Machine Vision and Applications (2018) 29:263–284
https://doi.org/10.1007/s00138-017-0895-6

ORIGINAL PAPER

Facial expression analysis and expression-invariant face recognition by manifold-based synthesis

Yao Peng[1] · Hujun Yin[1]

Received: 28 September 2016 / Revised: 26 October 2017 / Accepted: 30 October 2017 / Published online: 18 December 2017

Abstract
Over the last decades, expression classification and face recognition have received substantial attention in computer vision and pattern recognition with more recent efforts focusing on understanding and modelling expression variations. In this paper, we present an expression classification and expression-invariant face recognition method by synthesising photorealistic expression manifolds to expand the gallery set. By means of synthesising expression images from neutral faces, more within-subject variability can be obtained. Eigentransformation is utilised to generate both shape and expression details for novel subjects. Expression classification and face recognition are then performed on the extended training set with synthesised expressions. Experimental results on various datasets show that the proposed method is robust for recognising various expressions and faces with varying degrees of expression. Comprehensive experiments conducted and comparisons with the existing methods are reported. Cross-database synthesis and effect of landmark quality are also studied.

Keywords Face recognition · Expression classification · Expression manifold synthesis · Eigentransformation

1 Introduction

With the demand for secure access and human–computer interaction, a great deal of interest has been spurred in further enhancing and refining biometric technologies. Compared to other biometric means such as fingerprints, voice and iris, face recognition possesses the merit of being less intrusive and provides a more direct and convenient approach to identity and access management [54,61]. Face recognition is based on measuring and matching distinctive features of faces from images or videos for identification or verification of identities and has a growing number of applications in law enforcement, surveillance, security, system access, video indexing, social networking and so on [61]. Over the last two decades, despite intensive research and significant advances achieved in both academia and industry, there are still many challenges. One of the most challenging aspects of face recognition is handling large number of variations under unconstrained environments caused by lighting, background, pose, occlusion and facial expression. Unlike pose and illumination variations, less attention has been paid to modelling expression variations and recognising individuals across different expressions. In practical face recognition systems, expression variations can greatly affect recognition performance and cause severe problems [34,43,61]. Expression-invariant face recognition deals with changing human appearance due to expression and attempts to eliminate a broad range of discrepancies by considering dynamics of expression in recognition.

Facial expression analysis and synthesis also play an important role in computer vision and graphics. Although the subtlety, complexity and variability of facial expression make it difficult to recognise, its envisaged applications have fostered a great deal of research interest in automatic expression analysis [14,43]. Automatic recognition of facial expressions has become a key component in human–machine interfaces, emotion analysis, robotics and medical care [40].

To overcome the difficulty of identifying faces with expressions inconsistent to those in the training set and to improve expression classification accuracy, we generate synthetic expressions from neutral faces on both geometry attributes and expression details. These synthesised images are added to expand the gallery set. For a probe

✉ Hujun Yin
hujun.yin@manchester.ac.uk

Yao Peng
yao.peng@postgrad.manchester.ac.uk

[1] School of Electrical and Electronic Engineering, The University of Manchester, Manchester, UK

face with arbitrary expression, expression classification and face recognition are performed by representing each face as a weighted combination of shape and texture attributes on the extended training set. Since variation dynamics are important to interpreting facial expressions, we then further extend the synthesis to expression manifolds, which provide continuous expression image sequences. The effectiveness of the proposed method has been verified on the extended AU-Coded Cohn-Kanade (CK+) dataset [23,27], Bosphorus database [42], AR face database [30], Japanese female facial expression (JAFFE) database [28], MUG [3] and MMI [37,53] facial expression databases.

The rest of the paper proceeds as follows. Related work is reviewed in Sect. 2. Section 3 describes the eigentransformation algorithm for synthesising expressions and dynamic expression manifolds. In Sect. 4, we illustrate how to verify expression synthesis and how to use synthesised images to perform expression classification and invariant face recognition. Section 5 presents experimental results, followed by conclusions in Sect. 6.

2 Related work

Facial expression analysis plays a significant role in interpersonal communication since emotions are often conveyed through facial expression to make human behaviour more understandable. Originated in 1872, the earliest attempt in expression analysis was Darwin's demonstration of the general principles of expression and the link between human emotions and actions. In 1978, psychologists Ekman and Friesen [12] developed the Facial Action Coding System (FACS) based on anatomical analysis of facial muscle movements. They divided face into 44 anatomically separated action units (AUs) and used the combination of AUs to describe all possible expressions. FACS analyses the intensity of AUs and their effects on associated expressions, providing an objective measurement for facial expressions. FACS categorises the expression intensity into five levels as *trace*, *slight*, *marked* or *pronounced*, *extreme* or *severe*, and *maximum*. In 1992, Ekman also defined six basic expressions as *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* [13].

Recently, researchers from the computer graphics community have developed a number of techniques to generate realistic facial expressions [29,39,60]. Liu et al. [25] presented a novel facial expression mapping method by computing expression ratio images (ERIs) based on illumination changes of one person's expressions as to capture the subtle but visually important details of expressions. Together with geometric warping, more realistic facial expressions could be generated by mapping an ERI to different persons' faces. A shortcoming of this method was that it ignored the appearance changes from person to person. The requirements of ERIs from input subjects could also be difficult to obtain. Zhang et al. [60] proposed a geometry-driven facial expression system for generating photorealistic expressions with natural-looking expression details. Each face region was divided into 14 subregions. Given the geometry of an expression, the system automatically synthesised texture information by a combination of the corresponding regions in the training set. This system provided an effective solution to synthesising expression details, but it could not operate when there was only one target face image available. Wang and Ahuja [55] utilised a higher-order singular decomposition to decompose facial expression space into three subspaces (identity, expression and feature subspaces). This approach modelled the mapping between identities and expressions in order to synthesise expressions of novel faces, and the resulting subspaces could be used for simultaneous identity and facial expression recognition. However, this method cannot synthesise expressions of subjects with unforeseen characteristics, such as a beard if there was no similar images in the training set. Based on the bilinear kernel-reduced rank regression, Huang and De la Torre [21] proposed a facial expression synthesis method by learning a mapping function between neutral and various expressions. Combined with input neutral images, the method could output realistic expressions with person-specific expression details if enough training subjects were available.

In order to automatically analyse and recognise facial expressions, several approaches have been developed to derive and represent facial changes caused by expressions. Generally, they can be classified into two types: geometric feature-based and appearance-based [22]. Geometric feature-based methods extract geometric features, either shape [16] and locations of facial components [35,36,52] or the movements of facial landmarks [44]. In appearance-based methods, a bank of image filters such as Gabor wavelets [59] and local binary patterns (LBP) [43] are applied to either the entire image or specific regions to derive facial appearance changes. Song et al. [46] derived image ratio features to observe skin deformation parameters (SDPs) caused by expression in eight facial patches. Facial animation parameters (FAPs) were used to describe the facial feature movements and further improve expression recognition accuracy. Gu et al. [19] adopted radial encoding strategy to downsample the Gabor features derived from local patches. The local features were fed to local classifiers to be integrated into global features for representing facial expressions. Soyel and Demirel [47] presented a feature representation based on discriminative scale-invariant feature transform (D-SIFT) for facial expression classification. The Kullback–Leibler divergence and weighted majority voting-based classifier were employed to generate the overall

decision from the extracted facial feature vectors. Rahulamathavan et al. [40] performed expression recognition using local Fisher-discriminant analysis (LFDA). Mohammadi et al. [32] presented a sparse representation-based classification approach by modelling variations in difference images (expressive faces subtracted from neutral faces of the same subjects). In classification step, each test image was sparsely represented as a linear combination of the principal components of six universal facial expressions. Happy and Routray [20] derived features for recognising expressions on salient patches that stayed active during emotion elicitation.

The difficulty of recognising individuals across different expressions is that distances between faces with different expressions but the same identity can be greater than those with the same expression but different identities [34]. Existing approaches either synthesise or transform expressions to extend the gallery set [1,2,24] or treat the task as single sample per person problem to extract as much information as possible from the gallery images [31,49]. Dibeklioğlu et al. [10] proposed an automatic pose and expression-invariant 3D face recognition system, in which pose correction and nose segmentation were utilised. The results indicated an improvement in recognition accuracy for both expression and pose variations. Mohammadzade and Hatzinakos [34] introduced expression subspaces for synthesising new expressions from one image per subject. By projecting an arbitrary expression image into the expression subspaces, new expressions could be obtained to expand the original training set for discriminant analysis. Petpairote and Madarasmi [38] utilised modified thin-plate spline warping to remove expression from a probe image and performed face recognition using a gallery of neutral faces. Mohammadi et al. [33] addressed the recognition of expression and identity simultaneously using a sparse representation. Each facial appearance was coded as a sparse combination of identity and expression, and classification was conducted by estimating the two attributes for test subjects. Taigman et al. [48] developed a deep architecture and derived robust and generalised face representations. The system could also be applied to effectively align faces based on explicit 3D modelling and achieved state-of-the-art performances. Gao et al. [15] used supervised auto-encoders to build a deep architecture and extract features for image representations. Faces with expression and illumination variations were enforced to have similar features to the ones with neutral expressions and normal illumination. Zaman et al. [58] proposed a feature selection process by weighting local features through linear discriminant analysis. They then defined the contribution of each local classifier according to the corresponding weights and recognised faces using local vectors obtained from the locally lateral subspace (LLS) strategy. Experiments indicated that the contribution of facial features changed depending on the type of expressions.

3 Expression synthesis

One of the contributions of this work is to synthesise realistic expression images from only one image per subject. Based on the active appearance model (AAM) [9], each image is represented as a combination of shape and texture attributes. Under the assumption that similar persons have similar expression appearances and shapes [55], various expressions of a new subject can be synthesised on both normalised shape and expression details. Eigentranformation [50,51] has been used for face hallucination and photo-sketch synthesis, and we extend the approach to facial expression synthesis.

In this paper, six basic expressions (*anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*) are considered with five different intensity levels (*trace*, *slight*, *marked*, *extreme*, and *maximum*). Table 1 displays the mean faces in terms of mean shape for each category and intensity level. The general process of expression synthesis, shown in Fig. 1, includes shape alignment and texture normalisation, shape deformation and expression detail generation, and teeth retrieval. To further describe the dynamics of intensity variations, we extend the synthesis to expression manifolds.

3.1 Shape alignment and texture normalisation

To minimise the distribution variation of landmarks that describe the shape of a face, it is necessary to align all the shapes as closely as possible. In each image, shape is defined by a set of τ landmarks

$$S = \left[\left(l_x^{(1)}, l_y^{(1)}\right), \left(l_x^{(2)}, l_y^{(2)}\right), \ldots, \left(l_x^{(\tau)}, l_y^{(\tau)}\right)\right]. \quad (1)$$

Figure 2 gives an example of the arrangement of the landmarks on the face. These landmarks are located around key areas such as eyes, eyebrows, nose, mouse and contour. Generalised procrustes analysis (GPA) [18] is a common technique to align a set of shapes with respect to translation, rotation and scaling by calculating the least square estimation of those shape matrices. Although GPA is an effective and straightforward approach, the use of least square criterion suffers from the Pinocchio effect in which large variations limited to a single point or a few landmarks can be smeared out [8]. Consequently, GPA is not suitable for aligning shapes of faces in which expression variations account for major variation. Instead, we align all face images so that eyes are located at the same coordinates.

Furthermore, to remove geometry information in texture normalisation, each image is warped so that its landmarks

Table 1 Mean shapes of six expression categories (anger, disgust, fear, happiness, sadness and surprise) with different intensity levels (neutral, trace, slight, marked, extreme and maximum)

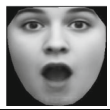
Expressions	Expression intensity level					
	Neutral	Trace	Slight	Marked	Extreme	Maximum
Anger						
Disgust						
Fear						
Happiness						
Sadness						
Surprise						

Fig. 1 General process of expression synthesis

match that of the mean shape of each category. Delaunay triangulation strategy is used to divide the face texture into small patches according to the distribution of landmarks, as shown on the left images in Fig. 3a, b, and a piece-wise affine transformation is applied to warp the texture from a triangle to its corresponding triangle. Figure 3 illustrates the warping results from original shapes to the mean shapes of each expression category.

3.2 Eigentransformation-based approach

Many studies have shown that a face image can be reconstructed by eigenfaces using the principal component analysis (PCA). Based on the eigenface approach, eigentransformation algorithm indicates that an image can also be represented by a weighted combination of training images instead of eigenfaces [51].

Fig. 2 An example of landmark distribution on the face

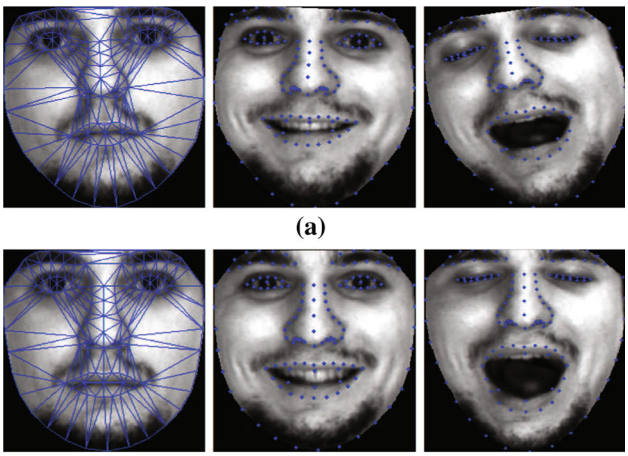


Fig. 3 **a** Original face images and **b** warped face images to the mean shape of each category, e.g. neutral, smile and surprise

Assume that the training set $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m]$ is a d-by-m matrix, where each column is an image vector, d is the number of pixels in each image, and m is the number of the training samples, an unseen image can be reconstructed by

$$\boldsymbol{x}_r = \boldsymbol{\psi} + \boldsymbol{E}\boldsymbol{\omega}, \tag{2}$$

where $\boldsymbol{\psi}$ is the mean image of the training set, $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_t]$ denotes a set of eigenvectors (also known as eigenfaces) of the covariance matrix $\boldsymbol{C} = \sum_{\alpha=1}^{m}(\boldsymbol{x}_\alpha - \boldsymbol{\psi})(\boldsymbol{x}_\alpha - \boldsymbol{\psi})^T = \boldsymbol{\phi}\boldsymbol{\phi}^T$, and $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_t]^T = \boldsymbol{E}^T(\boldsymbol{x} - \boldsymbol{\psi})$ is a set of weights by projecting $\boldsymbol{x}$ onto the eigenvectors. $\boldsymbol{\phi}$ is the zero-meaned training set. According to singular value decomposition [17],

$$\boldsymbol{E} = \boldsymbol{\phi}\boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}, \tag{3}$$

where $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_t]$ and $\boldsymbol{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_t]^T$ are the eigenvectors and eigenvalues of the matrix $\boldsymbol{\phi}^T\boldsymbol{\phi}$. Hence, the reconstructed image can also be represented as a linear combination of the training samples instead of eigenfaces by

$$\boldsymbol{x}_r = \boldsymbol{\psi} + \boldsymbol{\phi}\left(\boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\omega}\right). \tag{4}$$

To extend this principle for expressions, a novel neutral face, $\boldsymbol{x}^n$, can be represented by a training set containing a set of neutral faces,

$$\boldsymbol{x}^n = \boldsymbol{\psi}^n + \boldsymbol{\phi}^n\boldsymbol{A}^n, \tag{5}$$

where $\boldsymbol{A}^n = (\boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\omega})^n$. Similarly, for a corresponding expressive face such as smile, a training set consisting of corresponding smiling faces is used,

$$\boldsymbol{x}^e = \boldsymbol{\psi}^e + \boldsymbol{\phi}^e\boldsymbol{A}^e. \tag{6}$$

To generate expressive faces from a neutral face, an assumption to be made is that $\boldsymbol{A}^n \approx \boldsymbol{A}^e$ if $\boldsymbol{\phi}^n$ and $\boldsymbol{\phi}^e$ are from the same group of subjects. For the sake of simplicity, we assume that only neutral expression is used to synthesise other expressions, but this process can be generalised to other expressions being used as the original images.

3.3 Shape deformation and expression details generation

With the AAM, an unseen face can be interpreted as a combination of shape and grey-level variations learned from a training set [2]. To describe the expression variations contained in both shape deformation and subtle appearance changes, shape and texture information from the unseen face can be transformed by using statistical shape model [11] and shape-free texture [61], respectively.

Given a set of aligned shapes, we divide the landmarks of each shape into four groups corresponding to facial features: eyebrows and eyes, nose, mouth and face contour. For an unseen neutral shape $\boldsymbol{S}^n = \{\boldsymbol{S}_1^n, \boldsymbol{S}_2^n, \boldsymbol{S}_3^n, \boldsymbol{S}_4^n\}$, which is not contained in the training set, eigentransformation is applied to each group of landmarks to derive the weight matrix $\boldsymbol{A}_S^n = \{\boldsymbol{A}_1^n, \boldsymbol{A}_2^n, \boldsymbol{A}_3^n, \boldsymbol{A}_4^n\}$. Under the assumption that the weight matrix remains unchanged with different expressions, we can synthesise its corresponding expressive shapes $\boldsymbol{S}^e = \{\boldsymbol{S}_1^e, \boldsymbol{S}_2^e, \boldsymbol{S}_3^e, \boldsymbol{S}_4^e\}$.

To synthesise the subtle appearance variations, we warp a neutral expression $\boldsymbol{T}_\gamma^n$ into a non-neutral expressions $\boldsymbol{T}_\gamma^e$ and calculate the difference

$$\boldsymbol{D}_{T_\gamma} = \boldsymbol{T}_\gamma^e - \boldsymbol{T}_\gamma^n \tag{7}$$

Fig. 4 **a** Warped neutral face to the mean smile shape, **b** ground truth smile image warped to the mean smile shape and **c** absolute difference between (**a**) and (**b**)

for each pair of neutral and non-neutral expressions, where γ represents the subject index in the training set. To remove the effect of noise and small image misalignment, a 2-D Wiener filter is applied to each difference image $\boldsymbol{D}_{T_\gamma}$ by estimating local mean and variance around each pixel as

$$\boldsymbol{D}'_{T_\gamma} = \text{Wiener}(\boldsymbol{D}_{T_\gamma}), \tag{8}$$

$$\mu = \frac{1}{p^2} \sum_{i,j \in \eta} \boldsymbol{D}_{T_\gamma}(i, j), \tag{9}$$

$$\sigma^2 = \frac{1}{p^2} \sum_{i,j \in \eta} \left(\boldsymbol{D}^2_{T_\gamma}(i, j) - \mu^2 \right), \tag{10}$$

$$\boldsymbol{D}'_{T_\gamma}(i, j) = \mu + \frac{\sigma^2 - \nu^2}{\sigma^2} (\boldsymbol{D}_{T_\gamma}(i, j) - \mu), \tag{11}$$

where η is a p-by-p local neighbourhood of each pixel in the image $\boldsymbol{D}_{T_\gamma}$, ν^2 is the average of all the locally estimated variances.

Figure 4 illustrates the retrieval of expression details by computing the difference between warped neutral image and expression image.

With the eigentransformation, we can also obtain the weight matrix $\boldsymbol{A}^n_T$ for texture attributes. Based on the fact that similar persons perform similar expressions or expression in a large extent is universal, $\boldsymbol{A}^e_T = \boldsymbol{A}^n_T$ is used to synthesise expression appearance variations $\boldsymbol{D}'_T$ caused by skin deformation. With the weight matrix $\boldsymbol{A}^e_T$, the synthesised expression wrinkles $\boldsymbol{D}'_T$ is computed as a weighted sum of each $\boldsymbol{D}'_{T_\gamma}$ in the training set. For a novel face, the synthesised texture can be expressed by

$$\boldsymbol{T}^e = \boldsymbol{T}^n + \boldsymbol{D}'_T, \tag{12}$$

where $\boldsymbol{T}^n$ represents the warped neutral expressions, $\boldsymbol{D}'_T$ is the synthesised expression details and $\boldsymbol{T}^e$ is the synthesised expression appearance image for the unseen subject.

After obtaining the shape and texture attributes for a target subject, synthesised texture $\boldsymbol{T}^e$ is warped to the reconstructed shape $\boldsymbol{S}^e$ of the target face to obtain the final expression image.

Fig. 5 Teeth retrieval from training subject onto synthesised expression (both shape and texture)

3.4 Teeth retrieval

In certain expressions such as smile and surprise, teeth region is significant for photorealistic expression synthesis. However, due to that no landmarks are provided within the mouth region and that teeth region is person-specific, synthesised teeth by a linear combination of those from training set would be inaccurate and unrealistic. Instead, we retrieve teeth regions from the training set and choose the most similar one for the unseen face. Given a novel subject and its synthesised shape, we calculate the distances between synthesised inner mouth shape and those from training subjects and choose the teeth region from the subject having the minimum difference. As shown in Fig. 5, teeth region is retrieved from the training subject having the most similar mouth shape for the synthesised expression.

3.5 Expression manifold-based synthesis (EMS)

The above-described expression synthesis process can produce realistic standard expressions. However, it does not consider expression intensity. Expression dynamics are of great significance in interpreting facial behaviour and revealing human facial emotion [4,41]. On the one hand, emotions are often communicated by the dynamics of expressions conveyed in the small subtle changes. On the other hand, expression intensity information can be critical for applications such as patient monitoring, security surveillance and human–computer interaction.

Based on the observation that people change expressions continuously over time, various facial expressions of each

subject lie on smooth manifolds in the high-dimensional image space with neutral face being the central reference point [7]. In this paper, we propose to construct expression manifolds for synthesising dynamic expressions. We assume that different facial parts vary uniformly and continuously over expression changes.

In the above static expression synthesis process, we have obtained deformed shape and synthesised expression texture for a target subject. To extend it to expression manifolds for expressions with varying degrees, we need a series of shapes and expression details describing the variations of these expressions.

To capture the dynamics of expression, we first calculate the shape differences between neutral expression and its synthesised expression as

$$\boldsymbol{D}_S = \boldsymbol{S}^e - \boldsymbol{S}^n. \tag{13}$$

Then the proposed EMS generates a continuous expression manifold by interpolating both shape and texture attributes between neutral and a synthesised expression by

$$\boldsymbol{S}_k^e = \boldsymbol{S}^n + \boldsymbol{D}_S \times r_k, \tag{14}$$
$$\boldsymbol{T}_k^e = \boldsymbol{T}^n + \boldsymbol{D}_T' \times r_k, \tag{15}$$

where $k = 1, 2, \ldots, q$, q is the total number of synthesised image in the manifold for each subject, $\boldsymbol{S}_k^e$ and $\boldsymbol{T}_k^e$ represent the k-th shape and texture attributes in the expression manifold, respectively, and r_k is a ratio controlling the expression intensity. It is noted that r_k can be greater than 1. However, values larger than 1 may cause unnatural artefacts.

With a series of shapes and expression details obtained, we warp each $\boldsymbol{T}_k^e$ to its corresponding $\boldsymbol{S}_k^e$ to construct a synthesised image set $\boldsymbol{X}_\gamma = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q]$ in image space $\boldsymbol{X}_\gamma \in \mathbb{R}^d$ for each subject. We model the synthesised image set as an expression manifold $\mathcal{M}_\gamma \subset \boldsymbol{X}_\gamma$, $\dim(\mathcal{M}_\gamma) \ll d$, which is a function of $(\boldsymbol{S}_\gamma, \boldsymbol{T}_\gamma)$. The synthesised manifold is a topological space spanning from neutral to extreme expressions and is locally Euclidean.

3.6 Expression transfer

In the rare case where there is only one training subject with different expressions, we synthesis expressions for a new subject by transferring expressions from a source subject to the target subject. Assuming that the source subject performs a pair of neutral $\{\boldsymbol{S}^n_{\text{source}}, \boldsymbol{T}^n_{\text{source}}\}$ and non-neutral $\{\boldsymbol{S}^e_{\text{source}}, \boldsymbol{T}^e_{\text{source}}\}$ expressions, and the target subject has only neutral expression $\{\boldsymbol{S}^n_{\text{target}}, \boldsymbol{T}^n_{\text{target}}\}$.

To capture shape deformations, we first calculate the percentage of facial landmark movements between two expressions of the source subject and assume the ratio stays the same for the target subject

$$\text{ratio}_{\text{source}} = \frac{\boldsymbol{S}^e_{\text{source}} - \boldsymbol{S}^n_{\text{source}}}{\boldsymbol{S}^n_{\text{source}}}, \tag{16}$$
$$\text{ratio}_{\text{target}} = \text{ratio}_{\text{source}}. \tag{17}$$

Then the synthesised, non-neutral shape of the target subject can be interpreted as

$$\boldsymbol{S}^e_{\text{target}} = \text{ratio}_{\text{target}} \times \boldsymbol{S}^n_{\text{target}} + \boldsymbol{S}^n_{\text{target}}, \tag{18}$$

To simulate subtle wrinkles, we warp the neutral expressions of the source $\boldsymbol{T}^n_{\text{source}}$ and target $\boldsymbol{T}^n_{\text{target}}$ to the non-neutral expressions of the source subject $\boldsymbol{T}^e_{\text{source}}$. Similar to the proposed expression synthesis method, the difference image between $\boldsymbol{T}^n_{\text{source}}$ and $\boldsymbol{T}^e_{\text{source}}$ is computed and a 2-D Wiener filter is applied to the difference image for noise removal. Synthesised texture for non-neutral expressions of the target subject can be computed as,

$$\boldsymbol{T}^e_{\text{target}} = \boldsymbol{T}^n_{\text{target}} + \text{Wiener}\left(\boldsymbol{T}^e_{\text{source}} - \boldsymbol{T}^n_{\text{source}}\right). \tag{19}$$

Furthermore, we can synthesise an expression manifold for the target subject using the proposed EMS method by interpolating both shape and texture attributes.

4 Expression verification, classification and invariant Face Recognition

4.1 Expression verification on synthesised images

To evaluate the performance of the proposed method, synthesised images are compared with the ground truth images on both shape and texture attributes. In order to reduce the influence of image size, we calculate the normalised distance between the ground truth landmarks $\{(l^g_x, l^g_y)\}$ and the synthesised landmarks $\{(l^s_x, l^s_y)\}$ by

$$\text{dist} = \sum_{\beta=1}^{\tau} \left(\left| \frac{l_x^{g(\beta)} - l_x^{s(\beta)}}{\text{width}} \right| + \left| \frac{l_y^{g(\beta)} - l_y^{s(\beta)}}{\text{height}} \right| \right), \tag{20}$$

where the width and height of each face are computed as the largest distance between the ground truth landmarks in x and y axes, respectively.

For texture attributes, the correlation coefficient is used to measure the similarity between the ground truth texture $\boldsymbol{T}^g$ and the synthesised expression appearance $\boldsymbol{T}^s$ as

$$\rho = \frac{\sum_i \sum_j (\boldsymbol{T}^g(i,j) - \mu^g)(\boldsymbol{T}^s(i,j) - \mu^s)}{\sqrt{\sum_i \sum_j (\boldsymbol{T}^g(i,j) - \mu^g)^2 \sum_i \sum_j (\boldsymbol{T}^s(i,j) - \mu^s)^2}}, \tag{21}$$

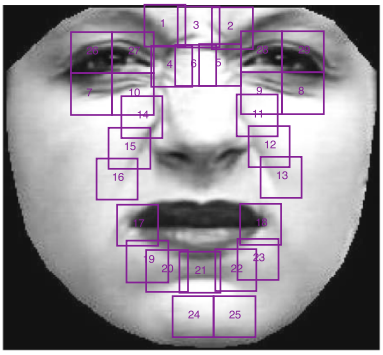


Fig. 6 Locations of facial patches for expression classification

where μ^g and μ^s are the mean values of $\boldsymbol{T}^g$ and $\boldsymbol{T}^s$, respectively.

4.2 Expression classification

For expression classification, we extract local patches from face images around facial landmarks. They contain local features of facial regions exhibiting considerable variations in expressions. Based on the prior knowledge of facial muscles and AUs, the extraction of discriminative regions around the areas of mouth, eye and eyebrows can further assist expression classification [62]. As shown in Fig. 6, we select 29 facial patches for both synthesised and ground truth images and apply scale-invariant feature transform (SIFT) [26] to each patch. In implementation, shapes $\boldsymbol{S}$ are concatenated with SIFT features $\boldsymbol{F}$ as

$$[\boldsymbol{W}_f \times \boldsymbol{S}, (\boldsymbol{I} - \boldsymbol{W}_f) \times \boldsymbol{F}], \tag{22}$$

where $\boldsymbol{I}$ is an identity matrix, and $\boldsymbol{W}_f = \text{diag}(w_f), 0 \leq w_f \leq 1$ is a diagonal matrix of weights controlling the ratio between $\boldsymbol{S}$ and $\boldsymbol{F}$.

Kernel discriminant analysis (KDA) [5] with Gaussian kernel width θ is used as the classifier to find nonlinear projection directions from which classes are separated.

4.3 Expression-invariant face recognition

For expression-invariant face recognition, we use only one neutral ground truth image in the gallery set for each subject. This setting represents the most difficult case. We use the proposed EMS method to synthesise various expression images for all the training subjects to expand the gallery set, and when a probe image with arbitrary expression is presented, we perform face recognition with the expanded training set. Each face is represented as a weighted combination of shape attributes $\boldsymbol{S}$ and texture attributes $\boldsymbol{T}$ as

$$[\boldsymbol{W}_t \times \boldsymbol{S}, (\boldsymbol{I} - \boldsymbol{W}_t) \times \boldsymbol{T}], \tag{23}$$

where $\boldsymbol{W}_t = \text{diag}(w_t)$ is a diagonal matrix of weights. Shape attributes are multiplied by appropriate weights to compensate for the difference between pixel distance values and pixel intensity values. In practice w_t is chosen to be equal to the sum of the normalised texture variance divided by the sum of the normalised shape variance. Linear discriminant analysis (LDA) [6] is used as the classifier and Euclidean distance as the similarity metric to find the match between the probe and training images.

In the cases where a great number of images for each known subject are available or have been collected from video sequences with expressions, face recognition is conducted with a set of probe images and the synthesised expression manifolds rather than single image. Apart from LDA and KDA, we also address face recognition based on image sets by means of manifold to manifold distance (MMD) [57] and manifold discriminant analysis (MDA) [56]. In the MMD and MDA methods, each image set is considered as a manifold represented by a collection of local linear models. MMD computes distances between the galley manifolds learned from the training image sets and the probe manifolds learned from the probe image set and recognition is achieved by seeking the minimum MMD. MDA seeks to learn a discriminant function to map the manifolds into an embedding space in which local models have better between-class separability and enhanced within-class compactness.

5 Experiments and discussion

This section describes various experiments that were conducted to evaluate the proposed expression synthesis method. The benchmark datasets used are first described in 5.1. Section 5.2 illustrates the synthesis results and effect, with measurements of synthesis performances in 5.3. Sections 5.4 and 5.5 present the experiments and comparisons on expression classification and expression-invariant face recognition, respectively. Section 5.6 evaluates the generalisation ability of the proposed method in cross-database synthesis and verification experiments. Then, the experiments conducted for evaluating the effect of landmark quality are described in Sect. 5.7.

5.1 Datasets

In order to test the performance of the proposed method, a number of experiments were conducted on several publicly available datasets including the CK+, AR, Bosphorus, JAFFE, MUG and MMI facial expression databases.

The CK+ database [23,27] contains 593 video sequences from 123 subjects displaying various expressions. Each subject was instructed to perform a series of one to seven

Fig. 7 On CK+ database. Top row (from left to right): input neutral expression; and synthesised expression of varying intensities (trace, slight, marked, extreme and maximum). Bottom row: corresponding ground truth images. **a** Anger, **b** disgust, **c** fear, **d** happiness, **e** sadness and **f** surprise faces

Fig. 8 On AR database. Top row (from left to right): input neutral expression in session 1; and synthesised anger, smile and surprise for sessions 1 and 2, respectively. Bottom row: corresponding ground truth images

Fig. 9 On Bosphorus database. Top row (from left to right): input neutral expression; and synthesised anger, disgust, fear, happiness, sad and surprise. Bottom row: corresponding ground truth images

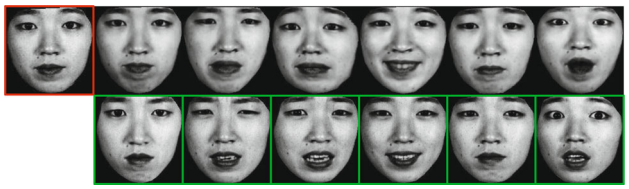


Fig. 10 On JAFFE database. Top row (from left to right): input neutral expression; and synthesised anger, disgust, fear, happiness, sad and surprise. Bottom row: corresponding ground truth images

expressions including happiness, surprise, disgust, fear, anger, contempt and sadness. Each video sequence starts with a neutral expression (first frames) and ends with the most expressive image (last frames). Image were digitised into either 640 × 480 or 640 × 490 pixel arrays.

The AR database [30] consists of over 4000 images from 126 individuals (70 men and 56 women) taken in two sessions (separated by two weeks). There are 26 images for each person with different facial expressions (neutral, anger, happiness and surprise), lighting conditions (right light on, left light on and all side lights on) and occlusions (wearing sunglasses or scarf). Each image in the database is of 768 × 576 pixels.

The Bosphorus database [42] contains 105 subjects (60 men and 45 women) in various poses, expressions and occlusion conditions. The majority of the subjects are Caucasian and aged between 25 and 35. The database consists of 4652

Fig. 11 On MUG database. First column: input neutral expressions. Top to bottom rows are synthesised anger, disgust, fear, happiness, sad and surprise faces of different intensities, with corresponding ground truth images shown in the last column

Fig. 12 On MMI database. First column: input neutral expressions. Top to bottom rows are synthesised anger, disgust, fear, happiness, sad and surprise faces of different intensities, with corresponding ground truth images shown in the last column

recordings of images with 31–54 samples per person. Images are under 13 different poses and cross-rotations, 34 expressions including 6 basic emotions, and partly occlusions by hands, glasses, etc.

The JAFFE [28] database is comprised of 213 facial expression images of 7 prototypic facial expressions posed by 10 Japanese female models. Each subject has three or four images for each expression and images are of the same size of 256×256.

The MUG [3] facial expression database consists of image sequences of 51 men and 35 women performing facial expressions. Among these 86 participants, images of 52 subjects are publicly available. All images from the videos have the same size of 896×896 recorded at 19 frames per second.

The MMI [37,53] facial expression database contains more than 2900 samples of video clips and static images collected from 75 subjects displaying various expressions. The video sequences are fully annotated for the presence of AUs. Among these sequences, 236 video clips of 31 subjects have been labelled with basic emotions. Each of the video sequences starts with the neutral expression and then goes to the apex state over the frames before returning to the offset. All images are of the same size of 720×576 recorded at 24 frames per second.

5.2 Results of expression synthesis and transfer

For each dataset, we adopted leave-one-subject-out strategy for expression synthesis due to limited number of subjects. For example, the synthesised smiling face of subject ξ was synthesised from a combination of smiling faces from all subjects except for ξ.

For the CK+ database, six images in the video sequences were used for each subject with the first frame being the neutral image and the last frame being the expression at the maximal level. Figure 7 illustrates the synthesised expressions at different levels using the proposed method and the corresponding ground truth images. The subplots of Fig. 7

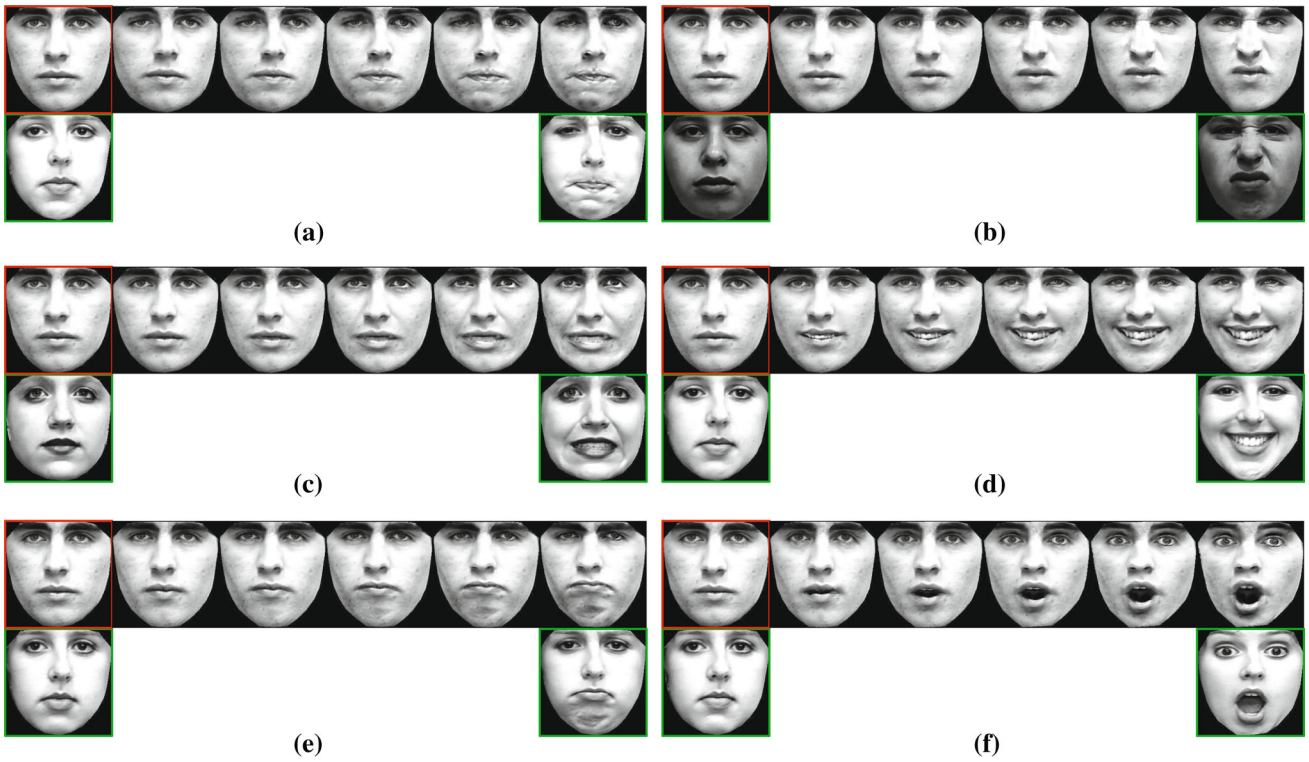


Fig. 13 Top rows: input neutral faces, followed by transferred expressions of varying intensities. Bottom rows: neutral and expressive images of the only training subject. **a** Anger, **b** disgust, **c** fear, **d** happiness, **e** sadness and **f** surprise faces

Table 2 Correlation coefficients, normalised distances between synthesised and ground truth texture, landmarks on CK+ database

Expressions	Expression verification [mean (SD)]	Expression intensity level				
		Trace	Slight	Marked	Extreme	Maximum
Anger	Correlation coefficients	0.990 (0.004)	0.984 (0.006)	0.978 (0.009)	0.970 (0.021)	0.968 (0.013)
	Normalised distance	1.218 (0.221)	1.554 (0.314)	1.946 (0.462)	2.309 (0.623)	2.480 (0.687)
Disgust	Correlation coefficients	0.989 (0.003)	0.984 (0.004)	0.978 (0.006)	0.974 (0.007)	0.970 (0.008)
	Normalised distances	1.460 (0.406)	1.877 (0.510)	2.085 (0.538)	2.309 (0.714)	2.661 (0.997)
Fear	Correlation coefficients	0.988 (0.004)	0.980 (0.007)	0.975 (0.010)	0.972 (0.009)	0.970 (0.009)
	Normalised distances	1.712 (0.461)	1.958 (0.498)	2.149 (0.504)	2.113 (0.410)	2.491 (0.673)
Happiness	Correlation coefficients	0.990 (0.005)	0.985 (0.005)	0.981 (0.007)	0.979 (0.008)	0.978 (0.008)
	Normalised distances	1.368 (0.322)	1.408 (0.331)	1.452 (0.406)	1.567 (0.418)	1.570 (0.330)
Sadness	Correlation coefficients	0.988 (0.006)	0.971 (0.018)	0.967 (0.022)	0.965 (0.015)	0.965 (0.013)
	Normalised distances	1.597 (0.423)	1.722 (0.480)	1.867 (0.480)	2.475 (0.809)	2.250 (0.705)
Surprise	Correlation coefficients	0.983 (0.009)	0.979 (0.014)	0.973 (0.019)	0.968 (0.020)	0.960 (0.026)
	Normalised distances	1.829 (0.466)	2.067 (0.512)	2.130 (0.574)	2.024 (0.502)	2.268 (0.552)

Table 3 Correlation coefficients, normalised distances between synthesised and ground truth texture, landmarks on AR database

Expression verification [mean (SD)]	Session 1			Session 2		
	Anger	Happiness	Surprise	Anger	Happiness	Surprise
Correlation coefficients	0.981 (0.016)	0.982 (0.012)	0.973 (0.017)	0.970 (0.023)	0.970 (0.022)	0.968 (0.019)
Normalised distances	1.639 (0.419)	1.500 (0.338)	2.784 (0.790)	1.628 (0.371)	1.469 (0.379)	2.734 (0.657)

Table 4 Correlation coefficients, normalised distances between synthesised and ground truth texture, landmarks on Bosphorus database

Expression verification [mean (SD)]	Expressions					
	Anger	Disgust	Fear	Happiness	Sad	Surprise
Correlation coefficients	0.974 (0.013)	0.964 (0.016)	0.972 (0.013)	0.966 (0.021)	0.977 (0.011)	0.976 (0.011)
Normalised distances	1.830 (0.753)	2.078 (0.613)	1.881 (0.478)	1.552 (0.506)	1.692 (0.599)	1.559 (0.485)

Table 5 Correlation coefficients, normalised distances between synthesised and ground truth texture, landmarks on JAFFE database

Expression verification [mean (SD)]	Expressions					
	Anger	Disgust	Fear	Happiness	Sad	Surprise
Correlation coefficients	0.980 (0.010)	0.967 (0.012)	0.975 (0.011)	0.977 (0.010)	0.982 (0.011)	0.974 (0.013)
Normalised distances	1.694 (0.663)	2.440 (0.795)	2.381 (0.473)	1.539 (0.389)	1.574 (0.441)	2.032 (0.737)

Table 6 Correlation coefficients, normalised distances between synthesised and ground truth texture, landmarks on MUG database

Expression verification [mean (SD)]	Expressions					
	Anger	Disgust	Fear	Happiness	Sad	Surprise
Correlation coefficients	0.983 (0.007)	0.973 (0.012)	0.980 (0.008)	0.975 (0.009)	0.979 (0.010)	0.975 (0.010)
Normalised distances	1.562 (0.495)	1.634 (0.503)	1.381 (0.348)	1.393 (0.296)	1.535 (0.483)	1.746 (0.429)

Table 7 Correlation coefficients, normalised distances between synthesised and ground truth texture, landmarks on MMI database

Expression Verification [mean (SD)]	Expressions					
	Anger	Disgust	Fear	Happiness	Sad	Surprise
Correlation Coefficients	0.989 (0.004)	0.982 (0.007)	0.981 (0.007)	0.984 (0.006)	0.979 (0.014)	0.983 (0.006)
Normalised Distances	1.780 (0.592)	1.204 (0.311)	1.520 (0.600)	1.518 (0.349)	1.349 (0.478)	1.797 (0.549)

Table 8 Average correlation coefficients and normalised distances on CK+, AR, Bosphorus, JAFFE, MUG and MMI datasets

Datasets	Expression verification [mean (SD)]	
	Correlation coefficients	Normalised distances
CK+	0.977 (0.011)	1.931 (0.572)
AR	0.974 (0.018)	1.959 (0.492)
Bosphorus	0.971 (0.014)	1.765 (0.572)
JAFFE	0.976 (0.011)	1.943 (0.583)
MUG	0.978 (0.009)	1.542 (0.426)
MMI	0.983 (0.007)	1.531 (0.480)
Average	0.977 (0.012)	1.782 (0.521)

correspond to (a) anger, (b) disgust, (c) fear, (d) happiness, (e) sadness and (f) surprise, respectively. The top-left image (in left most box) in each subplot represents the neutral expression and top rows (from left to right) are the synthesised expressions with increasing intensity (trace, slight, marked, extreme and maximum). The images in the bottom row represent their corresponding ground truth.

For the AR dataset, there are two sessions separated by two weeks. Taking the neutral image in session 1 as input, shown in Fig. 8 (in left most box), synthesised images are anger, happiness and surprise faces in sessions 1 and 2, respectively, shown in the top row of the figure. The corresponding ground truth images are displayed in the bottom row of Fig. 8.

In the Bosphorus database, we selected 62 subjects with each having seven images (one neutral + six basic expressions). Figure 9 displays synthesised six basic expressions: anger, disgust, fear, happiness, sadness and surprise, from left to right, and the ground truth images are at the bottom row.

In the JAFFE database, each subject has three or four images for each expression, and we took the mean value of these images for both shape and texture attributes to synthesise expression images. Figure 10 shows the synthesised six basic expressions against ground truth images.

Both of the MUG and MMI databases consist of video sequences from which multiple images can be collected to construct expression manifolds using the proposed EMS method. We took the first frame in each video as the neutral image and selected the expression images at its peak intensity.

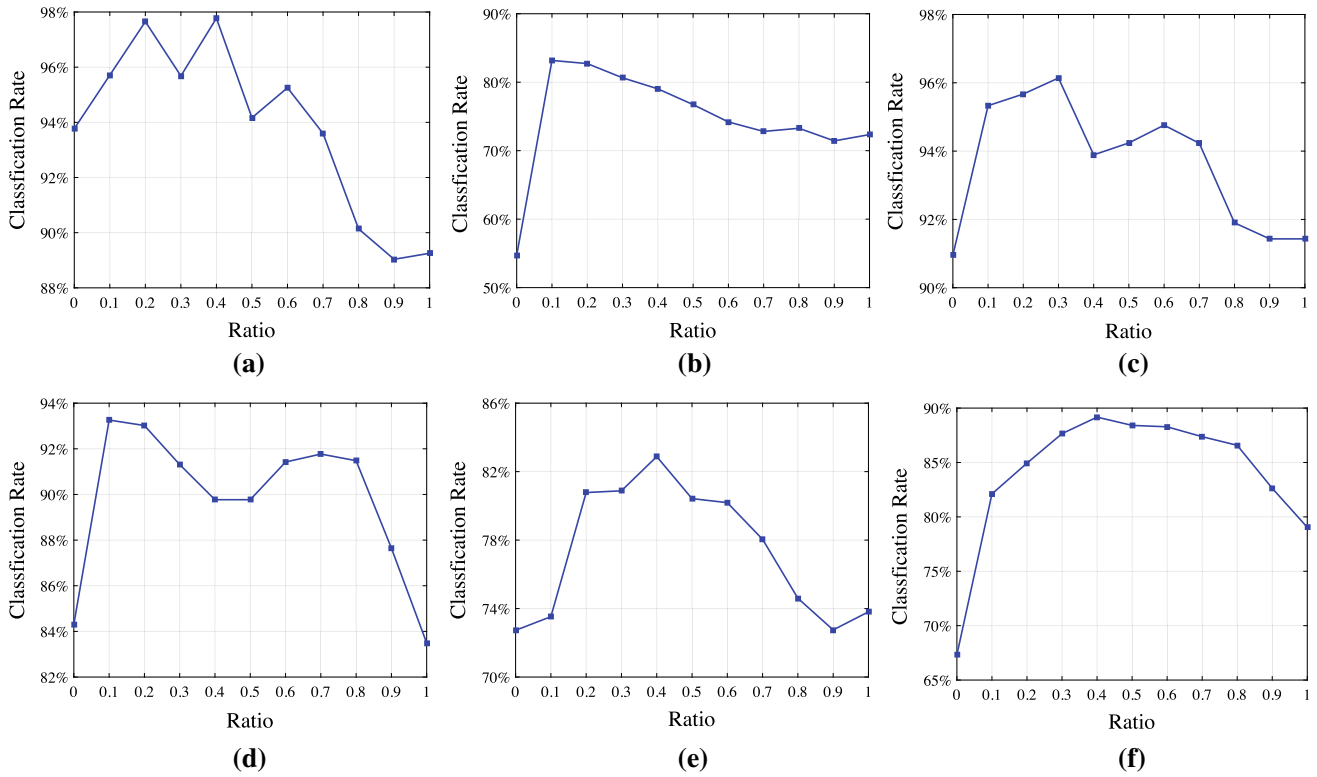


Fig. 14 Expression classification rate with different weights between shape and texture on **a** CK+, **b** Bosphorus, **c** JAFFE, **d** MUG, **e** MMI, and **f** combined databases

Table 9 Confusion matrix (%) on CK+ database with 10-fold cross-validation

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**99.37**	0	0.08	0	0	0.55	0
Anger	0	**100**	0	0	0	0	0
Disgust	3.54	0	**95.96**	0	0	0.50	0
Fear	4.35	0	0.72	**94.93**	0	0	0
Happiness	0	0	0	0	**100**	0	0
Sadness	2.78	0	0.93	0	0	**96.29**	0
Surprise	0.83	0	0	0	0	0	**99.17**

Bold numbers show classification accuracies for each expression, useful base ability for expression recognition

Figures 11 and 12 illustrate the input neutral images (in the first column), synthesised expression manifolds (samples) and the expression at peak intensity (in the last column). The synthesised expressions are anger, disgust, happiness, fear, sadness and surprise, from first to last rows.

When only one subject in a particular expression is available in training, we used expression transfer on the target subject. For example, in the CK+ database, some subjects have only one expression available. The results are illustrated in Fig. 13. Given the neutral and non-neutral expressions of the source subject (in the bottom rows) and the input neutral face of the target subject (in top-left boxes), non-neutral expressions were synthesised by transferring shape deformations and expression wrinkles from the source to the target subject.

5.3 Expression synthesis quality measure

To verify the closeness of the synthesised expression images to the ground truth, we calculated the correlation coefficients for texture attributes and normalised distances for shape attributes between the synthesised and ground truth images. To reduce the influence caused by different expression intensities, for each subject in the MUG and MMI databases, we computed the correlation coefficient, Eq. (21), and normalised distance, Eq. (20), between each synthesised image and selected ground truth image, and chose the ones with maximum correlation coefficient and minimum distance respectively. Results are presented in Tables 2, 3, 4, 5, 6 and 7. The average correlation coefficients and normalised distances from these tables are summarised in Table 8.

Table 10 Confusion matrix (%) on Bosphorus database with 10-fold cross-validation

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**85.95**	0.92	0	0	0	13.13	0
Anger	2.54	**83.64**	6.91	2.30	0	4.61	0
Disgust	0.23	8.52	**78.81**	1.61	5.07	5.53	0.23
Fear	8.06	0	0.92	**72.12**	0.24	5.53	13.13
Happiness	0	0	0	0	**99.77**	0.23	0
Sadness	12.67	4.15	7.38	0.69	0	**75.11**	0
Surprise	0.46	0	0	11.98	1.38	0	**86.18**

Bold numbers show classification accuracies for each expression, useful base ability for expression recognition

Table 11 Confusion matrix (%) on JAFFE database with person-independent cross-validation

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**90.00**	6.67	0	0	0	3.33	0
Anger	0	**93.33**	6.67	0	0	0	0
Disgust	0	0	**100**	0	0	0	0
Fear	0	0	0	**100**	0	0	0
Happiness	0	0	0	0	**100**	0	0
Sadness	0	0	0	0	3.33	**96.67**	0
Surprise	0	3.33	0	0	0	0	**96.67**

Bold numbers show classification accuracies for each expression, useful base ability for expression recognition

Table 12 Confusion matrix (%) on MUG database with 10-fold cross-validation

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**99.35**	0	0	0.65	0	0	0
Anger	1.63	**94.78**	1.96	1.31	0	0.32	0
Disgust	0	1.63	**98.04**	0	0.33	0	0
Fear	0.33	0	0	**92.16**	0	0	7.51
Happiness	1.63	0	0	0	**98.04**	0	0.33
Sadness	3.59	4.25	1.31	2.29	0	**88.56**	0
Surprise	2.61	0	0	15.04	0	0	**82.35**

Bold numbers show classification accuracies for each expression, useful base ability for expression recognition

It shows that with increase in expression intensity, the differences between the ground truth images and the synthesised expressions increase. Good agreements are found between the two sets with the averages of $\rho_{\text{avg}} = 0.977\ (0.012)$, $\text{dist}_{\text{avg}} = 1.782\ (0.521)$. The experiments on these various databases show that the proposed method can synthesise expressions well and preserve shape and most expressive details.

5.4 Expression classification results

We conducted cross-validation on the CK+, Bosphorus, JAFFE, MUG and MMI databases to evaluate the performance of the proposed method for expression classification. In these experiments, both synthesis and classification were conducted on the 10-fold cross-validation protocol. Only neutral and six prototypic expression images at its peak intensity were used as the ground truth training images. Expression classification was performed on the extended training set, containing both ground truth training images and synthesised expressive images.

In the 10-fold cross-validation, expressive images of one-fold of the subjects were synthesised based on that of the other ninefolds and were subsequently used to classify against all the ground truth expressive images. KDA classifier was used.

The ratio parameter w_f was optimised based on one-third of subjects from each database as the validation set and applied cross-validation on each validation set to optimise w_f and θ. Figure 14 reports the classification accuracies with different w_f on validation sets of the five databases and their combined, with optimal values as $(w_f)_{\text{CK+}} = 0.4$, $(w_f)_{\text{Bosphorus}} = 0.1$, $(w_f)_{\text{JAFFE}} = 0.3$, $(w_f)_{\text{MUG}} = 0.1$, $(w_f)_{\text{MMI}} = 0.4$ and $(w_f)_{\text{combined}} = 0.4$.

Tables 9, 10, 11, 12, 13 and 14 show the confusion matrices of seven emotions of these databases using Gaussian kernels with the overall average classification rates as 97.96% ($\theta =$

Table 13 Confusion matrix (%) on MMI database with 10-fold cross-validation

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**98.07**	1.16	0	0.77	0	0	0
Anger	12.50	**81.95**	0	0	5.55	0	0
Disgust	4.45	4.44	**87.22**	0	0	3.89	0
Fear	0	0	0	**92.43**	0	7.57	0
Happiness	9.65	0	0.88	5.26	**84.21**	0	0
Sadness	12.82	3.85	1.28	7.69	0	**74.36**	0
Surprise	21.88	0	0	7.29	0	3.12	**67.71**

Bold numbers show classification accuracies for each expression, useful base ability for expression recognition

Table 14 Confusion matrix (%) on the combined database with 10-fold cross-validation

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	**95.55**	1.13	0.23	0.64	0	2.45	0
Anger	3.56	**84.22**	5.33	0.78	0	6.11	0
Disgust	0.70	4.80	**90.10**	1.20	1.10	1.60	0.50
Fear	4.65	0.82	0.59	**83.88**	0	0.18	9.88
Happiness	0.48	0	0.48	0.67	**97.52**	0.57	0.28
Sadness	8.35	3.53	2.35	0.59	0.59	**84.59**	0
Surprise	0.40	0	0	11.20	1.00	0	**87.40**

Bold numbers show classification accuracies for each expression, useful base ability for expression recognition

Table 15 Expression classification rate (%) and comparisons with the state-of-the-art approaches

Methods	Features	Classes	Measures	CK+	Bosphorus	JAFFE	MUG	MMI
[46]	FAPs + SDPs	7	–	89.73	–	90.76	–	–
[59]	Patch-based Gabor	6	10-fold cross-validation	94.48	–	92.93	–	–
[19]	Radial encoded Gabor jets	7	10-fold cross-validation	91.51	–	–	–	–
			Leave-one-subject-out	–	–	89.67	–	–
[47]	D-SIFT	7	5-fold cross-validation	–	**86.20**	–	–	–
[40]	LFDA in the encrypted domain	6	Leave-one-out	–	–	94.37	93.35	–
[33]	Sparse representations	6	Leave-one-subject-out	94.50	–	–	–	72.73
[32]	PCA-based dictionary building	6	Leave-one-subject-out	97.19	72.41	–	–	78.51
[20]	Features from salient facial patches	6	10-fold cross-validation	94.09	–	91.80	–	–
[16]	Geometric features	6	10-fold cross-validation	97.80	–	–	**95.50**	77.22
[44]	Landmark-based pose-invariant descriptor	6	5-fold cross-validation	95.37	–	–	–	–
Proposed		7	10-fold cross-validation	**97.96**	83.08	–	93.32	**83.70**
			Leave-one-subject-out	–	–	**96.19**	–	–

Bold numbers indicate the best results of the various methods for all datasets

Table 16 Face recognition accuracy (%) on AR face database in sessions 1 and 2, respectively

Methods	Session1			Session2			Average
	Anger	Happiness	Surprise	Anger	Happiness	Surprise	
PCA	89	97	21	71	67	10	59.33
Proposed + PCA	98	99	50	99	96	47	81.50
Proposed + LDA	100	100	97	99	98	94	98.00

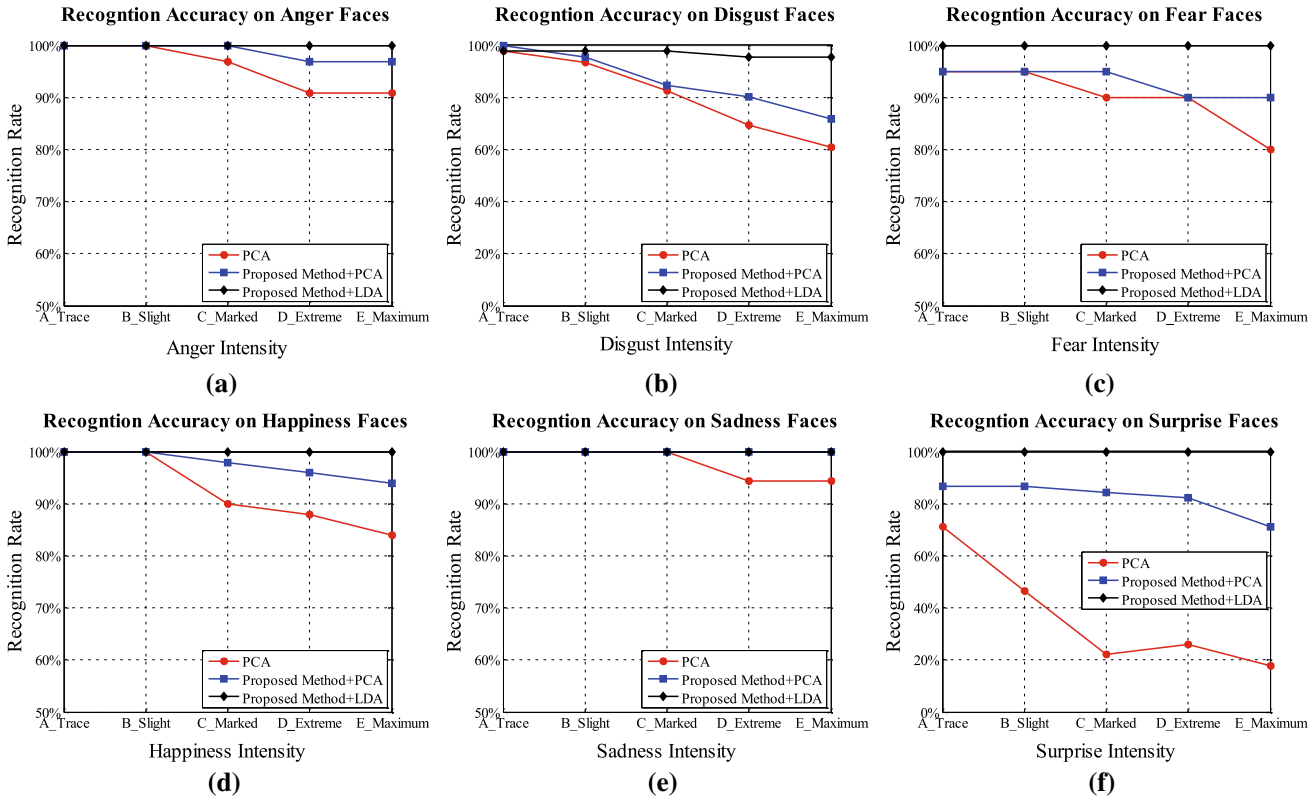


Fig. 15 Face recognition accuracy of **a** anger, **b** disgust, **c** fear, **d** happiness, **e** sadness and **f** surprise faces on CK+ database

Table 17 Face recognition accuracy (%) on Bosphorus database

Expressions	PCA	Proposed + PCA	Proposed + LDA
Anger	54.84	77.42	100
Disgust	59.68	69.35	98.39
Fear	56.45	77.42	100
Happiness	64.52	70.97	100
Sadness	77.42	87.10	100
Surprise	62.90	88.71	100
Average	62.64	78.50	99.73

Table 18 Face recognition accuracy (%) on JAFFE database

Expressions	PCA	Proposed + PCA	Proposed + LDA
Anger	80.00	70.00	96.67
Disgust	73.33	40.00	93.33
Fear	90.00	63.33	96.67
Happiness	90.00	80.00	100
Sadness	90.00	90.00	100
Surprise	56.67	56.67	96.67
Average	80.00	66.67	97.22

0.03), 83.08% ($\theta = 0.02$), 96.19% ($\theta = 0.12$), 93.32% ($\theta = 0.03$), 83.70% ($\theta = 0.08$) and 89.03% ($\theta = 0.06$), respectively.

Furthermore, the proposed method was compared to state-of-the-art approaches applied to these five databases with results shown in Table 15. The proposed method achieved considerably better results in most cases.

5.5 Face recognition results

With the gallery set containing only neutral images, synthesised expression images were added into the gallery set, and classifiers were built on the extended training set. In order to provide a comprehensive analysis, pairs of the expressions were tested separately as *neutral versus anger*, *neutral versus disgust*, *neutral versus fear*, *neutral versus happiness*, *neutral versus sadness* and *neutral versus surprise*. There was no overlap between the test faces and the training set.

As a baseline comparison, for the AR database, the extended training set contained original neutral images and synthesised anger, happiness and surprise images in sessions 1 and 2. The recognition performance is presented in Table 16. For the CK+, Bosphorus and JAFFE databases, compared to directly recognising the expressive faces using PCA, the proposed method with LDA classifier achieved better results as shown in Fig. 15, Tables 17 and 18. In these databases, the

Table 19 Face recognition accuracy (%) on MUG database

Expressions	PCA	EMS + PCA	EMS + LDA	EMS + KDA	EMS + MMD	EMS + MDA
Anger	92.16	89.28	99.24	98.18	97.73	100
Disgust	68.63	87.08	99.47	98.90	97.73	100
Fear	88.24	88.30	99.39	100	93.18	100
Happiness	92.16	90.11	99.73	99.89	100	100
Sadness	98.04	86.25	98.64	99.90	90.91	100
Surprise	50.98	79.20	98.90	99.81	88.64	100
Average	81.70	86.70	99.23	99.45	94.53	100

Table 20 Face recognition accuracy (%) on MMI database

Expressions	PCA	EMS + PCA	EMS + LDA	EMS + KDA	EMS + MMD	EMS + MDA
Anger	91.67	86.39	99.03	100	91.67	100
Disgust	93.33	92.56	100	99.44	100	100
Fear	90.91	94.24	98.48	100	100	100
Happiness	100	95.79	96.58	97.11	100	100
Sadness	92.31	100	100	100	100	100
Surprise	81.25	83.54	99.90	98.65	87.50	100
Average	91.58	92.09	99.00	99.20	96.53	100

Table 21 Comparisons of invariant face recognition (%) with the state-of-the-art approaches

Methods	Features	AR	CK+	Bosphorus	JAFFE	MUG	MMI
[31]	Eigenspace representation	84.67	92.62	–	–	–	–
[49]	Self-organising map projections	88	–	–	–	–	–
[10]	Pose correction and nose segmentation	–	–	91.65	–	–	–
[38]	Expression warping	97.3	–	–	–	**100**	–
[45]	MeshSIFT	–	–	97.7	–	–	–
[33]	Sparse representations	–	97.41	–	–	–	**100**
[15]	Deep supervised auto-encoders	85.21	–	–	–	–	–
[58]	Weighted LLS	**98**	98.12	–	96.91	–	–
Proposed		**98**	**99.28**	**99.73**	**97.22**	**100**	**100**

Bold numbers indicate the best results of the various methods for all datasets

dimensions of PCA and LDA subspaces were both set to the class number minus 1.

For the MUG and MMI databases, face recognition was performed by comparing synthesised expression manifolds and ground truth images. We conducted face recognition using two categories of methods: sample-based and set-based. For the sample-based methods such as PCA, LDA and KDA, we set the subspace dimensions to the class number minus 1 and used Gaussian kernels $\theta = 2$ in KDA. These three methods all determined the identity of probe set using the nearest neighbour classifier.

For the set-based methods such as MMD and MDA, PCA was first used to learn the linear subspaces of each image set with 90% information preserved. The canonical correlations were exploited to measure the set similarity in MMD, which the distance of linear spaces are measured by the correlation of the two exemplar samples. Tables 19 and 20 show the recognition results on the MUG and MMI databases. It is noted that both the sample-based and set-based methods achieved considerably good results on the two databases, with the proposed EMS method with MDA yielding perfect results.

The proposed method has also been compared with the state-of-the-art methods including MeshSIFT [45], sparse representations [33], deep auto-encoders [15] and weighted LLS [58]. Table 21 shows the overall face recognition results compared to the state-of-the-art methods on these six databases.

5.6 Cross-database generalisation results

To further test the generalisation ability of the proposed synthesis method, experiments were conducted across databases. That is, when trained on one database, how well is it able

to synthesise expressions on another database? The CK+, Bosphorus and MUG database were used.

The experimental scheme for the CK+ database was as follows:

- Bosphorus, MUG and combination of the two databases were used as the training set, respectively, and the CK+ dataset was regarded as the test set for synthesising expressions;
- Synthesised expressions were verified by calculating correlation coefficients and normalised distances between synthesised and ground truth expressions;
- Expression classification and face recognition were performed on the extended database with the same experimental set-up and parameters as in the previous sections.

For the Bosphorus or MUG database, experiments of similar cross-database scheme were conducted.

Examples of cross-synthesised expressions for the CK+, Bosphorus and MUG databases are displayed in Figs. 16, 17 and 18, respectively. Images of the first to the third rows show synthesised expressions from different training datasets, while the bottom row displays the corresponding ground truth expression images (anger, disgust, fear, happiness, sadness and surprise, from left to right). Table 22 lists the expression verification, classification and face recognition results of these cross-database generalisation tests.

The results show that the proposed method generalises well across datasets. Many of the expression classification and identity recognition accuracies were above 90%, despite that different illumination conditions exist across these datasets and subjects in these databases appeared to express the emotions not entirely in the same way. The proposed synthesis method still produces natural expressions and helps improve face recognition.

5.7 Effect of landmark quality

Shape synthesis replies on accurate extraction of facial landmarks. To evaluate the effect of landmark quality, we conducted experiments with various levels of inaccuracy added to landmarks on the Bosphorus database. For all subjects in the training and test set, random noise generated from normal distributions $\mathcal{N}(mu, sigma)$ was added to different groups of landmarks (face contour, nose, mouth, eyes and eyebrows, and all positions), which inevitably causes misalignments and shape deformations.

As illustrated in Sect. 3.1, faces were aligned to the eyes and distance between the two eyes in each face was fixed at 100. For each group of landmarks, four different levels of Gaussian noise were generated from four normal distributions with zero mean $mu = 0$ and standard deviation, $sigma = 0.5,\ 1,\ 2,\ 4$, respectively. Figure 19 shows exam-

Fig. 16 Synthesised expressions on CK+ database when trained using Bosphorus, MUG and the combined two databases from top to third rows, respectively. Images in bottom row represent corresponding ground truth expressions

Fig. 17 Synthesised expressions on Bosphorus database when trained using CK+, MUG and the combined two databases from top to third rows, respectively. Images in bottom row represent corresponding ground truth expressions

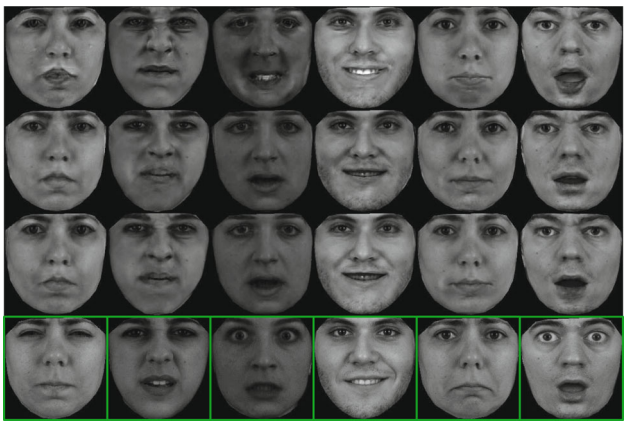


Fig. 18 Synthesised expressions on MUG database when trained using CK+, Bosphorus and the combined two databases from top to third rows, respectively. Images in bottom row represent corresponding ground truth expressions

Table 22 Results of generalisation experiments across databases

Train	Test	Correlation coefficients [mean (SD)]	Normalised distance [mean (SD)]	Expression classification (%) (Proposed + KDA)	Face recognition (%) (Proposed + LDA)
Bosphorus	CK+	0.950 (0.018)	2.446 (0.694)	94.94	92.07
MUG		0.950 (0.019)	2.467 (0.698)	94.19	94.28
Bosphorus, MUG		0.957 (0.018)	2.396 (0.682)	96.05	94.09
CK+	Bosphorus	0.915 (0.036)	2.510 (0.725)	79.38	91.67
MUG		0.950 (0.017)	2.401(0.716)	80.38	91.67
CK+, MUG		0.960 (0.016)	2.426 (0.723)	80.90	90.86
CK+	MUG	0.950 (0.023)	2.057 (0.517)	90.74	99.02
Bosphorus		0.960 (0.013)	2.046 (0.516)	90.06	96.73
Bosphorus, CK+		0.964 (0.013)	2.063 (0.504)	91.49	98.69

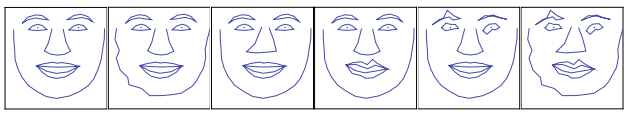


Fig. 19 Examples of landmark positions when random Gaussian noise generated from $\mathcal{N}(0, 4)$ was added on different group of landmarks. Images from left to right are the original landmarks, and landmark distributions when noise was added to face contour, nose, mouth, eyes and eyebrow, and all positions, respectively

ples of original landmarks, and disturbed landmarks with Gaussian noise $\mathcal{N}(0, 4)$ added to different groups. With these noisy landmarks, expressions were synthesised and verified, and expression classification and face recognition were then performed. Experimental set-up and parameters stayed the same as in the previous sections.

Figure 20b displays synthesised expressions with random Gaussian noise added on different group of landmarks (face contour, nose, mouth, eyes and eyebrows, and all, from the top to bottom rows). For each row, the standard deviation of the noise was varied from 0.5, 1, 2, to 4, from left to right. Figure 20a, c are the synthesised from the original landmarks and the ground truth, respectively. Table 23 lists the expression verification, classification and face recognition performances affected by these erroneous landmarks.

The results indicate that the proposed expression synthesis method can tolerate certain noise and misalignment errors. Small errors are automatically taken care of by the synthesis process, thereby generating photorealistic expressions. However, larger errors in landmarks will deteriorate recognition performance and synthesis quality.

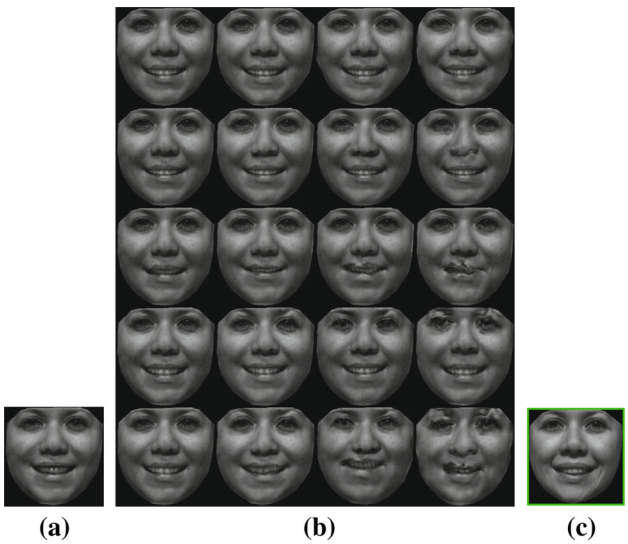


Fig. 20 Synthesised smile examples on Bosphorus database, **a** synthesised with original landmark, **b** synthesised with random Gaussian noise added on the landmarks located on face contour, nose, mouth, eyes and eyebrows and all positions from top to bottom rows; and in each row, the standard deviation of the noise was 0.5, 1, 2 and 4 from left to right, and **c** ground truth

6 Conclusions

In this paper, we have proposed an approach to expression classification and expression-invariant face recognition by synthesising photorealistic expression images to expand the gallery set. Eigentransformation has been extended to expression manifolds for synthesising shapes and expression appearance details with varying intensity or dynamics; this is important to precisely interpreting facial expression. The synthesised images along with the gallery set are used to train classifiers such as LDA and KDA, or to calculate manifold-based similarities in expression classification and expression-invariant face recognition.

The proposed method yields realistic synthesised expressions. The experimental results have shown marked improvements in performance for expression classification and invariant face recognition over those without synthesised

Table 23 Experimental results of landmark quality effect

Landmark groups	Gaussian noise (*mu*, *sigma*)	Correlation coefficients [mean (SD)]	Normalised distance [mean (SD)]	Expression classification (%) (Proposed + KDA)	Face recognition (%) (Proposed + LDA)
Contour	(0, 0.5)	0.970 (0.015)	2.056 (0.685)	81.08	97.27
	(0, 1)	0.970 (0.014)	2.076 (0.684)	79.72	97.04
	(0, 2)	0.968 (0.014)	2.187 (0.683)	79.03	96.77
	(0, 4)	0.961 (0.018)	2.489 (0.646)	78.57	95.16
Nose	(0, 0.5)	0.971 (0.014)	2.070 (0.694)	79.43	96.63
	(0, 1)	0.971 (0.014)	2.097 (0.683)	78.80	95.70
	(0, 2)	0.969 (0.015)	2.183 (0.682)	77.88	95.16
	(0, 4)	0.964 (0.016)	2.429 (0.660)	76.96	94.89
Mouth	(0, 0.5)	0.971 (0.014)	2.074 (0.699)	80.65	96.24
	(0, 1)	0.971 (0.014)	2.120 (0.724)	79.49	96.50
	(0, 2)	0.968 (0.014)	2.221 (0.714)	79.15	95.43
	(0, 4)	0.963 (0.016)	2.525 (0.676)	77.19	94.62
Eyes and eyebrows	(0, 0.5)	0.971 (0.014)	2.110 (0.674)	80.88	97.31
	(0, 1)	0.970 (0.014)	2.228 (0.662)	79.73	96.51
	(0, 2)	0.965 (0.015)	2.607 (0.657)	79.49	95.97
	(0, 4)	0.956 (0.016)	3.483 (0.735)	73.27	94.89
All	(0, 0.5)	0.970 (0.015)	2.118 (0.674)	79.95	96.56
	(0, 1)	0.968 (0.015)	2.311 (0.682)	79.37	96.24
	(0, 2)	0.958 (0.016)	2.785 (0.657)	76.04	94.09
	(0, 4)	0.932 (0.022)	4.008 (0.711)	72.12	90.32

images, especially when expressions are at extreme levels. The proposed method achieved consistently results with different expressions and different expression levels. The proposed method has also been shown to generalise well across databases and can tolerate certain misalignment errors and noise reflected in facial landmarks. Extensive comparisons with various recent methods have demonstrated the advantages of the proposed method. The future work will focus on development of real-time expression and face recognition methods for video sequences.

References

1. Abboud, B., Davoine, F.: Bilinear factorisation for facial expression analysis and synthesis. IEEE Proc. Vis. Image Signal Process. **152**(3), 327–333 (2005)
2. Abboud, B., Davoine, F., Dang, M.: Facial expression recognition and synthesis based on an appearance model. Signal Process. Image Commun. **19**(8), 723–740 (2004)
3. Aifanti, N., Papachristou, C., Delopoulos, A.: The MUG facial expression database. In: Proceedings of IEEE International Conference on Image Analysis for Multimedia Interactive Services Workshop, pp. 1–4 (2010)
4. Ambadar, Z., Schooler, J.W., Cohn, J.F.: Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. Psychol. Sci. **16**(5), 403–410 (2005)
5. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Comput. **12**(10), 2385–2404 (2000)
6. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces versus fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
7. Chang, Y., Hu, C., Turk, M.: Manifold of facial expression. In: Proceedings of IEEE International Conference on Analysis and Modeling of Faces and Gestures Workshops, pp. 28–35 (2003)
8. Chapman, R.E.: Conventional procrustes approaches. In: Proceedings of Michigan Morphometrics Workshop, pp. 251–267 (1990)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
10. Dibeklioğlu, H., Gökberk, B., Akarun, L.: Nasal region-based 3d face recognition under pose and expression variations. In: Proceedings of International Conference on Advances in Biometrics, pp. 309–318 (2009)
11. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis, vol. 4. Wiley, Chichester (1998)
12. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
13. Ekman, P., Rolls, E., Perrett, D., Ellis, H.: Facial expressions of emotion: an old controversy and new findings. Philos. T. Roy. Soc. B **335**(1273), 63–69 (1992)
14. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**(1), 259–275 (2003)
15. Gao, S., Zhang, Y., Jia, K., Lu, J., Zhang, Y.: Single sample face recognition via learning deep supervised auto-encoders. IEEE Trans. Inf. Forensic Sec. **10**(10), 2108–2118 (2015)

16. Ghimire, D., Lee, J., Li, Z.N., Jeong, S.: Recognition of facial expressions based on salient geometric features and support vector machines. Multimedia Tools Appl. **76**(6), 7921–7946
17. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. Numer. Math. **14**(5), 403–420 (1970)
18. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975)
19. Gu, W., Xiang, C., Venkatesh, Y.V., Huang, D., Lin, H.: Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern Recognit. **45**(1), 80–91 (2012)
20. Happy, S.L., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. **6**(1), 1–12 (2015)
21. Huang, D., De la Torre, F.: Bilinear kernel reduced rank regression for facial expression synthesis. In: Proceedings of European Conference on Computer Vision, pp. 364–377 (2010)
22. Jain, A.K., Li, S.Z.: Handbook of Face Recognition, vol. 1. Springer, Berlin (2005)
23. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
24. Lee, H., Kim, D.: Expression-invariant face recognition by facial expression transformations. Pattern Recogn. Lett. **29**(13), 1797–1805 (2008)
25. Liu, Z., Shan, Y., Zhang, Z.: Expressive expression mapping with ratio images. In: Proceedings of Conference on Computer Graphics and Interactive Techniques, pp. 271–276 (2001)
26. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of IEEE International Conference on Computer Vision, vol. **2**, pp. 1150–1157 (1999)
27. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshop, pp. 94–101 (2010)
28. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
29. Ma, W.C., Jones, A., Chiang, J.Y., Hawkins, T., Frederiksen, S., Peers, P., Vukovic, M., Ouhyoung, M., Debevec, P.: Facial performance synthesis using deformation-driven polynomial displacement maps. ACM Trans. Graph. **27**(5), 121:1–121:10 (2008)
30. Martinez, A., Benavente, R.: The AR face database. Technical Report, CVC Technical Report (1998)
31. Martinez, A.M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Anal. Mach. Intell. **24**(6), 748–763 (2002)
32. Mohammadi, M.R., Fatemizadeh, E., Mahoor, M.H.: PCA-based dictionary building for accurate facial expression recognition via sparse representation. J. Vis. Commun. Image R. **25**(5), 1082–1092 (2014)
33. Mohammadi, M.R., Fatemizadeh, E., Mahoor, M.H.: Simultaneous recognition of facial expression and identity via sparse representation. In: Proceedings of IEEE International Winter Conference on Applications of Computer Vision, pp. 1066–1073 (2014)
34. Mohammadzade, H., Hatzinakos, D.: Projection into expression subspaces for face recognition from single sample per person. IEEE Trans. Affect. Comput. **4**(1), 69–82 (2013)
35. Pantic, M., Patras, I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Trans. Syst. Man Cybern. B Cybern. **36**(2), 433–449 (2006)
36. Pantic, M., Rothkrantz, L.J.: Facial action recognition for facial expression analysis from static face images. IEEE Trans. Syst. Man Cybern. B Cybern. **34**(3), 1449–1461 (2004)
37. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp. 317–321 (2005)
38. Petpairote, C., Madarasmi, S.: Face recognition improvement by converting expression faces to neutral faces. In: Proceedings of International Symposium on Communications and Information Technologies, pp. 439–444 (2013)
39. Pyun, H., Kim, Y., Chae, W., Kang, H.W., Shin, S.Y.: An example-based approach for facial expression cloning. In: Proceedings of EG/SIGGRAPH Symposium on Computer Animation, pp. 167–176 (2003)
40. Rahulamathavan, Y., Phan, R.C.W., Chambers, J.A., Parish, D.J.: Facial expression recognition in the encrypted domain based on local fisher discriminant analysis. IEEE Trans. Affect. Comput. **4**(1), 83–92 (2013)
41. Sandbach, G., Zafeiriou, S., Pantic, M., Rueckert, D.: A dynamic approach to the recognition of 3d facial expressions and their temporal models. In: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition Workshops, pp. 406–413 (2011)
42. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: Proceedings of European Workshop on Biometrics and Identity Management, pp. 47–56 (2008)
43. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
44. Shojaeilangari, S., Yau, W.Y., Teoh, E.K.: Pose-invariant descriptor for facial emotion recognition. Mach. Vis. Appl. **27**(7), 1063–1070 (2016)
45. Smeets, D., Keustermans, J., Vandermeulen, D., Suetens, P.: meshSIFT: local surface features for 3D face recognition under expression variations and partial data. Comput. Vis. Image Und. **117**(2), 158–169 (2013)
46. Song, M., Tao, D., Liu, Z., Li, X., Zhou, M.: Image ratio features for facial expression recognition application. IEEE Trans. Syst. Man Cybern. B Cybern. **40**(3), 779–788 (2010)
47. Soyel, H., Demirel, H.: Localized discriminative scale invariant feature transform based facial expression recognition. Comput. Electr. Eng. **38**(5), 1299–1309 (2012)
48. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
49. Tan, X., Chen, S., Zhou, Z., Zhang, F.: Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble. IEEE Trans. Neural Netw. **16**(4), 875–886 (2005)
50. Tang, X., Wang, X.: Face photo recognition using sketch. In: Proceedings of IEEE International Conference on Image Processing, vol. **1**, pp. 257–260 (2002)
51. Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proceedings of IEEE International Conference on Computer Vision, pp. 687–694 (2003)
52. Tian, Y., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97–115 (2001)
53. Valstar, M., Pantic, M.: Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In: Proceedings of International Conference on Language Resources and Evaluation Workshop, pp. 65–70. Malta (2010)

54. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Toward a practical face recognition system: robust alignment and illumination by sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(2), 372–386 (2012)
55. Wang, H., Ahuja, N.: Facial expression decomposition. In: Proceedings of IEEE International Conference on Computer Vision, pp. 958–965 (2003)
56. Wang, R., Chen, X.: Manifold discriminant analysis. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 429–436 (2009)
57. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold–manifold distance with application to face recognition based on image set. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
58. Zaman, F.K., Shafie, A.A., Mustafah, Y.M.: Robust face recognition against expressions and partial occlusions. Int. J. Autom. Comput. **13**(4), 319–337 (2016)
59. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. IEEE Trans. Affect. Comput. **2**(4), 219–229 (2011)
60. Zhang, Q., Liu, Z., Quo, G., Terzopoulos, D., Shum, H.Y.: Geometry-driven photorealistic facial expression synthesis. IEEE Trans. Vis. Comput. Gr. **12**(1), 48–60 (2006)
61. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. **35**(4), 399–458 (2003)
62. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2562–2569 (2012)

Yao Peng received the B.Eng. degree in Detection Guidance and Control Techniques from Nanjing University of Science and Technology, China, in 2013, and the M.Sc. in Digital Image and Signal Processing from the University of Manchester, UK, in 2014. She is currently enrolled as a Ph.D. student in the School of Electrical and Electronic Engineering, the University of Manchester. Her current research interests include computer vision, face recognition, neural networks and deep learning.

Hujun Yin received the BEng degree in Electronic engineering and the MSc degree in Signal Processing from Southeast University and the PhD degree in Neural Networks from University of York, respectively. He has been with The University of Manchester, School of Electrical and Electronic Engineering, since 1998. His main research interests include neural networks, self-organizing systems, deep learning, image processing, face recognition, and bio-/neuro-informatics. He has studied and extended the self-organizing neural networks and related topics, such as manifolds, dimensionality reduction and data visualization, and proposed a number of methods for associated data analysis and modelling. He has published over 200 peer-reviewed articles in a range of topics, from neural networks, density modelling, image processing, face recognition, text mining and knowledge management, gene expression analysis and peptide sequencing, novelty detection, to financial time series modelling, and recently decoding neuronal responses. He is a Senior Member of the IEEE and a member of the UK EPSRC College. He is an Associate Editor for the IEEE Transactions on Cybernetics Networks and a member of the Editorial Board of the International Journal of Neural Systems. He had also served as an Associate Editor for the IEEE Transactions on Neural Networks between 2006 and 2010. He has been the Organising Chair, Programme Committee Chair, and General Chair for several international conferences, such as, International Workshop on Self-Organizing Maps (WSOM'01), International Conference on Intelligent Data Engineering and Automated Learning (IDEAL) since 2002, International Symposium on Neural Networks (ISNN'06).