstructed facial deformations are calculated. Pearson product-moment correlation measures how good the global match between the shapes of two signal sequences is. It is value range is [0 1]. The larger the coefficient, the better the estimated signal sequence matches with the original one. In our experiments, the coefficients of the training data and testing data are 0.8750 and 0.8743 respectively.

3. THE IFACE SYSTEM

We developed a face modeling and animation system, called the iFACE system [4]. The system provides functionalities for customizing a generic face model for an individual, text driven face animation, and off-line speech driven face animation. Based on the proposed integrated framework, we improve the iFACE system. A set of basic facial shapes is carefully and manually constructed so that their 2D projections are visually similar to the MUs. The real-time audio-to-visual mapping described in Section 2.3 is used to estimate MUPs from audio features. The face animation is obtained by linearly combining those basic facial shapes weighted by the estimated MUPs.

Figure 5 shows some typical frames in a real-time speech driven face animation sequence. The text of the sound track is "Dialog is an essential element."

Figure 5. A speech driven face animation example using nonlinear real-time audio-to-visual mapping.

4. CONCLUSIONS

In this paper, we propose a framework that integrates face modeling, facial motion analysis and synthesis. Within the framework, facial deformations are represented by Motion Units. We present methods for: (I) Learning MUs and using MU for face animation; (2) MU based facial motion tracking; and (3) Real-time speech driven face animation using MU as the visual representation. The proposed framework enables effective analysis and synthesis of facial movements for multimedia information delivery in distributed environments.

5. ACKNOWLEDGMENTS

This research is supported by USA Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0003.

6. REFERENCES

[1] K. Aizawa and T. S. Huang, "Model-based image coding," *Proc. IEEE,* vol. 83, pp. 259-271, Aug. 1995.

[2] D. DeCarlo and D. Mataxas, "Optical flow constraints on deformable models with applications to face tracking", *Int. Journal of Computer Vision,* 38(2), pp 99-127, 2000.

[3] I. A. Essa and A. Pentland, "Coding Analysis, Interpretation, and Recognition of Facial Expressions", *IEEE Transaction Pattern Analysis and Machine Intelligence,* vol. IO, no. 7, pp. 757 - 763, Jul. 1997.

[4] P. Hong, Z. Wen, T. S. Huang, iFACE: a 3D synthetic talking face. *International Journal Of Image and Graphics, vol.* 1, no. 1, pp. 1-8, 2001.

[5] 1. T. Jolliffe, *Principal Component Analysis,* Springer-Verlag, 1986.

[6] P. Kalra, A. Mangili, N. Magnenat Thalmann, D. Thalmann, "Simulation of Facial Muscle Actions Based on Rational Free Form Deformations," *Proc. Eurographics'92,* pp.59-69.

[7] Y. C. Lee, D. Terzopoulos and K. Waters, *"Realistic modeling for facial animation,"* SIGGRAPH 1995, pp. 55-62.

[8] H. Li, P. Roivainen and R. Forchheimer, "3-D motion estimation in model-based facial image coding", *IEEE Trans. On Pattern Analysis and Machine Intelligence* 15(6), pp. 545-555, 1993.

[9] D. W. Massaro, Perceiving Talking Faces, MIT Press, 1998.

[10] D. W. Massaro *et al.*, Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks, in Proc. AVSP'99, Aug. 1999, Santa Cruz, USA.

[11] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man-machine interface", *IEEE J. Selected Areas in Communications,* 4:594-599, 1991.

[12] F. 1. Parke, *A* parametric *model of human faces.* Ph.D. Thesis, University of Utah, 1974.

[13] D. Terzopoulos and K. Waters. "Analysis and synthesis of the facial image sequences using physical and anatomical models", *IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 15, no. 6,* pp. 569 579, Jun. 1993.

[14] K. Waters, "A muscle model for animating three-dimensional facial expressions," *Computer Graphics,* 21(4): 17-24, July 1987.

[15] K. Waters, J. M. Rehg, M. Loughlin, *et al.*, "Visual sensing of humans for active public interfaces," Cambridge Research Lab, Technical Report CRL 96-S.

[16] L. Williams, "Performance-driven facial animation", Computer Graphics no. 24, vol. 2, pp. 235-242, Aug. 1990.