Subtle Facial Expression Synthesis using Motion Manifold Embedding and Nonlinear Decomposable Generative Models

Chan-Su Lee¹, Yang Wang², Zhiguo Li¹, Atul Kanaujia¹, Ahmed Elgammal¹, Dimitris Samaras², Dimitris Metaxas¹, Xiangfeng Gu², Peisen Huang³

¹ Computer Science Department, Rutgers - The State University of New Jersey, NJ, USA
² Computer Science Department, State University of New York at Stony Brook, NY, USA
³ Mechanical Engineering Department, State University of New York at Stony Brook, NY, USA

Abstract

Facial motions convey personal characteristics and subtle emotional states. This paper presents a new framework to model facial motions of different people with multiple expression types from high resolution facial expression tracking data. We also provide a mechanism to animate subtle facial expressions based on video sequences. A conceptual motion manifold is used for a unified representation of facial motion dynamics. Subtle local motions in facial expressions are modeled by nonlinear mapping using empirical kernel map from an embedding manifold. We represent facial expressions in different people, as well as different expression type by a nonlinear decomposable generative model using multilinear analysis of the nonlinear mappings coefficient space. We can synthesize high resolution facial motions based on tracking of facial motions in video sequences by estimation of the model parameters.

1. Introduction

Human faces can express not only basic emotions such as anger, surprise, or happiness, but also subtle thoughts or emotions. In order to analyze subtle differences of facial expression between individuals, we captured high resolution smile expressions with subtle differences such as “soft affectionate”, “coy flirtatious”, and “devious smirk”. For each expression, the facial motion includes nonlinear deformations and unique temporal and geometric characteristics of the particular subject. We model these subtle differences in facial motions by a conceptual motion manifold embedding and nonlinear mappings from embedding to high resolution facial motion data.

A decomposable nonlinear generative model is proposed to model subtle motion characteristics of different people and expressions. An empirical kernel map along with an embedded manifold and a projection is used to represent the nonlinear mapping for each cycle of facial motions. Subtle differences of facial motions across different people and expressions are captured in the projection of each sequence from a common kernel space of the embedded manifold. By analyzing these linear projections using multilinear analysis, we can decompose the nonlinear mappings into two main factors: the personal style, individual characteristics of expressions, and the expression type, subtle variations in expressions. Consequently, new facial expressions can be generated by using different personal style and expression type factors, along with the common embedded manifold, which encodes the temporal information of facial expressions.

2. Synthesis High Resolution Facial Expressions from Video Sequence

Our high resolution facial expression synthesis system includes five main components: data acquisition, facial motion tracking, modeling facial expressions with a decomposable generative model, estimation of facial expression control parameters, and synthesis of high resolution facial expressions. In the data acquisition stage, we collect both high resolution dynamic range data of facial expressions [ZH04] and 2D video sequences separately for each subject. Both 2D and 3D facial expression sequences are captured at high frame rates from multiple people with several expression types. In
order to establish correspondences between different frames within one sequence and between different sequences, we employed a high resolution 3D tracking method using harmonic maps [WGZ*05] to extract detailed facial motions with subtleties from dynamic range data, while low resolution facial motions are extracted from video sequences using a 2D tracking method based on 2D contours and 3D deformable models.

We use a unit circle as a conceptual embedding for facial expressions, which is equivalent, i.e., homeomorphic to data-driven facial expression manifolds as we captured sequence of facial motions which start from a neutral expression and return to the neutral expression. We consider nonlinear mapping functions of the form

\[ f^{se}(x) = B^{se} \cdot \psi(x), \]  

where \( B \) is a \( d \times N \) coefficient matrix and \( \psi(\cdot) : R^d \rightarrow R^N \) is a nonlinear mapping where \( N \) kernel functions can be used to model the manifold in the embedding space. Multilinear tensor analysis of collected mappings coefficients from different people and expression types provide a decomposable nonlinear generative model for facial expressions in combination with the kernel map [LE05]. For a given personal style \( s \) and expression type \( e \), we can compute the mapping coefficient \( B^{se} \) with an expression style vector \( s \) and an expression type vector \( e \). For any given manifold point \( x_t \) we can reconstruct a facial motion as

\[ y_t^{se} = Z \times_1 s \times_2 e \times_3 \psi(x_t), \]  

Because the tracking results extracted from 2D video sequences and from dynamic range data have different levels of detail, we derive two generative models with different resolutions and the same kinds of state decomposition from these two sources of facial motion data. Although these expression bases associated to high resolution facial motions with subtleties and low resolution facial motions are different from each other because of different levels of detail, an important observation is that a new expression vector estimated from different resolution tracking data shares a similar distance to each expression basis in all levels of detail. It also implies that the normalized expression bases provided by our nonlinear decomposition algorithm are invariant to the levels of detail and the different vector dimensions after the kernel mapping.

We also present a performance-driven approach to create high resolution facial expressions based on an exemplar video sequence. We can infer style and expression parameters for the high resolution model using a low resolution facial model and personal style and expression type estimated from low resolution facial motion tracking for a given new actor’s video sequence. Therefore, we can synthesize high resolution subtle facial expressions using estimated style and expression parameters from video sequence.

Figure 2 shows synthesized high resolution images from two different subtle expressions. Even though two images are similar, the estimated personal style and expression type parameters distinguish subtle difference and we can generate high resolution facial expressions with subtle differences.

References

