Signal Processing 91 (2011) 2465–2477

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Expression transfer for facial sketch animation

Yang Yang [a,b,*], Nanning Zheng [a], Yuehu Liu [a], Shaoyi Du [a], Yuanqi Su [a], Yoshifumi Nishio [b]

[a] The Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
[b] The Department of Electrical and Electronic Engineering, The University of Tokushima, Tokushima 770-8506, Japan

ARTICLE INFO

Article history:
Received 13 May 2010
Received in revised form
10 April 2011
Accepted 15 April 2011
Available online 22 April 2011

Keywords:
Facial sketch
NET model
Hierarchical animation

ABSTRACT

This paper presents a hierarchical animation method for transferring facial expressions extracted from a performance video to different facial sketches. Without any expression example obtained from target faces, our approach can transfer expressions by motion retargetting to facial sketches. However, in practical applications, the image noise in each frame will reduce the feature extraction accuracy from source faces. And the shape difference between source and target faces will influence the animation quality for representing expressions. To solve these difficulties, we propose a robust neighbor-expression transfer (NET) model, which aims at modeling the spatial relations among sparse facial features. By learning expression behaviors from neighbor face examples, the NET model can reconstruct facial expressions from noisy signals. Based on the NET model, we present a hierarchical method to animate facial sketches. The motion vectors on the source face are adjusted from coarse to fine on the target face. Accordingly, the animation results are generated to replicate source expressions. Experimental results demonstrate that the proposed method can effectively and robustly transfer expressions by noisy animation signals.

1. Introduction

As a high-level abstraction of human faces, facial sketch uses a small number of strokes to represent basic facial characteristics. It has wide applications in the entertainment industry and the public security. For example, by rendering facial sketches, suspect can be recognized easily based only on the witness description. For these applications, many methods [1–4] have been proposed recently to render static facial sketches. Chen et al. [1] provided an example-based composite method to synthesize sketches. Xu et al. [3] presented a hierarchical method to draw sketches according to a face image from the low to high resolutions.

However, how to dynamically generate a facial sketch animation is still a challenging research area. This is because facial sketch only contains sparse features of face shape and lacks spatial self-constraints necessary for animation. Some animation methods [5–7] require the expression examples of target faces. For example, Buck et al. [5] proposed the performance-driven method to drive hand-drawn characters, and Sucontphunt et al. [6] utilized the geometry-driven approach to interactively edit portraits. The problem associated with these methods is that the target expression examples are difficult to prepare in practice. On the other hand, motion retargetting methods [8–10] have been suggested to animate novel faces by transferring expressions without any example, in which motion vectors are adjusted directly from the source face onto target faces. But these motion retargetting methods are only available for dense facial features.

To deal with the above difficulties, an effective face model is crucial. Traditional face models, such as the

* Corresponding author at: The Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China.
E-mail address: yyang@aiar.xjtu.edu.cn (Y. Yang).

0165-1684/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2011.04.020

active shape model (ASM) [11] and the morphable model (MM) [12], represent face in a general-face-parameter space to guarantee a natural face shape. However, they are hard to distinguish face features from noisy signals. In the practical animation process, noise always exists: (1) the image noise reduces the accuracy of expression extraction; (2) the transmission noise perturbs animation signals; (3) due to the difference between two faces, source expression signals are also too noisy to be directly applicable to target faces. Therefore, a more robust face model is required in order to present the personal attribute from noise. Otherwise, the animation result may not look like the original target face.

In this paper, first, a neighbor-expression transfer (NET) model is built by employing the expression behaviors from neighbor examples. The NET model can reconstruct target expression faces from noisy signals robustly. Second, based on the NET model, we propose a method to transfer facial expressions extracted from a source performance video to facial sketches, without requiring the expression examples of target faces. A hierarchical method is designed for expression transfer. Different from the hierarchical sketch rendering method [3], we utilize the correspondence map provided from lower level to higher level, to adjust motion vectors from coarse to fine to target faces. Finally, the NET model is evaluated by comparing to other models under different experimental environments, and the evaluation of the animation method is performed on the realistic expression data. The experimental results verify that our method is robust and effective to animate facial sketches.

The remainder of this paper is organized as follows. We review the related work of facial animation in Section 2, and briefly introduce the MM in Section 3. In Section 4, we present the proposed NET model, including the model overview, model computation and the analysis about two special cases. Following that is Section 5, in which the hierarchical expression transfer method is presented. We discuss the experimental results in Section 6, and finally in Section 7 conclude the paper with a summary.

2. Related work

During the past few years, there has been a large amount of research on the facial animation. The animation methods can be roughly divided into two categories: one requires expression examples of target face; and the other generates expressions with no target examples.

2.1. First category

The blendshape [13–15] is an intuitive method which has been widely used in the computer animation industry. By varying the blending weights, a full range of facial expressions can be represented by a set of key examples. However, it is an annoying work to adjust blend weights by a trial and error process even for professional animators. So some research [13,15] efforts attempt to reduce the blendshape interference, like segmenting the face model into smaller regions for manipulating conveniently. Some performance-driven methods [5,16–19] have greatly improved the efficiency of the animation process. Their methods combine the motion capture data and the blendshape interpolation, to drive facial animations by reusing the blend weights from existing motion data.

Pyun et al. [20] proposed an example-based expression cloning approach. By building sufficient correspondence maps between source and target examples, it produces satisfactory results for transferring expressions between very different faces. Similar idea has been used in some related work. For example, Buck et al. [5] used 2D facial videos to hand-drawn characters, Na et al. [21] used human face to ox and gorilla, and Sucontphunt et al. [6] gave a 3D face posing through 2D portrait. Nevertheless, without prepared target expression examples, they are invalid to animate novel faces. Our animation method is also based on the examples. But the examples are not limited to the target face. We learn the target's neighbor examples, which are easy to acquire in practice.

2.2. Second category

On the other hand, to generate expressions for novel target faces Blanz et al. [22] applied a simple vector space operation to add the 3D displacements of the surface points onto another person's neutral face. Similar approach was used by Williams [23]. While they ignored the expression variations across individuals, hence they are effective only for two similar face models. Some facial expression parameters were tried to define a general-face model, like the FAPs in MPEG-4 [24] for synthesizing simple expressions. Saveran et al. [25] generated visual speech by synthesizing 3D face points to drive the MPEG-4 facial animation.

A class of learning-based methods [26–29] is based on different people's expression examples that allow modeling of a full population. Abboud et al. [26,27] assumed the linear and bilinear relationships among neutral and some specific expressions in the active appearance model [30] (AAM) parameter space. Zhou et al. [28] used a kernel-based bi-factor factorization model to represent expression faces. Ghent et al. [29] applied several artificial neural networks to learn the transfer relationships among expressions. These methods are effective to synthesize some known expressions.

In order to transfer arbitrary expressions between different faces, Tao et al. [31] analyzed a 3D face data by the probabilistic model for tensors in the Bayesian framework, which has been applied for expression transfer. Liu et al. [32] proposed the expression ratio image (ERI) for transferring texture details between images of people. Some motion retargetting methods [8–10] were used for expression transfer, in which source face's motion vectors were adapted to the target face model directly. Noh et al. [8] built dense correspondences between model vertices, and transferred the motion vectors by adjusting magnitudes and directions according to local coordinate systems. Sumner et al. [9] reused the source animation data by transferring the affine deformation for each corresponding triangle mesh, and solved a global optimization problem with continuity constraints on the target surface. Song et al. [10] proposed a generic method for high-quality

expression transfer between both 3D face meshes and 2D face images. However, these methods require the dense surface features for building the face model. They also need to capture accurate motion features on the source face. The idea of our animation method is similar as the motion retargetting. But based on the proposed NET model, the expression transfer task can be performed on the sparse features under noisy conditions. We allow the expression data to be captured by the commonly used video device, like a camera.

3. Morphable model overview

The morphable model [12] (MM) represents an object class as shape and texture vectors using independent model parameters. It reconstructs an object by linearly combining examples. In this paper, we focus on the face shape object. A face shape vector S can be represented by the MM as:

$$S=\sum_{i=1}^{N}\alpha_i S_i+\varepsilon, \tag{1}$$

where $\{S_i|i=1,2,\ldots,N\}$ stand for face examples defined by some geometric features, e.g. a set of feature points. Their vector correspondences are established in a pre-processing phase to make sure all the faces have the same scale, pose and centroid position. In order to minimize the reconstruction error ε, a set of model parameters α_i is optimized as the least square problem:

$$\min_{\alpha_i}\left(\left\|S-\sum_{i=1}^{N}\alpha_i S_i\right\|^2\right), \tag{2}$$

with the constraint $\sum_{i=1}^{N}\alpha_i=1$.

Every object face is parameterized in the face vector space to avoid a non-face result. However, in the MM, it is hard to distinguish noise from the model parameters. When the object face contains noise, the reconstructed result may not look like the original face. Fig. 1 shows an example. For the comparison purpose, we also present the object face with neutral expression. We can see that when the random noise is added to the expression face, the MM has turned the original thick eyebrow into a different thin one.

To robustly maintain the target face attribute in the face model, especially its expression characteristics, is important for the expression transfer task. Because, besides the noise in real applications, the specific characteristics of the source face will also influence the target face to represent expressions. But the MM performs inadequately on this aspect.

4. Neighbor expression transfer model

In this section, the NET model is proposed. We give the details of model computation and the analysis of two special cases.

4.1. NET model

Our idea is to use prior knowledge learned from face examples to overcome the noise sensitivity. In particular, we assume that similar neutral faces also have similar expression faces. Therefore, we improve the importance of the similar examples for reconstructing target expression faces.

Before introducing the model details, we clarify some notations. S_{tar}^{neu} and S_{tar}^{exp} denote the object's neutral face and expression face, respectively. S_i^{neu} stands for one of the examples with neutral expression and S_i^{exp} is the corresponding expression example belonging to the face S_i^{neu}. In the expression transfer task, S_{tar}^{neu} is often given accurately, and S_{tar}^{exp} is the input to be reconstructed by the model.

Based on the above assumption, the NET model is proposed by modifying the MM on three aspects:

(1) Similarity constraint: A regularization term $\|S_{tar}^{neu}-\sum_{i=1}^{N}\alpha_i S_i^{neu}\|^2$ is added to Eq. (2), according to the similarity of neutral faces S_{tar}^{neu} and S_i^{neu}. This term penalizes the dissimilar examples for having smaller weights in the reconstruction.

(2) Expression function transfer: There is a reconstruction error ε in Eq. (1), since the examples fall short of representing the face details. For example, in Section 6.1.2, we will show an eyebrow's ridge is lost by the limited examples. However, it is reasonable to understand that the detail parts on the face are transformed according to their main components during making expressions. Based on this observation, we construct a set of synthesis expression examples to replace the examples S_i^{exp} to construct S_{tar}^{exp}. We learn the expression transform function F_i from each example face. The synthesis expression examples $F_i(S_{tar}^{neu})$ are acquired by transferring F_i to the target neutral face. More detail parts of the target face can be preserved in the synthesis expression examples.

(3) Neighbor example candidates: However, to transfer an expression function between largely different faces will generate improper expression examples. For example, a big eye's motion trace is not appropriate to a small eye. But based on our assumption, only similar examples occupy the main weights in the reconstruction. Therefore, in practice, we select S_{tar}^{neu}'s neighbor examples as candidates to learn F_i. And the expression function transfer

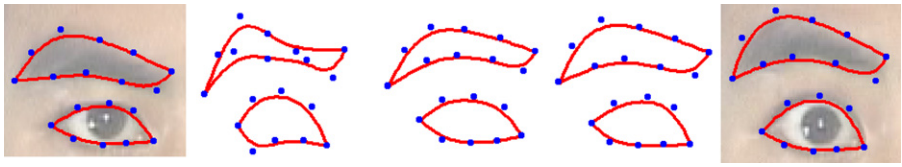

Fig. 1. Reconstruct noise expression face. (a) and (b) are the given neutral face and input noise expression facial; (c) and (d) are the reconstructed results of (b) by the MM and NET model, respectively. (e) gives the corresponding real expression face as a reference for comparison.

between neighbor examples will not cause improper synthesis results.

Accordingly, the formulation for the NET model is:

$$\min_{\alpha_i}\left(\left\|S_{tar}^{exp}-\sum_{i=1}^{K}\alpha_i F_i(S_{tar}^{neu})\right\|^2+\lambda\left\|S_{tar}^{neu}-\sum_{i=1}^{K}\alpha_i S_i^{neu}\right\|^2\right)$$
$$s.t.\ \alpha_i\geq 0,\sum_{i=1}^{K}\alpha_i=1, \tag{3}$$

where i is the notation for the S_{tar}^{neu}'s K-nearest neighbor (KNN) examples. Parameter λ is used to balance two terms. The parameters α_i must be non-negative and sum to one for rotational and scale invariance.

4.2. Model computation

The computation framework is illustrated in Fig. 2, which contains two components: off-line training and on-line computation.

The off-line training is a single routine for a given neutral face S_{tar}^{neu}. The expression database is composed by example pairs belong to different individuals. Each example pair $\{S_i^{neu},S_i^{exp}\}$ includes one neutral face and one corresponding expression face with an arbitrary expression. For every example whose $S_i^{neu}\in \mathrm{KNN}(S_{tar}^{neu})$:

1. Find expression transform function F_i such that $S_i^{exp}=F_i(S_i^{neu})$.
2. Transfer F_i for synthesis expression examples $F_i(S_{tar}^{neu})$.

With the constraint of neighbor example, the functions F_i could be performed flexibly, such as the geometric deformation function on the face.

The on-line computation consists of three steps:

1. Use expression animation signals to acquire the target expression face S_{tar}^{exp}. S_{tar}^{exp} is the reconstructed object in the NET model, which usually contains noise as illustrated in Fig. 1(b).
2. Calculate the weights α_i by Eq. (3). This is a quadratic problem with linear constraints, where the objective function is positive semi-definite, which could be solved by quadratic programming [33,34].
3. Reconstruct the target expression face as $\hat{S}_{tar}^{exp}$ by the weights α_i,

$$\hat{S}_{tar}^{exp}=\sum_{i=1}^{K}\alpha_i F_i(S_{tar}^{neu}). \tag{4}$$

There are two free parameters in the NET model: λ and K. When the examples are adequate, the model is not sensitive to these parameters. But when the examples are inadequate, usually big λ and small K are used to ensure S_{mod}^{exp} still looks like the same person as S_{tar}^{neu}.

4.3. Model analysis

We analyze two special cases of the NET model. One is to model known face with unknown expression; the other is to model known expression for unknown face. These special cases are also utilized by our expression transfer method presented in Section 5.2.

*Case*1: When a set of example expressions S_{tar}^{i} of the target face have been prepared in the database, the target face is known. Hence, Eq. (3) is simplified as S_i^{neu} are all same as S_{tar}^{neu}:

$$\begin{aligned}E_{case1}&=\left\|S_{tar}^{exp}-\sum_{i=1}^{K}\alpha_i F_i(S_{tar}^{exp})\right\|^2+\lambda\left\|S_{tar}^{neu}-\sum_{i=1}^{K}\alpha_i S_{tar}^{neu}\right\|^2\\&=\left\|S_{tar}^{exp}-\sum_{i=1}^{K}\alpha_i F_i(S_{tar}^{neu})\right\|^2\\&=\left\|S_{tar}^{exp}-\sum_{i=1}^{K}\alpha_i S_{tar}^{i}\right\|^2.\end{aligned} \tag{5}$$

Then the target expression face is reconstructed by S_{tar}^{i} according to weights α_i:

$$\hat{S}_{tar}^{exp}=\sum_{i=1}^{K}\alpha_i S_{tar}^{i}. \tag{6}$$

Based on Eq. (6), we can see that the special case 1 of the NET model is same as the blendshapes approach, which utilizes the key examples to generate new expressions.

*Case*2: When all the expression faces in the database are with the same expression, like smile, here we want to model this known expression for an arbitrary target face.

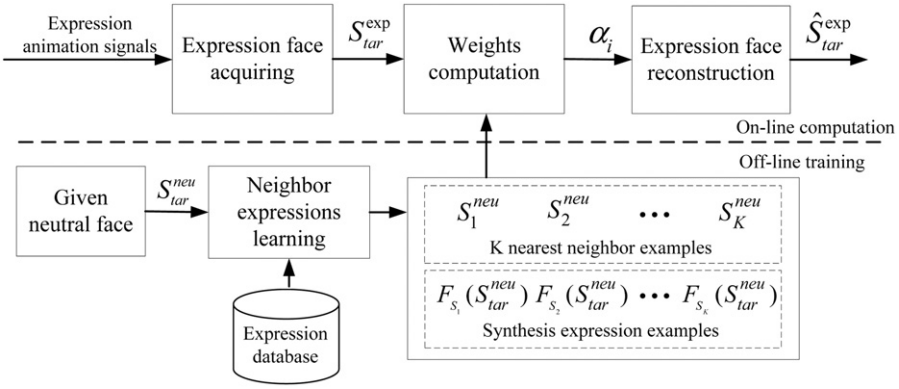

Fig. 2. The NET model computation framework.

The NET model is based on the assumption that the more similar neutral faces are, the more similar expressional faces they will have. Therefore, when the latter term of Eq. (3) is minimized, the former term of Eq. (3) will be minimized too, vice versa:

$$\min\left\|S_{tar}^{exp}-\sum_{i=1}^{K}\alpha_i F_i(S_{tar}^{neu})\right\|^2 \Leftrightarrow \min\left\|S_{tar}^{neu}-\sum_{i=1}^{K}\alpha_i S_i^{neu}\right\|^2. \quad (7)$$

We compute the weights by the similarity term of the neutral faces:

$$E_{case2}=\left\|S_{tar}^{neu}-\sum_{i=1}^{K}\alpha_i S_i^{neu}\right\|^2, \quad (8)$$

and use the weights to reconstruct the target face for the known expression. This idea has been achieved in the previous work [35] for synthesizing some basic expressions.

The above two special cases show that the NET model is a generic model in the facial expression animation field. Besides the applications in the special cases, its most value is to transfer arbitrary expressions to arbitrary target faces introduced in the next section.

5. Hierarchical expression transfer

A set of feature points are extracted on the source face shown in Fig. 3. The purpose of expression transfer is to transfer motion vectors of these features to the target face. We propose a hierarchical method. The motion vectors on the higher level are transferred based on correspondence maps provided by the lower level, and refined for the target face by the NET model.

5.1. Hierarchical facial sketch representation

We also extract the corresponding feature points on the target facial sketch. The hierarchical representation for the facial sketch is shown at the bottom of Fig. 3. There are three levels. The bottom level 1 is defined by a group of curves. It exhibits the global facial structure with location and scale information. The middle level 2 is the facial feature points to describe the local facial component shape. The top level 3 is the rendering details of the facial sketch, including other accessories on the face, e.g. glasses.

Correspondingly, the hierarchical facial sketch contains three-level parameters:

$$W=\left\{W^S,\bigcup_{i=1}^{7}W^{C_i},W^R\right\}.$$

W^S is the parameter of structure curves built by parabola models: single parabola fits eyebrow, nose and face contour; double parabolas fit eye, mouth outer and inner contours. W^S is calculated by fitting the subset of feature points. The parameter definitions are illuminated on the upper right in Fig. 3.

W^{C_i} is the parameter for each facial component, including right and left eyes, right and left eyebrows, nose, mouth and face contour, defined by point coordinates of their local features. W^{C_i} are independent of each other, in this case, at this level, different facial components will not impact each other.

W^R is the parameter for the rendering vector of the sketch, which is related to artist styles, e.g. line width and rendering layer.

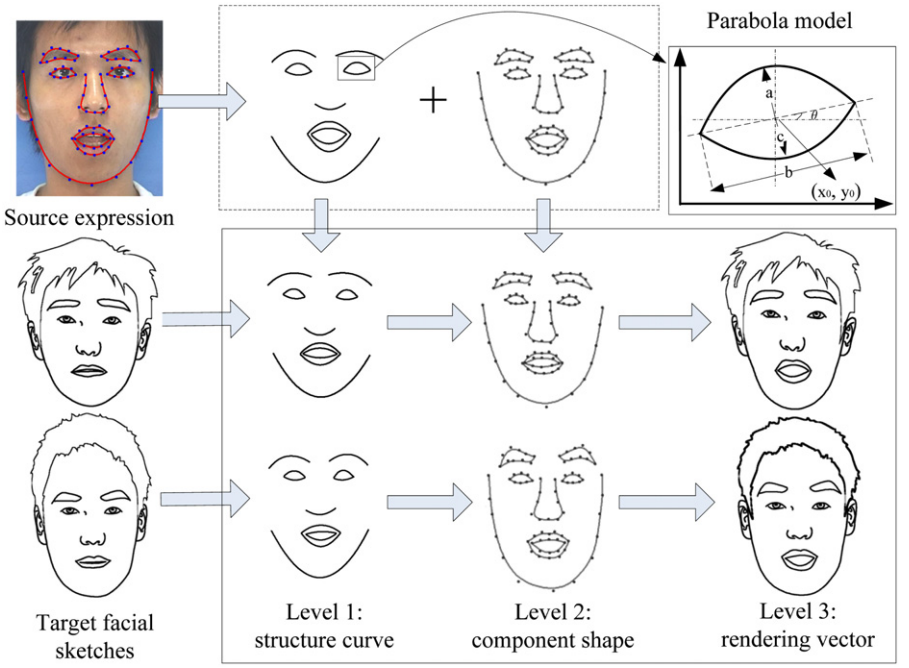

Fig. 3. The hierarchical expression transfer process.

5.2. Expression transfer method

The three-level expression transfer process is demonstrated in Fig. 3. Based on the feature points, the source face features are represented as the first two level parameters. After transferring expressions at the first two levels, accurate feature points can be acquired for the target expression face. Therefore, the facial sketch with expressions could be drawn by warping its rendering vectors according to the positions of these feature points. In this section, we mainly discuss the expression transfer method at the first two levels.

5.2.1. Level 1: structure curve animation

We utilize the special cases of the NET model to animate the structure curves. As in the special case 2, we select several example expressions as known expressions, and synthesize these expressions for the target face. We take the above synthesized results as the expression examples of target face in the special case 1, and generate new expressions by blending shapes.

Correspondingly, the frames containing the example expressions are chosen from the source expression video. The weights β_i for other source expressions are computed by Eq. (6). Now, only the weights need to be directly transferred to the target face.

More specifically, three steps for the expression transfer on level 1 are shown as follows:

1. Synthesize the target examples with known expressions and make pairs for source and target facial examples with the same expressions:

   $W_0^{S_{sou}} \Leftrightarrow W_0^{S_{tar}}, \ldots, W_M^{S_{sou}} \Leftrightarrow W_M^{S_{tar}}$,

   where W_0^S denote the structure curve of neutral face, M is the number of the selected known expressions.
2. Get weights β_i such that $W_{new}^{S_{sou}} = \sum_{i=1}^{M} \beta_i W_i^{S_{sou}}$.
3. Get $W_{new}^{S_{tar}}$ by weights β_i that $W_{new}^{S_{tar}} = \sum_{i=1}^{M} \beta_i W_i^{S_{tar}}$.

In the first step, expression transform function F_i is acquired by calculating the differences of parabola parameters between neighbor's neutral and expression examples. The synthesized expression examples $W_i^{S_{tar}} = F_i(W_0^{S_{tar}})$ are constructed by adding these differences to the structure curve parameters of the target neutral face.

The above expression transfer method is only suited for the level 1, because a small number of examples are inadequate to produce expression details for the higher level, whereas a large number of example expressions are difficult to prepare. In our applications, we choose 5–7 basic expressions in this level, like smile and anger.

5.2.2. Level 2: Component shape animation

Suppose that the structure curves of the target face have been obtained in the level 1, we build the correspondence map between the source and target facial components.

Specifically, the thin-plate spline [36] (TPS) mapping is used. We sample corresponding point sets $U = (u_1, u_2, \ldots, u_m)^T$ from the source structure curves and $V = (v_1, v_2, \ldots, v_m)^T$ from the target structure curves. The TPS fits a mapping function T in the form of

$$T(u_i) = v_i \cdot d + \phi(v_i) \cdot w, \tag{9}$$

by minimizing the following energy function:

$$E_{TPS}(T) = \sum_{i=1}^{m} \|v_i - T(u_i)\|^2 + \eta \int\int \left[\left(\frac{\partial^2 T}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 T}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 T}{\partial y^2}\right)^2 \right] dx\, dy, \tag{10}$$

where d and w represent the affine transformation and non-affine deformation parameters. The vector $\phi(\cdot)$ is related to the TPS kernel. η is a model parameter. An analytical solution of T can be obtained in the research [37].

For transferring the motion vectors at the level 2, we need to build two set of correspondence maps. One is T_{neu} between the source and target neutral faces, such that $W_{neu}^{S_{sou}} = T_{neu}(W_{neu}^{S_{tar}})$. The other is T_{exp} between the current source and target expression faces, such that $W_{exp}^{S_{sou}} = T_{exp}(W_{exp}^{S_{tar}})$.

Motion vectors are transferred according to T_{neu} and T_{exp} from the source face to target face. The facial component shape vectors are calculated by

$$W_{exp}^{C_{tar}} = W_{neu}^{C_{tar}} + T_{exp}(W_{exp}^{C_{sou}}) - T_{neu}(W_{neu}^{C_{sou}}). \tag{11}$$

Fig. 4 shows the expression transfer process for the level 2. The correspondence maps based on lower level's structure curves provide a coarse transferring result $W_{exp}^{C_{tar}}$ by Eq. (11). With the accurate target neutral face $W_{neu}^{C_{tar}}$, $W_{exp}^{C_{tar}}$ can be reconstructed by the NET model for the

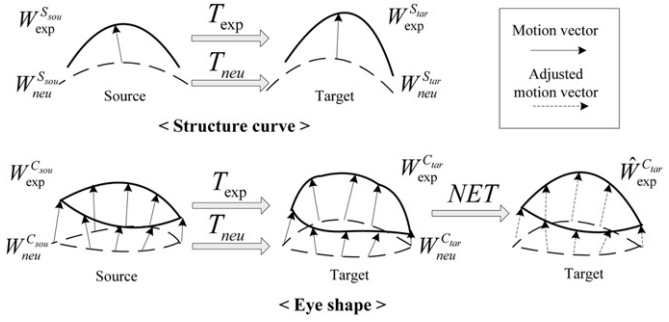

Fig. 4. Level 2: component shape animation.

final target expression face:

$$\hat{W}_{exp}^{C_{tar}} = NET(W_{neu}^{C_{tar}}, W_{exp}^{C_{tar}}). \quad (12)$$

In Eq. (12), the expression function F_i of the NET model is a kind of deformation function. Among neighbor examples, deformation parameters do not need to be adjusted for the target face. Many deformation methods like radius base function interpolation or the TPS could be used flexibly.

6. Experiments and results

Training examples are obtained from the Cohn–Kanade database [38] and the AIAR database [39], including the western and eastern peoples. The former contains several expression sequences for each person. The latter includes a set of expression face images of ten Asians. A total of 438 expression images are selected from 89 persons as examples. To learn the expression function in the NET model, each expression face and its neutral face compose an example pair. And the source and target faces for expression transfer are not restricted to these examples. Eighty feature points are tracked from the source expression videos by the ASM.

All the example images are aligned according to their eye corner's positions to 300×300 pixels. The parameters of example faces for structure curve and each facial component need to be normalized at the same time. Note that, we use neutral face's normalizing functions, including the scale, rotation and translation transformation, to align corresponding expression faces. In this way, the NET model would not lose the rigid transformation on the local facial components during the expression function transfer.

Our experiments are designed to contain two parts: model evaluation and animation method evaluation. In practice, the image noise and the transmission noise will influence the accuracy of the animation signals in any type. Without loss of generality, to test the robustness of the NET model, in the following experiments, we add the Gaussian noise and manually click-drag feature points to perturb ground truth data. To evaluate the animation method objectively, we add the Gaussian noise to realistic animation signals for the quantitative comparison with other methods.

6.1. Model evaluation

The model evaluation is designed both objectively and subjectively under different experimental conditions. Our first set of experiments is conducted to test the model's effectiveness and robustness under noisy conditions. In the second set of experiments, we compare the NET model with the MM and the ASM using a small amount of examples. The third set of experiments present the expression synthesis under different parameters using a special case of the NET model.

6.1.1. Model's effectiveness and robustness test under noisy condition

We use the mean square error (MSE) to quantitatively evaluate the reconstruction accuracy. A set of feature points are manually labeled on real expression faces as the ground truth data. The MSE is calculated in Eq. (13) according to the coordinate vectors between the reconstructed feature points f_i and the ground truth feature points r_i, where N is the number of feature points:

$$MSE = \frac{\sum_{i=1}^{N} \|f_i - r_i\|^2}{N}. \quad (13)$$

We randomly choose 60 example pairs in the database. A group of zero means Gaussian noises, whose variance ranged from 0 to 16, are added to their feature points on expression faces. Then these noise data are reconstructed by the MM and the NET models. Fig. 5 exhibits their MSE curves. From the figure, it can be observed that the NET model's MSE curve is still stable when the noise variance increases, and performs more robust than the MM model.

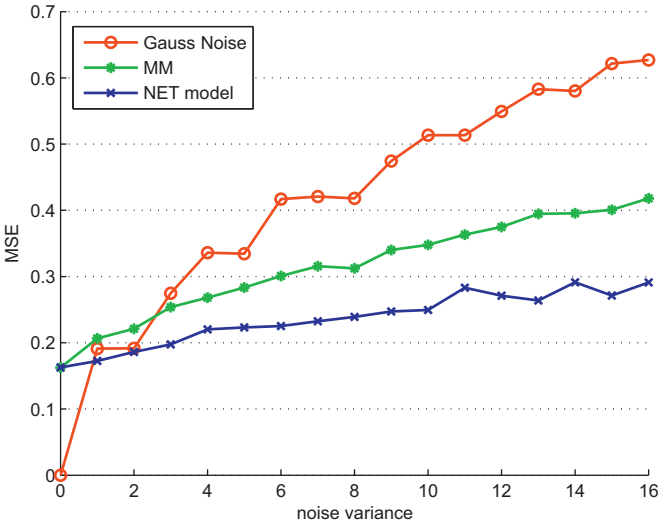

Fig. 5. MSE curves with different noise variances.

In the subjective way, we display the visual effect for reconstructing noisy faces and compare the results with other models. The ASM represents the face shape by applying the PCA. We remain 95% prime components in the ASM. In Fig. 6, we use the real face images as reference. Taking the mouth as an example, we manually move the ground truth feature points, as shown in the first row in Fig. 6. And two models take these noisy faces as the objects to reconstruct. From the second and third rows in Fig. 6, we can see that the ASM and the NET models can both acquire a natural mouth shape, but compared with the original images, the NET model outputs the most matching results. Although only with a little change on the reconstruction effect, it is crucial to improve the facial personalized attribute subjectively.

In Fig. 7, we directly add the noisy animation signals to the target neutral faces to acquire the noisy expression faces. All of the models reconstruct the animated results to some expression faces. But compared with their neutral faces, the NET model produces the most satisfactory results, which still look like the original faces. We notice that to use other two models on the first and fourth target

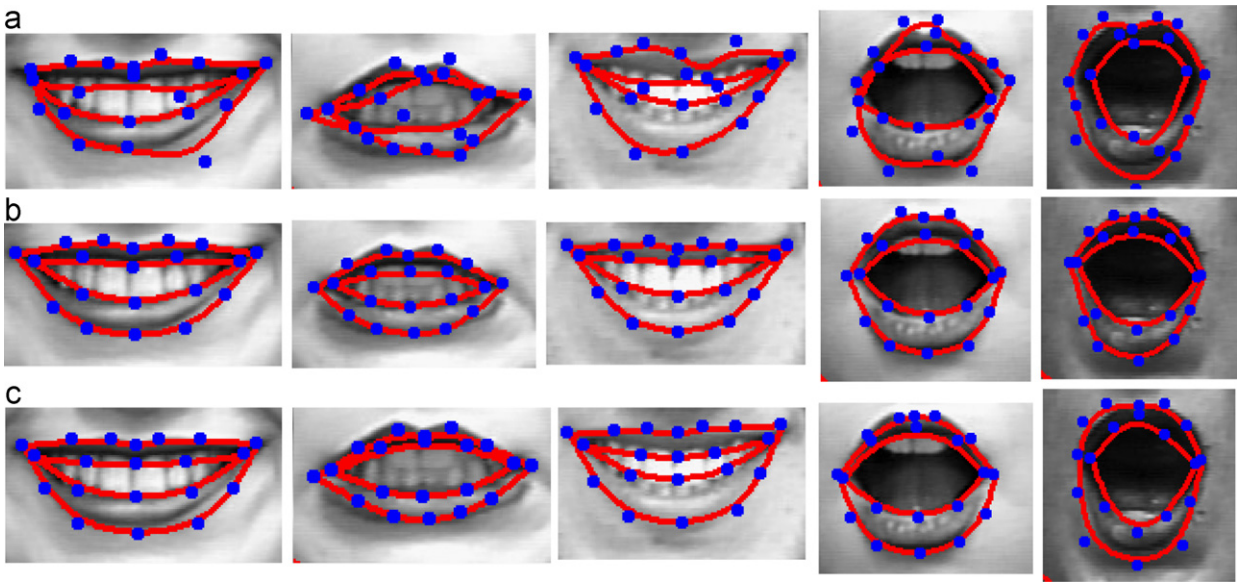

Fig. 6. Noisy mouth reconstruction results by the NET model and the ASM. (a) Input noisy data. (b) Reconstructed results by the NET model. (c) Reconstructed results by the ASM. Here mouth images are used as the reference for reader.

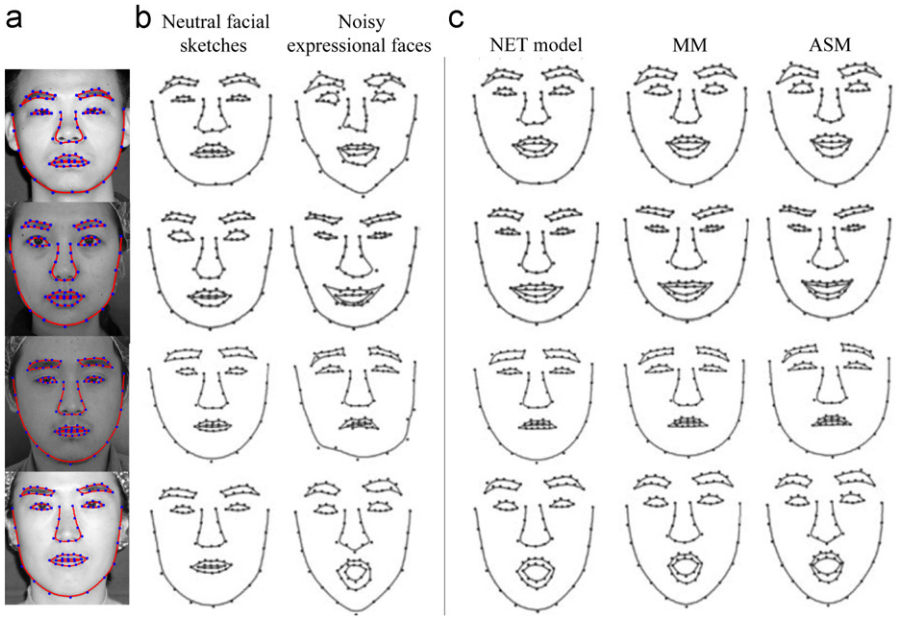

Fig. 7. Comparison reconstruction results by the NET model with the MM and the ASM. (a) Neutral facial images. (b) Input faces. (c) Reconstructed results.

faces, their shapes of face contour have been changed; and on the second and third target faces, the eyebrows have been distorted. So, we can say our NET model is superior to the MM and ASM in the model effectiveness and robustness under noisy conditions.

6.1.2. Model representation ability test under a small amount of examples

We construct an eye and eyebrow database with a small amount of examples. A target object's neutral face is shown in Fig. 8(a). There is an obvious brow ridge on the eyebrow. We reconstruct his anger face by the NET model, MM and ASM, respectively. Both the results by the MM and ASM miss the brow ridge, since all the eyebrows in the database are only with smooth shapes. But the NET model can recover it very well using the same database. The reason is that the NET model utilizes the expression function transfer to synthesize examples.

6.1.3. Expression synthesis test under different parameters

Some expression synthesis methods [26,27,29] need a large amount of expression training examples to learn a general expression transform function. But the NET model is more flexible. In Fig. 9, we synthesize a smile expression for different faces by different numbers of neighbor examples. Under the NET model's special case 2, both the parameters $K=1$ and 10 can synthesize satisfactory expression faces. It seems that they produce a more general smile for every person with $K=10$, and generate more personalized smiles with $K=1$.

6.2. Animation method evaluation

In this part, we compare the animation results by different expression transfer methods. We also provide two kinds of evaluation indexes. One is the recognition rate, which is based on the expression-invariant face recognition method to recognize the generation faces. The other is the square difference rate (SDR) to quantitatively compare the local animation details.

6.2.1. Intuitive animation results and their comparisons by different methods

In Fig. 10, the first row exhibits three kinds of source expression faces (smile, anger and surprise). The target face is shown on the left column together with her face image. By transferring expressions from the source face to

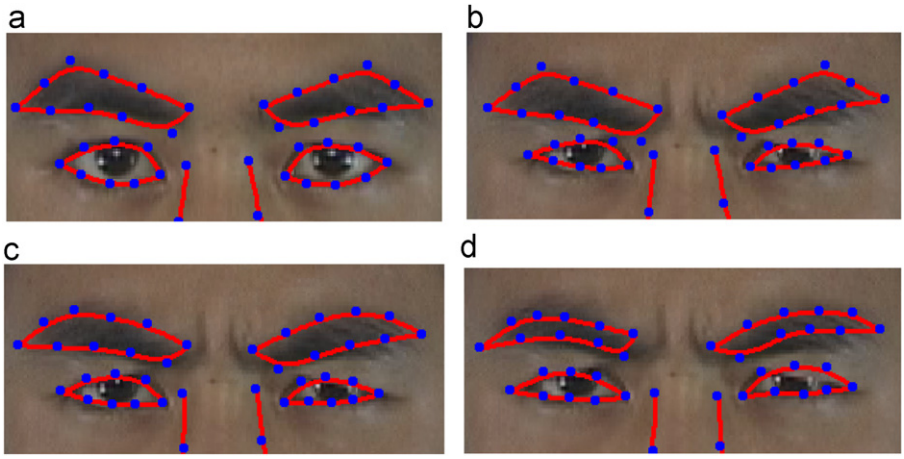

Fig. 8. Comparison of the reconstruction results by the NET model with the MM and the ASM. Here, real facial images are used as reference. (a) Original neutral face. (b) Reconstructed expression face by the NET model. (c) Reconstructed expression face by the MM. (d) Reconstructed expression face by the ASM.

Fig. 9. Smile expression synthesis by the NET model with different numbers of neighbor examples.

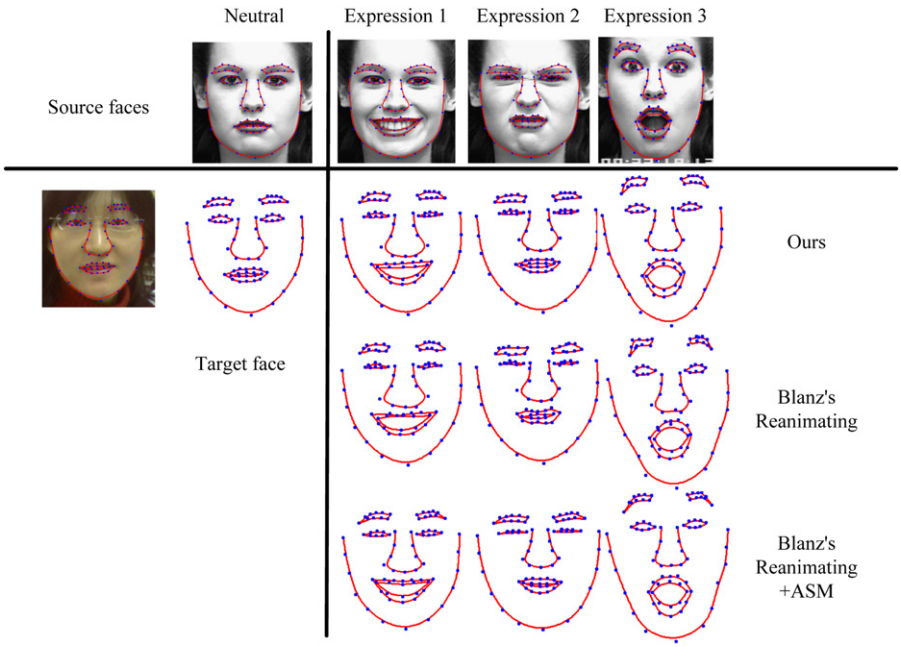

Fig. 10. Animation results by different methods.

the target face, the second row presents the animation results by our proposed method. The third row shows the animation results by the Blanz's reanimating method [22]. The Blanz's method assumes that the displacements of surface points are the same for all individuals, so that the expressions can be transferred to another person by the simple operations in the vector space. This assumption is only good for the similar source and target faces, and will cause unnatural appearance results otherwise. Although be able to refine the above results to natural faces, the ASM can not distinguish personalized facial features from noise, making the animation results different from target faces. The ASM refined results are shown in the bottom row, in which the jaw part in the smile expression becomes fat, the width of the eye changes in the anger expression, and the eyebrow's shape deforms rigidly in the surprise expression. By the intuitive comparison, our proposed method generates more realistic results.

In Fig. 11, we display more results by our expression transfer method. The source face images from the performance video are exhibited in the top row. Here, target facial sketches are rendered by the AIAR cartoon face generation system [41]. Their corresponding face images are given on the left column. We synthesize five basic expressions for structure curves at level 1. We can see that the animation results exhibit natural expressions and replicate the actions on the source face.

6.2.2. Animation result evaluation by expression-invariant face recognition

Motivated by the research on the expression-invariant face recognition technique, we inversely evaluate the animation method by recognizing the generation results. The higher recognition rate by our method than the others means that our results can keep the target face's personalized attribute better. Xu et al. [40] proposed a face recognition method by expression-driven sketch matching. They selected several local parts of the sketch graphs which are relatively invariant to the expression changes for matching two sketches. We use Xu et al.'s methods to recognize the animation results.

Four kinds of expression faces (smile, surprise, sadness, and anger) are considered in these recognition experiments. We prepare each kind of expression database by the corresponding example pairs. The test faces in this section come from the Cohn–Kanade database including 87 individuals. We randomly choose one example pair from one kind of expression database to transfer this expression on the test faces. Table 1 displays the face recognition results generated by our proposed method, the Blanz's reanimating method, and the Blanz's+ASM method, respectively. To test the robustness of expression transfer, we add Gaussian noise $G(0,4)$ to the source expressions. We can see that the expression transfer results by our method acquire the highest recognition rates.

6.2.3. Animation result evaluation by local details

The eyebrow is an important part for recognizing faces from the facial sketches. We notice that the eyebrows often move rigidly during making expressions. Eyebrow's shape only has a little change. So another index to evaluate the expression transfer performance is to measure the variations on the eyebrow. We define a square difference rate (SDR) by calculating of eyebrow's squares before and after being animated:

$$SDR = (Square(X_{exp}) - Square(X_{neu}))/Square(X_{neu}), \quad (14)$$

where $Square(X)$ calculates the square of shape X.

Fig. 11. Expression transfer results for facial sketches. The original facial images are given on the left column, whose corresponding facial sketches are generated as the target face by AIAR cartoon generation systems.

Table 1
Compared results by recognition rate (%) on different expressions with noise.

	G (0,0)				G (0,4)			
	Smile	Surprise	Sadness	Anger	Smile	Surprise	Sadness	Anger
Proposed	93.10	87.36	93.1	90.80	97.70	80.46	97.7	94.25
Blanz's	77.01	66.66	86.21	77.01	59.77	41.38	68.97	59.77
ASM	54.02	34.48	42.53	34.48	54.02	28.74	49.43	35.63

The example pairs are classified into six basic expression databases. In each expression database, we use leave-one-out method to animate each neutral face by transferring expression from one randomly chosen example pair. Table 2 shows the SDR on the animation results by our proposed methods, the Blanz's methods, and the Blanz's+ASM methods, respectively. We can see that our method has the minimum SDR. With the Gaussian noise $G(0,0-16)$ added to the source expression signals, the average SDR curves for 60 target faces by different methods are illustrated in Fig. 12. And our proposed method also performs best under the noise. From the observation of statistical results on the manually labeled ground truth data, about 10% SDR is an acceptable result which is close to the reality. However, the average result achieved by the ASM is close to 20% SDR, which influence the visual effect subjectively.

Table 2
Compared results by SDR (%) on six basic expressions.

	Smile	Angry	Surprise	Sadness	Fear	Disgust	Avg.
Proposed	10.72	11.36	13.73	10.99	13.93	10.04	11.80
Blanz's	11.74	14.48	16.67	10.53	17.17	11.98	13.76
ASM	17.08	21.98	18.59	16.58	23.28	20.85	19.73

7. Conclusions

In this paper, we proposed an expression transfer method for facial sketch animation. Given a target facial sketch with the neutral expression, our method can generate vivid expressions by retargetting motion vectors from the source face. And our method is useful in the practical application. Facial expressions, captured by the

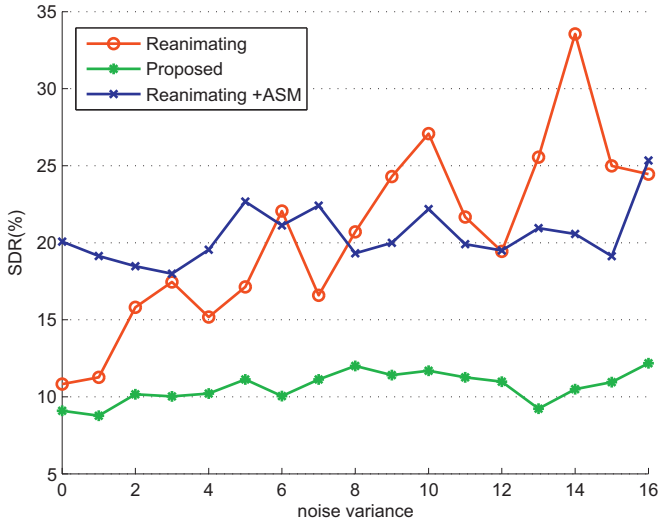

Fig. 12. Eyebrow reconstruction SDR curves with different noise variances.

common vision device, are the inputs for animating very different facial sketches. A robust face model is critical in our expression transfer task. By learning a class of expression behaviors from the neighbor examples, the proposed NET model can reconstruct the target face from noisy signals. From the analysis of the model special cases, we can see that it is also a generic model for facial animation. Moreover, in this paper, a hierarchical expression transfer method is presented based on the NET model. It contains two main steps: facial structure animation and component shape animation. The former removes the global structural differences for the expression transfer between very different faces; and the latter adjusts the motion vectors for the local face shapes. Our face model and the expression transfer method guarantee the animation results while maintaining the target face's attributes to replicate the source facial expressions.

The NET model pays more attention to the face shape rather than the texture, to meet the requirement for controlling facial sketches. It deals with the critical problem for the facial animation. Currently, the expression transfer method is tested on the frontal faces, since the face model is built on the frontal face database. One way to solve this limitation is to diversify the current approach with 3D face model. In that case, non-frontal facial expressions can be generated by the project transformation. With the help of 3D face model, we believe that the proposed model is also effective for the 3D facial sketch animation. In future work, we plan to transfer expressions for 3D facial sketches.

Acknowledgments

The authors are grateful to the helpful comments and suggestions from the anonymous reviewers. This work was supported by the National Natural Science Foundation of China (60775017, 90920008, 61005014).

References

[1] H. Chen, Z. Liu, C. Rose, et al., Example-based composite sketching of human portraits, in: Proceedings of International Symposium on Non-Photorealistic Animation and Rendering, Annecy, France, 7–9 June 2004, pp. 95–153.
[2] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, California, USA, 20–26 June 2005, pp. 1005–1010.
[3] Z. Xu, H. Chen, S. Zhu, J. Luo, A hierarchical compositional model for face representation and sketching, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (6) (2008) 955–969.
[4] B. Xiao, X. Gao, D. Tao, X. Li, A new approach for face recognition by sketches in photos, Signal Processing 89 (8) (2009) 1576–1588.
[5] I. Buck, F. Adam, Performance-driven hand-drawn animation, in: Proceedings of International Symposium on Non-Photorealistic Animation and Rendering, Annecy, France, 5–7 June 2000, pp. 101–108.
[6] T. Sucontphunt, Z. Mo, U. Neumann, Z. Deng, Interactive 3d facial expression posing through 2d portrait manipulation, in: Proceedings of Graphics Interface, Windsor, Ontario, Canada, 28–30 May 2008, pp. 177–184.
[7] Q. Zhang, Z. Liu, et al., Geometry-driven photorealistic facial expression synthesis, IEEE Transactions on Visualization and Computer Graphics 12 (1) (2006) 48–60.
[8] J.Y. Noh, U. Neumann, Expression cloning, in: Proceedings of ACM SIGGRAPH, Los Angeles, California, USA, 12–17 August 2001, pp. 277–288.
[9] R.W. Sumner, J. Popovic, Deformation transfer for triangle meshes, in: Proceedings of ACM SIGGRAPH, Los Angeles, California, USA, 8–12 August 2004, pp. 399–405.
[10] M. Song, Z. Dong, et al., A generic framework for efficient 2D and 3D facial expression analogy, IEEE Transactions on Multimedia 9 (7) (2007) 1384–1395.
[11] T.F. Cootes, C.J. Taylor, Active shape models—their training and application, Computer Vision and Image Understanding 61 (1) (1995) 38–59.
[12] M.J. Jones, T. Poggio, Multidimensional morphable models, in: Proceedings of International Conference on Computer Vision, Bombay, India, 4–7 January 1998, pp. 683–688.
[13] J. Lewis, J. Mooser, Z. Deng, U. Neumann, Reducing blendshape interference by selected motion attenuation, in: Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, Los Angeles, California, USA, 31 July–4 August 2005, pp. 25–29.
[14] Z. Deng, P. Chiang, P. Fox, U. Neumann, Animating blendshape faces by cross mapping motion capture data, in: Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, Redwood City, California, USA, 14–17 March 2006, pp. 43–48.

[15] P. Joshi, W. Tien, M. Desbrun, F. Pighin, Learning controls for blend shape based realistic facial animation, in: Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation, San Diego, California, USA, 26–27 July 2003, pp. 35–42.
[16] E. Chuang, C. Bregler, Performance driven facial animation using blendshape interpolation, Standford University Computer Science Technical Report, CSTR-2002-02, Standford University, April 2002.
[17] X. Liu, T. Mao, S. Xia, Y. Yu, Z. Wang, Facial animation by optimized blendshapes from motion capture data, Computer Animation and Virtual Worlds 19 (3–4) (2008) 235–245.
[18] A. Khanam, M. Mufti, Intelligent expression blending for performance driven facial animation, IEEE Transactions on Consumer Electronics 53 (2) (2007) 578–584.
[19] X. Ma, B. Le, Z, Deng, Style learning and transferring for facial animation editing, in: Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation, New Orleans, USA, 1–2 August 2009, pp. 114–123.
[20] H. Pyun, Y. Kim, W. Chae, H. Kang, S. Shin, An example-based approach for facial expression cloning, in: Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation, San Diego, California, USA, 26–27 July 2003, pp. 167–176.
[21] K. Na, M.R. Jung, Hierarchical retargetting of fine facial motions, Computer Graphics Forum 23 (3) (2004) 687–695.
[22] V. Blanz, C. Basso, et al., Reanimating faces in images and video, Computer Graphics Forum 22 (3) (2003) 641–650.
[23] L. Williams, Performace-driven facial animation, in: Proceedings of the ACM SIGGRAPH, Dallas, Texas, USA, 6–10 August 1990, pp. 235–242.
[24] J. Ostermann, Animation of synthetic faces in MPEG-4, in: Proceedings of IEEE Computer Society Conference on Computer Animation, Philadelphia, Pennsylvania, USA, 8–10 June 1998, pp. 49–55.
[25] A. Savran, L. Arslan, L. Akarun, Speaker-independent 3D face synthesis driven by speech and text, Signal Processing 86 (10) (2006) 2932–2951.
[26] B. Abboud, F. Davoine, M. Dang, Facial expression recognition and synthesis based on an appearance model, Signal Processing: Image Communication 19 (8) (2004) 723–740.
[27] B. Abboud, F. Davoine, Bilinear factorization for facial expression analysis and synthesis, Vision, Image and Signal Processing 152 (3) (2005) 327–333.
[28] Y. Du, X. Lin, Facial expressional image synthesis controlled by emotional parameters, Pattern Recognition Letters 26 (6) (2005) 2611–2627.
[29] J. Ghent, J. McDonald, Photo-realistic facial expression synthesis, Image and Vision Computing 23 (12) (2005) 1041–1050.
[30] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.
[31] D. Tao, M. Song, et al., Bayesian tensor approach for 3D face modelling, IEEE Transactions on Circuits and Systems for Video Technology 18 (10) (2008) 1397–1410.
[32] Z. Liu, Y. Shan, Z. Zhang, Expressive expression mapping with ratio images, in: Proceedings of ACM SIGGRAPH, Los Angeles, California, USA, 12–17 August 2001, pp. 271–276.
[33] D.G. Luenberger, Linear and nonlinear programming, Addison-Wesley, 1984.
[34] Y. Ye, Interior point algorithms: theory and analysis, John Wiley, 1997.
[35] L. Xiong, N. Zheng, S. Du, et al., Facial expression synthesis based on facial component model, International Journal of Pattern Recognition and Artificial Intelligence 23 (3) (2009) 637–657.
[36] F.L. Bookstein, Principal warps: thin-plate splines and the decomposition of deformations, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (6) (1989) 567–585.
[37] H. Cui, A new point matching algorithm for non-rigid registration, Computer Vision and Image Understanding 89 (2–3) (2002) 114–141.
[38] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Processing of Automatic Face and Gesture Recognition, Grenoble, France, 26–30 March 2000, pp. 46–53.
[39] N. Zheng, Y. Fu, T. Zhang, F. Zhuo, Facial expression transformation, aging and invisible view reconstruction (1), Chinese Journal of Electronics 31 (12) (2003) 1955–1962.
[40] Z. Xu, J. Luo, Face recognition by expression-driven sketch graph matching, in: Proceedings of International Conference on Pattern Recognition, Hong Kong, 20–24 August 2006, pp. 1119–1122.
[41] Y. Liu, Z. Wu, Y. Shao, D. Jia, Face cartoon rendering: a unified cartoon rendering approach based on sample learning, in: Processing of AsiaGraph, Tokoy, Japan, 23–26 October 2008, pp. 22–25.