L

Learning 3D Face-Animation Model

Zhen Wen
University of Illinois at Urbana-Champaign, USA

Pengyu Hong
Harvard University, USA

Jilin Tu
University of Illinois at Urbana-Champaign, USA

Thomas S. Huang
University of Illinois at Urbana-Champaign, USA

INTRODUCTION

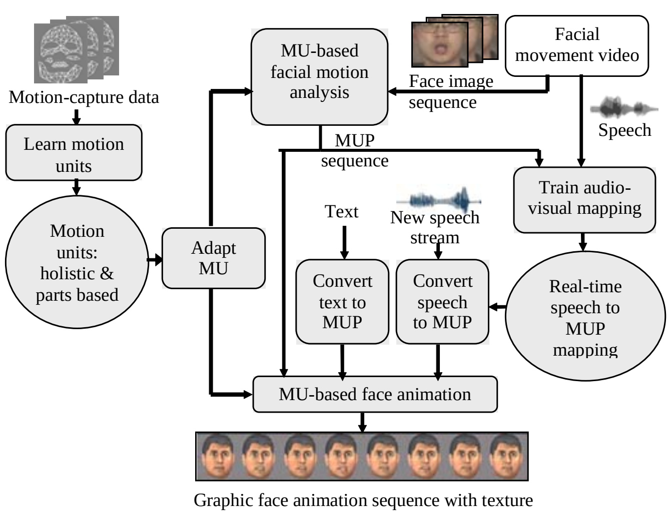
A synthetic human face is useful for visualizing information related to the human face. The applications include visual telecommunication (Aizawa & Huang, 1995), virtual environments and synthetic agents (Pandzic, Ostermann, & Millen, 1999), and computer-aided education.

One of the key issues of 3D face analysis (tracking) and synthesis (animation) is to model facial deformation. The facial deformation is complex, which often includes subtle expressional variations to which people are very sensitive. Therefore, traditional models usually require extensive manual adjustment. Recently, the advance of motion-capture techniques sparked data-driven methods (e.g., Guenter et al., 1998). They achieve realistic animation by using real face-motion data to drive 3D face animation. However, the basic data-driven methods are inherently cumbersome because they require a large amount of data.

More recently, machine learning techniques have been used to learn *compact* and *flexible* face-deformation models from motion-capture data. The learned models have been shown to be useful for realistic face-motion synthesis and efficient face-motion analysis. A unified framework on facial deformation analysis and synthesis is demanded to address in a systematic way the following problems: (a) how to learn a compact 3D face-deformation model from data, and (b) how to use the model for robust facial-motion analysis and flexible animation.

In this article, we present a unified machine-learning-based framework on facial deformation modeling, and facial motion analysis and synthesis. The framework is illustrated in Figure 1. In this framework, we first learn from

Figure 1. Machine-learning-based framework for facial deformation modeling, and facial motion analysis and synthesis



extensive 3D facial motion-capture data a compact set of *Motion units* (MUs), which are chosen as the quantitative visual representation of facial deformation. Then, arbitrary facial deformation can be approximated by a linear combination of MUs, weighted by coefficients called *motion unit parameters* (MUPs). Based on interpolation, the MUs can be adapted to the face model with new geometry topology. MU representation is used in both robust facial motion analysis and efficient synthesis. We also utilize MUs to learn the correlation between speech and facial motion. A real-time audio-to-visual mapping is learned using an artificial neural network (ANN). Experimental results show that our framework achieved natural face animation and robust nonrigid tracking.

Figure 2. (a) Generic model in iFACE; (b) Personalized face model based on the Cyberware scanner data; (c) Feature points defined on a generic model for MU adaptation

(a) (b) (c)

BACKGROUND

Facial Deformation Modeling

Representative 3D spatial facial deformation models include free-form interpolation models (e.g., affine functions, splines, radial basis functions), parameterized models (Parke, 1974), physics-based models (Waters, 1987), and more recently, machine-learning-based models. Because of the high complexity of natural motion, these models usually need extensive manual adjustments to achieve plausible results. To approximate the space of facial deformation using simpler units, people (Tao, 1998) proposed to describe arbitrary facial deformation as a combination of action units (AUs) based on the facial action coding system (FACS; Ekman & Friesen, 1977). Because AUs are only defined qualitatively, they are usually manually customized for computation. Recently, people turned to apply machine learning techniques to learn models from data (Hong, Wen, & Huang, 2002; Kshirsagar, Molet, & Thalmann, 2001; Reveret & Essa, 2001).

Facial Motion Analysis

Human facial motion analysis is the key component for many applications, such as model-based very-low-bitrate video coding, audiovisual speech recognition, and expression recognition. High-level knowledge of facial deformation must be used to constrain the possible deformed facial shapes. For 3D facial motion tracking, people have used various 3D deformable model spaces, such as the 3D parametric model (DeCarlo, 1998), B-spline surface (Eisert, Wiegand, & Girod, 2000), and FACS-based models (Tao, 1998). These models, however, are usually manually defined, thus may not capture characteristics of the real facial motion well. Therefore, people have recently proposed to use subspace models trained from real motion data (Reveret & Essa, 2001}.

Facial Motion Synthesis

In this article, we focus on real-time-speech face animation. The core issue is the audio-to-visual mapping. HMM-based methods (Brand, 1999) utilize long-term contextual information to generate smooth motion trajectory, but they can only be used in off-line scenarios. For real-time mapping, people proposed various methods such as vector quantization (VQ; Morishima & Harashima, 1991), the Gaussian mixture model (GMM), and ANN (Morishima & Harashima). To use short-time contextual information, people used the concatenated audio feature over a short time window (Massaro et al., 1999) or a time-delay neural network (TDNN; Lavagetto, 1995).

Figure 3. Markers; (a) markers are shown as white dots, (b and c) the mesh is showin in two different viewpoints

(a) (b) (c)