Facial Expression Transfer Method Based on Frequency Analysis

Wei Wei[a,*], Chunna Tian[b], Stephen John Maybank[c], Yanning Zhang[a]

a. School of Computer Science, Northwestern Polytechnical University, Xi'an, China

b. VIPS Lab, School of Electronic Engineering, Xidian University, Xi'an, China

c. Department of Computer Science and Information Systems, Birkbeck College, University of London, UK

Abstract. We propose a novel expression transfer method based on an analysis of the frequency of multi-expression facial images. We locate the facial features automatically and describe the shape deformations between a neutral expression and non-neutral expressions. The subtle expression changes are important visual clues to distinguish different expressions. These changes are more salient in the frequency domain than in the image domain. We extract the subtle local expression deformations for the source subject, coded in the wavelet decomposition. This information about expressions is transferred to a target subject. The resulting synthesized image preserves both the facial appearance of the target subject and the expression details of the source subject. This method is extended to dynamic expression transfer to allow a more precise interpretation of facial expressions. Experiments on Japanese Female Facial Expression (JAFFE), the extended Cohn-Kanade (CK+) and PIE facial expression databases show the superiority of our method over the state-of-the-art method.

Key words: Expression transfer; warping technique; facial feature location; frequency domain analysis

1. Introduction

Emotions are often conveyed through body gestures or facial expressions rather than verbal communication [1][2]. Thus, automatic facial expression analysis is an interesting research topic. There have been many recent achievements in related research sub-areas such as facial landmark localization [3][4][5][6][7][8], tracking and recognition [9][10]. Realistic facial expression synthesis is useful for affective computing, human computer interaction [11][12], realistic computer animation [13] and facial surgery planning [14][15], etc. There are more than 20 groups of facial muscles innervated by facial nerves [16], which control actions of the face (eg. opening or closing of eyes or mouth) and variations in local appearance (eg. facial wrinkles and furrows).

Psychologists Ekman and Friesen developed a facial action coding system (FACS) based on the movements of facial muscles and their effects on the expression [17]. They divided the face into 44 action units (AU) and analyzed the motion characteristics and their effects on associated expressions. Many expression synthesis approaches had concentrated on capturing expressions through AUs. Platt and Badler proposed a model of the muscles to simulate FACS to synthesize facial expression [18]. Waters extended this model to a hierarchical one [19], in which facial muscles were divided into linear muscles and sphincter muscles to control the skin stretch and

shrinkage, respectively. Based on Water's method, Koch et al. used a finite element method to simulate the physical structure of the human face [20]. These models emphasized the simulation of muscle movements. They are relatively simple compared with the one developed by Lee et al. [21]. Lee's model has a three-level structure of skin, bone and muscle, based on physiology. Thus, it can synthesize much more realistic expressions. But the complicated structure and heavy calculation load prevented its application in practice. FACS was used to define a non-continuous and non-uniform scale for scoring the strength of facial activities. It was hard to identify subtle facial activities. Thus, the valuable information contained in subtle expression changes can be lost [22]. The subtle local appearance variations are usually the main components of micro-expressions, which are important clues to distinguish different expressions. For example, without the micro-expressions, 'fear' and 'surprise' are hard to distinguish. In addition, micro-expressions reflect the emotion and inner working of people. As a result, photo-realistic facial expression synthesis is an active research area.

Darwin revealed the consistency of expressions among different races and genders [23]. This is because the shape deformations are similar for the same expression. However, the local appearance deformations, viz., the expression details, are quite person-specific. Therefore, a person-specific expression transfer, which clones the expression of a source subject to a target subject, has a wide range of applications. However, the subtle appearance deformations are difficult to synthesize. In this study, we locate the facial landmarks automatically in order to describe the shape deformation, and extract and transfer the subtle local expressions using frequency analysis. Since dynamic variations are important in interpreting facial expression [24][25][26], we extend our work to dynamic expression transfer. The effectiveness of the proposed algorithm is verified on Japanese Female Facial Expression (JAFFE), the extended Cohn-Kanade (CK+) [27] and PIE databases.

The remainder of this paper is structured as follows. Section 2 provides an overview of the related work. Section 3 presents the facial landmark localization and face alignment methods. Section 4 proposes the static expression transfer method based on frequency analysis. We extend the expression transfer method to dynamic expression synthesis in Section 5. Section 6 details the evaluation of the proposed method as well as the experimental setup. Finally, Section 7 concludes the paper.

2. Related Work

Our synthesis method involves local facial landmark detection and expression transfer. Below, we give a concise overview of relevant prior work on these two topics.

The Active Shape Model (ASM) [28] used a parametric deformable model to fit the shape of

human face. It learned the variation modes of a shape from a set of training examples. Transformation and shape parameters were estimated iteratively to fit the mean shape of the observed object. The Active Appearance Model (AAM) [29][30] integrated facial texture with the ASM to fit the facial shape better. Matthews and Baker [31] proposed a computationally efficient AAM algorithm with rapid convergence to improve the fitting. The shapes of multi-view faces can be fitted through a gradient-descent search [32]. A vectorial regression function was learned from the training image with an Explicit Shape Regression (ESR) model to locate the facial landmarks. A two-level boosted regression, shape-indexed features and a correlation-based feature selection method were combined with an ESR model to locate the facial landmarks more accurately [5]. In [6], local binary features were learned using a regression tree to preserve the most discriminative information contained in the local texture around each facial landmark. The local binary features improved the efficiency of facial landmark detection.

Zhu and Ramanan [7] used a multi-tree model to handle different expressions. Each expression was modeled by a deformable model of the joint distribution of parts, consisting of local patches around facial landmarks. The Histogram of Oriented Gradient (HoG) features were used to describe the local patches. Each branch of the multi-tree model corresponds to one expression, but different trees share a pool of parts. The deformable model was trained by a latent support vector machine (LSVM). Since the human face is non-rigid and large nonlinear deformations occur in extreme expressions, [4] used a Haar-like feature based Adaboost face detector [33] to initialize the face location, then adopted a Supervised Descent Method (SDM) to refine the locations of facial landmarks. [8] presented a state-of-the-art method for facial landmark detection with super-real time performance, which used an ensemble of regression trees to estimate the positions of facial landmark accurately from a sparse subset of pixel intensities. The ensemble of regression trees was learned based on gradient boosting. The appropriate priors exploiting the structure of image data is helpful to efficient feature selection.

To transfer expression, the facial texture of the target subject was warped to the shape of the source subject, given the correspondence of the key facial landmarks [34]. The warped expression captures the shape motions of different expressions but ignores the micro-expressions. Liu et al. [35] represented the grey level of each point on the face image with the Lambertian model. Then, they calculated the ratio between the neutral and the non-neutral expression at each point to obtain an expression ratio image. They used this expression ratio image to transfer the non-neutral expression to other neutral faces. In [36], [37] and [38] , multi-expression faces were arranged as a tensor data. In [36], a High Order Singular Value Decomposition (HOSVD) [39] was applied to the AAM [40] coefficients of the training faces to obtain a generative model. In this model,

AAM coefficients of training images were factorized into identity and expression subspaces. The test face image was represented by the AAM coefficients, which were reconstructed by estimating its identity and expression coefficients in the generative model. The solved identity coefficients were combined with the remaining expression coefficients in the expression subspace of the generative model to synthesize multi-expression face images. In [37], a generative model of shape and texture is built in order to obtain the identity and expression coefficients, separately. The expression transfer was realized by swapping the expression coefficients of the source and target subjects. [38] proposed a tensor-based AAM, in which texture is aligned with the normalized shape of the AAM. The expression coefficients of the test face were synthesized by linearly combining the expression coefficients of training faces in the latent expression subspace. A texture variation ratio between the neutral and non-neutral expressions was used to transform the expression of the test face. However, the expression variations are often strongly nonlinear, thus the expression coefficient estimation for the test image may not be accurate. The expression variation ratio is not adaptive to the nonlinear variation of extreme expressions. [41] incorporated the expression manifold with the Tensor-AAM model to synthesize dynamic expressions of the training face. The Bilinear Kernel Reduced Rank Regression (BKRRR) method for static general expression synthesis was proposed in [15]. It synthesizes general expressions on the face of a target subject. Zhang and Wei [2] used TensorFace combined with an expression manifold to synthesize the dynamic expressions of a training face, then extracted and transferred the dynamic expression details of the training face to the target face. For extensive reviews on facial expression synthesis, we refer the reader to [42] and [43].

Since the facial landmarks around eyes, nose, mouth and eyebrows *etc.* are required to describe the shape correspondence between different expressions, we use state-of-the-art method proposed in [8] to detect the facial landmarks (See Fig.1) in this study. We use AAM to separate and align the shapes and texture of multi-expression faces. Since the expression details are more easily observed using frequencies, we adopt the wavelet transform to divide images into four frequency bands and then transfer the details of expressions. We extend this work to dynamic expression transfer, to obtain highly realistic and natural looking facial animations. In summary, our main contributions include: 1) a proposal (See Fig.1) for a static expression transfer method based on frequency analysis; 2) the extension of the static expression transfer to dynamic expression transfer; 3) a unified automatic dynamic expression transfer framework including facial landmark localization, expression alignment, dynamic shape synthesis, expression warping, expression detail extraction and transferring.

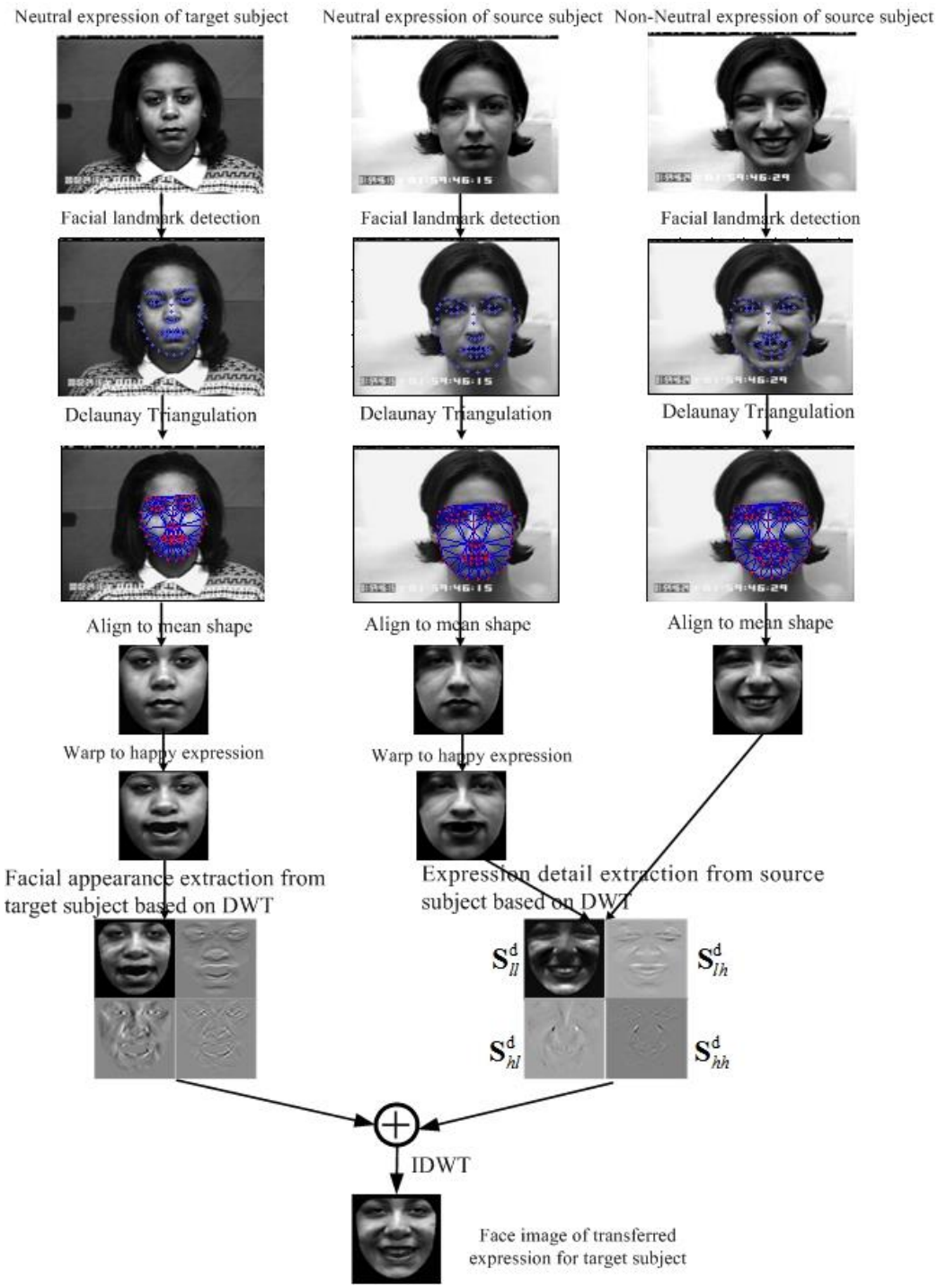


Fig.1 The whole framework for static expression transfer method based on frequency analysis.

3. Automatic Facial landmark Extraction and Face Alignment

In this study, we use the method proposed in [8] to detect facial landmarks (See Fig.1). Given the facial landmarks, we use AAM to separate and align the shapes and texture of multi-expression faces. Since human face is not homogeneous, we cannot use a global uniform transformation to warp the whole facial texture to the aligned shape. Thus, we divide the facial texture into small

homogenous patches by Delaunay Triangulation strategy (See Fig.2 (a), (b)) according to the facial landmarks. The landmarks build the correspondence of triangular patches under different expressions. We warp the texture from triangle T_i to its corresponding triangle $T_{j.}$ (See Fig.2 (c), (d)) through affine transformation. Thus, the robustness of our method to landmark localization errors is not good. It is especially sensitive to localization errors around eyes and mouth. Fortunately, the method in [8] can provide accurate facial landmarks in most cases.

Fig.2 (c) and (d) demonstrate the affine transformation [28], which maps the point (x_i, y_i) from the triangle T_i to (x_j, y_j) in the corresponding triangle T_j. The affine transformation is defined by equation (1).

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_j \\ y_j \\ 1 \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{1}$$

The transformation parameters $a_1 \sim a_6$ is determined by the mapping between the corresponding vertex coordinates (V_x, V_y) of T_i and (v_x, v_y) of T_j. The texture value at $\boldsymbol{t}_j(x_j, y_j)$ is decided by that of $\boldsymbol{t}_i(x_i, y_i)$, which means the texture mapping is determined by the warping between the corresponding triangles. The points in the triangles of Fig.2 (b) which do not have corresponding points in the triangles of Fig.2 (a) are interpolated through bicubic interpolation.

$$\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} v_{x_1} & v_{x_2} & v_{x_3} \\ v_{y_1} & v_{y_2} & v_{y_3} \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} V_{x_1} & V_{x_2} & V_{x_3} \\ V_{y_1} & V_{y_2} & V_{y_3} \\ 1 & 1 & 1 \end{bmatrix} \tag{2}$$

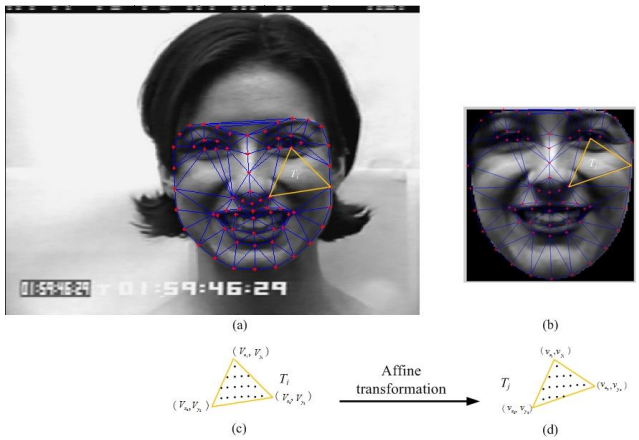


Fig.2 Demonstration of the triangular mapping to align faces to the normalized shape.

4. Static Expression Transfer Based on Frequency Analysis

Given the aligned multi-expression face images, we transfer the static expression based on frequency analysis. The expression variations include the deformations of global shape and local texture. For example, when people are happy facial-feature movements include: lifting up the corners of mouth and eyelid contraction. Subtle local appearance variations include: cheek

wrinkles and "crow's feet" at corners of the eyes. To simulate the global shape deformation, we warp the neutral expressions of the source and target subjects to non-neutral expressions. The synthesized expression of the target subject is determined by the given non-neutral expression of the source subject. The automatic expression classification of the source subject can be realized by available methods, e.g. the Computer Expression Recognition Toolbox (CERT) and methods reviewed in [48]. The subtle wrinkles of local texture are transferred using frequency analysis. Since wavelets have good spatial-frequency localization characteristics and can be used to detect multi-scale, multi-directional texture changes, 2D Discrete Wavelet Transformation (DWT) is applied to the non-neutral expression and the warped expression of the source subject. Then we calculate the differences of them in each frequency band, which are used to obtain the synthesis coefficients for transferring the expression details to the target subject. With the synthesis coefficients, we use a simple linear combination at each channel of the frequency images. Finally, a 2D inverse DWT (IDWT) is applied on those frequency bands to synthesize the transferred expression in the image domain. To determine how many bands of wavelet transform should be used, we analyze the frequency of non-neutral expressions through Fast Fourier Transform (FFT) and we find most expression details lie in the high frequency components while the middle frequency components contain less expression details. More decomposition layers of wavelet transform give more middle and low frequency information. Since the high frequency bands of one layer wavelet transform contain most expression details, we use only one layer wavelet transform for expression transfer in this study, which is also effective in computation compared with more bands. The whole procedure is shown in Fig.3.

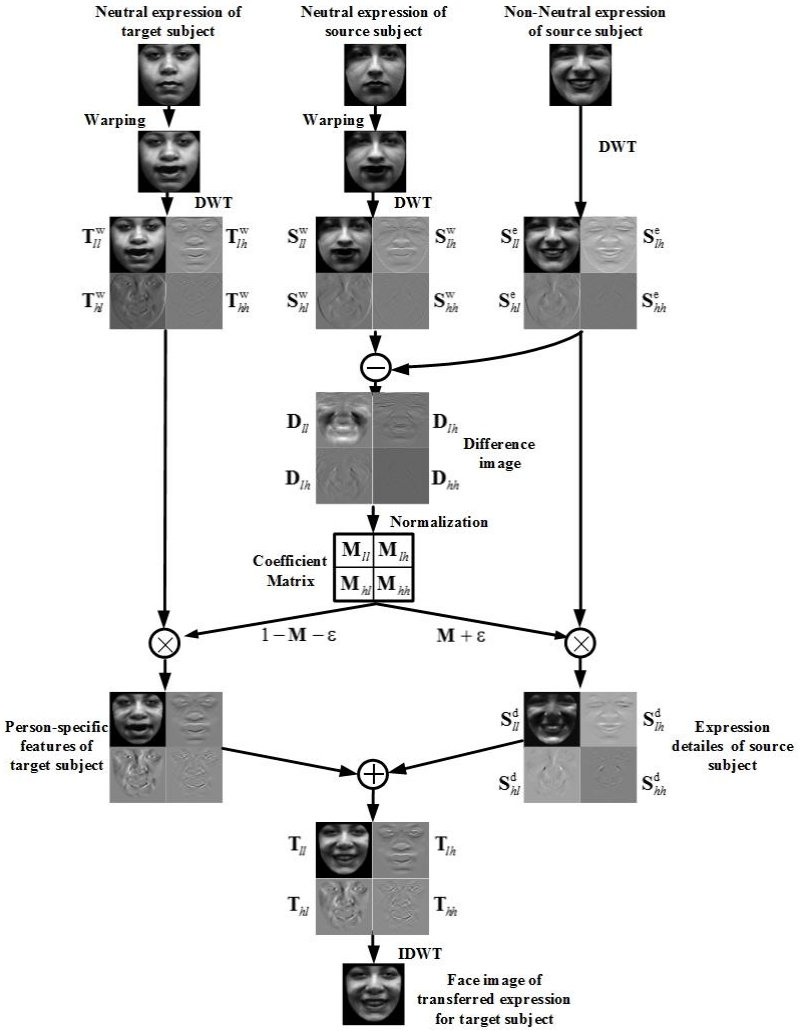


Fig.3 The flowchart of the proposed static facial expression transfer method

The neutral facial expressions of source and target subjects are $\mathbf{S}^n$ and $\mathbf{T}^n$, respectively. The non-neutral expression of the source subject is $\mathbf{S}^e$. $\mathbf{S}^n$ and $\mathbf{T}^n$ are warped to yield $\mathbf{S}^w$ and $\mathbf{T}^w$. We apply 2D DWT to $\mathbf{S}^e$ and $\mathbf{S}^w$ to obtain their frequency images $\{\mathbf{S}^e_{ll},\mathbf{S}^e_{lh},\mathbf{S}^e_{hl},\mathbf{S}^e_{hh}\}$ and $\{\mathbf{S}^w_{ll},\mathbf{S}^w_{lh},\mathbf{S}^w_{hl},\mathbf{S}^w_{hh}\}$ in four frequency bands. In this study, the subscripts ll,lh,hl,hh denote the low-frequency, horizontal high frequency, vertical high frequency and diagonal high frequency, respectively. The subtle expression differences of $\mathbf{S}^e$ and $\mathbf{S}^w$ are more obvious in some frequency bands. We first calculate the differences $\mathbf{D}_i$ between $\mathbf{S}^e$ and $\mathbf{S}^w$ in the four frequency bands as follows.

$$\mathbf{D}_i = \mathbf{S}^e_i - \mathbf{S}^w_i, \quad i \in \{ll, lh, hl, hh\} \tag{3}$$

We normalize the difference images to obtain the weighting coefficient matrix $\mathbf{M}$ in each frequency band, as follows.

$$\mathbf{M}_i = (\mathbf{D}_i - \mathbf{D}_{i_{\min}}) / (d_{i_{\max}} - d_{i_{\min}}), i \in \{ll, lh, hl, hh\} \quad (4)$$

In equation (4), $\mathbf{D}_{i_{\min}}$ is the constant matrix with the same size as $\mathbf{D}_i$. The value in $\mathbf{D}_{i_{\min}}$ represents the minimum value of matrix $\mathbf{D}_i$, which is denoted as $d_{i_{\min}}$. The largest value in $\mathbf{D}_i$ is denoted as $d_{i_{\max}}$. The denominator in equation (4) gives the frequency range of each frequency band. The subtle expression details $\mathbf{S}_i^d, i \in \{ll, lh, hl, hh\}$ of the source subject are extracted in different frequency bands as follows

$$\mathbf{S}_i^d = \mathbf{S}_i^e \circ (\mathbf{M}_i + \boldsymbol{\varepsilon}_i), \quad i \in \{ll, lh, hl, hh\} \quad (5)$$

where $\circ$ is the Hadamard product. $\boldsymbol{\varepsilon}_i$ is a constant matrix, which has the same size with $\mathbf{M}_i$. The value of $\boldsymbol{\varepsilon}_i$ is empirically set in the range of $[0, \ 0.4]$. Since the expression details of $\mathbf{S}^e$ are more salient in some frequency bands, we emphasize this saliency with different weights $\boldsymbol{\varepsilon}$ in different frequency bands.

To transfer the extracted expression details of the source subject to the warped face of the target subject, we need to generate the person-specific features of the target subject from $\mathbf{T}^w$, in which some information should be eliminated to incorporate the extracted expression details of the source subject. Since the expression details are extracted in four frequency bands, we apply 2D DWT to the warped image $\mathbf{T}^w$ of the target subject to obtain $\{\mathbf{T}_{ll}^w, \mathbf{T}_{lh}^w, \mathbf{T}_{hl}^w, \mathbf{T}_{hh}^w\}$. Then, transfer the extracted expression details $\mathbf{S}_i^d$ in equation (5) to the warped image $\mathbf{T}_i^w$ of the target subject in each frequency band as equation (6), in which $\mathbf{T}_i^w$ is multiplied with $(\mathbf{1} - \mathbf{M}_i - \boldsymbol{\varepsilon}_i)$ to smooth the blending of $\mathbf{S}_i^d$ and $\mathbf{T}_i^w$.

$$\mathbf{T}_i = \mathbf{S}_i^e \circ (\mathbf{M}_i + \boldsymbol{\varepsilon}_i) + \mathbf{T}_i^w \circ (\mathbf{1} - \mathbf{M}_i - \boldsymbol{\varepsilon}_i), \quad i \in \{ll, lh, hl, hh\} \quad (6)$$

In equation (6), $\mathbf{1}$ is a matrix of ones with the same size as $\mathbf{M}_i$. The final transferred expression of the target subject is obtained by applying the 2D IDWT on the four frequency images $\{\mathbf{T}_{ll}, \mathbf{T}_{lh}, \mathbf{T}_{hl}, \mathbf{T}_{hh}\}$.

5. The Extension to Dynamic Expression Transfer

In the static expression transfer method (See Fig. 1), we need the warped expressions from the

neutral expression of the source and target subjects to the shape of target expression. Then, extract the expression details based on the difference of the warped expression and the peak expression of the source subject in wavelet domain. To extend this work to dynamic expression transfer, we first generate the dynamic facial shapes of each expression from the training data, then warp the neutral expression of the source and target faces to those shapes to yield their warped dynamic expressions (See Fig.4). Secondly, we have to synthesize dynamic expressions with vivid details for source subjects using a continuous expression manifold in the K-TensorFace model. Finally, we extract and transfer the synthesized dynamic expression details of the source subject to the warped dynamic expressions of the target subject by the static expression transfer method (See Fig.4). The details of the first and second steps are given in subsection 5.1 and 5.2, respectively.

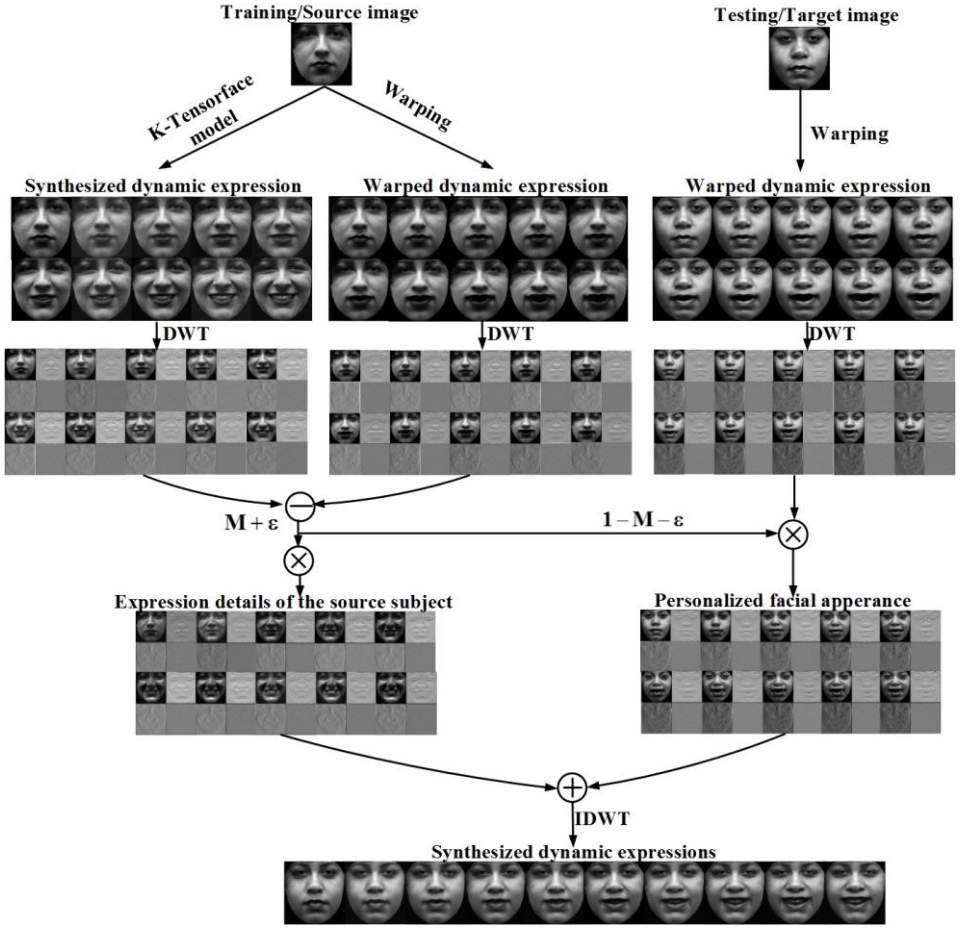


Fig.4 Illustration of the dynamic expression transfer method

5.1 Generating warped dynamic expressions for the source and target subjects

To capture the dynamic variation of certain expression, we need several training faces (Say 3 to 5 images) of each source subject which can describe the discrete variation of this expression. The training faces under the same expression are aligned to the mean shape of this expression by AAM. We sort the discrete shapes of training faces in the order of expression variation (See Fig.5), then

use spline fitting between them to generate the dynamic shapes of this expression. Finally, we warp the neutral expressions of the source and target subjects to the dynamic shapes to yield their warped dynamic expressions (See Fig.4).

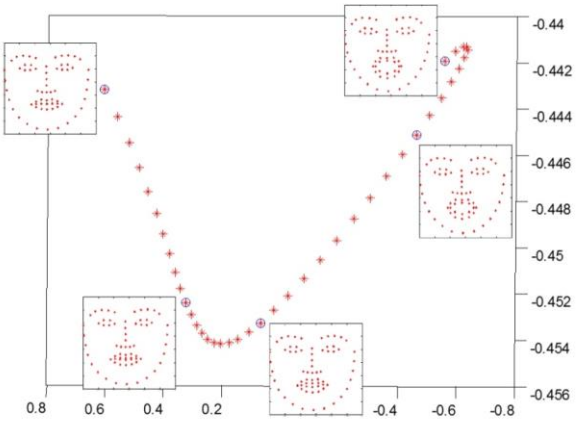


Fig.5 The illustration of the dynamic shapes of surprise expression where only the first three dimensions are shown.

5.2 Synthesizing dynamic expressions with vivid details for the source subject

In this step, we need construct the expression manifold first to capture the dynamic nature of expression variation in the low dimensional subspace obtained through HOSVD. Then, we use nonlinear mapping to map the continuous expression coefficients in the manifold to the dynamic expression images.

We use $\mathbf{y}_{1:N}^{1:K}$ to represent the training faces of N discrete expressions of K source subjects. We arrange $\mathbf{y}_{1:N}^{1:K}$ into a 3-order tensor $\boldsymbol{Y} \in R^{N\times K\times d}$. Then HOSVD [39] is applied to $\boldsymbol{Y}$ (See equation (7)) to obtain the mode matrices $\mathbf{U}_{\text{exp.}} \in R^{N\times N}$, $\mathbf{U}_{id.} \in R^{K\times K}$ and $\mathbf{U}_{pixel} \in R^{d\times d}$ to represent the variations of expression, identity and pixel intensity of the facial texture, respectively.

$$\boldsymbol{Y} = \boldsymbol{Z} \times_1 \mathbf{U}_{\text{exp.}} \times_2 \mathbf{U}_{id.} \times_3 \mathbf{U}_{pixel} \tag{7}$$

In equation (7), $\boldsymbol{Z}$ is the core tensor, which governs the interactions among different modes. $\times_i$ is the mode i product. Fig.6 illustrates the decomposition of multi-expression faces. In Fig.6, each row of $\mathbf{U}_{\text{exp.}}$ is a coefficient vector of one training expression. Each row of $\mathbf{U}_{\text{id.}}$ is a coefficient vector of one training identity. $\mathbf{U}_{pixel}$ contains the basis vector of pixels in the whole images. Since expressions in $\boldsymbol{Y}$ are discrete, we build an expression manifold as follows to

describe the continuous expression variation (See Fig.7). We sort the expression coefficient vectors in $\mathbf{U}_{\text{exp.}}$ (the blue circles in Fig.7) according to the order of expression variation. Then use spline fitting between them to generate a continuous expression manifold *E*.

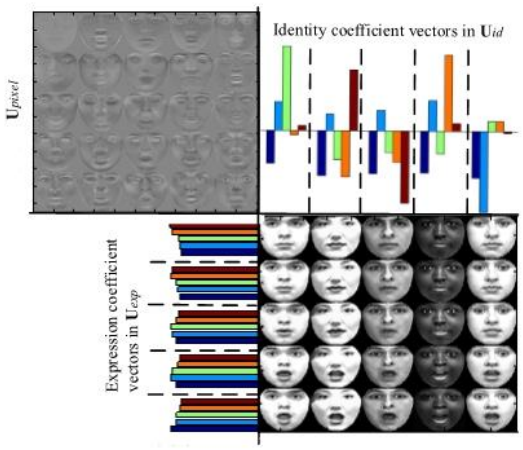


Fig.6 The illustration of the tensor decomposed multi-expression faces

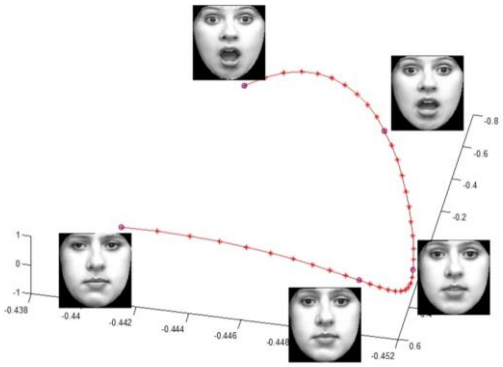


Fig.7 The illustration of the generated expression manifold for multi-expression face images where only the first three dimensions are shown.

Since most expression coefficient vectors (See the red stars in Fig.7) in manifold *E* are interpolated, we should map the new vectors to the image space to synthesize more expressions of the training identity, which is realized by the K-TensorFace model [44] as follows. For faces $\mathbf{y}_{1:N}^{1:K}$ in the training dataset, we use $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}$ to denote their expression coefficient vectors in the manifold *E*, where $\mathbf{e} \in R^N$. We use Gaussian kernels $\phi(\cdot)$ to build the nonlinear mapping between $\mathbf{e}$ and the *j*-th pixel y^j of image $\mathbf{y} \in R^d$ as

$$y^j(\mathbf{e}) = \sum_{i=1}^{n} w_i^j \phi(\| \mathbf{e} - \mathbf{z}_i \|) + [1, \mathbf{e}] g^j \tag{8}$$

where $\mathbf{z}_i$ is the *i*-th kernel center of $\phi(\cdot)$ sampled from the expression manifold and w_i^j is the weight of kernel $\phi(\| \mathbf{e} - \mathbf{z}_i \|)$. g^j is the mapping coefficient of the linear polynomial [1, **e**]. We denote $\varphi(\mathbf{e}_i) = [\phi(\| \mathbf{e}_i - \mathbf{z}_1 \|), \cdots, \phi(\| \mathbf{e}_i - \mathbf{z}_n \|), 1, \mathbf{e}_i]$ and $\mathbf{A} = [(1, \mathbf{e}_1)^{\mathrm{T}}, (1, \mathbf{e}_2)^{\mathrm{T}}, \cdots, (1, \mathbf{e}_N)^{\mathrm{T}}, \mathrm{O}_{(N+1)\times(N+1)}]$. $\varphi(\mathbf{e}_i)$ is a vector of dimension $1 \times (n + N + 1)$. O is a zero matrix. The mapping between multi-expression faces of the *k*-th person { $\mathbf{y}_1^k$, $\mathbf{y}_2^k$, … , $\mathbf{y}_N^k$ } and their low-dimensional expression coefficient vectors {$\mathbf{e}_1$, $\mathbf{e}_2$, … , $\mathbf{e}_N$} can be represented by

$$\begin{pmatrix} \varphi(\mathbf{e}_1) \\ \varphi(\mathbf{e}_2) \\ \vdots \\ \varphi(\mathbf{e}_N) \\ \mathbf{A} \end{pmatrix} \mathbf{D}^k = \begin{pmatrix} \mathbf{y}_1^k \\ \mathbf{y}_2^k \\ \vdots \\ \mathbf{y}_N^k \\ \mathbf{O}_{(N+1)\times d} \end{pmatrix}. \tag{9}$$

In equation (9), $\mathbf{D}^k$ is a $(n + N + 1) \times d$ matrix. Its *j*-th column is $[w_1^j, w_2^j, \cdots, w_N^j, g^j]$. In equation (9), only $\mathbf{D}^k$ is unknown. $\mathbf{D}^k$ can be solved from equation (9) to get the linear mapping coefficients g^j and nonlinear mapping coefficients $w_1^j, w_2^j, \cdots, w_N^j$. For the *k*-th person under expression *i*, the mapping can be represented by

$$\mathbf{y}_i^k = \varphi(\mathbf{e}_i)\mathbf{D}^k. \tag{10}$$

In equation (9) and (10), $\mathbf{y}_i^k$ is identity dependent. However, $\varphi(\mathbf{e}_i)$ is identity free. We deduce that the identity information is embedded in the linear matrix $\mathbf{D}^k$ in this mapping. To extract the identity information, we stack $\mathbf{D}^{1:K}$ to form a tensor $\boldsymbol{D}$ = [$\mathbf{D}^1$, . . . , $\mathbf{D}^K$]. Then, we apply HOSVD to $\boldsymbol{D}$ to abstract the low-dimensional identity coefficient vectors $\mathbf{p}^k \in R^K$ and the core tensor $\mathcal{C}$. This results in the K-TensorFace model for multi-expression faces. In this model, $\mathbf{y}_i^k \in R^d$ can be synthesized by identity $\mathbf{p}^k$ and expression coefficient vector $\mathbf{e}_i$ as

$$\mathbf{y}_i^k = \mathcal{C} \times_1 \mathbf{U}_{pixel} \times_2 \mathbf{p}^k \times_3 \varphi(\mathbf{e}_i) \tag{11}$$

The localized Gaussian kernels $\phi(\cdot)$ in $\varphi(\mathbf{e}) \in R^{(N+n+1)}$ guarantee the interpolated expression points are localized to their nearest $\mathbf{z}$. Thus, the mapping of equation (11) preserves the latent structure of the expression manifold in the image space of **y**. Therefore, nearby points in the expression manifold result in face images with similar expressions. If we vary the expression coefficients over the manifold, we can synthesize the dynamic expressions for the *k*-th source

subject (See Fig.4) by equation (11) with a small size training set.

6. Experimental Results and Analysis

In this study, we verify the performance of the proposed method on the JAFFE, CK+ and PIE datasets. The JAFFE database contains 213 face images of 10 persons. Each person has several expressions and each expression has 3 to 4 images. We selected several extreme expressions from the JAFFE database. The CK+ database is based on AU coding. People are aged from 18 to 30 and are of different genders and races. The CK+ database contains 486 sequences of 97 subjects. Each sequence contains images with expressions ranging from neutral to extreme. In CK+ database, the sequences which contain faces polluted by digits or contain un-typical expressions are not chosen. If one person has several sequences under the same expression type, we choose only one sequence for this person. Finally, we select a subset containing 392 sequences. The selected sequences include 69 subjects in happy (Ha), 83 subjects in surprise (Su), 69 subjects in disgust (Di), 52 subjects in fear (Fe), 62 subjects in sadness (Sa), 44 subjects in angry (An) and 13 subjects in contempt (Co). In each sequence, each subject has three images that vary from neutral to extreme expressions.

6.1 Precision evaluation of the detected facial landmarks

We use the method in [8] to locate facial landmarks. We also compare its results with the combination method of [7] and [4]. The latter method uses multi-tree model with a shared pool of parts [7] to roughly locate faces then use SDM [4] to fit facial landmarks accurately. The training datasets of the latter landmark detection algorithm include the CMU MultiPIE [45], LFPW [46] and part of LFW [47]. Since CK+ database provides the ground truth facial landmarks labeled by hand, we calculate the pixel-wise absolute distance between the detected landmarks and the ground truth landmarks to evaluate the localization error. Each face has 68 facial landmarks for all methods. The faces in CK+ are in different size. To reduce the influence of image size on the localization errors, we have to normalize the absolute distance. We use the biggest distances between the ground truth facial landmarks in x axis and y axis as the width and height of each face. Since the aspect ratio of each face is different, we normalize the distance in x axis and y axis with the width and height of each face, respectively. The normalized Distance Precision (*DP*) between the ground truth landmark (x^1_g, y^1_g) and the detected one (x^1_d, y^1_d) by [8] is formulated as equation (12).

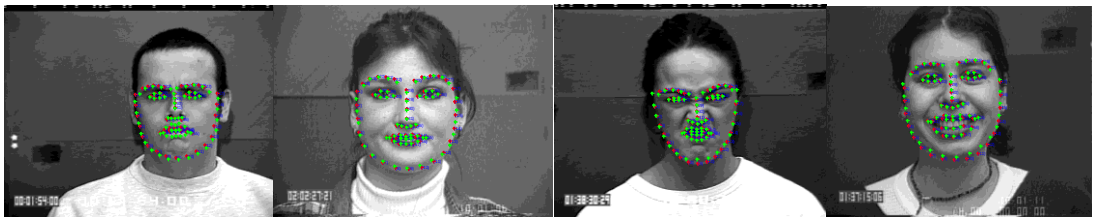
$$DP_1 = \left| \frac{x^1_g - x^1_d}{width} \right| + \left| \frac{y^1_g - y^1_d}{height} \right| \tag{12}$$

The average Distance Precision (avgDP), the standard deviation of Distance Precision (stdDP), the

maximum Distance Precision (maxDP) and the minimum Distance Precision (minDP) are calculated based on the *DP* of facial landmarks in each type of expression. The results are given in Table 1. We can see the method in [8] works better than the combination of methods in [7] and [4]. The avgDP and stdDP of both methods are very small for every expression, which indicate the robustness of facial landmark detection methods. We also illustrate the landmarks detected by [8] which have maxDP for each expression in Fig.8. In Fig.8, the ground truth and detected landmarks are illustrated in red and green, respectively. Fig.8(e) and Fig.8 (f) show that the ground truth landmarks are not always accurate.

Table 1. The normalized distance between the ground truth and the detected landmarks

Precision	**avgDP**		**stdDP**		**maxDP**		**minDP**	
Methods / Expressions	Ref[4]+ [7]	Ref[8]	Ref[4]+ [7]	Ref[8]	Ref[4]+ [7]	Ref[8]	Ref[4]+ [7]	Ref[8]
Anger	0.0233	0.0216	0.0102	0.0038	0.1024	0.0385	0.0142	0.0127
Contempt	0.0325	0.0274	0.0032	0.0020	0.0414	0.0311	0.0243	0.0231
Disgust	0.0271	0.0214	0.0150	0.0042	0.1314	0.0401	0.0145	0.012
Happy	0.0236	0.022	0.0078	0.0043	0.0571	0.0379	0.0138	0.0139
Sad	0.0250	0.0253	0.0150	0.0142	0.1058	0.105	0.0132	0.0136
Surprise	0.0216	0.0211	0.0063	0.0043	0.0790	0.0402	0.0128	0.013
Fear	0.0229	0.0218	0.0089	0.0047	0.1178	0.038	0.0112	0.0129
Average	0.0251	**0.0229**	0.0094	**0.0053**	0.0907	**0.0472**	0.0148	**0.0145**



(a) (b) (c) (d)

(e) (f) (g)

Fig.8 Facial landmarks accuracy comparison between the ones detected by method [8] and the ground truth under the expressions of **(a)** Anger, (b) Contempt, (c) Disgust, (d) Happy, (e) Sad, (f) Surprise and (g) Fear. The ground truth landmarks provided by the CK+ database are in red. The landmarks detected by [8] are in green.

Fig. 9 shows of the results of face alignment under different expressions. The samples of raw texture under the happy shape are given in Fig.9(a). We calculate the mean shape of each

expression. Then, facial texture under this expression is aligned to the mean shape through affine transformation. The normalized results are given in Fig.9 (b).

(a)

(b)

Fig. 9 Illustration of the face alignment under different expressions. (a) Facial texture extracted according to the detected facial landmarks. (b)The normalize face under the mean shape of the expression of happy

6.2 The experimental results on static expression synthesis

Fig.10 and Fig.11 illustrate the synthesized expressions using the proposed static expression synthesis method on the JAFFE and CK+ databases, respectively. The subfigures of Fig.10 correspond to the expressions: (a) Surprise, (b) Fear, (c) Disgust, (d) Sad, (e) Happy, (f) Anger. The subfigures of Fig.11 correspond to the expressions: (a) Anger, (b) Contempt, (c) Disgust, (d) Fear, (e) Happy, (f) Surprise, (g) Sad. The images in each subfigure in the 1st row represent the neutral expression (left), warped expression (middle) and non-neutral expression (right) of the source subject. The images in each subfigure in the 2nd row represent the neutral expression of the target subject (left), warped expression of the target subject (middle) and the transferred expression (in green rectangle). From Fig. 10 and Fig. 11, we can see the synthesized images preserve both the facial appearance of the target subject and the local and global expression deformations of the source subject. The deformations look realistic and natural. Some of the synthesized results of the target subject look like the source subject in the JAFFE database. This is because the people in JAFFE database are quite similar. The CK+ database contains people from different races, with large difference between their faces.

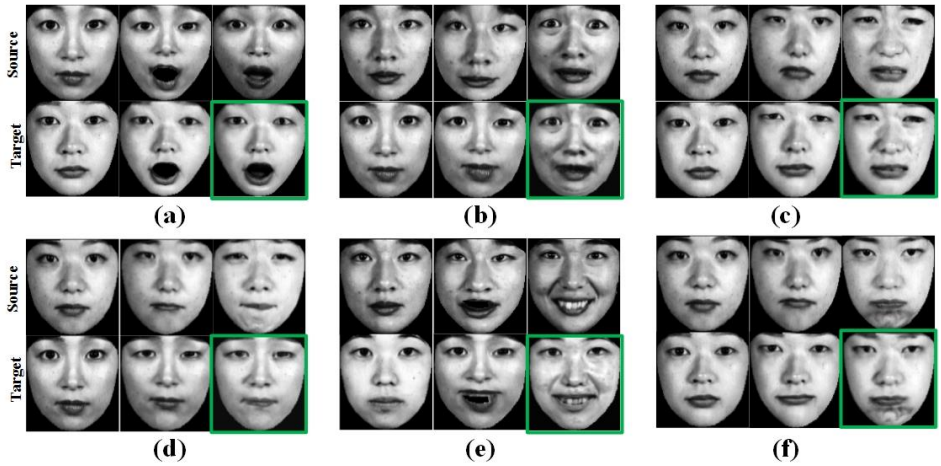


Fig.10 The synthesized static expressions on JAFFE database

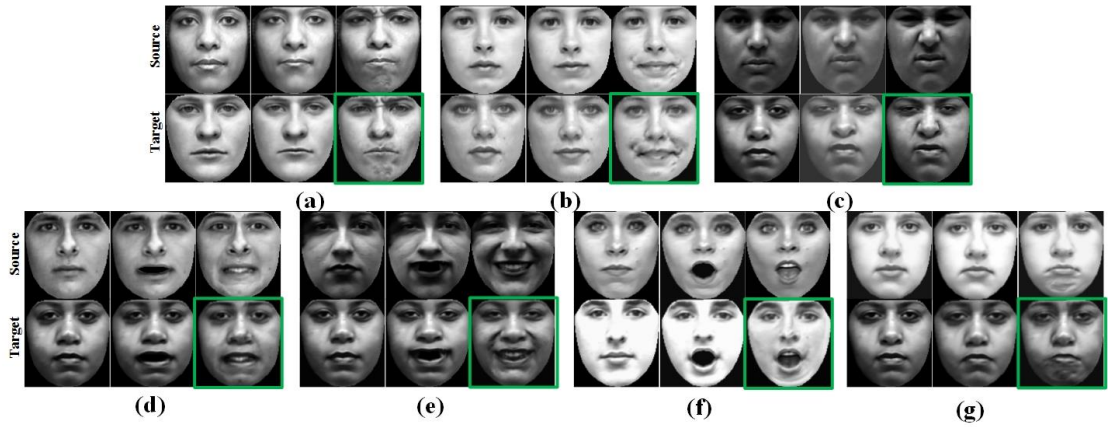


Fig.11 The synthesized static expressions on CK+ database

To further verify that the proposed method can synthesize person-specific expressions, we transfer the happy and fear expressions of different persons to the same neutral face. Partial results are illustrated in Fig.12. The 1st row of sub-figures (a)-(h) includes the neutral face (left) and non-neutral face (right) of the source subject. The 2nd row of sub-figures (a)-(h) includes the neutral face (left) of the target subject and synthesized face (right) of the target subject with the transferred expression of the source subject. Taking the happy expression as an example, we can see the expression details of different source subjects are quite person-specific. The transferred expression of the target subject retains the person-specific expression details of the source subject very well. The proposed method only needs one neutral and one non-neutral expression of the source subject and one neutral expression of the target subject for static expression synthesis. This efficient use of data facilitates the application of our method. In addition, the proposed method takes advantage of the frequency domain to describe the expression details. Thus the transferred expression retains the details better and is less affected by changes in illumination compared with the traditional methods.

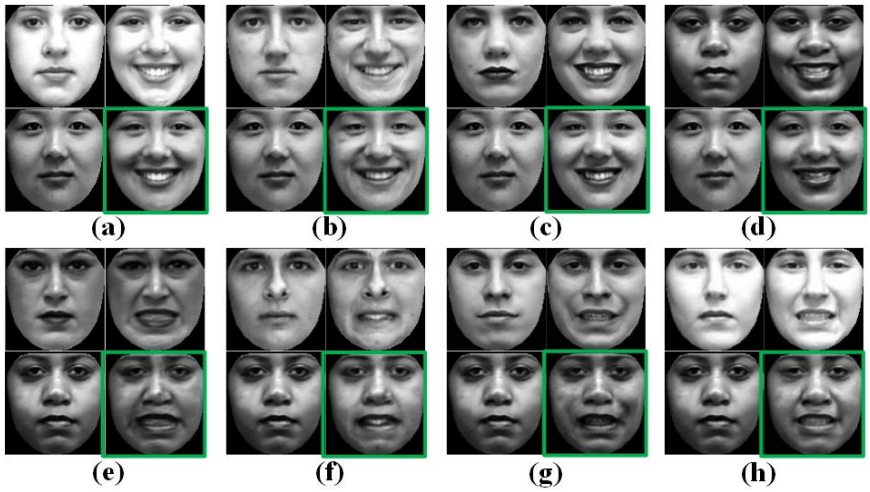


Fig. 12 Person-specific expression synthesis by the proposed static expression transfer method

We use the faces from the PIE dataset to verify the robustness of our method to pose variation and side illumination. The results in Fig.13 demonstrate the transferred expression from a non-neutral expression in profile face, which proves the robustness of our method to pose variation. The faces in Fig.14 illustrate the transferred expression where half face of the target subject is shaded. From the results we can see the transferred expression in Fig.14 (f) is slightly influenced by the side light. The brightness of the synthesized face is influenced by the low frequency component of the source subject somehow.

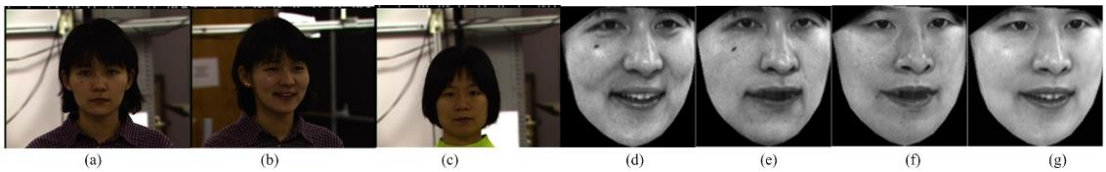
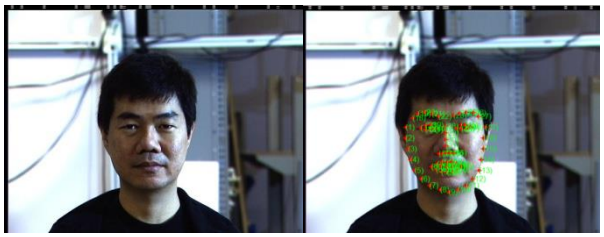


Fig. 13. Demonstration of our method to the robustness of pose variation. (a) illustrates neutral expression of the source person. (b) illustrates non-neutral expression of the source person, which is profile. (c) is the neutral expression of the target person. (d) is the facial texture extracted from (b). (e) is the warped face of (a) to the shape of (e). (f) is the warped face of (c) to the shape of (e). (f) is the synthesized expression of the target person in (c) in profile.



(a) (b)

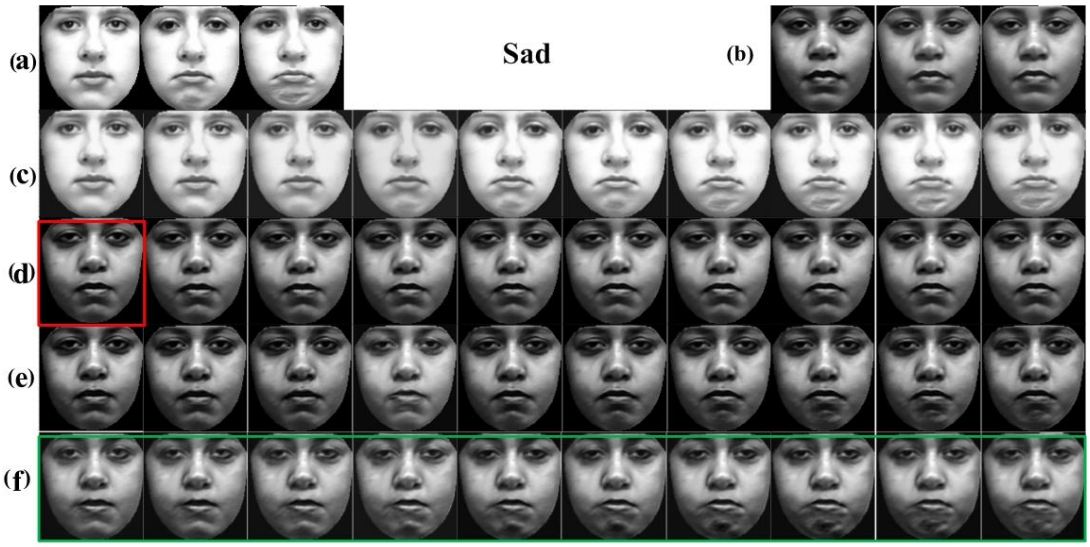
(c) (d) (e) (f)

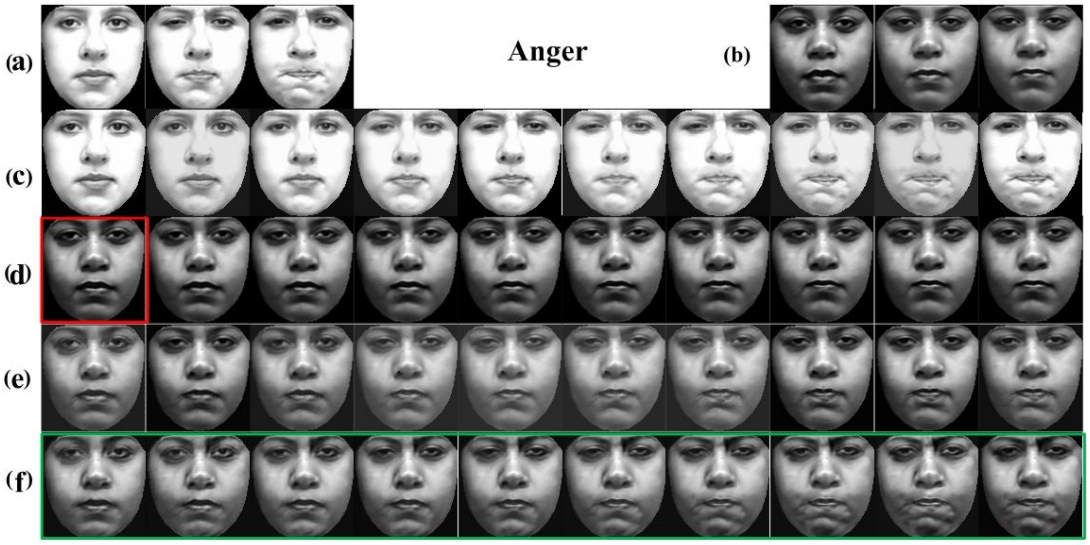
Fig. 14. Demonstration of our method to the robustness of illumination variation. (a) illustrates neutral expression of the target person, whose half face is shaded. (b) illustrates the facial landmarks automatically detected from (a). (c) is the surprise expression of the source subject. (d) is the warped surprise face from the neutral expression of the source subject. (e) is the warped surprise face of the target subject from the neutral expression in (a). (f) is the synthesized expression of the target person in (a).

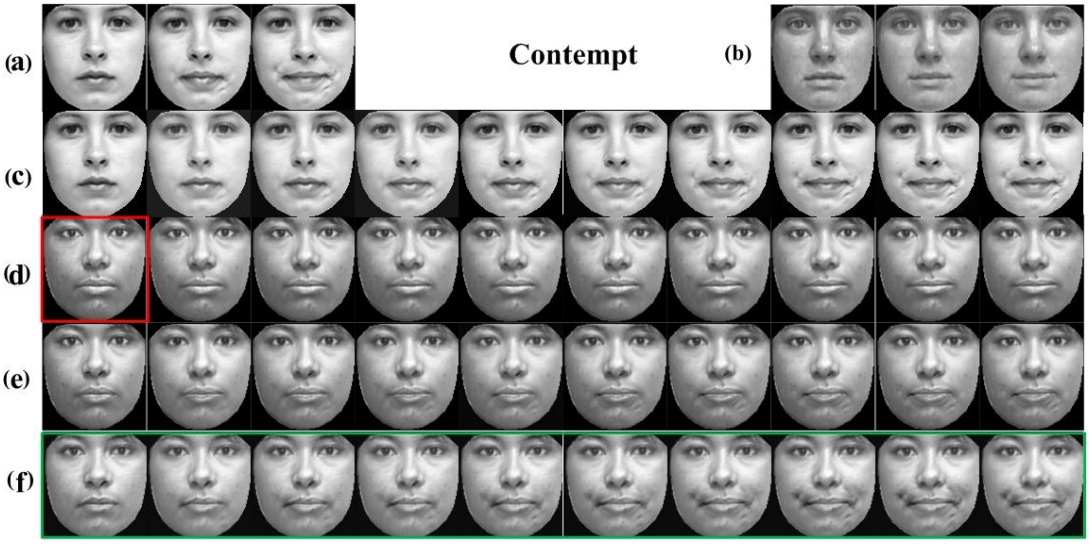
6.3 The experimental results on dynamic expression synthesis

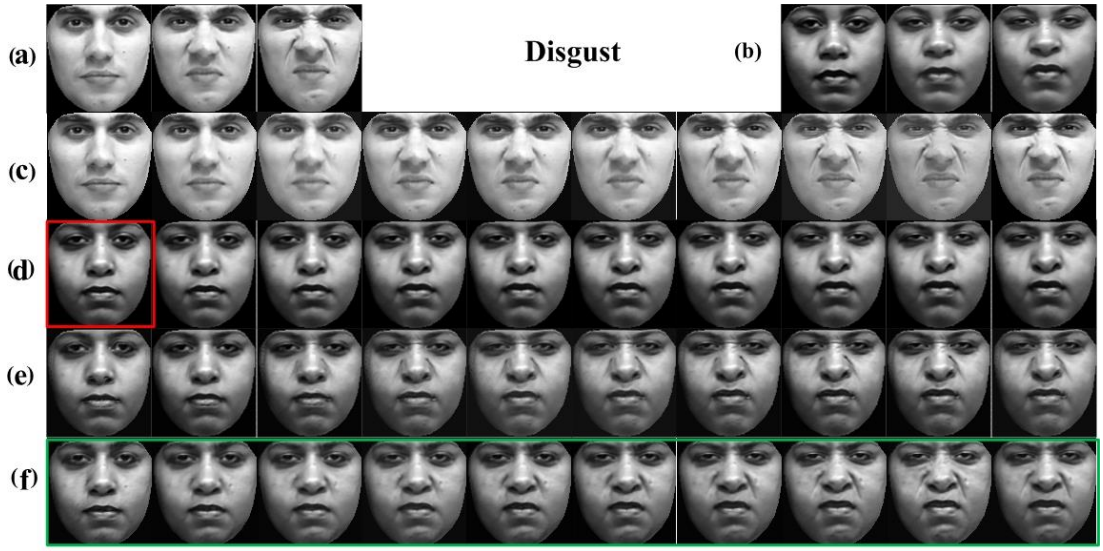
Since only the CK+ database has continuously varied expressions, we synthesize the dynamic expressions of different persons on the CK+ database. We compare our method with the BKRRR, geometric warping based method and dynamic expression transfer method in image domain [2]. The results are illustrated in Fig.15. To transfer the dynamic expression of the source subject to the target subject under neutral expression, we synthesize the dynamic expressions of the source subject (Fig.15 (c)) by the K-TensorFace model, then warp the neutral expression of the target face (in the red rectangle in Fig.15 (d)) to the dynamic shapes of the source subject, to get the geometric warped dynamic expressions of the target subject in Fig.15 (d). The dynamic expression details of the source subject in Fig.15 (c) are transferred to the warped expressions in Fig. 15 (d) to synthesize the images in Fig.15 (e) by the dynamic expression transfer method in image domain [2]. Fig.15 (f) shows the expression transfer results of the proposed dynamic expression transfer method.

The K-TensorFace based method can only synthesize the natural dynamic expressions of the source subject. Though the BKRRR method can synthesize the expressions of the target subject, the synthesized expressions are limited to the expression types of the training images, and the synthesized expressions are not person-specific. The dynamic expression transfer method in the image domain [2] can transfer the person-specific dynamic expressions, which are much closer to the source expression than the warped dynamic expression. Our method has better results than [2] in keeping the tiny creases and wrinkles of the source subject. This is because the expression details can be better extracted and transferred by our method in frequency domain. Thus, the experimental results on CK+ database also prove the effectiveness of our method.

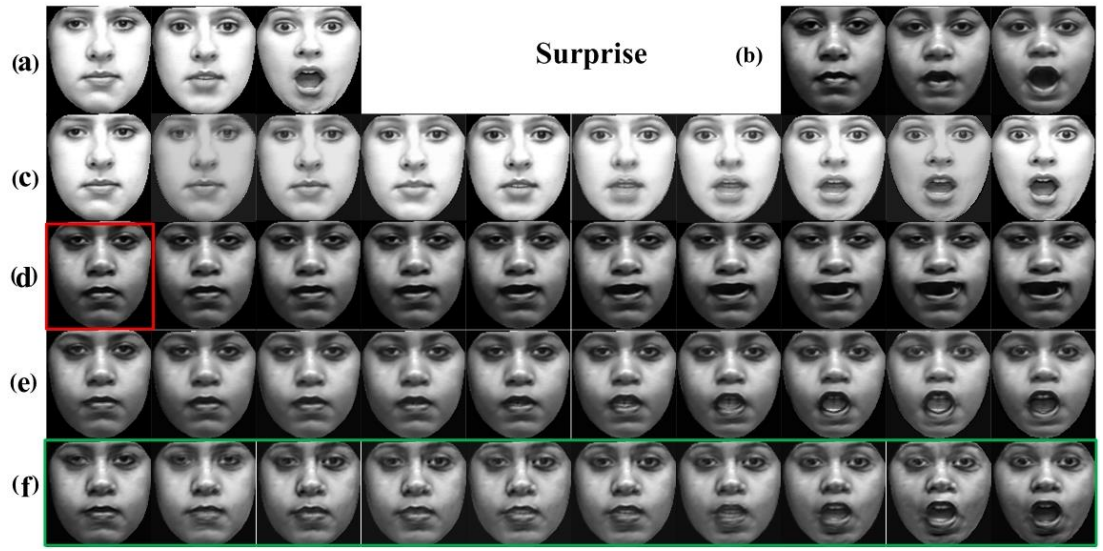

(a)
Sad
(b)
(c)
(d)
(e)
(f)


(a)
Anger
(b)
(c)
(d)
(e)
(f)


(a)
Contempt
(b)
(c)
(d)
(e)
(f)


(a)
Disgust
(b)
(c)
(d)
(e)
(f)


(a)
Fear
(b)
(c)
(d)
(e)
(f)


(a)
Surprise
(b)
(c)
(d)
(e)
(f)

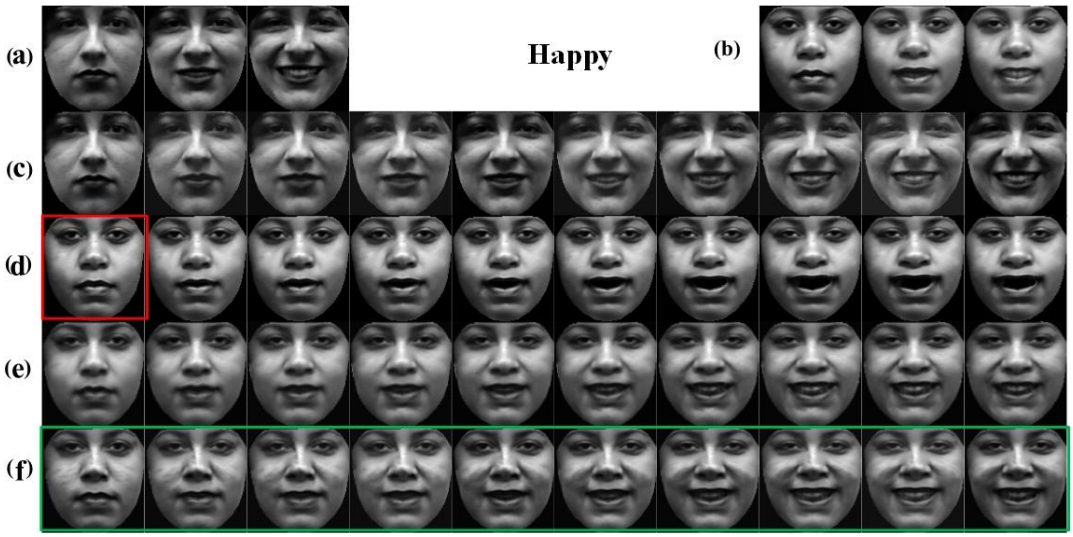


Fig.15 The comparison of faces synthesized by different methods in different expressions. (a) The training examples of corresponding expressions. (b) Faces synthesized by the BKRRR method in each expression. (c) The synthesized dynamic expressions of the training examples in (a) by the K-TensorFace model. (d) Faces warped from the target face (in the red rectangle) into the dynamic shapes of expressions in (c). (e) The synthesized dynamic expressions by transferring the expressions in (a) to (d) with the method in [2]. (f) The synthesized the dynamic expressions by transferring the expressions in (a) to (d) with the proposed method.

A table exhibiting the time intervals of computations concerning the landmark detection by method of [8], static expression transfer and dynamic expression transfer is given as Table 2. The testing is conducted on a computer with an Intel Xenon 2 core, 2.66 GHz CPU with 8 GB RAM.

Table 2. The speed test of each stage of our expression synthesis method (second/frame).

Stages	Landmark detection by method of [8]	Static expression transfer	Dynamic expression transfer
frame/second	56	16	7

7. Conclusions

A static facial expression transfer method based on frequency domain analysis is proposed. We locate the facial landmarks automatically and use them to describe the shape deformation between neutral expressions and the non-neutral expressions. We adopt a wavelet transform to divide images into four frequency bands and transfer the fine details of the expression. Our method is insensitive to illumination variation. Since the dynamic variations are important in interpreting facial expression precisely, we extend our work to dynamic expression transfer. The experiments on the CK+, JAFFE and PIE databases show that our method can synthesize images, while preserving both the facial landmarks of the target subject and the expression details of the source subject. Experiments show that this dynamic expression transfer is superior to the state-of-the-art

methods. The proposed method can be widely applied in practice because it only needs a small training set.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.61231016, 61301192, 61571354 and 61201291), Natural Science Basis Research Plan in Shaanxi Province of China (No.2013JQ8032), the NPU Foundation for Fundamental Research (No. 3102015JSJ0006) and Basic Science Research Fund in Xidian University (No.JB140222).

References

[1] M. Albert, Communication without words, Psychology Today 2(4) (1968) 53-56.

[2] Y. Zhang, W. Wei, A realistic dynamic facial expression transfer method, Neurocomputing 89 (2012) 21-29.

[3] P. Perakis, T. Theoharis, I. A. Kakadiaris, Feature fusion for facial landmark detection, Pattern Recognition 47(9) (2014) 2783-2793

[4] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 532-539.

[5] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, 2012, pp. 2887-2894.

[6] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 FPS via regressing local binary features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1685-1692.

[7] X. Zhu, D. Ramanan, Face detection, pose estimation and landmark localization in the wild, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, 2012, pp. 2879-2886.

[8] V. Kazemi, J. Sullivan. One millisecond face alignment with an ensemble of regression trees. in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition . Columbus, OH, 2014, pp. 1867 - 1874.

[9] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimisionality: High dimensional feature and its efficient compression for face verification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 3025-3032.

[10] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, Technical Report, arXiv: 1406.4773, 2014.

[11] K. Yu, Z. Wang, L. Zhuo, J. Wang, Z. Chi, D. Feng, Learning realistic facial expressions from web images, Pattern Recognition 46(8) (2013) 2144-2155.

[12] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D.H. Salesin, Synthesizing realistic facial expressions from photographs, in: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, New Orlando, FL, 1998, pp. 75-84.

[13] J. Y. Noh, U. Neumann, Expression cloning, in: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, 2001, pp. 277-288.

[14] E. Keeve, S. Girod, R. Kikinis, B. Girod, Deformable modeling of facial tissue for craniofacial surgery simulation, Computer Aided Surgery 3(1998) 223-228.

[15] D. Huang, F. De la Torre, Bilinear kernel reduced rank regression for facial expression synthesis, in: Proceedings of the European Conference on Computer Vision, Crete, Greece, 2010, pp. 364-377.

[16] K. W. Chung, H. M. Chung, Gross anatomy (Board Review), Lippincott Williams & Wilkins, Hagerstown, 2005.

[17] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion, Journal of Personality Social Psychology 2(17) (1971) 124-129.

[18] S. M. Platt, N. I. Badler, Animating facial expressions, in: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, Dallas, TA, 1981, pp. 245-252.

[19] K. Waters, A muscle model for animating three dimensional facial expression, in: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, Anaheim, CA, 1987, pp.17-24.

[20] R. Koch, M. H. Gross, F. R. Carls, D. F. Von Buren, G. Fankhauser, Y. I. H. Parish, Simulating facial surgery using finite element methods, in: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, 1996, pp.421-428.

[21] Y. Lee, D. Terzopoulos, K. Waters, Realistic modeling for facial animation, in: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, 1995, pp.55-62.

[22] M. T. Eskil, K. S. Benli, Facial expression recognition based on anatomy, Computer Vision and Image Understanding, 119 (2014) 1-14.

[23] C. Darwin, The expression of the emotions in man and animals, London: John Murray, 1st edition, 1872.

[24] Z. Ambadar, J. W. Schooler, J. F. Cohn, Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions, Psychological Science 16(5) (2005) 403-410

[25] H. Fang, N. M. Parthalain, A. J. Aubrey, G. K. L. Tam, R. Borgo, P. L. Rosin, P. W. Grant, D. Marshall, M. Chen, Facial expression recognition in dynamic sequences: An integrated approach, Pattern Recognition 47(3) (2014) 1271-1281

[26] T. Wang, J. J. Lien, Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation, Pattern Recognition 42(5) (2009) 962-977

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Mattews, The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression, in: Proceedings of the IEEE CVPR Workshop on Human Communicative Behavior Analysis, San Francisco, CA, 2010, pp. 94-101.

[28] T. F. Cootes, C. J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Computer Vision and Image Understanding 61 (1) (1995) 38-59.

[29] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, in: Proceedings of the European Conference on Computer Vision, Freiburg, Germany, 1998, pp. 484-498.

[30] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active appearance models, IEEE Transaction on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681-685.

[31] I. Matthews, S. Baker, Active appearance models revisited, International Journal on Computer Vision 60 (2) (2004) 135-164.

[32] K. Ramnath, S. Koterba, J. Xiao, C. Hu, I. Matthews, S. Baker, J. Cohn, T. Kanade, Multi-view AAM fitting and construction. International Journal on Computer Vision 76 (2008) 183-204.

[33] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2001, pp. 511-518.

[34] F. I. Parke, K. Waters, Computer facial animation, Wellesley, Massachusetts, 1996.

[35] Z. Liu, Y. Shan, Z. Zhang, Expressive expression mapping with ratio images, In: Proceedings of ACM Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, 2001, pp. 271-276.

[36] H. Wang, N. Ahuja, Facial expression decomposition, in: Proceedings of the International Conference on Computer Vision, Beijing, China, 2003, pp. 958-965.

[37] I. Macedo, E. Vital, B. L. Velho, Expression transfer between photographs through multilinear AAM's, in: Proceedings of Brazilian Symposium on Computer Graphics and Image Processing, Amazonas, Brazil, 2006, pp. 239-246.

[38] H. Lee, D. Kim, Tensor-based AAM with continuous variation estimation: application to variation robust face recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence 31 (6) (2009) 1102-1116.

[39] M. A. O. Vasilescu, D. Terzopoulos, Multilinear analysis of image ensembles: tensorfaces, in: Proceedings of the Seventh European Conference on Computer Vision, Copenhagen, Denmark, 2002, pp. 447-460.

[40] G. J. Edwards, C.J. Taylor, T.F. Cootes, Interpreting face images using active appearance models, in: Proceedings of the International Conference on Face and Gesture Recognition, Nara, Japan, 1998, pp. 300-305.

[41] C. Lee, A. Elgammal, Nonlinear shape and appearance models for facial expression analysis and synthesis, in: Proceedings of the International Conference on Pattern Recognition, Hong Kong, China, 2006, pp. 497-502.

[42] Z. Deng, J. Noh, Computer facial animation: A survey, In Data-Driven 3D Facial Animation, Springer-Verlag, (2007) 1-28.

[43] M. Pantic, L. J. M. Rothkrantz, Automatic analysis of facial expressions: The state of the art, IEEE Transaction on Pattern Analysis and Machine Intelligence 22 (12) (2000)1424-1445.

[44] C. Tian, G. Fan, X. Gao, Q. Tian. Multi-view face recognition: From TensorFace to V-TensorFace and K-TensorFace, IEEE Transaction on Systems, Man, and Cybernetics: Part B: Cybernetics 42(2)(2012)320-333.

[45] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker. Multi-pie. Image and Vision Computing, 2010.

[46] http://www.kbvt.com/LFPW/

[47] http://vis-www.cs.umass.edu/lfw/

[48] V. Bettadapura. Face expression recognition and analysis: The state of the art. Tech Report, arXiv:1203.6722, 2012.