



Cascade Regression-Based Face Frontalization for Dynamic Facial Expression Analysis

Yiming Wang¹ · Xinghui Dong² · Gongfa Li³ · Junyu Dong² · Hui Yu¹

Received: 18 September 2020 / Accepted: 20 January 2021
© The Author(s) 2021

Abstract

Facial expression recognition has seen rapid development in recent years due to its wide range of applications such as human–computer interaction, health care, and social robots. Although significant progress has been made in this field, it is still challenging to recognize facial expressions with occlusions and large head-poses. To address these issues, this paper presents a cascade regression-based face frontalization (CRFF) method, which aims to immediately reconstruct a clean, frontal and expression-aware face given an in-the-wild facial image. In the first stage, a frontal facial shape is predicted by developing a cascade regression model to learn the pairwise spatial relation between non-frontal face-shape and its frontal counterpart. Unlike most existing shape prediction methods that used single-step regression, the cascade model is a multi-step regressor that gradually aligns non-frontal shape to its frontal view. We employ several different regressors and make an ensemble decision to boost prediction performance. For facial texture reconstruction, active appearance model instantiation is employed to warp the input face to the predicted frontal shape and generate a clean face. To remove occlusions, we train this generative model on manually selected clean-face sets, which ensures generating a clean face as output regardless of whether the input face involves occlusions or not. Unlike the existing face reconstruction methods that are computationally expensive, the proposed method works in real time, so it is suitable for dynamic analysis of facial expression. The experimental validation shows that the ensembling cascade model has improved frontal shape prediction accuracy for an average of 5% and the proposed method has achieved superior performance on both static and dynamic recognition of facial expressions over the state-of-the-art approaches. The experimental results demonstrate that the proposed method has achieved expression-preserving frontalization, de-occlusion and has improved performance of facial expression recognition.

Keywords Face frontalization · Facial expression recognition · Cascade regression · Facial analysis · Person-independent

Introduction

Facial expression recognition (FER) has a wide range of applications including human–computer interaction (HCI) [38, 6], animation [1, 36, 21] and security [24]. Healthcare is one of the most important applications of FER. FER-based research has been conducted on traditional mental

care of Autism Spectrum Disorder (ASD) and Parkinson patients who has difficulty in recognizing and performing affective facial expressions. Recently, researchers in this field has started to care more about the healthcare cases in recognizing the state of pain, fatigue/confusion, boredom/stressful, sleepiness, inattention and facial palsy [19].

Considerable FER performance has been achieved by using deep learning technologies [32, 48]. Currently, most existing FER approaches focus on static facial images processing [10]. However, facial expressions are inherently dynamic actions which can be better described as several sequential pieces of facial motions in a time interval. Although static FER methods have achieved impressive results, they completely ignored discriminative features conveyed by subtle facial muscle movements. Dynamic FER on the whole image sequence is more natural and reasonable.

The aim of dynamic FER is to predict facial expression categories from an image sequence. A common sequence

✉ Hui Yu
hui.yu@port.ac.uk

¹ School of Creative Technologies, University of Portsmouth, Portsmouth, England, UK

² Ocean University of China, Qingdao, China

³ Key Laboratory of Metallurgical Equipment and Control Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan, China
Email: ligongfa@wust.edu.cn

of the expression evolutionary process often evolves initial neutral state, onset, apex phase, offset and final neutral state. The existing dynamic facial feature descriptors can be classified into two categories: low-level feature representation and high-level motion-aligned representation. The best known low-level feature representations are LBP-TOP [47] and LPQ-TOP [15] which capture the local gradient features over both spatial neighborhoods in one frame and temporal neighborhoods between adjacent frames. Recent research on deep learning FER methods [13, 33, 46] often combines Convolutional Neural Network (CNN) with Long-Short Term Memory (LSTM) to capture both spatial and temporal facial features. Although deep learning methods could extract features from very deep levels, they still follow the end-to-end protocol without specifically considering time-alignment. High-level semantic methods aim to derive a meaningful facial motion representation in which temporal alignment is often presented to different sequences with different time intervals into a uniform temporal space (five states of the expression evolutionary process) [9]. As both categories address the problem of temporal modelling, they have been proved to be effective on the posed facial expression data sets but poor on unconstrained image sequences where there are large head-pose changes and occlusions. Currently, there is no attempt that focuses on head-pose normalization for dynamic facial expression recognition.

Intuitively, the problem of head-pose changes and occlusions from unconstrained facial images can be well normalized through face frontalization [8, 49]. The main objective of generic face frontalization is to automatically recover the non-frontal face to its frontal view. Traditional face frontalization methods often include two key components: frontal face-shape estimation and frontal face-texture fitting. Frontal shape estimation localizes facial key point positions and aligns them to the frontal positions. The objective of frontal shape estimation is to align the non-frontal facial landmarks to their frontal positions. The task of texture fitting is to fit textures to the predicted shape by texture warping and rectification. It has been reported that frontal shape estimation is quite challenging. The mainstream approaches focus on hard frontalization that an unmodified frontal shape template (often made in neutral) is used as the reference and facial textures of all the query faces will be fitted to the template [50]. By this strategy, the reconstructed faces usually lose facial expression-related information. Recent face frontalization methods are mainly based on deep generative models, especially generative adversarial network (GAN). Most GAN-based methods [37, 45] use a structure of an encoder-decoder network connecting to a CNN-like classification network to simultaneously generate a face and identify whether it is in frontal view or not. Obviously, there is no need for these kinds of generative methods to consider shape alignment, which may also lose subtle expression cues in the generated faces.

There are only a few approaches so far that are able to achieve expression preserving face frontalization.

In this paper, a novel cascade regression-based face frontalization (CRFF) method is proposed for dynamic analysis of facial expressions. The key issue is how to predict the frontal position of key facial points given a detected non-frontal shape. Inspired by the idea of Supervised Decent Method (SDM) [20, 42] and residual neural network [11], we propose a novel cascade regression model for 2D frontal shape estimation. To begin with, a set of facial images in different viewpoints is collected and each non-frontal face is associated with its frontal counterpart. Several regression approaches are chosen to learn the pairwise relations between the non-frontal shape and frontal shape. It is obvious that pairwise changes in head-pose, expressions and individual differences are nonlinearly coupled in 2D shapes. Thus, one step regression is not able to well model this relation. As is shown in Fig. 1, we propose an adaptive cascade regression model in which the input is designed to be mapped to the difference between target and input instead of directly being mapped to the target, so that the non-frontal shape can be gradually approximated to its frontal shape. In the training stage, the cascade regression model aims to minimize the difference between the predicted facial shape and groundtruth position. During testing, the cascade regression model will be used as a frontal face-shape predictor that estimates the key point positions in frontal view. The obtained frontal shape will be viewed as the based mesh which is used to feed the texture fitting in the next step. For texture-fitting, we employ active appearance model (AAM) instantiation [3, 22] to reconstruct the facial appearances. To remove occlusions, AAM model is trained on manually selected clean-face sets, which ensures that the AAM model only learns how to generate a clean face. Therefore, the output of the AAM model is expected to be always clean faces regardless of whether the input face involves occlusions or not. The reconstructed face will be: (a) in frontal view, (b) remaining deformations of detailed facial expressions, (c) no occlusions. The contributions of the paper can be summarized as follows:

- 1) We propose a cascade regression-based face frontalization (CRFF) approach for head-pose normalization. Different from computationally expensive 3D solutions which suffer from the one-minute-per-frame problem, this method is based on 2D face reconstruction and works in real time.

- 2) Different from the existing 2D face frontalization methods that usually ignore facial deformation containing expression information, the proposed frontalization method is expression-aware. The vivid expression changes are preserved and the occlusions are effectively removed.

- 3) The experiment shows that dynamic FER accuracy can be improved by using our expression-preserving face frontalization method.

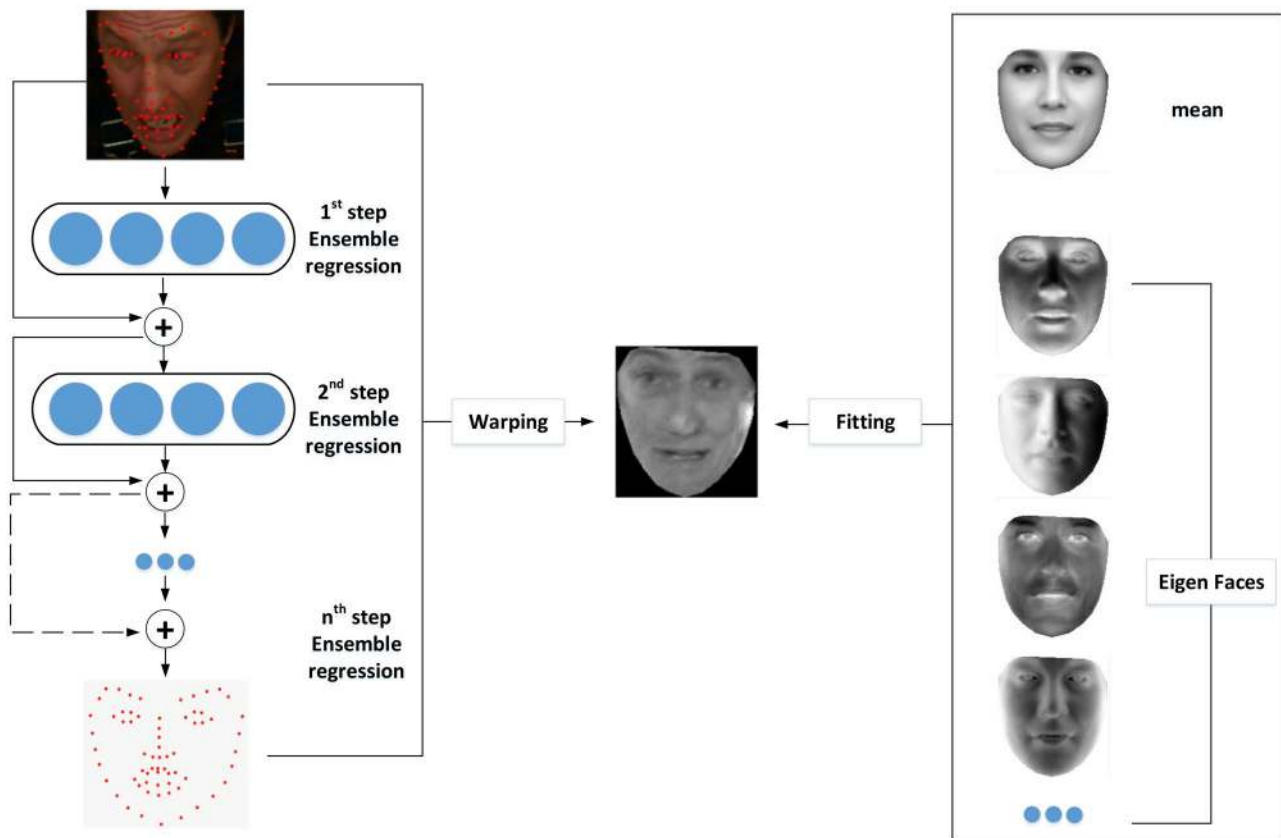


Fig. 1 Diagram of proposed face frontalization

The rest of this paper is organized as follows. Section 2 surveys the state-of-the-art FER and face frontalization methods. Section 3 presents the detailed implementation of the proposed method. We conduct qualitative experiment and quantitative visualization in Section 4. Finally, we summarize this paper with a discussion of contribution and future work in Section 5.

Related Works

Facial Expression Recognition

The main focus of this paper is on traditional mental care based on recognition of facial expressions of emotions. Dynamic FER models capture spatiotemporal features which represent a range of frames within a time interval. As is mentioned above, the existing methods can be divided into low-level feature-based and motion-based approaches.

Low-level spatiotemporal representations can be seen as an extension of low-level spatial representations. Shape features are described by tracked fiducial facial points. The location of each point, as well as the length and angle of pairwise points connection, forms the basic shape features.

Till now, shape representations are less common because it has been reported and well validated that appearance models outperform shape models [30]. Appearance representation are the mainstream for dynamic FER. LBP-TOP is a popular method that extracts Local Binary Pattern (LBP) features from Three Orthogonal Planes (TOP) [47]. Many existing works on dynamic FER are based on LBP-TOP [23, 40]. The original LBP features extracted from a single spatial plane are extended to two more spatiotemporal planes, which enable extracting gradient features between frames. LPQ-TOP follows the same principle and is used in Action Unit (AU) recognition [15].

Obviously, low-level features do not consider the specific knowledge in the facial expression domain. Recent research focuses more on capturing high-level semantic features represented by facial motions. It is commonly accepted that the facial expression process of human beings includes five states: initial neutral, onset, apex, offset and final neutral. This standard process can be seen as a template and motion-based methods are actually a time alignment strategy that normalizes the input sequence to the five reference states. Koelstra et al [47] used free-form deformations (FFD) [28] based non-rigid registration to capture motions for AU recognition. Guo et al [9] used diffeomorphic transformation

for time alignment and proposed atlas construction to capture facial appearance movements. Wang et al. [41] assumed each local facial points movement as a local facial event and learned the motion dependency by modelling temporally overlapping facial events and their temporal relations by an interval temporal Bayesian network. In [17], a universal manifold model (UMM) is learned to statistically unify each input video (modelled as spatiotemporal manifold via low-level features) to the standard expression evolutionary process.

All the methods mentioned above only address the problem of temporal state alignment, but ignore spatial texture alignment. If the subjects of the video clips move their head frequently, the appearance changes caused by head-pose will be much larger than subtle expressions changes. This is why most dynamic FER methods perform well on the posed expression data sets but poor on the unconstrained image sequences. A well designed dynamic descriptor should consider both time alignment and facial appearance normalization. The problem of facial appearance normalization for suitable expression registration is quite challenging that has not been addressed yet.

Face Frontalization

Recently, face frontalization has attracted wide attentions due to its effectiveness in facial analysis [29, 39]. It is commonly accepted that more robust features can be captured from the frontal face rather than the profile face. Thus, the main objective of face frontalization is to recover the frontal faces from non-frontal viewpoints. This is a comprehensive research topic which is often associated with face alignment, face deformation and texture rectification.

If a facial image is captured from the non-frontal view, one half side of the face contains abundant facial texture while the other side is occluded. Direct interpolation will cause large distortions on the reconstructed images. Therefore, the main problem of face frontalization is how to fill the invisible part. Traditional methods for face frontalization includes 2D-based models and 3D-based models.

In [14] and [26], two approaches of person-specific 3D model reconstruction were proposed, in which several images captured from one person in different poses and expressions are used to reconstruct the 3D face. The main drawback of these methods is they are unable to reconstruct 3D surface of novel faces. In order to deal with this problem, many methods were proposed based on 3D Morphable Model (3DMM), which is, theoretically, capable of reconstructing a full 3D facial surface from a single input image [8, 49, 50]. Although 3D-based methods can implement frontalization, they are not practical since a) it is usually computationally expensive to build 3D models, b) a massive training data is

required to learn shape models, and c) it is very challenging to reconstruct the 3D models of novel subjects.

In [43] and [31], two effective 2D frontalization approaches were presented and both of them belong to hard frontalization which employ a single 2D/3D reference template as base shape. The query images are then used to fit the facial textures to the reference template. Soft symmetry [43] fills the invisible regions by the corresponding symmetric visible parts of face. Apparently, this approach is sensitive to occlusions and it only enables tilt head rotation recovery but fails in recovering the faces in pan angles. The texture-fitting strategy of Robust Statistical face Frontalization (RSF) [29] is based on active appearance model (AAM) instantiation [22] which reconstructs the appearances by combining a group of eigen faces. It is robust to occlusions and is capable of recovering the faces in whatever pan or tilt angles. Therefore, RSF is more stable than soft symmetry.

Recently, deep learning methods are also used in face frontalization. Yim et al. [44] used multi-task deep neural network (DNN) to generate a facial image of any query head-pose from a single input image. Multi-task DNN included a main DNN that generated a face with desired head-pose and an auxiliary DNN for the secondary task of identity maintenance. The output of this model is an identity-preserving face of the desired head-pose. Tran et al. [37] proposed a GAN-based framework that used an encoder-decoder structured generator and a multi-task CNN as discriminator for facial identity classification task and pose classification task. The output of the generator is the synthesized identity-preserving face of the desired pose. Li et al. [16] presented in-painting and frontalization GAN (IF-GAN) which was the only method that was specifically designed to remove occlusions. However, this method could only work on artificial occlusions, in which the occlusion mask must be given in advance as prior knowledge. Therefore, IF-GAN cannot be used in real-world conditions and occlusion remains a big challenge in this field.

These methods have been proved to be effective in face recognition tasks. However, the reconstructed faces are expected to approximate real frontal faces while the identity and expressions are more or less removed. There is only one published work on facial expression-aware face frontalization (FEAF) [39]. In that approach, multiple emotional shape templates are designed instead of a single shape template and it achieves good results in static FER. Inherently, it is still a kind of hard frontalization and is not suitable for dynamic FER because all the frames of the image sequence are arbitrarily normalized to the templates so that the dynamic information of subtle expression and shape changes will be lost. Dynamic FER requires a novel face frontalization method that is able to not only recover

appearances to frontal view, but also distinguish subtle changes in terms of facial shapes and appearances.

Methodology

We propose a new face frontalization method which synthesizes clean, frontal and expression-preserving faces. It mainly includes two stages: cascade regression-based frontal face-shape estimation and AAM-based frontal face-texture fitting. In the first stage, the non-frontal facial shape features (positions of landmark points) are detected by SDM, and then sequentially fed to a four-step cascade regression model which gradually maps the non-frontal shape into its frontal view. We have adopted four different regression methods as the base regressor. They work independently and their outputs are combined to make a comprehensive decision via an ensemble learning framework. The whole system in the first stage receives a facial image with an arbitrary head-pose and outputs its predicted frontal shape. In the second stage, the facial textures are roughly warped into this frontal shape. The distortion caused by warping, as well as occlusions, is rectified by fitting a pre-trained AAM model which linearly combines a group of clean eigen faces. The final output is a realistic expression-preserving facial images in frontal view without any occlusions.

Regression Approaches

In this stage, we address the problem of learning an associated pattern between the non-frontal facial shape and its corresponding frontal counterpart in a regression manner. Given a pair of shape vectors:

$$\begin{aligned}
 x &= [x^1, y^1, x^2, y^2, \dots, x^n, y^n] \\
 \mathbf{x}_0 &= [x_0^1, y_0^1, x_0^2, y_0^2, \dots, x_0^n, y_0^n]
 \end{aligned}
 \tag{1}$$

representing non-frontal and frontal facial annotations, respectively, the problem is to learn a regression function \mathcal{R} that makes $\mathcal{R}(x)$ most approximate to \mathbf{x}_0 . We implement four popular regressors for this task: Linear Regression, linear Support Vector Regression (SVR), Radial Basis Function (RBF) kernel-based SVR and Gaussian Process Regression (GPR).

Linear Regression: Linear regression is one of the most commonly used machine learning methods. The geometric view of linear regression model is that all sample points are constrained on a hyperplane. It makes the prediction

by a weighted combination of all input features that can be referred to as:

$$y = \langle \omega, x \rangle + b \tag{2}$$

where ω is a weight and b is a bias term. Consider the training data $\{(x^1, y^1), \dots, (x^l, y^l)\}$, where $x = x^1, \dots, x^l \in \mathbb{R}^n$ is the n-dimensional feature vector and $y = y^1, \dots, y^l \in \mathbb{R}$ is the response, linear regression solves a least square problem by minimizing:

$$\operatorname{argmin}_{\omega} \|y - \omega^T x\|^2 + \lambda \|\omega\|^2 \tag{3}$$

where $\lambda > 0$ is a regularization coefficient used to adjust the balance between under-fitting and over-fitting. The solution for ω is given as:

$$\omega = yx^T (xx^T + \lambda I)^{-1} \tag{4}$$

where I is the identity matrix.

Support Vector Regression: SVR also considers the linear function $y = \langle \omega, x \rangle + b$. Unlike linear regression that learns a hyperplane by blindly minimizing the prediction error throughout all sample points, SVR specifies a maximum margin hyperplane and minimize the margin by only considering the points that lies far away from hyperplane. SVR leaves much flexibility to the researchers to dynamically determine how much error can be tolerated in the model. This makes SVR less sensitive to noise and reduces the risk of overfitting. The SVR function can be expressed as:

$$\begin{aligned}
 \min \quad & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l (\xi_i^- + \xi_i^+) \\
 \text{s.t.} \quad & \begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \epsilon + \xi_i^+ \\ \langle \omega, x_i \rangle + b - y_i \leq \epsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \end{cases}
 \end{aligned}
 \tag{5}$$

where maximum margin ϵ is the distance from hyperplane to the constrained farthest sample points. It allows the model to contain some small errors within a strictly restricted boundary. This restriction is loosened by the slack variable ξ . It is a soft margin which further increases the tolerance of the points outside the maximum margin. This soft margin is also dynamically controlled by parameter C which adjusts the balance between accuracy and tolerance of deviation.

After introducing the dual problem and Lagrangian multipliers. The optimization problem becomes:

$$\max \begin{cases} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \end{cases} \quad (6)$$

$$\text{s.t. } \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]$$

where α_i and α_i^* are the Lagrangian multipliers. Sequential Minimal Optimization (SMO) [25] is commonly used to optimize the parameters in the SVR training stage. The obtained SVR model can predict the result by Equation (2).

The SVR method mentioned above is still linear. Kernel function is a trick in SVR to solve non-linear problems. The model is obtained by replacing the dot product $\langle x_i, x_j \rangle$ in Equation (6) by a kernel function $K(x_i, x_j) = \phi(x_i), \phi(x_j)$, where $\phi(x)$ is a human-designed transformation rule that maps x to a transformed feature space. RBF is a commonly used kernel where $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. Both linear SVR and RBF kernel-based SVR will be used for the facial spatial alignment task.

Gaussian Process Regression: Similar to non-linear SVR, GPR also starts from the linear function $y = \langle \omega, x \rangle + b$ and progresses into the kernel-based model. GPR is a probabilistic model that specifies ω as a prior distribution $p(\omega)$ which is assumed to be Gaussian. The bias b is amended to comply with the Gaussian distribution $\epsilon \sim N(0, \sigma_n^2)$. For convenience, we use $f(x)$ instead of $\langle \omega, x \rangle$ and define GPR model as:

$$y = f(x) + \epsilon \quad (7)$$

Accordingly, $f(x)$ is assumed to be a Gaussian Process (GP).

A GP is defined as a set of random variables in which any arbitrary subsets comply with the same joint Gaussian distribution. Given any arbitrary random variables x and x' , a GP prior is specified as:

$$f(x) \sim GP(m(x), K(x, x')) \quad (8)$$

where $m(x)$ is the mean of $f(x)$ and $K(x, x')$ is the covariance matrix. The collection of regression targets (the true landmark positions) are also assumed to be jointly Gaussian distributed, denoted as $y \sim GP(m(x), K(X, X)) + \sigma_n^2 I_n$, where I_n is an identity matrix.

In the training stage, the popular RBF kernel is used as the covariance function, showing as follows:

$$K(x, x') = \mu_f^2 \exp\left(-\frac{1}{2l^2} \|x - x'\|^2\right) \quad (9)$$

where l is a length scale. There are three hyperparameters $\theta = \{l, \mu_f^2, \mu_n^2\}$ that can be optimized by minimizing the negative log-likelihood:

$$L(\theta) = -\log p(y|X; \theta) \quad (10)$$

We skip over the optimization process as it is irrelevant to our topic.

In the testing stage, a new instance of x^* can be incorporated into the Gaussian distributed collection. The joint distribution of true targets and prediction of new instant f_* can be expressed as:

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu^* \end{pmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 I_n & K(X, x^*) \\ K(x^*, X) & K(x^*, x^*) \end{bmatrix}\right) \quad (11)$$

The posterior probability of f^* can be computed as:

$$P(f^* | x^*, X, f) \sim N(\mu^*, \Sigma^*) \quad (12)$$

where

$$\mu^* = \mu(x^*) + K(x^*, X)[K(X, X) + \sigma_n^2 I_n]^{-1}(y - \mu(X)) \quad (13)$$

The estimated result f_* is then specified by the mean value μ^* and variance Σ^* .

Cascade Regression Model

The problem of facial spatial alignment is obviously non-linear. Although the above mentioned regression approaches are able to solve the non-linear problem, their single-step mapping strategies are still inadequate for the frontalization problem. The presented problem should model facial changes under various head-poses and expressions, which is too complex to be solved in one-step regression. Therefore, we develop an adaptive cascade regression model that learns the frontal-profile associations in a cascade manner and gradually approximates to the optimum in several steps of regression rather than only one step.

In the frontalization task, the landmark positions are normalized through Procrustes analysis in which the in-plane rotation and size of the face are adjusted. Given M annotated facial image pairs of non-frontal and corresponding frontal faces, the linear function can be defined as $\mathbf{x}_0 + \Delta\mathbf{x} = \mathcal{R}(\mathbf{x}_0)$, where \mathbf{x}_0 and \mathbf{x} represents the shape vector for non-frontal and frontal images, respectively. And $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$ is known. Consequently, the final objective of regression can be expressed as:

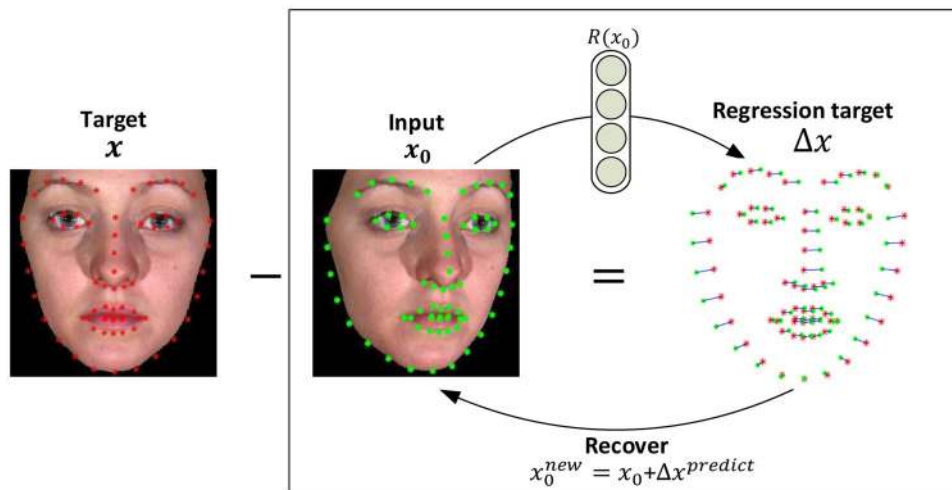
$$\Delta\mathbf{x} \leftarrow \mathcal{R}(\mathbf{x}_0) \quad (14)$$

This equation can be referred to as the linear function of all the three regression methods mentioned above. This function is illuminated in Fig. 2.

Then we introduce the cascade regression manner regarding $\Delta\mathbf{x}^i$ representing the obtained $\Delta\mathbf{x}$ in the i th cascade. In each cascade, we revise Eq. (14) into:

$$\Delta\mathbf{x}^i \leftarrow \mathcal{R}^i(\mathbf{x}_0^i) \quad (15)$$

Fig. 2 Cascade regression training process



and train a regression model at the current stage. In the new round, $x_0^{i+1} = \Delta x^i + x_0^i$, $\Delta x^{i+1} = x - x_0^{i+1}$ and they will be used to train a regression model in the current round. The cascade regression model keeps doing iteration until these parameters and Δx^i become zero. Empirically, it converges in 4 or 5 steps.

During testing, the non-frontal facial landmarks should be localized first. There are many existing facial landmark detection methods that have been proved to be effective. With the obtained facial landmarks, frontal face-shape will be estimated using Eq. (15) sequentially.

By performing cascade on four regression approaches (linear regression, linear SVR, RBF SVR and GPR), we achieve four independent cascade regression models. We further introduce ensemble to combine all the regression models together and make a comprehensive decision. Ensemble learning is a model combination strategy that combines multiple models and then computes the average or weighted summation for the final prediction. We apply the ensemble and employ both the average and weighted summation strategy to predict the final result.

Ensemble Regression: Ensemble learning is also known as classifier or regressor combination. It simultaneously processes multiple independent learners and strategically combines them to improve prediction accuracy. Naturally, an ensemble learning model would expect base models to be very different from each other. The most common way to increase between-model variations is to use the bootstrap aggregating (bagging) strategy in which each base learner is fed with a random subset of the whole training dataset. But bagging is not suitable for our model because our base learners are four different regressors which has already involved large variations.

For our ensemble regression model, there are four base regressors: $[\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, \mathcal{R}_4]$. Given an input x , the output $E(x)$ is computed by a linear function:

$$E(x) = \lambda_1 \mathcal{R}_1(x) + \lambda_2 \mathcal{R}_2(x) + \lambda_3 \mathcal{R}_3(x) + \lambda_4 \mathcal{R}_4(x) \quad (16)$$

where λ is the weight. Suppose the groundtruth landmark position is y , our goal is to minimize:

$$\operatorname{argmin}_{\lambda} \|x + E(x) - y\|^2 \quad (17)$$

The optimal value of the weight λ_i can be easily computed through the least square method. Beside calculating the optimal weight value, we also introduce an averaging strategy as comparison, in which λ value is always fixed to 2.5. The comparison of two strategies will be made in the experimental section.

In the training stage, the four regressors are trained independently without using the ensemble process. Then, the training data $[x, y]$ is used again to compute the weight value λ . In each cascade, we compute a group of weight parameters. Finally, the whole ensemble model, as well as obtained weight values, is used in validation and the testing set to predict more accurate frontal facial shape by Eq. 16 and $y_{predict} = x + E(x)$.

Texture Reconstruction

The cascade regression model outputs a predicted frontal facial shape which is seen as a base mesh. The next step is to reconstruct realistic facial textures using this mesh model. This is achieved by using a two-step image warping: 1) a warp function $W(x;p)$ is computed to associate the each pixel position from the base mesh with the pixel positions of the

input image I , and 2) the value of each pixel x in the warped image $I(W(x;p))$ is obtained by sampling the image I at the corresponding position. We employ the piecewise affine warping method [27] to calculate the warp function $W(x;p)$. Piecewise affine warping is based on an assumption that image warping on a small local region can be seen as a linear transformation although whole face warping is nonlinear.

Given a base shape, Delaunay triangulation is used to create multiple non-overlapping triangles formed by facial landmark points. All these triangles make up the mesh. Each triangle accounts for a fairly small region such that it is reasonable to use linearly affine warping.

Let s_0 denote the base mesh whose pixels are denoted as $x = (x, y)$. Assume a pixel (x^0, y^0) in the base mesh falls into a triangle whose vertices are (x_i^0, y_i^0) , (x_j^0, y_j^0) and (x_k^0, y_k^0) , this pixel can be uniquely expressed as:

$$(x^0, y^0) = (x_i^0, y_i^0) + \alpha[(x_j^0, y_j^0) - (x_i^0, y_i^0)] + \beta[(x_k^0, y_k^0) - (x_i^0, y_i^0)]$$

$$\text{where } \begin{cases} \alpha = \frac{(x^0 - x_i^0)(y_k^0 - y_j^0) - (y^0 - y_i^0)(x_k^0 - x_j^0)}{(x_j^0 - x_i^0)(y_k^0 - y_j^0) - (y_j^0 - y_i^0)(x_k^0 - x_j^0)} \\ \beta = \frac{(y^0 - y_i^0)(x_j^0 - x_i^0) - (x^0 - x_i^0)(y_j^0 - y_i^0)}{(x_j^0 - x_i^0)(y_k^0 - y_j^0) - (y_j^0 - y_i^0)(x_k^0 - x_j^0)} \end{cases} \quad (18)$$

Let s denote the shape of the input face where there is a unique triangle (x_i, y_i) , (x_j, y_j) and (x_k, y_k) associated with the triangle (x_i, y_i) , (x_j, y_j) and (x_k, y_k) from base mesh. The results of α and β are used to calculate the associated pixel position (x, y) in the input image:

$$W(x;p) = (x, y) = (x_i, y_i) + \alpha[(x_j, y_j) - (x_i, y_i)] + \beta[(x_k, y_k) - (x_i, y_i)] \quad (19)$$

As the location of each pixel of the base mesh is assigned to the corresponding position in the input image, the pixel value of the base mesh is obtained by sampling pixel values from the input image I at the corresponding position. The most commonly used sampling strategy is bilinear interpolation.

After piece-wise affine warping, the selected template is filled in textures from input images. However, the warped image may not be realistic due to the self-occlusion caused by out-of-plane head rotation, as is illuminated in Fig. 3. Thus, we design a further process to rectify the textures. This is achieved by mainly implementing AAM model instantiation for texture fitting.

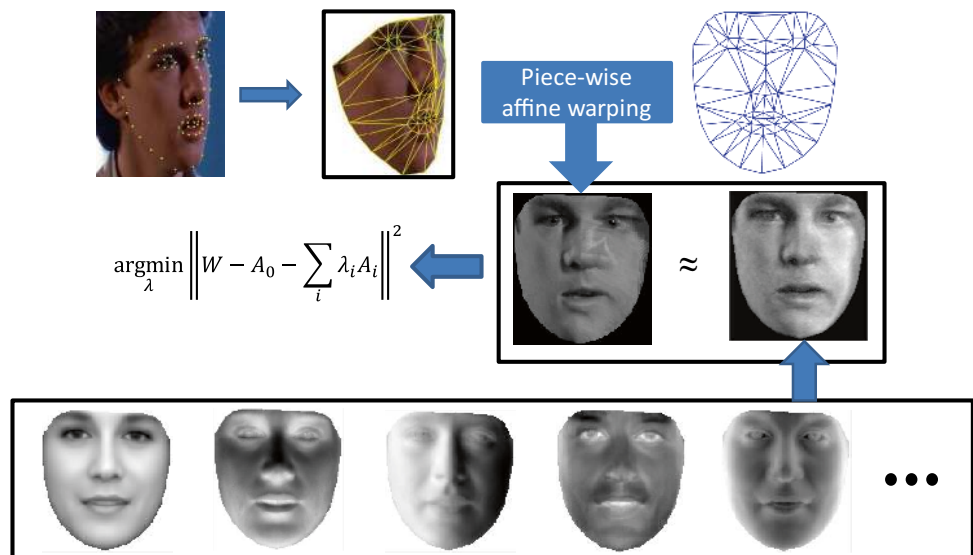
Active Appearance Model: AAM [12] is well known for facial landmark detection. AAM model fitting is employed to reconstruct facial textures. For each query image $I \in \mathbb{R}^{m \times n}$, AAM model instantiation minimizes an objective function:

$$\operatorname{argmin}_{\lambda} \| A(x) - I(W(x;p)) \|^2 \quad (20)$$

where $A(x) = A_0(x) + \sum_{i=1}^m \lambda_i A_i(x)$ is the required frontal face in which $A_0(x)$ is the mean face and $\sum_{i=1}^m \lambda_i A_i(x)$ is a linear combination of a set of pre-defined eigen faces $M_A = [A_1(x)|A_2(x)| \dots |A_m(x)]$, parameterized by λ . The eigen faces are computed by applying Principal Component Analysis (PCA) to a set of warped training images. The original training images should normally contain clean (no occlusion) and frontal faces. They are then normalized by piece-wise affine warping their facial shapes and appearances onto a base mesh (selected template). By applying PCA, A_i is manually set to be m eigen faces with regards to m largest eigenvalues.

In our experiment, the shape of the input face is automatically detected by SDM. The detected landmarks may be sometimes incorrect, which may directly lead to a failure

Fig. 3 Texture fitting



of the frontalization result. Therefore, we continue employing the AAM gradient search strategy to enhance the landmark detection results. The Eq. (21) is modified by minimizing:

$$\operatorname{argmin}_{\lambda, \Delta p} \| A(x) - I(W(x;p + \Delta p)) \|^2 \tag{21}$$

where p is updated by $p \leftarrow p + \Delta p$. The linear approximation is given by a Taylor series expansion:

$$I(W(x;p + \Delta p)) = I(W(x;p)) + \nabla I \frac{\partial W}{\partial p} \Delta p \tag{22}$$

where ∇I is the gradient image, $\frac{\partial W}{\partial p}$ is the warp Jacobian evaluated by p , and p is the parameter of current shape referred to the equation $s = s_0 + \sum_{i=1}^n p_i s_i$ defined by the active shape model (ASM) [35]. The base shape s_0 is the mean shape of all shapes of training images and the eigenvectors s_i represent shape variance computed by applying PCA to the training shapes. Let shape denotes the landmark positions as $s = (x_1, y_1, x_2, y_2, \dots, x_v, y_v)$. The warp Jacobian is computed by applying the chain rule:

$$\frac{\partial W}{\partial p} = \sum_{i=1}^v \left[\frac{\partial W}{\partial x_i} \frac{\partial x_i}{\partial p} + \frac{\partial W}{\partial y_i} \frac{\partial y_i}{\partial p} \right] \tag{23}$$

where

$$\begin{aligned} \frac{\partial W}{\partial x_i} &= (1 - \alpha - \beta, 0) \quad \text{and} \quad \frac{\partial W}{\partial y_i} = (0, 1 - \alpha - \beta) \\ \frac{\partial x_i}{\partial p} &= (s_1^{x_i}, s_2^{x_i}, \dots, s_n^{x_i}) \quad \text{and} \quad \frac{\partial y_i}{\partial p} = (s_1^{y_i}, s_2^{y_i}, \dots, s_n^{y_i}) \end{aligned} \tag{24}$$

The solution of Equation (22) for λ is given as:

$$\lambda = (I(W(x;p + \Delta p)) - A_0(x))M_A^T \tag{25}$$

where $M_A^T = M_A^{-1}$ since they are orthonormal eigenvectors. Then the solution to Δp is to use Gaussian Newton approximation as:

$$\Delta p = \sum_x \left(\left[\nabla I \frac{\partial W}{\partial p} \right]^T \left[\nabla I \frac{\partial W}{\partial p} \right] \right)^{-1} \left[\nabla I \frac{\partial W}{\partial p} \right]^T [A(x) - I(W(x;p))] \tag{26}$$

The algorithm works iteratively with the update rule $p \leftarrow p + \Delta p$ until reaching the stop criterion. AAM fitting starts with the mean face and gradually searches for the optimal landmark locations. But the slow convergent rate will break out the requirement of real-time settings. Thus, we take the SDM output as initialization and only conduct 3 iterations to reduce computational time. By this strategy, the whole system could work well in real-time with satisfied results. The final parameter λ is used to calculate the frontal facial image $A_0(x) + \sum_{i=1}^m \lambda_i A_i(x)$.

Generally, AAM fitting is able to well amend the distortion caused by warping and remove occlusions. Empirically, the method works well for the head-pose within ± 45 degree. Both SDM face alignment and AAM reconstruction methods, as well as most face alignment and frontalization methods, were reported ineffective when the head-pose angles exceed this range. Considering that the head-pose of in-the-wild facial images seldom goes beyond this range, our method could still work effectively in most cases.

Experiment

The proposed method has been validated with the following tasks: 1) regression error of spatial alignment, 2) static FER in the wild, 3) dynamic FER in the wild.

Training Data Collection

Binghamton University 3D Facial Expression (BU3DFE) is a static 3D facial expression database which includes 100 subjects with 2500 3D facial expression models. We generate the training data by projecting the rendered 3D models from the BU3DFE database to the 2D image space. Images are captured at 7 pan angles ($-45^\circ, -30^\circ, -15^\circ, 0^\circ, 15^\circ, 30^\circ, 45^\circ$) and 5 tilt angles ($-30^\circ, -15^\circ, 0^\circ, 15^\circ, 30^\circ$), which results in totally 35 different viewpoints, as is shown in Fig. 4. Each training instance includes the position landmark points in one of the 34 non-frontal rotations and the corresponding points in frontal pose.

To train the cascade regression model, we need to detect facial keypoints in each 2D images. We use both SDM and OpenFace [2] for our landmark detection task. Most landmarks can be well detected using these two methods. Misaligned points were manually corrected or removed. This process generates 1887 training examples in pair denoted by $[x, x_0]$. The 1887 generated images are used to train the frontal-shape prediction model in the first stage. For the second stage, we manually picked 302 frontal view images without any occlusions from FER dataset, Static Facial Expressions In The Wild (SFEW) and Acted Facial Expressions In The Wild (AFEW). They are used to train the AAM model. Finally 200 eigen faces are obtained and used for texture fitting.

Unlike many deep learning methods that require millions of training examples, our method can work with a small volume of training data. Although we used over a thousand external training images in the first stage, it is still a small number compared with millions of external images used by deep learning methods. Furthermore, the 1887 images are easy to obtain without much manual intervention. For the recognition stage, there is no need to use any external data to train the recognition model.

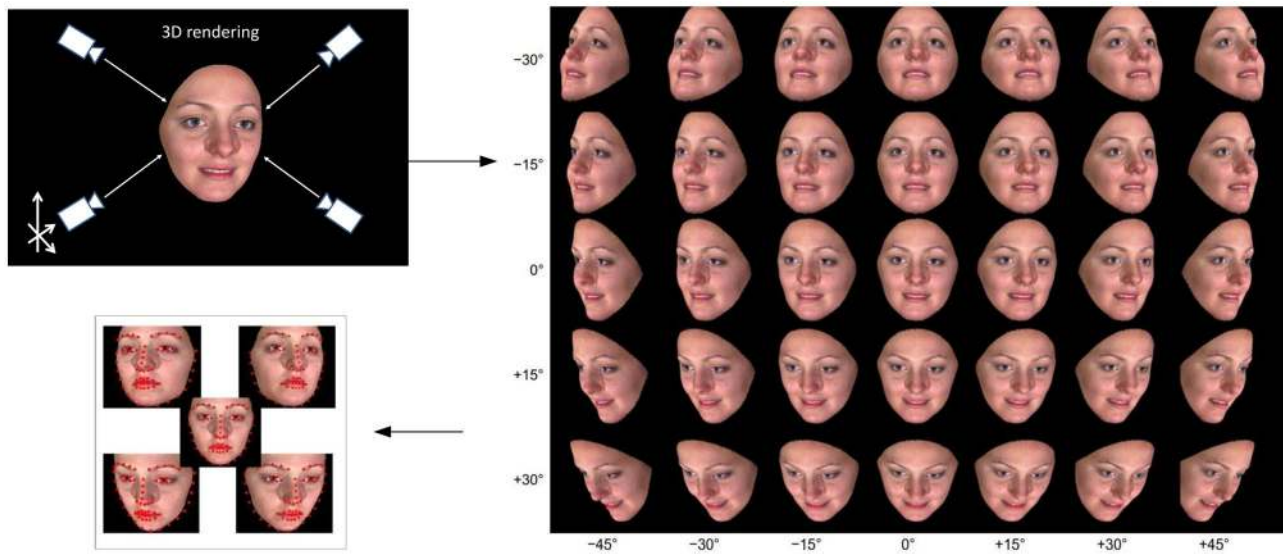


Fig. 4 3D rendering and training data generating

Spatial Alignment

In this section, we measure the accuracy of the frontal facial shape prediction based on the generated 1887 example pairs. We perform a 10-folder cross-validation strategy to evaluate the performance of different regression models. The alignment error is measured by:

$$err = \sqrt{(x - x_0)^2 + (y - y_0)^2} / l \quad (27)$$

where (x, y) and (x_0, y_0) are aligned position and groundtruth, respectively, and l is the width of facial bounding box which is calculated by $(\max(x_0) - \min(x_0) + \max(y_0) - \min(y_0)) / 2$.

Table 1 shows the comparison of alignment errors using the regression approaches mentioned in Section 3.1. We can have some clues from this table:

1) Non-linear regression approaches (GPR and SVR with RBF kernel) perform better than linear models (linear regression and linear SVR). GPR shows significant superiority than other regressors.

2) Each cascade regression model has an around 5% superior performance over the corresponding single-step regression model. Furthermore, the error of the cascade

linear model is smaller than single-step RBF-based SVR models, which indicates that a linear model embedded in a cascade manner can also achieve an effective non-linear solution.

3) Ensemble of four cascade regressors can improve the performance over each base regressor. For the two ensemble methods, the weighted summation strategy performs better than the simple averaging strategy. On the contrary, the combination of four single-step regression models has a higher error than GPR showing the superiority of the GPR method over the other regressors. Ensemble learning can boost the performance only when the base learners have similar generalization abilities.

This result demonstrates that combining multiple models can provide a better solution to facial spatial alignment than using a single model. In this experiment, the best result is obtained by the ensemble of four cascade regression models and this approach will be used to facilitate the next step of texture fitting.

Static FER on SFEW

The proposed method is used to solve dynamic FER problems. It is also appropriate to be applied to static FER.

Table 1 Alignment error (%) of different regression methods

| | Linear | SVR(linear) | SVR(RBF) | GPR | Ensemble(average) | Ensemble(weighted) |
|------------------------|--------|-------------|----------|-------|-------------------|--------------------|
| Single-step Regression | 18.87 | 18.44 | 16.99 | 11.13 | 14.88 | 13.26 |
| Cascade Regression | 13.82 | 12.92 | 12.40 | 10.84 | 10.53 | 9.84 |

Table 2 Comparison of recognition rate (%) of the state-of-the-art methods on SFEW database

| | Angry | Disgust | Fear | Happy | Neutral | Sadness | Surprise | Overall |
|----------|-------|---------|-------|-------|---------|---------|----------|---------|
| Baseline | 23.00 | 13.00 | 13.90 | 29.00 | 23.00 | 17.00 | 13.50 | 18.90 |
| [34] | 25.89 | 28.24 | 17.17 | 42.98 | 14.00 | 33.33 | 10.99 | 24.70 |
| [7] | 24.11 | 14.12 | 20.20 | 50.00 | 23.00 | 23.23 | 21.98 | 26.14 |
| Ours | 40.18 | 25.88 | 48.48 | 55.26 | 37.00 | 36.36 | 37.36 | 40.71 |

Statistical Facial Expression in the Wild (SFEW) [17] is a static spontaneous facial expression database which contains 700 images captured from movies labelled in seven categories: six universal emotions and neutral.

For FER, there is a standard evaluation protocol provided by the authors of SFEW. The evaluation is strictly person-independent. In this experiment, Local Binary Pattern (LBP) and Support Vector Machine (SVM) are used for feature extraction and emotion classification, respectively. In Table 2, the methods of [34] and [7] are the state-of-the-art approaches. It is worth mentioning that the method in [7], which is also based on frontalization, only performs frontal shape prediction without texture reconstruction. Our method has achieved over 14% superior accuracy than [7]. This suggests that texture fitting is also necessary for facial expression analysis and can significantly boost its performance.

In this comparison, we did not mention deep learning methods because our work focus on the small sample learning task which is quite different from deep learning. Meanwhile, deep learning methods usually require a large volume of external training data, which do not comply with the evaluation protocol of SFEW.

Dynamic FER on AFEW

In this section, experiments of frontalization and expression analysis on sequential facial images are conducted which are followed by both quantitative and qualitative analysis. In order to evaluate the performance on video sequences, Acted Facial Expressions In The Wild (AFEW) [5] dataset is applied. AFEW is an unconstrained facial expression database in which video clips are collected from movies. It contains 1368 video clips which are divided into three parts: 578 for training, 383 for validation and 407 for testing. Considered that the groundtruth of testing images is still unreleased, we follow the evaluation protocol of Emotion Recognition in the Wild Challenge 2014 (EmotiW 2014) [4] but only compare the performance on validation data.

The intuitive visualization of frontalization results is shown in Fig. 5. From this figure, we can see three main advantages of the proposed method. Firstly, the occlusions are well removed. As is displayed, for example, the hand covering on the face is removed and its occluded regions are well recovered. Secondly,

all the non-frontal faces are normalized and reconstructed in the frontal view. Finally, facial expressions are still aware and identifiable in the reconstructed images. These visual results indicate the effective performance of the proposed methods in terms of recognizable expression reconstruction, frontalization and de-occlusion.

Based on the reconstructed facial image sequences, we applies LBP-TOP and SVM for feature extraction and emotion classification, respectively. In Table 3, the baseline result is achieved by the database creators who also used the traditional LBP-TOP + SVM strategy. The difference between our method and the baseline is that we have introduced a novel frontalization process. The accuracy of our method is 15% higher than the baseline, which indicates the effectiveness of face frontalization. The winner of EmotiW 2014 competition is [18] which used both audio and video signals, and combined SIFT, HOG and DCNN with external training data. Due to the very large amount of external training examples, [18] still keeps the record of EmotiW 2014 competition. The algorithm in [10] is a typical representation of time alignment which aims to model the variations of time extent.

We firstly compare the two traditional methods between ours and [10]. Our method applies spatial alignment without temporal alignment while [10] does the opposite. Our method has a minor superiority than [10], which demonstrates that both temporal alignment and spatial alignment are important to dynamic analysis of facial expressions. Although many researchers focus on modelling temporal relations, the importance of spatial relations problems caused by head-pose and occlusions still remain.

The result of the proposed method and [10] are both based on reasonable heuristics without using any external training data. The method in [18] achieves 0.1% higher accuracy than our method, but it used a large amount of external training data. Although deep learning methods usually outperform the small sample learning methods, it may be restricted in some application cases due to the requirement for a large training dataset.

Conclusion

In this paper, we have presented a novel CRFF method for facial expression-preserving face frontalization and applied it to dynamic FER. It successfully fills the gap that there is no dynamic FER approaches for spatial alignment. To address the challenging problem of accurate frontal facial



Fig. 5 Visualization of frontalization

Table 3 Comparison of recognition rate (%) of the state-of-the-art methods on AFEW database

| | Angry | Disgust | Fear | Happy | Neutral | Sadness | Surprise | Overall |
|-------------|-------|---------|-------|-------|---------|---------|----------|---------|
| Baseline[4] | 50.00 | 25.00 | 15.21 | 57.14 | 34.92 | 16.39 | 21.73 | 33.15 |
| [18] | 84.75 | 17.95 | 27.27 | 82.54 | 70.49 | 22.03 | 6.52 | 48.52 |
| [9] | - | - | - | - | - | - | - | 48.3 |
| Ours | 65.63 | 27.50 | 32.61 | 74.07 | 57.14 | 32.79 | 36.96 | 48.40 |

shape prediction, we have developed an adaptive cascade regression learning method and an ensemble learning method to boost the prediction performance. Our method have successfully achieved frontal face generation and de-occlusion, while preserving subtle facial expression cues. Different from existing 3D methods and deep learning methods, our method can generate realistic faces even with a very small training dataset and the whole model works

in real-time. Experimental results demonstrate that the proposed method can boost the performance of FER and it is suitable for in-the-wild facial analysis. In the future, we plan to further improve the dynamic FER performance by combining time alignment methods with the proposed spatial alignment methods. Meanwhile, we would apply this method to a wider applications, such as Boredom, inattention and pain detection, to further support the research of healthcare.

Funding Information This study was funded by Engineering and Physical Sciences Research Council (Grant No. EP/N025849/1)

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aneja D, Colburn A, Faigin G, Shapiro L, Mones B. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*. Springer, 2016. p. 136–153.
- Baltrušaitis T, Robinson P, Morency LP. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV)*. IEEE Winter Conference on 2016. p. 1–10.
- Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. In *European conference on computer vision*. Springer, 1998. p. 484–498.
- Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014. p. 461–466.
- Dhall A, Goecke R, Lucey S, Gedeon T, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*. 2012;19(3):34–41.
- Dureha A. An accurate algorithm for generating a music playlist based on facial expressions. *Int J Comput Appl*. 2014;100(9):33–9.
- Eleftheriadis S, Rudovic O, Pantic M. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Trans Image Process*. 2015;24(1):189–204.
- Ferrari C, Lisanti G, Berretti S, Del Bimbo A. Effective 3d based frontalization for unconstrained face recognition. In *Pattern Recognition (ICPR)*, 23rd International Conference on. IEEE, 2016. p. 1047–1052.
- Guo Y, Xia Y, Wang J, Yu H, Chen R-C. Real-time facial affective computing on mobile devices. *Sensors*. 2020;20(3):870.
- Guo Y, Zhao G, Pietikäinen M. Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Trans Image Process*. 2016;25(5):1977–92.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proc IEEE Conf Comput Vis Pattern Recognit*. 2016. p. 770–778.
- Heisele B, Ho P, Poggio T. Face recognition with support vector machines: Global versus component-based approach. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, IEEE, 2001. vol. 2, p. 688–694.
- Jaiswal S, Valstar M. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016. p. 1–8.
- Jeni LA, Cohn JF, Kanade T. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG)*, 2015 11th IEEE International Conference and Workshops on, IEEE, 2015. vol. 1, p. 1–8.
- Jiang B, Valstar MF, Martinez B, Pantic M. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Trans. Cybernetics*. 2014;44(2):161–74.
- Li K, Zhao Q. If-gan: Generative adversarial network for identity preserving facial image inpainting and frontalization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, p. 158–165.
- Liu M, Shan S, Wang R, Chen X. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proc IEEE Conf Comput Vis Pattern Recognit*, 2014. p. 1749–1756.
- Liu M, Wang R, Li S, Shan S, Huang Z, Chen X. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM, 2014. p. 494–501.
- Liu X, Xia Y, Yu H, Dong J, Jian M, Pham TD. Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation. *IEEE Trans Neural Syst Rehabil Eng*. 2020;28(10):2325–32.
- Lou J, Cai X, Wang Y, Yu H, Canavan S. Multi-subspace supervised descent method for robust face alignment. *Multimed Tools Appl*. 2019;78(24):35455–699.
- Lou J, Wang Y, Nduka C, Hamed M, Mavridou I, Wang F-Y, Yu H. Realistic facial expression reconstruction for vr hmd users. *IEEE Trans Multimedia*. 2019;22(3):730–43.
- Matthews I, Baker S. Active appearance models revisited. *Int J Comput Vis*. 2004;60(2):135–64.
- Mattivi R, Shao L. Human action recognition using as sparse spatio-temporal feature descriptor. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 2009. p. 740–747.
- Pfister T, Li X, Zhao G, Pietikäinen M. Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV)*, 2011 IEEE International Conference on. IEEE, 2011. p. 1449–1456.
- Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
- Roth J, Tong Y, Liu X. Unconstrained 3d face reconstruction. In *Proc IEEE Conf Comput Vis Pattern Recognit*. 2015. p. 2606–2615.
- Rudovic O, Pantic M, Patras I. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Trans Pattern Anal Mach Intell*. 35(6):1357, 1369–2013
- Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Trans Med Imaging*. 1999;18(8):712–21.
- Sagonas C, Panagakis Y, Zafeiriou S, Pantic M. Robust statistical face frontalization. In *Proc IEEE Int Conf Comput Vis*. 2015. p. 3871–3879.
- Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans Pattern Anal Mach Intell*. 2015;37(6):1113–33.
- Shan C, Gong S, McOwan PW. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis Comput*. 2009;27(6):803–16.
- Shao J, Qian Y. Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing*. 2019;355:82–92.

33. Sun B, Wei Q, Li L, Xu Q, He J, Yu L. Lstm for dynamic emotion and group emotion recognition in the wild. In Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016. p. 451–457.
34. Taheri S, Qiu Q, Chellappa R. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Trans Image Process.* 2014;23(8):3590–603.
35. Tariq U, Yang J, Huang TS. Multi-view facial expression recognition analysis with generic sparse coding feature. In European Conference on Computer Vision. Springer, 2012. p. 578–588.
36. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M. Face2face: Real-time face capture and reenactment of rgb videos. In Proc IEEE Conf Comput Vis Pattern Recognit. 2016. p. 2387–2395.
37. Tran L, Yin X, Liu X. Disentangled representation learning gan for pose-invariant face recognition. In CVPR, p. 7, 2017.
38. Wang S, Wang J, Wang Z, Ji Q. Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions. *IEEE Trans Multimedia.* 2015;17(12):2185–97.
39. Wang Y, Yu H, Dong J, Stevens B, Liu H. Facial expression-aware face frontalization. In Asian Conference on Computer Vision. Springer, 2016. p. 375–388.
40. Wang Y, Yu H, Stevens B, Liu H. Dynamic facial expression recognition using local patch and lbp-top. In 2015 8th International conference on human system interaction (HSI). IEEE, 2015. p. 362–367.
41. Wang Z, Wang S, Ji Q. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In Proc IEEE Conf Comput Vis Pattern Recognit. 2013. p. 3422–3429.
42. Xiong X, De la Torre F. Supervised descent method and its applications to face alignment. In Proc IEEE Conf Comput Vis Pattern Recognit. 2013. p. 532–539.
43. Xue M, Liu W, Li L. Person-independent facial expression recognition via hierarchical classification. In Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on. IEEE, 2013. p. 449–454.
44. Yim J, Jung H, Yoo B, Choi C, Park D, Kim J. Rotating your face using multi-task deep neural network. In Proc IEEE Conf Comput Vis Pattern Recognit. 2015. p. 676–684.
45. Yin X, Yu X, Sohn K, Liu X, Chandraker M. Towards large-pose face frontalization in the wild. In Proc. ICCV 2017. p. 1–10.
46. Yu Z, Liu G, Liu Q, Deng J. Spatio-temporal convolutional features with nested lstm for facial expression recognition. *Neurocomputing.* 2018;317:50–7.
47. Zhao G, Pietikainen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans Pattern Anal Mach Intell.* 2007;29(6):915–28.
48. Zhou F, Kong S, Fowlkes CC, Chen T, Lei B. Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing.* 2020.
49. Zhu X, Lei Z, Liu X, Shi H, Li SZ. Face alignment across large poses: A 3D solution. In Proc IEEE Conf Comput Vis Pattern Recognit. 2016. p. 146–155.
50. Zhu X, Lei Z, Yan J, Yi D, Li SZ. High-fidelity pose and expression normalization for face recognition in the wild. In Proc IEEE Conf Comput Vis Pattern Recognition. 2015. p. 787–796.