Facial Action Transfer with Personalized Bilinear Regression

Dong Huang and Fernando De La Torre

Robotics Institute, Carnegie Mellon University

Abstract. Facial Action Transfer (FAT) has recently attracted much attention in computer vision due to its diverse applications in the movie industry, computer games, and privacy protection. The goal of FAT is to "clone" the facial actions from the videos of one person (source) to another person (target). In this paper, we will assume that we have a video of the source person but only one frontal image of the target person. Most successful methods for FAT require a training set with annotated correspondence between expressions of different subjects, sometimes including many images of the target subject. However, labeling expressions is time consuming and error prone (i.e., it is difficult to capture the same intensity of the expression across people). Moreover, in many applications it is not realistic to have many labeled images of the target. This paper proposes a method to learn a personalized facial model, that can produce photo-realistic person-specific facial actions (e.g., synthesize wrinkles for smiling), from **only** a neutral image of the target person. More importantly, our learning method does not need an explicit correspondence of expressions across subjects. Experiments on the Cohn-Kanade and the RU-FACS databases show the effectiveness of our approach to generate video-realistic images of the target person driven by spontaneous facial actions of the source. Moreover, we illustrate applications of FAT to face de-identification.

Keywords: Facial action transfer, Bilinear regression.

1 Introduction

Facial Action Transfer (FAT) from a source person to a target person (from which we only have one image) has attracted much attention in computer vision and computer graphics due to its increasing number of applications. Beyond the film making industry, such technology is also applicable to preserve privacy of subjects in video (i.e., face de-identification) [1,2], online image and video collections [3], and virtual avatars [4]. A major challenge of FAT is to transfer subtle facial actions from the source to the target to create video-realistic outputs. Throughout the paper we will refer to this problem as FAT instead of facial expression transfer to emphasize that our method is able to deal with subtle and spontaneous facial movement, rather than only imitating some predefined expressions (e.g., happy, sad, disgust) [5,6].

There are four major challenges on FAT: (1) Typically, the facial structure (shape) and texture (appearance) of the source and target subjects are quite different, as well as the dynamics of the facial actions. There are person-specific facial features in the source images (e.g., glasses, freckles, wrinkles, eyelashes) that the target person might not have. Directly copying the shape and appearance changes from the source to the target (e.g., [7]) can result in artifacts in the rendered target person. (2) Although

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 144–158, 2012.

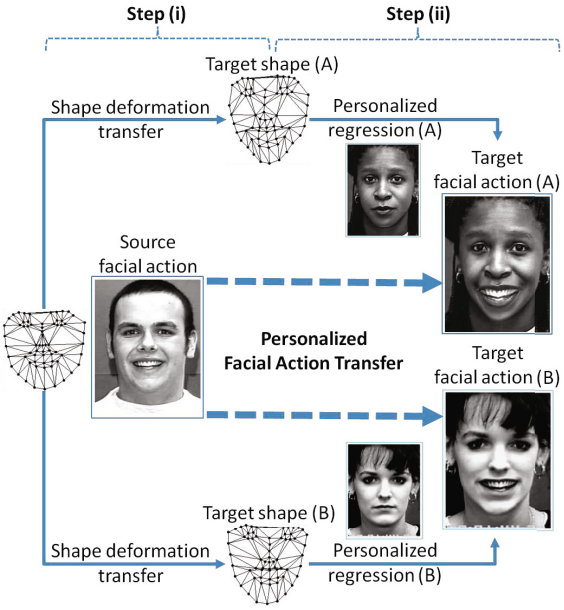

Fig. 1. Overview of our personalized FAT method. The facial action of the source person (a smile) is transferred to both subjects (A) and (B) in two steps: (i) Shape deformation transfer, where the shape change of the source w.r.t. his neutral face is transferred to the targets; (ii) Generate the appearance changes using personalized bilinear regression. Observe how the wrinkles of the smile are adapted (i.e. personalized) to subject (A) and (B).

shape deformation transfer has been successfully applied to computer graphics problems [8,9,10,11], transferring the appearance/texture remains an unsolved problem due to the high dimensionality of the image data and potential nonlinearity of the facial actions. (3) An ideal algorithm should be able to factorize identity, expression and pose changes. However, this factorization is very difficult in practice, because the facial motions are combinations of the movement of facial muscles, head motion and person-specific features. Moreover, existing methods that decouple identity from expression (e.g., tensor and regression based methods [12,13,14,5,6]) require correspondences across predefined expressions (e.g., happy, sad, disgust) for different subjects. Labeling the correspondence of facial expressions across subjects is time-consuming and error prone. Even if all subjects are instructed to pose with the same expression, they cannot perform the facial action with exactly the same style and intensity. (4) In real world scenarios (e.g., interviews, movies), facial behaviors are spontaneous and usually combined with pose changes. The complexity of these facial actions are beyond the representation ability of predefined expressions (e.g., happy, sad, disgust) that could be posed and labeled in controlled environments.

To solve these problems, this paper proposes a two-step approach for FAT (see Fig. 1). In the first step, our method transfers the shape of the source person to the target subject (A) and (B) using the triangle-based deformation transfer method [10]. In the second step, our method generates the appearance of the target person using a

personalized mapping from shape changes to appearance changes. Our main hypothesis is that the physical structure of the face (e.g., bones, muscles and skin) defines a consistent and measurable pattern that relates the movements (shape changes) of facial components to the appearance changes [15] for a particular person. Based on this intuition, our method learns a bilinear regression between the shape and appearance changes from training images. A major contribution of this work is to personalize the shape-to-appearance regression with **only** one sample of the target subject. More importantly, unlike previous methods [12,13,14,4,5,6], **our learning method does not require the correspondence of expressions across training subjects**. Fig. 1 illustrates the main idea of our method.

2 Previous Works

Existing approaches to FAT can be broadly grouped into two categories: direct transfer methods and learning-based transfer methods.

Direct transfer methods copy the shape and/or appearance changes of the source person to the target face image. [8,9] represents the face by a densely partitioned triangle mesh, usually containing 10^4 triangles. The shape changes under a given expression are transferred to the target face as a set of local affine transformations while preserving the connectivity of the target triangles. These methods do not transfer appearance changes. Liu et al. [7] proposed a geometric warping algorithm in conjunction with the Expression Ratio Image (ratio between the appearance of the neutral image and the image of a given expression) to copy subtle appearance details such as wrinkles and cast shadows to the target. This method tends to produce artifacts on the target face image since the appearance details to be transferred are not adapted to the target subject.

Learning-based FAT methods learn a transformation from a training set of face images that have been labeled across expressions. The correspondence is determined manually or semi-automatically [5,6,4,16]. Existing learning-based FAT can be broadly classified into two major approaches: the regression-based and tensor-based methods.

Regression-based methods include two modalities: (1) Regression between expressions [5,6] that learns a mapping from a reference expression (e.g., neutral) to the expressions to be transferred (e.g., smile). Given a reference face of a target person, the smile face of the target person can be predicted with the regression specifically learned for the smile expression. A major limitation of this method is its inability to represent untrained expressions. (2) Regression between subjects [4,16]. This method learns a mapping between multiple pairs of corresponding expressions performed by both the source and target subjects, and then uses the learned regression to transfer new expressions. In the case, that there are no corresponding images between expressions of different people, [4] generates the correspondent images by learning a regression from the neutral face to the pre-defined expression, similar to [5,6], and apply this mapping to the neutral of the target subject. In addition, [4] learns a generic regressor from the shape to the appearance. In our work, we extend this approach by personalizing the regressor using only one training sample, achieving a highly photo-realistic result. [16] learns two Active Appearance Models (AAMs), one for the source and one for the target. It performs FAT by learning a mapping between AMMs' coefficients. This method

also requires solving for the correspondence between the expressions of the target and source, which is not possible in many realistic applications.

Tensor-based approaches [12,13,14] perform Higher-Order Singular Value Decomposition (HOSVD) to factorize the facial appearance into identity, expression, pose and illumination. Given the factorization, expression transfer [17,18,19] is done by first computing the identity coefficients of the new testing person, and then reassembling the identity factor with expression factors learned by the HOSVD. A major drawback of tensor-based approaches is the need of carefully labeled correspondences across expression, pose and illumination. [20,21] generalize the tensor-based approaches to build non-linear manifolds of human body actions and facial expressions. Similar to the standard tensor-based approaches, these methods require to solve for the correspondence of states on the manifold (content) across different subjects (style).

The existing learning-based FAT methods rely on the the availability and labeling accuracy of the similar expression in faces of different subjects. However, labeling expression is time consuming and error prone (i.e., it is hard to capture and solve correspondence for expressions under different intensities). In addition, in many applications it is not possible to have labeled training samples for the target. This paper proposes a more practical approach to FAT that does not require expression correspondence across subjects. Moreover, we are able to generate photo-realistic rendering using a personalized bilinear regression that only requires one frontal image of the target.

3 Personalized FAT

This section describes how to transfer the shape and appearance changes to achieve a personalized FAT method.

3.1 Transferring Shape Changes

Let $\mathbf{x}_i^{neu} \in \Re^{v\times 1}$ (see notation [1]) be a vector containing the two-dimensional coordinates of 66 landmarks ($v = 2 \times 66 = 132$) for the i^{th} subject under the neutral expression ($i = 1, \cdots, p$) (see Fig. 1). By performing Delaunay triangulation using these landmarks, the face region is partitioned into 106 triangles. $\mathbf{X}_i^e = \{\mathbf{x}_i^{e_1}, \cdots, \mathbf{x}_i^{e_{n_i}}\} \in \Re^{v\times n_i}$ is a matrix that contains the landmark coordinates of the n_i face images for the i^{th} subject performing different facial actions (i.e., subscript "e").

The first step to transfer shape changes is to compute the shape deformation from the neutral to the facial expression of the source subject. We used an affine transformation between triangles [10] (See Fig. 2). In the second step, we will transfer the triangle-wise shape transformation to the target subject.

The mapping between the neutral and a given facial expression of the source person (first step) is done as follows. Let the vectors $\{\mathbf{v}_{s_j}^{(1)}, \mathbf{v}_{s_j}^{(2)}, \mathbf{v}_{s_j}^{(3)}\} \in \Re^{2\times 3}$ contain the

[1] Bold capital letters denote matrices $\mathbf{X}$, bold lower-case letters a column vector $\mathbf{x}$. $\mathbf{x}_j$ represents the j^{th} column of the matrix $\mathbf{X}$. All non-bold letters represent scalar variables. $diag$ is an operator that transforms a vector to a diagonal matrix or takes the diagonal of the matrix into a vector. $vec(\cdot)$ vectorizes a matrix into a vector. $\mathbf{I}_k \in \Re^{k\times k}$ denotes the identity matrix. $\mathbf{1}_n \in \Re^n$ is a vector of all 1s. $vec(\mathbf{A})$ rearrange the elements of $\mathbf{A}$ in a vector. $\|\mathbf{x}\|_2$ denotes the L2-norm of the vector $\mathbf{x}$. $\|\mathbf{A}\|_F^2$ designates the Frobenious norm of matrix $\mathbf{A}$.

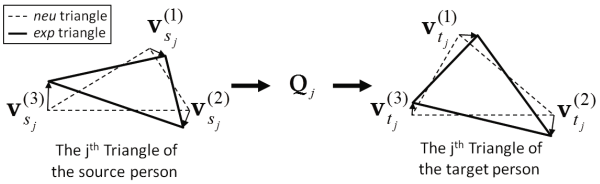

Fig. 2. Transfer shape changes for the j^{th} triangle of the source person (with vertex $\{\mathbf{v}_{s_j}^{(1)}, \mathbf{v}_{s_j}^{(2)}, \mathbf{v}_{s_j}^{(3)}\}$) to the corresponding triangle of the target person (with vertex $\{\mathbf{v}_{t_j}^{(1)}, \mathbf{v}_{t_j}^{(2)}, \mathbf{v}_{t_j}^{(3)}\}$). The dash and thick edged triangles represent the neutral expression and other expression, respectively. $\mathbf{Q}_j$ is the affine transformation matrix to be transferred.

two dimensional coordinates of the three vertices for the j^{th} triangle of the source person ($j = 1, \cdots, 106$). The matrix $\mathbf{V}_{s_j} = [\mathbf{v}_{s_j}^{(2)} - \mathbf{v}_{s_j}^{(1)}, \mathbf{v}_{s_j}^{(3)} - \mathbf{v}_{s_j}^{(1)}] \in \Re^{2\times 2}$ is the translation-free representation of this triangle. We compute the affine transformation $\mathbf{Q}_j \in \Re^{2\times 2}$ for the triangle from neutral to an expression by minimizing: $\min_{\mathbf{Q}_j} \left\| \mathbf{V}_{s_j}^{exp} - \mathbf{Q}_j \mathbf{V}_{s_j}^{neu} \right\|_F^2$, where $\mathbf{V}_{s_j}^{exp}$ and $\mathbf{V}_{s_j}^{neu}$ represent the j^{th} triangle containing the landmarks of source facial expression to be transferred and the source neutral face, respectively.

After computing all triangle-wise shape changes between the neutral and a different expression of the source person, the next step is to transfer the shape changes to the target person. Let $\widetilde{\mathbf{X}}_t^e \in \Re^{2\times 66}$ be a rearranged version of the target shape vector $\mathbf{x}_t^e = vec(\widetilde{\mathbf{X}}_t^e) \in \Re^{132\times 1}$. Applying $\mathbf{Q}_j$'s individually to the target neutral shape might result in disconnected vertexes. To solve this problem, we jointly transfer the transformations by minimizing:

$$\min_{\mathbf{x}_t^e} \sum_{j=1}^{106} w_j \left\| \mathbf{V}_{t_j}^{exp} - \mathbf{Q}_j \mathbf{V}_{t_j}^{neu} \right\|_F^2, \tag{1}$$

where $\mathbf{V}_{t_j}^{exp} = \widetilde{\mathbf{X}}_t^e \mathbf{S}_j \in \Re^{2\times 2}$, $\mathbf{S}_j \in \Re^{66\times 2}$ is a matrix of elements $\{-1, 0, 1\}$ that transforms $\widetilde{\mathbf{X}}_t^e$ to a translation-free representation $\mathbf{V}_{t_j}^{exp}$ for the j^{th} triangle, and w_j is the weighting coefficient proportional to the number of pixels within the j^{th} triangle. Eq. (1) is a least-square problem and has a closed-form solution as:

$$\mathbf{x}_t^e = vec\left[\sum_{j=1}^{106} w_j \mathbf{Q}_j \mathbf{V}_{t_j}^{neu} \mathbf{S}_j^T \left(\sum_{l=1}^{106} w_l \mathbf{S}_l \mathbf{S}_l^T \right)^{-1} \right]. \tag{2}$$

3.2 Estimating Appearance Changes from Shape Changes

Once we have transferred the shape changes, our next step is to transfer appearance changes. We normalized the pixel intensity (appearance) of a face image by warping the pixels within each triangle to their corresponding pixels on a common template. $\mathbf{y}_i^{neu} \in \Re^{d\times 1}$ is the normalized appearance vector (d pixels) for the i^{th} subject. $\mathbf{Y}_i^e = \{\mathbf{y}_i^{e_1}, \cdots, \mathbf{y}_i^{e_{n_i}}\} \in \Re^{d\times n_i}$ contains the appearance vectors for the n_i face images. Directly transferring appearance changes is extremely difficult due to the high

dimensionality of the image data and potentially nonlinearity of the facial actions. This process typically produces unwanted artifacts (see Fig 5 "Copy"). A key observation is that there is a strong correlation between the shape changes and the appearance changes performed by the same person. Fig. 3 shows the projection onto the first three principal components of the shape and appearance for 4 sequences of the same person. The shape and appearance are projected independently, and then an affine transformation is computed to align shape to appearance. As we can observe (also numerically validated in Sec.4.1.), there is a strong correlation between shape and appearance, that allows us to predict the appearance from the shape. See caption of Fig. 3 for more details.

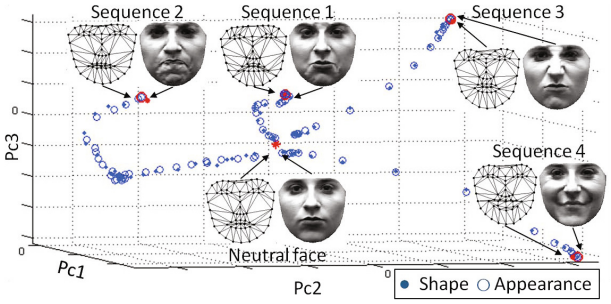

Fig. 3. Low dimensional embedding that shows the correlation between shape ("•") and appearance changes ("○") for four image sequences (82 face images in total) of subject "S014" in the Cohn-Kanade (CK) database [22]. The circles ("○") are the projections of appearance changes of each face image with respect to the neutral face ($\mathbf{y}^e - \mathbf{y}^{neu}$) along the first three principal components (denoted by axis Pc1-Pc3). The dots ("•") represent shape changes ($\mathbf{x}^e - \mathbf{x}^{neu}$) projected in the three principal components after linear regression (alignment) to ($\mathbf{y}^e - \mathbf{y}^{neu}$).

After transferring the shape changes to the target, and assuming that we have training samples of the target person, we could build a person-specific model by explicitly learning a mapping $\mathbf{T}$ between shape ($\mathbf{x}$) and appearance ($\mathbf{y}$) that minimizes:

$$\min_{\mathbf{T}} \sum_{e \in \Omega_i} \|(\mathbf{y}_i^e - \mathbf{y}_i^{neu}) - \mathbf{T}\,(\mathbf{x}_i^e - \mathbf{x}_i^{neu})\|_2^2\,, \tag{3}$$

where Ω_i is the index set of available instances of the i^{th} subject performing different facial actions, $\mathbf{T} \in \Re^{d \times v}$ is the regression matrix that maps the displacement of facial landmarks to the change of pixel intensity. Solving Eq. (3) leads to a person-specific regression matrix $\mathbf{T}$. The column space of $\mathbf{T}$ correlates with the appearance changes and its row space with the shape changes of the i^{th} subject.

In realistic scenarios, we will not have training samples of the target subject. An alternative is to learn a generic regression using images from the training subjects but ignoring images from the target subjects. However, as we will show in the experimental part, the learned generic regression can only predict the averaged appearance changes for all training subjects but it fails to model specific facial features from the target image. In addition, training instances would need to be well-sampled in order to generalize to all facial actions to be transferred. In the following section, we propose to personalize the mapping $\mathbf{T}$ given only a neutral face image of the target.

3.3 Personalizing the Regression

For the i^{th} person ($i = 1, \cdots, p$), the person-specific mapping $\mathbf{T}_i \in \Re^{d \times v}$ minimizes

$$E(\mathbf{T}_i) = \|\mathbf{Y}_{\Delta i} - \mathbf{T}_i \mathbf{X}_{\Delta i}\|_F^2 = \left\|\mathbf{Y}_{\Delta i} - \mathbf{B}_i \mathbf{A}_i^T \mathbf{X}_{\Delta i}\right\|_F^2, \tag{4}$$

where $\mathbf{Y}_{\Delta i} = \mathbf{Y}_i - \mathbf{y}_i^{neu} \mathbf{1}_{n_i}^T \in \Re^{d \times n_i}$ and $\mathbf{X}_{\Delta i} = \mathbf{X}_i - \mathbf{x}_i^{neu} \mathbf{1}_{n_i}^T \in \Re^{v \times n_i}$ contain the appearance and shape changes w.r.t. the neutral face, for all images belonging to the i^{th} person. To avoid the curse of dimensionality and reduce the amount of parameters to estimate, the row and column subspaces of $\mathbf{T}_i$ are explicitly modeled by the outer product of two matrices $\mathbf{T}_i = \mathbf{B}_i \mathbf{A}_i^T$ [5], so that the column space of $\mathbf{B}_i \in \Re^{d \times k}$ and $\mathbf{A}_i \in \Re^{v \times k}$ are respectively correlated with the shape and appearance subspace of this person ($k = rank(\mathbf{T}_i)$). A necessary condition for the existence of a minimum of E (in Eq. (4)) w.r.t. $\mathbf{B}_i$ and $\mathbf{A}_i$ has to satisfy $\frac{\partial E}{\partial \mathbf{B}_i} = \mathbf{0}$ and $\frac{\partial E}{\partial \mathbf{A}_i} = \mathbf{0}$, which leads to

$$\mathbf{B}_i = \mathbf{Y}_{\Delta i} \mathbf{X}_{\Delta i}^T \mathbf{A}_i (\mathbf{A}_i^T \mathbf{X}_{\Delta i} \mathbf{X}_{\Delta i}^T \mathbf{A}_i)^{-1}, \tag{5}$$

$$\mathbf{A}_i = (\mathbf{X}_{\Delta i} \mathbf{X}_{\Delta i}^T)^{-1} \mathbf{X}_{\Delta i} \mathbf{Y}_{\Delta i}^T \mathbf{B}_i (\mathbf{B}_i^T \mathbf{B}_i)^{-1}. \tag{6}$$

Eq. (5) and (6) imply that the columns of $\mathbf{B}_i$ and $\mathbf{A}_i$ are in the subspaces spanned by the appearance changes $\mathbf{Y}_{\Delta i}$ and the shape changes $\mathbf{X}_{\Delta i}$ respectively. If we solve Eq. (4) over all facial actions performed by the i^{th} subject, the column spaces of $\mathbf{B}_i$ and $\mathbf{A}_i$ are optimized to capture the specific shape and appearance changes for this subject, respectively. However, as mentioned before the person-specific model is not available in many applications, and the generic model is not accurate enough. Our goal in this section is to personalize $\mathbf{B}_i$ and $\mathbf{A}_i$ using only one neutral face.

A key aspect to build a personalized model from one sample, is to realize that from a neutral image, we can predict many different facial expressions [5,6,4,16]. Observe that the neutral image has enough information to generate an approximation of the texture of a particular face under several expressions. That is, $\mathbf{Y}_{\Delta i} \approx [\mathbf{R}_1 \mathbf{y}_i^{neu} \ \cdots \ \mathbf{R}_n \mathbf{y}_i^{neu}]$, where $\mathbf{R}_i$ is a regressor for a particular expression ($i = 1, \cdots, n$). However, learning this expression-specific regressions requires carefully posed expressions and labeling expression across all subjects [5,6,4,16].

In this paper, we overcome this limitation by explicitly learning the mapping from the appearance and shape of a neutral face to the person-specific matrices $\mathbf{B}_i$ and $\mathbf{A}_i$ for all training subjects ($i = 1, \cdots, p$). Following recent work on subspace regression [23], we learn a mapping between a neutral face and a subspace of shape and appearance. That is, the person-specific subspace is parameterized as: $\mathbf{B}_i \approx [\mathbf{W}_1 \mathbf{y}_i^{neu} \cdots \mathbf{W}_k \mathbf{y}_i^{neu}]$, where $\mathbf{W}_k \in \Re^{d \times d}$. In practice the number of parameters for each $\mathbf{W}_k$ is large and some rank constraints have to be imposed to avoid overfitting. Alternatively, to solve this problem we kernealize the previous expression as: $vec(\mathbf{B}_i) \approx \mathbf{B}\varphi_y(\mathbf{y}_i^{neu})$, where $\mathbf{B} \in \Re^{dk \times w_y}$ is a transformation matrix, and $\varphi_y(\cdot) \in \Re^{w_y}$ is a kernel mapping of the neutral appearance from the d spaces to a w_y (possible infinite) dimensional space. Observe that, the columns of matrix $\mathbf{B}$ span the dk dimensional space of person-specific matrices $\mathbf{B}_i$'s that model possible appearance changes for all subjects ($i = 1, \cdots, p$). Similarly, $vec(\mathbf{A}_i) \approx \mathbf{A}\varphi_x(\mathbf{x}_i^{neu})$ where $\mathbf{A} \in \Re^{vk \times w_x}$ spans the vk dimensional spaces of person-specific matrices $\mathbf{A}_i$'s that model possible shape changes for all subjects ($i = 1, \cdots, p$), and $\varphi_x(\cdot) \in \Re^{w_x}$ is kernel mapping of the neutral shape.

As in traditional kernel methods, we will assume that the rows of the matrix $\mathbf{B}$ can be expanded as the combination of $\varphi_y(\mathbf{Y}^{neu})$, i.e., $\mathbf{B} = \mathbf{R}_B\varphi_y(\mathbf{Y}^{neu})^T$, where $\mathbf{R}_B \in \Re^{dk\times p}$ is a coefficient matrix and $\mathbf{Y}^{neu}$ contains all neutral appearances for $i = 1, .., p$. Similarly, the row vectors of $\mathbf{A}$ can be spanned by $\varphi_x(\mathbf{X}^{neu})$, i.e., $\mathbf{A} = \mathbf{R}_A\varphi_x(\mathbf{X}^{neu})^T$, where $\mathbf{X}^{neu}$ contains all neutral shapes. Using the kernel trick, we can re-write $\mathbf{B}_i$ and $\mathbf{A}_i$ in a more compact form as:

$$\mathbf{B}_i \approx \mathbf{T}_B(\mathbf{k}_{\mathbf{y}_i^{neu}} \otimes \mathbf{I}_k), \quad \mathbf{A}_i \approx \mathbf{T}_A(\mathbf{k}_{\mathbf{x}_i^{neu}} \otimes \mathbf{I}_k), \tag{7}$$

where $\mathbf{T}_B \in \Re^{d\times kp}$ contains re-organized elements of $\mathbf{R}_B$, $\mathbf{T}_A \in \Re^{v\times kp}$ contains re-organized elements of $\mathbf{R}_A$. $\mathbf{k}_{\mathbf{y}_i^{neu}} = \varphi(\mathbf{Y}^{neu})^T\varphi(\mathbf{y}_i^{neu}) \in \Re^p$ is the kernel vector measuring the similarity between the i^{th} person with other subjects for the neutral appearance. Similarly, $\mathbf{k}_{\mathbf{x}_i^{neu}} = \varphi(\mathbf{X}^{neu})^T\varphi(\mathbf{x}_i^{neu}) \in \Re^p$ is the kernel vector measuring the similarity between the i^{th} person with other subjects for neutral shapes.

Now we can rewrite the error in Eq. (4) by combining it with Eq. (7) as:

$$\min_{\mathbf{T}_B,\mathbf{T}_A} E(\mathbf{T}_B, \mathbf{T}_A) = \sum_{i=1}^{p} \left\|\mathbf{Y}_{\Delta i} - \mathbf{T}_B\mathbf{M}_i\mathbf{T}_A^T\mathbf{X}_{\Delta i}\right\|_F^2, \tag{8}$$

where $\mathbf{M}_i = (\mathbf{k}_{\mathbf{y}_i^{neu}} \otimes \mathbf{I}_k)(\mathbf{k}_{\mathbf{x}_i^{neu}} \otimes \mathbf{I}_k)^T \in \Re^{kp\times kp}$. To estimate $\mathbf{T}_B$ and $\mathbf{T}_A$, we use an alternated least square (ALS) method to monotonically reduce the error of E. ALS alternates between optimizing for $\mathbf{T}_A$ while $\mathbf{T}_B$ is fixed, and vice versa. This is guaranteed to converge to a critical point of E. The update equations for ALS are:

$$\mathbf{T}_B = \left(\sum_{i=1}^{p} \mathbf{Y}_{\Delta i}\mathbf{X}_{\Delta i}^T\mathbf{T}_A\mathbf{M}_i^T\right)\left(\sum_{i=1}^{p} \mathbf{M}_i\mathbf{T}_A^T\mathbf{X}_{\Delta i}\mathbf{X}_{\Delta i}^T\mathbf{T}_A\mathbf{M}_i^T\right)^{-1}, \tag{9}$$

$$vec(\mathbf{T}_A) = \left(\sum_{i=1}^{p} \mathbf{H}_i \otimes \mathbf{G}_i\right)^{-1} vec\left(\sum_{i=1}^{p} \mathbf{X}_{\Delta i}\mathbf{Y}_{\Delta i}^T\mathbf{T}_B\mathbf{M}_i\right), \tag{10}$$

where $\mathbf{H}_i = \mathbf{M}_i^T\mathbf{T}_B^T\mathbf{T}_B\mathbf{M}_i$ and $\mathbf{G}_i = \mathbf{X}_{\Delta i}\mathbf{X}_{\Delta i}^T$. Given an initial guess of $\mathbf{T}_B$ and $\mathbf{T}_A$, Eq. (9) and (10) are alternated until convergence.

For a new target person t, we represent the neutral face by the shape $\mathbf{x}_t^{neu}$ and appearance $\mathbf{y}_t^{neu}$, and compute the personalized regression matrices as $\mathbf{B}_t = \mathbf{T}_B(\mathbf{k}_{\mathbf{y}_t^{neu}} \otimes \mathbf{I}_k)$ and $\mathbf{A}_t = \mathbf{T}_A(\mathbf{k}_{\mathbf{x}_t^{neu}} \otimes \mathbf{I}_k)$. Given the target shape change transferred from the source (Section 3.1), the target appearance change $\mathbf{y}_{\Delta t}^e$ is predicted using the personalized bilinear regression as $\mathbf{y}_{\Delta t}^e = \mathbf{B}_t\mathbf{A}_t^T(\mathbf{x}_t^e - \mathbf{x}_t^{neu})$. Finally, the appearance vector of the target person under expression "e" is computed as $\mathbf{y}_t^e = \mathbf{y}_t^{neu} + \mathbf{y}_{\Delta t}^e$.

4 Experiments

This section provides quantitative and qualitative evaluation of our method on two challenging databases: **(1) Cohn-Kanade (CK) database** [22]: This database contains posed facial action sequences for 100 adults. There are small changes in pose and illumination in the sequences. Each person has several sequences performing different

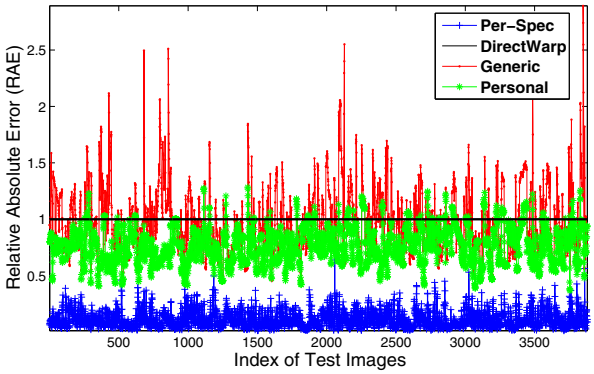

Fig. 4. Comparison of appearance reconstruction error over a random testing set from the CK database. RAE is computed for 4 methods: "Per-Spec", "Direct warp", "Generic" and our method "Personal". Note "Personal" uses only one neutral image per target subject.

expressions for about 20 frames, beginning at neutral and ending in the peak of the expression. We used 442 sequences posed by 98 persons for which the 66 landmarks per frame were available. The total number of images used in our experiments is 8036. **(2) RU-FACS database** [24]: This database consists of videos of 34 young adults recorded during an interview of approximately 2 minutes, duration in which they lied or told the truth in response to an interviewer's questions. Head pose was mostly frontal with small to moderate out-of-plane head motion. Image data from five subjects could not be analyzed due to image artifacts. Thus, image data from 29 subjects was used.

4.1 Reconstruction Ability of the Shape-to-Appearance Mapping

In the first experiment, we evaluated how well several shape-to-appearance mapping methods (Section 3.2 and 3.3) can reconstruct the facial appearance of subjects in the CK database. We compared the performance of four approaches: (1) "Generic" mapping, computed for all subjects similar to [4] as explained in Section 3.2; (2) Person-specific mapping ("Per-Spec"), where we learn a different regressor $\mathbf{T}_i$ for each subject (Eq. (4)); (3) Our personalized bilinear mapping ("Personal") estimated given only a neutral face of each person (Eq. (8)); (4) "Direct warp", which is the baseline method where we directly warped the pixels of neutral appearances $\mathbf{y}^{neu}$ to a common template. We computed the Relative Absolute Error (RAE) between the estimated and the ground truth appearances on the common template as: $RAE = \frac{|\mathbf{y}^{exp} - \widetilde{\mathbf{y}}^{exp}|}{|\mathbf{y}^{exp} - \mathbf{y}^{neu}|}$, where $\widetilde{\mathbf{y}}^{exp}$ is the

Table 1. Comparison of averaged RAE of appearances for four methods on CK database over 20 repetitions. Note "Per-Spec" reproduces the appearances of training images for each target subject while "Personal" predicts the appearances using only one neutral face per subject.

	Per-Spec	Direct warp	Generic	Personal (Our method)
RAE	$0.11 \pm 0.07\%$	$1 \pm 0\%$	$0.93 \pm 0.29\%$	$0.74 \pm 0.15\%$

estimated appearance and $\mathbf{y}^{exp}$ is the ground truth appearances. Observe that the baseline method ("Direct warp") produces the same appearance as the neutral face on the common template (i.e. $\widetilde{\mathbf{y}}^{exp} = \mathbf{y}^{neu}$) and its RAE is 1 for all images.

For each subject, we consider the neutral face as the average of the first frame over all the sequences (between 2 and 6) for each subject. We randomly selected half of the subjects (49) for training and cross-validation, and the other 49 subjects are used for testing. We used Gaussian kernels in Eq. (7) to respectively measure the similarity among neutral shapes and appearances. The bandwidth parameters for the neutral shape and appearance Gaussian kernel were chosen by cross-validation. We repeated the experiments 20 times. The average and standard deviation are summarized in Table 1. Fig. 4 shows an instance of the RAE on a randomly selected test set.

As shown in Fig. 4 and Table 1, the "Generic" mapping produces the largest error because it computes the averaged appearance changes which, in many cases, are not the appearances the target (test) subject can produce. As expected from Fig. 3, the "Per-Spec" method achieves the least error because it learns the person-specific regression using the data of this particular subject. This further validates the strong correlation between shape-appearance used in Sec.3.2. Note this is the ideal scenario, where we have images of the subject to train and test. Unfortunately, the "Per-Spec" mapping is typically not available in practice because of the lack of training samples for the target person (only one frontal image is available). Finally, our method ("Personal") produces lower RAE than the generic mapping using only a neutral face of each person.

4.2 Facial Action Transfer

This section presents qualitative (visual) evaluation for FAT on the CK database. Note that this is a different experiment from the one in the previous section. Now our goal is to transfer the expression to a different subject, and not to reconstruct the appearance from the shape of the same subject.

We used 49 subjects to learn the regressors as explained in the previous section. The other 49 subjects are used as target subjects. The source for these 49 targets are randomly selected from the other 49 training subjects. We transfer the facial actions of the source (Fig. 5 the "Source" column) to the target subjects. Fig. 5 shows the results for four methods: (1) Copying directly the shape and appearance changes from the source to the target similar to [7], the "Copy" column; (2) Person-specific mapping learned from available instances of each target person (for evaluation only, instances usually are not available in practice), the "Per-Spec" column; (3) The generic mapping computed for all training subjects [4], the "Generic" column; and (4) our personalized regression, estimated from a neutral face of the target, the last column: "Personal". Here we do not compare with "Direct warp" as in the previous subsection because it produces no appearance changes. The pixels within the mouth (e.g., teeth) in Fig. 5-7 are directly warped from the images of the source subject.

As shown in Fig. 5, the direct coping method ("Copy") does not adapt the expression changes of the source subjects to the specific facial features of the targets. It produced strong artifacts around eye brows, cheeks and jaws on the target faces. The person-specific method (the "Per-Spec" column) performed very well in reconstructing the appearance in the last experiment; however, it behaved poorly in the experiment of

Fig. 5. Facial action transfer from the source persons (the "Source" column) to the target persons (the neutral faces in the "Target" column) using the direct copying ("Copy"), personal specific mapping ("Per-Spec"), the generic mapping ("Generic") and our method ("Personal")

transferring the expression. In this case, we learn the regression from shape to appearance using the video of the target person, but there is no guarantee that the expression performed by the source will be represented in the available data of the target person. This is the reason why the person-specific method performs poorly. In fact, it is usually difficult to get well-sampled instances of the target subject. The generic mapping ("Generic") generates averaged appearance changes, which in many cases does not fit the specific facial features of the target subjects. Our method ("Personal") estimates the personalized regression from only a neutral face of the target person, and it produces video-realistic personalized facial actions. Although the ground truth for the transferred expressions are not available, the result of our personal mapping is the one that is visually more video-realistic [2].

4.3 Face De-Identification

Advances in camera and computing equipment hardware in recent years have made it increasingly easy to capture and store extensive amounts of video data. This, among

[2] See more results at http://www.andrew.cmu.edu/user/dghuang/fat.htm

other things, creates many opportunities for the sharing of video sequences. In order to protect the privacy of subjects visible in the scene, automated methods to de-identify the images, particularly the face region, are necessary [1]. So far the majority of privacy protection schemes currently used in practice rely on ad-hoc methods such as pixelation or blurring of the face. This section explores the use of our method for face de-identification in real interviews when there is subtle, spontaneous facial expressions combined with pose changes using the RU-FACS database.

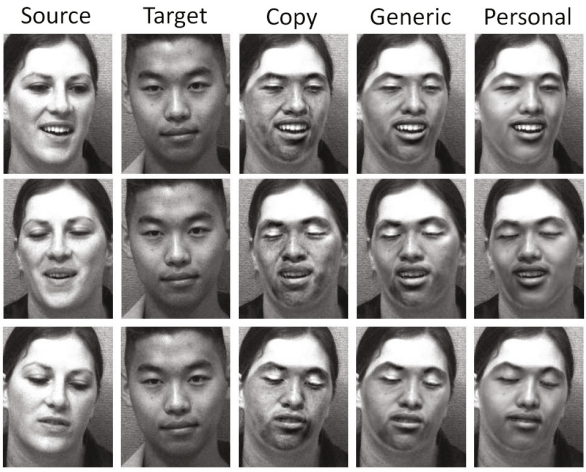

Fig. 6. Application of FAT to de-identification: replacing the identity of the source persons (the "Source" column) with that of the target person (the "Target" column) performing the same facial actions. Three methods compared: Direct copying ("Copy"), the generic mapping ("Generic") and our method ("Personal"). The "Copy" and "Generic" columns produce artifacts on the cheek, eyelids and around the mouth. To clearly illustrate the de-identified/replaced face region on the source images, we do not adapt the target skin to the source skin.

As shown in Fig. 6-7, we transferred the facial actions of the source person (the "Source" column) to the target person (the "Target" column), and then warped the transferred target facial image patch back to the image of the source person. In this way, we replaced the identity of the source person with the target person performing the same facial actions. Three methods compared are: Direct copying ("Copy"), the generic mapping ("Generic") and our method ("Personal"). For each pair of source and target persons in Fig. 6-7, we show three instances of de-identification including various head poses and subtle facial actions. In all the instances, both "Copy" and "Generic" generate exceptional bright or dark texture in cheeks, eyelids, jaw and eyebrows. This is because the shape and appearance changes imposed in those regions are not suitable for the target facial features. Using the personalized regression estimated from one target neutral face, our method ("Personal") produces the realistic personalized facial actions compatible with the target facial features such as eyes, nose, lips and the skin. We ex-

Fig. 7. Application of FAT to face de-identification (continued with Fig. 6). The "Copy" and "Generic" columns produce artifacts on the cheek, eyebrows and the jaw.

plicitly use a person of different color skin as target to illustrate the de-identification process.

5 Conclusion

This paper presents a personalized supervised bilinear regression method for FAT. We have illustrated how our algorithm can outperform state-of-the-art methods for generating video-realistic face images of a target subject from a source video. Moreover, we illustrated how our method can also be used for face de-identification in the challenging RU-FACS database. The main reason for the superior performance of our method is that is able to personalize the relation between the shape and appearance changes using only one target neutral face. This brings two main advantages: (1) Our learning method does not rely on expression correspondences which are difficult to obtain; (2) To transfer the facial actions to a new person, our method only requires the neutral face of the person, which is a realistic assumption in many scenarios. Further improvements of our method can be achieved by using local models for different face regions and better regression methods to model the relations between the bilinear regression matrices and the neutral face. Finally, although we have illustrated the benefits of our approach on the problem of FAT, our method is more general and can be applied to other appearence/texture synthesis problems.

References

1. Gross, R., Sweeney, L., De la Torre, F., Baker, S.: Model-based face de-identification. In: CVPR Workshop on Privacy Research in Vision (2006)
2. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M.: B. Wu, A.B., Adam, H., Neven, H., Vincent, L.: Largescale privacy protection in google street view. In: ICCV (2009)
3. Senior, A. (ed.): Protecting Privacy in Video Surveillance. Springer (2009)
4. Saragih, J., Lucey, S., Cohn, J.: Real-time avatar animation from a single image. In: AFGR (2011)
5. Huang, D., De la Torre, F.: Bilinear Kernel Reduced Rank Regression for Facial Expression Synthesis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 364–377. Springer, Heidelberg (2010)
6. Zhang, Q., Liu, Z., Guo, B., Shum, H.: Geometry-driven photorealistic facial expression synthesis. In: ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 177–186 (2003)
7. Liu, Z., Shan, Y., Zhang, Z.: Expressive expression mapping with ratio images. In: Ann. Conf. on Computer Graphics and Interactive Techniques (2001)
8. Noh, J., Neumann, U.: Expression cloning. SIGGRAPH, 277–288 (2001)
9. Chai, J., Xiao, J., Hodgins, J.: Vision-based control of 3d facial animation. In: Eurographics (2003)
10. Sumner, R., Popovic, J.: Deformation transfer for triangle meshes. SIGGRAPH 23, 399–405 (2004)
11. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. SIGGRAPH 24, 408–416 (2005)
12. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear Analysis of Image Ensembles: TensorFaces. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 447–460. Springer, Heidelberg (2002)

13. Tenenbaum, J., Freeman, W.: Separating style and content with bilinear models. Neural Computation 12, 1247–1283 (2000)
14. Wang, H., Ahuja, N.: Facial expression decomposition. In: ICCV (2003)
15. Chung, K.: Gross Anatomy (Board Review). Lippincott Williams & Wilkins, Hagerstown (2005)
16. De La Hunty, M., Asthana, A., Goecke, R.: Linear facial expression transfer with active appearance models. In: ICPR (2010)
17. Abboud, B., Davoine, F.: Appearance factorization for facial expression analysis. In: BMVC (2004)
18. Vlasic, D., Brand, M., Pfister, H., Popovic, J.: Face transfer with multiliner models. ACM Trans. Graphics 24, 426–433 (2005)
19. Macedo, I., Brazil, E., Velho, L.: Expression transfer between photographs through multilinear aam's. In: SIBGRAPI, pp. 239–246 (2006)
20. Elgammal, A., Lee, C.: Separating style and content on a nonlinear manifold. In: CVPR, pp. 478–485 (2004)
21. Wang, Y., Huang, X., Lee, C., Zhang, S., Li, Z., Samaras, D., Metaxas, D., Elgammal, A., Huang, P.: High resolution acquisition, learning and transfer of dynamic 3D facial expressions. Computer Graphic Forum 23, 677–686 (2004)
22. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: AFGR (2000)
23. Kim, M., Zhang, Z., De la Torre, F., Zhang, W.: Subspace regression: Predicting a subspace from one sample. In: Asian Conference on Computer Vision, ACCV (2010)
24. Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Automatic recognition of facial actions in spontaneous expressions. Journal of Multimedia 1, 22–35 (2006)