Nonlinear Dynamic Shape and Appearance Models for Facial Motion Tracking

Chan-Su Lee, Ahmed Elgammal, and Dimitris Metaxas

Rutgers University, Piscataway, NJ, USA
{chansu,elgammal,dnm}@cs.rutgers.edu

Abstract. We present a framework for tracking large facial deformations using nonlinear dynamic shape and appearance model based upon local motion estimation. Local facial deformation estimation based on a given single template fails to track large facial deformations due to significant appearance variations. A nonlinear generative model that uses low dimensional manifold representation provides adaptive facial appearance templates depending upon the movement of the facial motion state and the expression type. The proposed model provides a generative model for Bayesian tracking of facial motions using particle filtering with simultaneous estimation of the expression type. We estimate the geometric transformation and the global deformation using the generative model. The appearance templates from the global model then estimate local deformation based on thin-plate spline parameters.

Keywords: Nonlinear Shape and Appearance Models, Active Appearance Model, Facial Motion Tracking, Adaptive Template, Thin-plate Spline, Local Facial Motion, Facial Expression Recognition.

1 Introduction

Recently there has been extensive research on modeling and analyzing dynamic human motions for human computer interaction, visual surveillance, autonomous driving, computer graphics, and virtual reality. Facial motions intentionally or unintentionally display internal emotional states explicitly through facial expressions. Accurate facial motion analysis is required for affective computer interaction, stress analysis of users or vehicle drivers, and security systems such as deception detection. However, it is difficult to accurately model global facial motions since they undergo through nonlinear shape and appearance deformations, which varies across different people and expressions. Local facial motions are also important to detect subtle emotional states for stress analysis, and recognizing deception.

Active Shape Models (ASMs) is a well known statistical model-based approach that uses point distribution models in linear subspace [1]. By constraining shape deformation into the linear subspace of training shapes, the model achieves robust estimation of shape contour [1]. Active Appearance Models (AAMs) [2] combine the shape model and the linear appearance subspace after aligning the

D. Mery and L. Rueda (Eds.): PSIVT 2007, LNCS 4872, pp. 205–220, 2007.

appearance into a normalized shape. It employs an iterative model refinement algorithm based on a prediction model, that is learned as a regression model. A variety of approaches have been proposed to improve the update schemes such as compositional update schemes [3,4], direct AAMs [5], and adaptive gradient methods [6], are proposed. However, all these methods have limitations in modeling facial shape and appearance deformations since they approximate nonlinear deformations in shape and appearance using a linear subspace. As a result of the linear approximation, the model requires high dimensional parameters to model nonlinear data. The high dimensionality makes it difficult to find the optimal shape and appearance parameters. In addition, it is difficult to generate accurate facial animations using linear approximations since linear subspace requires large amount of data in order to model shape and appearance variations accurately [7].

Template based approaches are frequently used for estimating geometric transformation [8,9,10,11] that are invariant to shape and appearance variations. Recently templates based on nonlinear warping parameter estimation have been used for tracking nonrigid shape deformation [12]. Although, the method provides effective facial motion tracking under small facial deformations, it loses track for large deformations.

We propose a nonlinear facial motion tracking framework that can accurately estimate the local and global shape deformation in addition to the geometric transformation. We estimate the geometric transformation and global facial motions based on a global nonlinear appearance model. The global nonlinear appearance model provides a compact low dimensional representation of the facial motion state using an embedded representation of the motion manifold. Our system also factorizes the shape variations into different expressions. We achieve tracking of large facial motions using particle filter within the Bayesian framework based on the global nonlinear appearance model.

The global model is not enough for accurate tracking of local deformation and shape deformation that are limited in training. The global nonlinear appearance model, however, provides accurate appearance templates according to the estimated embedding state and the expression type. The local facial deformation estimation using single template TPS parameter estimation fails to track large facial motion deformation. The global model that supports large shape deformations provides normalized-appearance models for local deformation estimation. By combining the global appearance model and local deformation, we can achieve accurate estimations of facial motions in large deformations.
Our contributions are as follows:

Modeling nonlinear shape and appearance deformations: We propose a nonlinear shape and appearance model of facial expressions that factorizes facial expression type and facial motion state. A low dimensional representation for facial motion state is achieved using an embedded representation of the motion manifold. For accurate facial appearance model, we employ nonlinear warping of the appearance templates based on TPS (Sec. 3.1) warping.

Tracking global facial motions using particle filter: Using the global nonlinear shape and appearance model in conjunction with low dimensional facial motion separations, we estimate the geometric transformation and facial deformations. We use the global model for global facial motion tracking within the Bayesian particle tracking based framework (Sec. 4.1).

Local facial motion estimation using adaptive appearance templates: We extend tracking of the motion deformations using single template in [12] by adaptive templates to cover large facial deformation. After estimating the state of the global shape and appearance, local nonrigid deformation is estimated using TPS warping control (Sec. 4.2). The local facial deformation is directly estimated using shape landmark points from the adaptive normalized-appearance templates.

2 Framework

We develop a facial shape and appearance model for large facial deformations with different expressions. A dynamic facial deformation for a given facial expression is a nonlinear function of the facial configuration; as the facial configuration varies over time, the corresponding observed facial shape and appearance changes according to the given facial configuration. In addition, the facial deformation is variant in different expressions. Different facial expressions undergo different facial deformations in shape and appearance.

Let facial configuration at time t be $\boldsymbol{x}_t$, and the corresponding observed nonlinear shape and appearance at the same time be $\boldsymbol{y}_t$ for given expression sequence k, then the nonlinear facial shape and appearance can be represented by

$$\boldsymbol{y}_t = f^k(\boldsymbol{x}_t) = g(\boldsymbol{e}^k, \boldsymbol{x}_t), \tag{1}$$

where $f^k(\cdot)$ is a nonlinear function variant in different expression type $\boldsymbol{e}^k$, $g(\cdot)$ is a nonlinear mapping with factorization of expression type parameter $\boldsymbol{e}^k$ in addition to facial configuration $\boldsymbol{x}_t$. Hence, to develop nonlinear shape and appearance model, we need to find a representation of facial configuration $\boldsymbol{x}_t$ and a factorization of nonlinear function $f^k(\cdot)$ in different expressions. In Sec. 3.2 we present a nonlinear generative model with low dimensional embedded representation of the motion manifold and a factorization of nonlinear mapping using empirical kernel map and decomposition.

For a given nonlinear shape and appearance model, tracking of facial motion is estimating facial configuration and expression type, and geometric transformation, which match generated shape and appearance to the observed image frame. For a given observation $\boldsymbol{z}_t$ and state $\boldsymbol{s}_t$, we can represent the Bayesian tracking as a recursive update of the posterior $P(\boldsymbol{s}_t|\boldsymbol{z}^t)$ over the object state $\boldsymbol{s}_t$ given all the observation $\boldsymbol{z}^t = \boldsymbol{z}_1, \boldsymbol{z}_2, .., \boldsymbol{z}_t$ up to time t:

$$P(\boldsymbol{s}_t|\boldsymbol{z}^t) \propto P(\boldsymbol{z}_t|\boldsymbol{s}_t) \int_{\boldsymbol{s}_{t-1}} P(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) P(\boldsymbol{s}_{t-1}|\boldsymbol{z}^{t-1})$$

Rao-Blackwellized particle filtering is applied for efficient state estimation in Sec. 4.1.

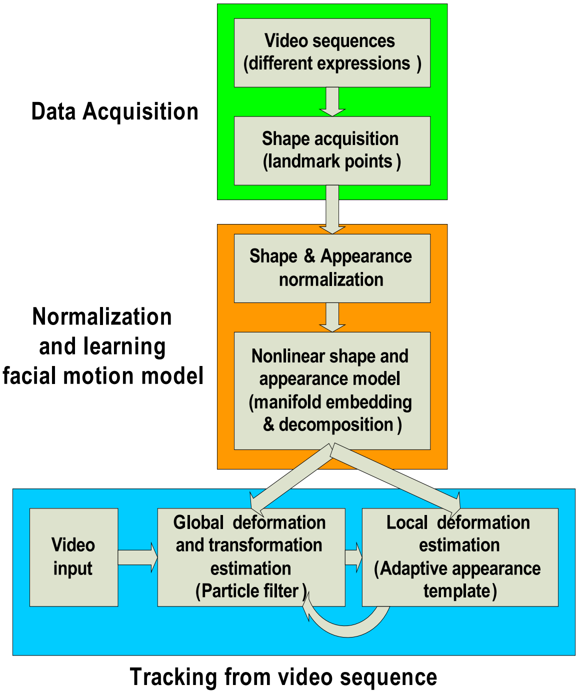
Data Acquisition
Video sequences (different expressions)
Shape acquisition (landmark points)
Normalization and learning facial motion model
Shape & Appearance normalization
Nonlinear shape and appearance model (manifold embedding & decomposition)
Video input
Global deformation and transformation estimation (Particle filter)
Local deformation estimation (Adaptive appearance template)
Tracking from video sequence

Fig. 1. The block diagram of nonlinear facial motion tracking system

Facial state estimation based on the global facial shape and appearance model provides global facial motion tracking for the training data. The estimated states, however, are not sensitive to local deformations: small misalignment in geometric transformation. Hence we enhance our facial motion tracking system based on local nonrigid deformation estimation using TPS-warping parameter estimation and adaptive appearance templates in Sec. 4.2. The appearance templates for TPS-warping parameter estimation are provided from the global facial shape and appearance model, which support different appearance model according to facial motion state and expression type. For accurate estimation of local nonrigid deformation from the appearance template, we need accurate appearance representation of the global shape and appearance model. We use TPS warping for accurate shape-normalized appearance template (Sec. 3.1). The local estimation of facial motion is used to update global shape model by linear combination of expression weights (Sec. 4.3).

Our facial motion tracking system consist of three stages: data acquisition, normalization and learning nonlinear facial motion model, and tracking facial motion from the video sequence. Fig. 1 shows the block diagram of our facial motion tracking system. First, we collect multiple video sequences with different expressions and manually mark some of the frames. Prior to learning nonlinear facial shape and appearance model, we collect normalized-shape and -appearance using similarity transformation and TPS warping respectively (Sec. 3.1). Collected normalized-shapes and corresponding normalized-appearances in different expression sequences are used for learning the nonlinear shape and appear model. We use particle filtering and the nonlinear shape and appearance model to estimate global deformation and geometric transformation using particle filter. Based on the estimated global state, we generate the appearance template for local nonrigid deformation estimation. The estimated local deformation is used to refine global model state estimation as shown in Fig. 1.

3 Nonlinear Global Shape and Appearance Models

In this section, we explain how to achieve accurate and appearance normalization for a normal shape, and how to learn nonlinear shape and appearance model using an embedded representation of the motion manifold.

3.1 Facial Shape and Appearance Normalization

Facial shape normalization: We align collected landmark shape points using weighted similarity transformation for shape normalization. The ith shape with n landmark points is represented by a vector $\boldsymbol{p}_i = (x_{i1}, y_{i1}, x_{i2}, y_{i2}, \cdots, x_{in}, y_{in})$. Given two shapes $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, we can find the similarity transformation for shape j, $S(\delta_j)$ that minimizes the weighted sum $E_j = (\boldsymbol{p}_i - S(\delta_j)\boldsymbol{p}_j)^{\mathsf{T}}\boldsymbol{D}(\boldsymbol{p}_i - S(\delta_j)\boldsymbol{p}_j)$, where $\boldsymbol{D}$ is a weighting diagonal matrix. The mean shape, represented by $\boldsymbol{p}_0$, is computed by averaging shape landmark points after shape normalization. This mean shape is used as a normal shape for normalized-appearance representation.

Facial appearance normalization: Normalized-appearance is a vector representation for appearance of the normal shape. It is important to have precise normalized-appearance as we use the normalized-appearance as an adaptive appearance template for the local deformation estimation (Sec. 4.2) in addition to the observation model in Bayesian tracking using particle filtering (Sec. 4.1). We use TPS warping [13] for non-rigid registration of appearance image to the mean shape that is estimated after shape normalization. The TPS warping leads to smooth deformations of shape by control points. Though piecewise-affine warping are frequently used in linear appearance models [14,4], the piecewise-affine warping can cause artifacts around the boundaries in non-rigid deformation of shape due to facial motion [15]. The normalized-appearances, which are computed by the TPS warping of the given landmark points to the mean shape, are used to represent appearance variations in vector space.

We compute the normalized facial appearance precisely using TPS backward warping. Given the image frames $\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_{N_K}$, we collect manually marked corresponding shape vectors $\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{N_K}$, where N_K is the number of image frame for training. A normalized-appearance template for training image j, $\boldsymbol{I}_j^{'}$, is generated from the image $\boldsymbol{I}_j$ with a corresponding shape vector $\boldsymbol{p}_j$ by TPS warping the shape $\boldsymbol{p}_j$ to the mean shape $\boldsymbol{p}_0$. We denote this normalized-appearance computation for the given image $\boldsymbol{I}_j$, and shape vector $\boldsymbol{p}_j$ by $\boldsymbol{I}_j^{'} = \boldsymbol{I}_j(\mathcal{W}(\boldsymbol{p}_0, \boldsymbol{p}_j))$, where $\mathcal{W}(\cdot)$ denotes a TPS warping from control landmark point $\boldsymbol{p}_j$ to $\boldsymbol{p}_0$. In actual computation, we apply a backward warping due to discrete nature of the raster images and computational efficiency. In case of backward warping we need to warp output image coordinate into input image coordinate and interpolate the intensity values. The TPS warping $\mathcal{W}(\cdot)$ needs to be computed once for the mean shape $\boldsymbol{p}_0$.

Normalized shape-appearance representation: In image sequences, the kth image $\boldsymbol{I}_k$ can be represented by its aligned shape $\boldsymbol{p}_k$ and the TPS warped normalized-appearance $\boldsymbol{a}_k$. We combine the normalized-shape vector and the normalized-appearance vector as a new shape-appearance vector $\boldsymbol{y}_k = [\boldsymbol{p}_k^\mathsf{T}\ \boldsymbol{a}_k^\mathsf{T}]^\mathsf{T}$. We extract the normalized-appearance vector $\boldsymbol{a}_k$ as pixels which are inside the contour of the mean shape $\boldsymbol{p}_0$ after the TPS warping of the original image $\boldsymbol{I}_k$ from the original shape $\boldsymbol{p}_k$ to the mean shape $\boldsymbol{p}_0$. We denote this procedure as

$$\boldsymbol{a}_k = \underset{\boldsymbol{\xi}\in\boldsymbol{p}_0}{\boldsymbol{I}_k}(\mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}_k)) = \Upsilon(\boldsymbol{I}_k, \boldsymbol{p}_k) \tag{2}$$

So, $\Upsilon(\boldsymbol{I}_j, \boldsymbol{p}_i)$ returns a normalized-appearance vector for the given image $\boldsymbol{I}_j$ with the TPS warping from a shape vector $\boldsymbol{p}_i$ to the mean shape $\boldsymbol{p}_0$. If the pixel number within the mean shape is N_a, then the dimension of the shape-appearance vector $\boldsymbol{y}_k$ is $N_{as} = 2n + N_a$.

3.2 Nonlinear Generative Models with Manifold Embedding and Factorization

Facial motion embedding and nonlinear mapping: We propose to use the nonlinear facial shape and appearance model based on low-dimensional manifold embedding and empirical kernel mapping to track accurate nonlinear appearance deformations in different facial motions. Since dynamic facial expressions lie on low dimensional manifolds, we use a conceptual unit circle as an embedded representation of the facial motion manifold for each of the facial expression cycle [16]. Sets of image sequences, which represent a full cycle of facial expressions, are used for the embedded representation of the motion manifold. We denote each expression sequence by $\boldsymbol{y}^e = \{\boldsymbol{y}_1^e \cdots \boldsymbol{y}_{N_e}^e\}$ where e denotes the expression type and N_e the number of frames for a given expression sequence. Each sequence is temporally embedded on a unit circle at equal distance. Given a set of distinctive representative embedding points $\{\boldsymbol{x}_i \in \mathbb{R}^2, i = 1 \cdots N\}$, we can define an empirical kernel map[17] as $\psi_N(\boldsymbol{x}) : \mathbb{R}^2 \to \mathbb{R}^N$ where $\psi_N(\boldsymbol{x}) = [\phi(\boldsymbol{x}, \boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}, \boldsymbol{x}_N)]^\mathsf{T}$, given a kernel function $\phi(\cdot)$.

For each input $\boldsymbol{y}^e$ and its embedding $\boldsymbol{x}^e$, we learn a nonlinear mapping function $f^e(\boldsymbol{x})$ that satisfies $f^e(\boldsymbol{x}_i) = \boldsymbol{y}_i^e, i = 1 \cdots N_e$ and minimizes a regularized risk criteria. Such function admits a representation of the form $\psi(\boldsymbol{x}) = \sum_{i=1}^N w_i \phi(\boldsymbol{x}, \boldsymbol{x}_i)$, i.e., the whole mapping can be written as

$$f^e(\boldsymbol{x}) = \boldsymbol{B}^e \cdot \psi(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{B}$ is a $d \times N$ coefficient matrix. The mapping coefficient can be obtained by solving the linear system $[\boldsymbol{y}_1^e \cdots \boldsymbol{y}_{N_e}^e] = \boldsymbol{B}^e[\psi(\boldsymbol{x}_1^e) \cdots \psi(\boldsymbol{x}_{N_e}^e)]$. Using this nonlinear mapping, we capture nonlinearity of facial expressions in each sequence.

Expression type factorization: Given learned nonlinear mapping coefficients $\boldsymbol{B}^1$, $\boldsymbol{B}^2$, $\cdots$, $\boldsymbol{B}^K$ of K different expression type sequences, the nonlinear mappings are factorized by fitting an asymmetric bilinear model to the coefficient space [18]. As a result, we can generate a nonlinear shape and appearance instance $\boldsymbol{y}_t^k$ for a particular expression type k at any configuration $\boldsymbol{x}_t$ as

$$\boldsymbol{y}_t^k = \boldsymbol{\mathcal{A}} \times \boldsymbol{e}^k \times \psi(\boldsymbol{x}_t) = g(\boldsymbol{e}^k, \boldsymbol{x}_t), \tag{4}$$

where $\boldsymbol{\mathcal{A}}$ is a third order tensor, $\boldsymbol{e}^k$ is an expression type vector for the expression class k. We can analyze and represent nonlinear facial expression sequences by estimating the facial motion state vector $\boldsymbol{x}_t$, and expression type $\boldsymbol{e}$ in this generative model.

4 Tracking Global and Local Facial Motions

In order to track nonrigid local facial deformations as well as global large facial deformations in different expression type, we first estimate the global facial motion and the geometric transformation. We then apply local nonrigid facial deformation estimation using the appearance template generated from the global facial motion estimation. Estimated global facial motion parameters are updated to reflect local facial deformation.

4.1 Global Facial Motion Estimation

Our global facial motion tracking routine incorporates two components: the geometric transformation, and the global deformation. The geometric transformation explains the rigid movement of face due to head motion. The global deformation motion captures the nonlinear facial deformation in different expression types and motion states (configurations). If we describe the geometric transformation parameters by $T_{\boldsymbol{\alpha}_t}$, the global shape and appearance deformation as $\boldsymbol{y}_t, i.e. \boldsymbol{a}_t, \boldsymbol{p}_t$, then the goal of our global tracking algorithm for a given image $\boldsymbol{I}_t$ is to estimate sub state vector $\boldsymbol{\alpha}_t^*$, $\boldsymbol{p}_t^*$ and $\boldsymbol{a}_t^*$ that minimize

$$\begin{aligned} E(\boldsymbol{\alpha}_t^*, \boldsymbol{p}_t^*, \boldsymbol{a}_t^*) &= \min_{\boldsymbol{\alpha}_t, \boldsymbol{p}_t, \boldsymbol{a}_t} (\Upsilon(\boldsymbol{I}_t, T_{\boldsymbol{\alpha}_t} \cdot \boldsymbol{p}_t) - \boldsymbol{a}_t) \\ &= \min_{\boldsymbol{\alpha}_t, \boldsymbol{p}_t, \boldsymbol{a}_t} (\Upsilon(\boldsymbol{I}_t, T_{\boldsymbol{\alpha}_t} \cdot (q(\boldsymbol{y}_t)) - a(\boldsymbol{y}_t)) \end{aligned} \tag{5}$$

where $a(\boldsymbol{y}_t^*) = \boldsymbol{a}_t^* = \boldsymbol{y}_t^*(2n+1 : N_{as})$ is an appearance sub-vector and $q(\boldsymbol{y}_t^*) = \boldsymbol{p}_t^* = \boldsymbol{y}_t^*(1 : 2n)$ is a shape sub-vector from shape-appearance vector $\boldsymbol{y}_t^*$. The shape-appearance vector $\boldsymbol{y}_t^* = \boldsymbol{\mathcal{A}} \times \boldsymbol{e}^* \times \psi(\boldsymbol{x}_t^*)$ is computed for the estimated expression type $\boldsymbol{e}^*$ and facial motion state $\boldsymbol{x}_t^*$ by Eq. 4. Therefore, tracking of the global deformation of facial motion essentially invloves estimating $\boldsymbol{e}^*$, and $\boldsymbol{x}_t^*$, which are the best fitting global shape-appearance template after the geometric transformation $\boldsymbol{\alpha}_t$.

Global facial motion tracking:particle filtering Given the nonlinear generative shape and appearance model, we can describe the observation of shape and appearance instance $\boldsymbol{z}_t$ by geometric transformation and global shape-appearance vector, i.e., state parameters $\boldsymbol{\alpha}_t$ and $\boldsymbol{y}_t$. The global shape-appearance vector is defined by expression type $\boldsymbol{e}_t$ and facial configuration $\boldsymbol{x}_t$ in Eq. 4. Therefore, tracking facial motion is effectively inferring the configuration $\boldsymbol{x}_t$, facial expression type parameter $\boldsymbol{e}_t$, and global transformation $T_{\boldsymbol{\alpha_t}}$ given the observation $\boldsymbol{z}_t$ at time t.

In our model, the state $\boldsymbol{s}_t$ $[\boldsymbol{\alpha}_t, \boldsymbol{x}_t, \boldsymbol{e}_t]$ uniquely describe the state of the tracking facial deformation. The observation $\boldsymbol{z}_t$ is composed of shape vector $\boldsymbol{z}_{\boldsymbol{p}_t}$ and appearance vector $\boldsymbol{z}_{\boldsymbol{a}t}$ for the given image at time t. The global transformation parameter is independent of the global deformation state as we can combine any shape and appearance model with any geometrical transformation to synthesize a new shape and appearance in the image space. However, they are dependent on the given observation. We approximate the joint posterior distribution $P(\boldsymbol{\alpha}_t, \boldsymbol{x}_t, \boldsymbol{e}_t|\boldsymbol{z}^t) = P(\boldsymbol{\alpha}_t, \boldsymbol{y}_t|\boldsymbol{z}^t)$ by two marginal distribution $P(\boldsymbol{\alpha}_t|\boldsymbol{y}_t^*, \boldsymbol{z}^t)$ and $P(\boldsymbol{y}_t|\boldsymbol{\alpha}_t^*, \boldsymbol{z}^t)$, where $\boldsymbol{\alpha}_t^*$, and $\boldsymbol{y}_t^*$ are representative values like MAP (maximum a posteriori).

We estimate the likelihood of the observation $\boldsymbol{z}_t$ for given state $\boldsymbol{s}_t = (\boldsymbol{\alpha}_t, \boldsymbol{y}_t)$ by

$$P(\boldsymbol{z}_t|\boldsymbol{\alpha}_t, \boldsymbol{y}_t) \propto \exp\left(-\frac{||\Upsilon(\boldsymbol{I}_t, T_{\boldsymbol{\alpha}_t} \cdot \boldsymbol{p}_t) - \boldsymbol{a}_t||}{\sigma}\right) \tag{6}$$

where $\boldsymbol{p}_t = \boldsymbol{y}_t(1 : 2n)$, $\boldsymbol{a}_t = \boldsymbol{y}_t(2n + 1 : N_{as})$, and σ is the scaling factor for the measured image distance. In particle filtering, the state $\boldsymbol{s}_t$ is updated by estimating the weight $\pi_t^{(i)}$ using the observation likelihood:

$$\pi_t^{(i)} \propto P(\boldsymbol{z}_t|\boldsymbol{s}_t^{(i)}) = P(\boldsymbol{z}_t|\boldsymbol{\alpha}_t^{(i)}, \boldsymbol{y}_t^{(i)}).$$

Particle filter for the geometric transformation: We estimate geometric transformation using particle filter based on the predicted global shape and appearance. We assume that expression state varies smoothly, and predicted configuration explains temporal variation of the estimated expression state. The estimated global shape and appearance at time t, $\boldsymbol{y}_t^*$, is estimated from the previous expression state $\boldsymbol{e}_{t-1}$, and predicted configuration $\boldsymbol{x}_t^{'}$. The prediction of configuration, $\boldsymbol{x}_t^{'}$, is estimated from previous estimated embedding $\boldsymbol{x}_{t-1}^*$ using dynamics of the configuration along the embedded representation of the motion manifold [19]. This predicted shape and appearance $\boldsymbol{y}_t^{'}$ is used as the representative value $\boldsymbol{y}_t^*$. Given this global shape and appearance template, we estimate the best geometric transformation $\boldsymbol{\alpha}_t$ for the given observation at time t, $\boldsymbol{z}_t$.

The geometric transformation state $\boldsymbol{\alpha}_t$ consists of geometric transformation parameters γ_t,θ_t, and $\boldsymbol{\tau}_t$ for scaling, rotation, and translation. The marginal probability distribution is represented by N_α particles $\{\boldsymbol{\alpha}_t^{(i)}, {}^{\boldsymbol{\alpha}}\pi_t^{(i)}\}_{i=1}^{N_\alpha}$. We update weights ${}^{\boldsymbol{\alpha}}\pi_t^{(i)}, i = 1, 2, \cdots, N_\alpha$ with $\boldsymbol{y}_t^{'}$ using Eq. 6.

Rao-Blackwellized particle filtering for global deformation tracking: For the state estimation of the global deformation, we utilize Rao-Blackwellized particle filtering. In order to estimate global deformations using generative model in Eq. 4, we need to estimate the state vector $\boldsymbol{x}_t$, and $\boldsymbol{e}_t$ whose dimensions are 2, and N_e. The dimension of the expression state N_e depends on the number of expression types which can be high. When we know the configuration vector $\boldsymbol{x}_t$, we can achieve approximate solution for the expression vector as explained in the following. The original Rao-Blackwellized particle filtering for dynamic Bayesian networks [20] assumes accurate solution for the state that is not represented by particle state. We utilize an approximate solution for the expression type vector to avoid sampling for high dimensional state density estimation, which requires large number of particles for accurate approximation.

The facial motion state $\boldsymbol{x}_t$ is embedded in two dimensional space with one constraint for unit circle embedding. So, the embedding dimension is actually one-dimensional and we can represent the embedding parameter β_t as one-dimensional state vector. We represent the distribution of facial motion embedding β by N_β particles $\{\beta_t^{(i)}, {}^{\beta}\pi_t^{(i)}\}_{i=1}^{N_\beta}$. If we represent the approximate estimation of expression vector as $\boldsymbol{e}_t^*$, we can approximate the marginal distribution as

$$\begin{aligned} P(\boldsymbol{e}_t^*|\boldsymbol{y}_t) &= \sum_{\beta} P(\boldsymbol{e}_t^*|\beta_t, \boldsymbol{y}_t) P(\beta_t|\boldsymbol{y}_t) = \sum_{\beta} P(\boldsymbol{e}_t^*|\beta_t, \boldsymbol{y}_t) \sum_{i=1}^{N_\beta} {}^{\beta}\pi_t^{(i)} \delta(\beta_t^{(i)}, \beta_t) \\ &= \sum_{i=1}^{N_\beta} {}^{\beta}\pi_t^{(i)} P(\boldsymbol{e}_t^*|\beta_t^{(i)}, \boldsymbol{y}_t), \end{aligned}$$

where $\delta(x, y) = 1$ if $x = y$ and 0 otherwise.

We represent the estimated expression vector by a linear weighted sum of known expression vectors. We assume that the optimal expression vector can be represented by a linear combination of the expression classes in the training data; we can generate the global deformations as the configuration changes along the manifold through the linear weighted sum of expression classes. Now, we need to solve linear regression weights κ such that $\boldsymbol{e}^{new} = \sum_{k=1}^{K_e} \kappa_k \boldsymbol{e}^k$ where each $\boldsymbol{e}^k$ is one of K_e expression classes. For a given configuration β_t, that is $\boldsymbol{x}_t = h(\beta_t)$, we can obtain expression conditional class probability $p(\boldsymbol{e}^k|\boldsymbol{y}_t, \boldsymbol{x}_t)$ proportional to the observation likelihood $p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{e}^k)$. Such likelihood can be estimated as a Gaussian density centered around $\boldsymbol{\mathcal{A}} \times \boldsymbol{e}^k \times \psi(\boldsymbol{x}_t)$, i.e.,

$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{e}^k) \approx \mathcal{N}(\boldsymbol{\mathcal{A}} \times \boldsymbol{e}^k \times \psi(\boldsymbol{x}_t), \Sigma^{\boldsymbol{e}^k}).$$

Given expression class probabilities, we can set the weights of expression classes to $\kappa_k^{(i)} = p(\boldsymbol{e}^k \mid \boldsymbol{y}_t, \boldsymbol{x}_t^{(i)})$. The estimated expression vector is the weighted sum of each expression type $\boldsymbol{e}_t^* = \frac{\sum_{i=1}^{N_\beta} \sum_{k=1}^{N_e} \kappa_k^{(i)} \boldsymbol{e}^k}{\sum_{i=1}^{N_\beta} \kappa_k^{(i)}}$.

4.2 Local Facial Motion Estimation

We perform local facial motion tracking for estimating local deformations that differs from global facial model, and to refine inaccurate estimation of the geometric transformation. The estimated global facial motion state using particle filter, with limited number of particle samples shows misalignment of geometric transformations and inaccurate estimations of the global deformations sometimes. In addition, facial deformation of the new sequence can be different from the learned global model even for the same person with the same expression type. Therefore, we need local facial motion tracking to refine the global tracking result.

We propose template-adaptive local facial motion tracking with shape description using thin plate splines (TPS) warping. We utilize landmark points in the facial shape description as control points in TPS. The shape-normalized appearance is used as a template for local facial motion tracking. The proposed local facial motion tracking is similar to the non-rigid object motion tracking using TPS parameters and image gradients [12]. In our case, the tracking result of global deformation using the nonlinear shape and appearance model provides a new appearance template for each frame. In addition, the landmark shape estimated from the global deformation after applying geometric transformations provides initial shape for local facial motion tracking.

Let the estimated global shape and appearance be $\boldsymbol{y}^g_{t0}$, its shape vector be $\boldsymbol{p}^g_{t0}$, appearance vector be $\boldsymbol{a}^g_{t0}$, and current input image be I_t, the objective of the local deformation fitting is to minimize the error function

$$\begin{aligned} E(\delta \boldsymbol{p}_t) &= \sum \|\Upsilon(\boldsymbol{I}_t, \boldsymbol{p}^g_{t0} + \delta \boldsymbol{p}_t) - \boldsymbol{a}^g_{t0}\| \\ &= \sum_{\boldsymbol{\xi} \in \boldsymbol{p}_0} \|\boldsymbol{I}_t(\mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}^g_{t0} + \delta \boldsymbol{p})) - \boldsymbol{I}^g_{t0}(\xi)\|^2 \end{aligned} \tag{7}$$

where $\boldsymbol{I}^g_{t0}$ is an image in normalized shape with global appearance vector $\boldsymbol{a}^g_{t0}$. Since we use shape normalized appearance as the template in the local tracking, the TPS warping $\mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}^g_{t0} + \delta \boldsymbol{p})$ is determined by the coordinate control points $\boldsymbol{p}^g_{t0} + \delta \boldsymbol{p}$. For the given $\boldsymbol{p}_{t0}$ from the global deformation tracking, the warping function is solely determined by the local deformation $\delta \boldsymbol{p}$.

Gradient descent technique is applied to find the local deformation parameter $\delta \boldsymbol{p}$ which minimize Eq. 7 similar to [12,8]. Linearization is carried out by expanding $\boldsymbol{I}_t(\mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}^g_{t0} + \delta \boldsymbol{p}))$ in the Taylor series about $\delta \boldsymbol{p}$,

$$\boldsymbol{I}_t(\mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}^g_{t0} + \delta \boldsymbol{p})) = \boldsymbol{I}_t(\mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}^g_{t0})) + \delta \boldsymbol{p}^\mathsf{T} \boldsymbol{M}_t + \text{h.o.t}, \tag{8}$$

where $\boldsymbol{M}_t = [\frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{p}_1} | \frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{p}_2} | \cdots | \frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{p}_{2n}}]$. Each term $\frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{p}_k}$ can be computed using warped image coordinate $\boldsymbol{\xi} = \mathcal{W}(\boldsymbol{\xi}, \boldsymbol{p}_0; \boldsymbol{p}^g_{t0})$ by applying chain rule: $\frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{p}_k} = \frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{\xi}} \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{p}_k}$. The $\frac{\partial \boldsymbol{I}_t}{\partial \boldsymbol{\xi}}$ is the gradient of current input image $\boldsymbol{I}_t$ after TPS warping to the mean shape. The warping coefficients are fixed and can be pre-computed since we use the common mean shape in all the normalized-appearance templates. The solution for Eq. 7 can be computed when the higher-order terms in Eq. 8 is ignored:

$$\delta \boldsymbol{p} = (\boldsymbol{M}_t^{\top} \boldsymbol{M}_t)^{-1} \boldsymbol{M}_t^{\top} \delta \boldsymbol{I}_t, \tag{9}$$

where $\delta \boldsymbol{I}_t$ is the image difference between template appearance image and current image warped to the template shape. We achieve better fitting of the shape to local image features by iterative updating the local shape model. This local fitting provides better alignment of shape and normalized appearance for a given input image.

4.3 Combining Global Facial Motion Estimation and Local Facial Motion Estimation

We update the global deformation state using the new shape normalized appearance image after local fitting. As a result of accurate local fitting, the new shape-normalized appearance vector will represent appearance more accurately than the one estimated by the global facial motion tracking.

Using the new shape-normalized appearance vector estimated from local deformation, we update the expression state. First, we estimate the new expression weight $\boldsymbol{\kappa}^l$ based on the new appearance vector after local fitting. Then, the combined new estimated expression weight is computed by linear combination of the local expression weight $\boldsymbol{\kappa}^l$ and global expression weight $\boldsymbol{\kappa}^g$,

$$\boldsymbol{\kappa}^{new} = (1 - \varepsilon)\boldsymbol{\kappa}^g + \varepsilon \boldsymbol{\kappa}^l \tag{10}$$

This process enhances the robustness in the expression parameter estimation. The combining parameter ε, which is empirically estimated, depends on the reliability of local fitting. For example, local fitting is less reliable for unknown subject and we assign small value for ε. Even though the combination is in linear interpolation, the overall system preserve nonlinear characteristics of the facial motions. This refined global state estimation improves accuracy of geometry transformation in the subsequent frames.

5 Experimental Results

In order to build global shape and appearance model for different expressions, we use Cohn-Kanade AU coded facial expression database [21]. The landmarks have 38 points in each frame image ($n = 38$). The appearance vector was represented by 35965 pixels ($N_a = 35965$) inside landmark shape contour in the mean shape. This appearance vector size depend on mean shape size. By reducing the mean shape size, we can reduce the appearance vector dimension. We manually marked the shape landmarks of every other frame to learn the shape and appearance model. As the database has expression sequences from the neutral expression and to the peak expression, we embed each frame on the half circle with equal distance for each sequence.

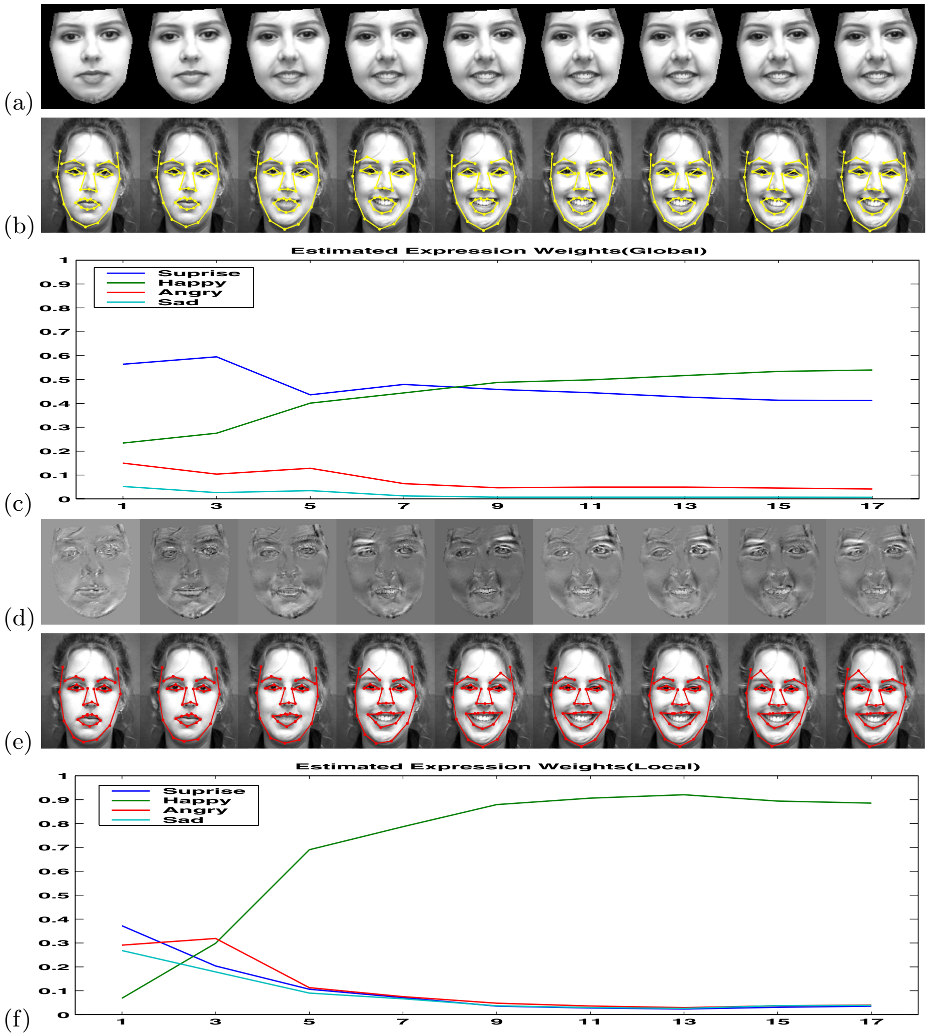

Fig. 2. Facial expression tracking with global and local fitting: (a) Best fitting global appearance in normalized shape. (b) Global shape tracking facial motion. (c) Expression weights in global facial motion estimation. (d) Image error in the local fitting. (e) Local tracking facial motion with adaptive template provided by global appearance model. (f) Expression weights in local facial motion estimation.

Facial motion tracking with expression type estimation: Estimated expression type shows how well the facial motion tracking discriminate variations in facial motion of different expressions. Fig. 2 shows tracking a smile expression sequence with the local fitting. At each frame, global facial motion tracking is estimated expression weights (c) and facial shape after global transformation estimation. The best fitting shape-appearance parameter provided shape-normalized

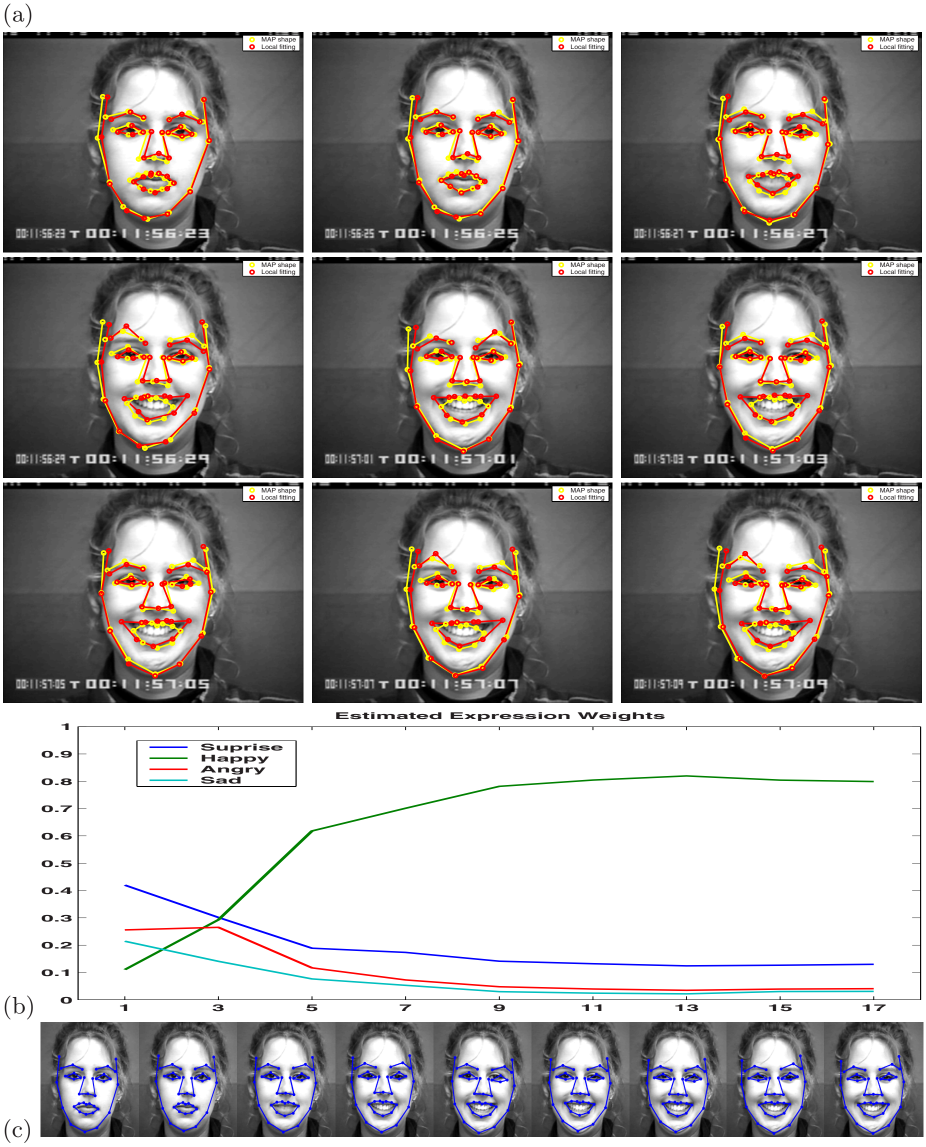

Fig. 3. Comparison of Facial expression tracking: (a) Comparison of tracking result: yellow-global fitting, red-local fitting. (b) Update of estimated expression weights by combination of local and global expression estimation. (c) Best fitting global model using updated expression state.

appearance template (a) and facial shape tracking after global deformation (b). After local nonrigid deformation estimation, tracking result (e) shows better estimation of shape deformation to the input image and better estimation of facial

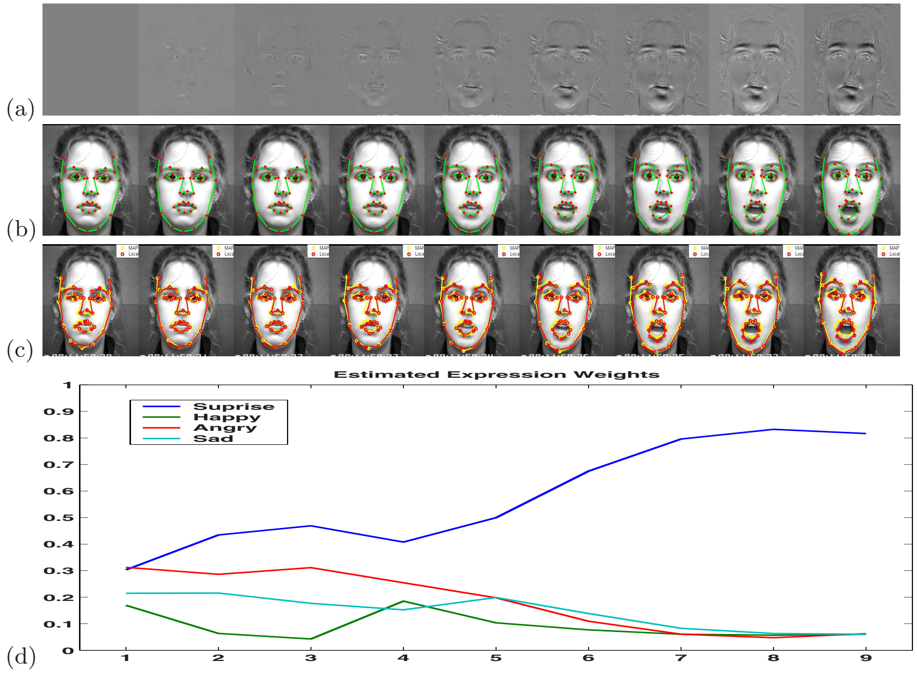

Fig. 4. Tracking surprise expression : (a) Error image based on a template after local fitting. (b) Tracking result by the local deformation estimation with an initial frame as a template. (c) Tracking result with adaptive template by global shape and appearance model: yellow-global fitting, red-local fitting. (d) Estimated global expression weights.

expression type (f). Facial expression weight in global deformation had similar weights between 'surprise' and 'happy (smile)'. After the local deformation estimation, the estimated expression type got higher weight for happy expression correctly. However, some points like left eyebrow show inaccurate local fitting. Fig. 3 shows comparison of tracking accuracy. After updating estimated expression type by combining global deformation and local deformation, we got new estimation of expression weight Fig. 3(a). Based on the new expression weight, we accurately estimated global facial motion tracking(c).

Tracking large facial deformations: We compared tracking accuracy with a single template and adaptive templates in large facial deformations. Fig. 4 (b) is facial motion tracking result based on single frame. It shows appropriate tracking of facial motion in small deformations. However, it fails to track large facial deformations around mouth region. Fig. 4 (c) shows facial motion tracking result using adaptive templates at each frame. As the global deformation model provides updated appearance template in addition to initial shape for tracking, it achieved more accurate tracking of large facial deformations.

6 Conclusion and Future Works

We have proposed a new framework for facial motion tracking for handling large facial deformations. The global deformation tracking based on nonlinear shape and appearance model provides appearance adaptive template in large facial deformation. The local fitting with the appearance adaptive templates enables accurate fitting of global, coarse estimation of facial motion.

Here, we count facial motion deformation by combination of expression type in addition to configuration change. We plan to extend our system to consider variations of facial shape and appearance in different people by applying multilinear analysis. The TPS warping is expensive computationally. We may efficiently program this computation using general-propose computing on graphics processing units(GPGPU), which provides efficient parallel processing.

References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models: Their training and applications. CVIU 61(1), 38–59 (1995)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Proc. of ECCV, vol. 2, pp. 484–498 (1998)
3. Baker, S., Matthews, I.: Equivalence and efficiency of image alignment algorithms. In: Proc. of CVPR, vol. 1, pp. 1090–1097 (2001)
4. Matthews, I., Baker, S.: Active appearance models revisited. IJCV 60(2), 135–164 (2004)
5. Hou, X., Li, S., Zhang, H., Cheng, Q.: Direct appearance models. In: Proc. of CVPR, vol. 1, pp. 828–833 (2001)
6. Batur, A.U., Hayes, M.H.: A novel convergence scheme for active appearance models. In: Proc. of CVPR, vol. 1, pp. 359–366 (2003)
7. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH 1999, pp. 187–194. ACM Press/Addison-Wesley Publishing Co., New York (1999)
8. Hager, G.D., Belhumeur, P.N.: Efficient region tracking with parametric models of geometry and illumination. IEEE Trans. PAMI 20(10) (1998)
9. Black, M.J., Jepson, A.D.: Eigentracking: Robust matching and tracking of articulated objects using view-based representation. Int.J. Compter Vision, 63–84 (1998)
10. Ho, J., Lee, K.-C., Yang, M.-H., Kriegman, D.: Visual tracking using learned linear subspaces. In: Proc. of CVPR, pp. 782–789 (2004)
11. Elgammal, A.: Learning to track: Conceptual manifold map for closed-form tracking. In: Proc. of CVPR, pp. 724–730 (2005)
12. Lim, J., Yang, M.H.: A direct method for modeling non-rigid motion with thin plate spline. In: Proc. of CVPR, vol. 1, pp. 1196–1202 (2005)
13. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Trans. PAMI 11(6), 567–585 (1989)
14. Stegmann, M.B.: Analysis and segmentation of face images using point annotations and linear subspace techniques. Technical Report TMM-REF-2002-22, Technical University of Denmark (2002)

15. Cootes, T.F.: Statistical models of appearance for computer vision. Technical report, University of Manchester (2004)
16. Lee, C.S., Elgammal, A.: Facial expression analysis using nonlinear decomposable generative models. In: Zhao, W., Gong, S., Tang, X. (eds.) AMFG 2005. LNCS, vol. 3723, pp. 17–31. Springer, Heidelberg (2005)
17. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2002)
18. Elgammal, A., Lee, C.S.: Separating style and content on a nonlinear manifold. In: Proc. CVPR, vol. 1, pp. 478–485 (2004)
19. Lee, C.-S., Elgammal, A.: Style adaptive bayesian tracking using explicit manifold learning. In: Proc. of British Machine Vision Conference (2005)
20. Murphy, K., Russell, S.: 24 Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks. In: Sequential Monte Carlo Methods in Practice, pp. 499–515. Springer, Heidelberg (2001)
21. Kanade, T., Tian, Y., Cohn, J.F.: Comprehensive database for facial expression analysis. In: Proc. of FGR., pp. 46–53 (2000)