Face alignment based on fusion subspace and 3D fitting

Jiahui Zhang1 | Lan Di1 | Jiuzhen Liang2

1 School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
2 School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, China

Correspondence
Lan Di, School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China.
Email: dilan@jiangnan.edu.cn

Abstract
The traditional face alignment approaches based on cascade regression have achieved satisfactory result on the frontal face, but for the face with large changes in posture and expression, a single initial shape will lead to the result falling into local optimum. In order to solve this problem, a two-stage cascade regression model for face alignment is proposed, which generates coarse initial shape from the aligned salient shape. The first stage is used to align the salient shape that contains some prominent landmarks. To enhance the robustness of authors’ method, the fusion subspace is used to divide the samples, and each subset trains cascade regression model separately. The alignment results of the first stage are used to generate the coarse initial shapes for the second stage through 3D fitting. The second stage is still based on cascade regression, which is used to further predict the full shape. The experimental results demonstrate the proposed method can achieve state-of-art performance, especially in unconstrained conditions with various poses.

1 INTRODUCTION

Face alignment, which is also called facial landmark localisation, aims to locate the facial landmarks (the points around eyebrows, eyes, nose, mouth and contour) based on face detection. It is also an essential part of face image processing, and plays an important role in face recognition [1], expression analysis [2], 3D face reconstruction [3] and so on. At present, most of the face alignment methods have achieved satisfactory results on the frontal face images. However, it is still a challenge to achieve high-precision alignment for unconstrained face images with various changes in expression and pose.

There are two main categories of face alignment methods: (1) methods based on generation model. The representative algorithm is active appearance model (AAM) [4] proposed by Cootes et al. Firstly, use principal component analysis (PCA) to construct parameter model in shape and texture feature space, and then match the model with face image by optimizing parameters. Although this method has made various improvements [5–7], it cannot deal with subtle shape changes due to the limited expression ability of parameter model. (2) Methods based on discriminative model. The goal of these methods is to directly map the image texture features to face shape. Cascaded regression model (CRM) [8] proposed by Dollá et al. The method based on cascade regression regards face alignment as a nonlinear optimization problem between image texture features and face shape. By learning a series of feature-to-shape mappings, the model can gradually update the initial shape to the final shape. In recent years, in addition to the cascade regression model, the method based on deep network [9–16] has also been widely used. The earliest application of deep network in face alignment is the method based on deep convolution neural network (DCNN) [9] proposed by Sun et al. In this method, the cascade method is still used to achieve coarse to fine alignment. 3D face model has also been used to achieve face alignment. In these methods [12–14], the 3D model is fitted to the face image by adjusting the parameters, and then the fitting result is projected to the 2D plane to obtain the alignment result. Due to the large number of 3D face model parameters, most of these methods also need to use deep network. Although the performance of face alignment methods has been greatly improved by the addition of deep network, there are still many excellent methods based on cascaded regression [17, 18], which have lower training cost, and in some cases have better performance than those based on deep learning.

Shape initialization is the first step of face alignment method based on cascaded regression. Many traditional cascaded regression methods [19–21] take a frontal neutral face shape or the
mean ground truth shape of all samples in the training set as the initial shape. Then, based on the initial shape, the final prediction result is achieved through multiple iterations. When the initial shape is much different from the ground truth shape, the final result may not be ideal due to the limited number of iterations or falling into local optimum. Therefore, the initial shape selection will greatly affect the final result. The supervised descent method (SDM) [19] used gradient descent method to learn the mapping function from image features to face shapes, but it used only a single initial shape, which made the results fall into local optimum. To address the issue, Xiong et al. [22] proposed to divide the samples into multiple sub-sets according to domains of homogeneous descent (DHD), and then establish regression models, respectively, on the basis of SDM, so as to avoid the results into local optimum. However, the partition method proposed by them needs to use the ground truth shapes of the samples, which is not realistic in the test phase. The coarse to fine shape searching (CFSS) [23] solved the initialization problem from another perspective. It took the ground truth shape of all samples as the solution space to explore. Although it solves the limitation of initial shape, it also reduces the computation speed.

On the basis of previous work, we propose a new solution to solve the problem of shape initialization, which is to generate appropriate initial shape for each sample. After observation, we found that the facial pose changes can be expressed by a small number of key landmarks. Therefore, we propose a two-stage cascade regression model (2-SCRM). The first stage is a salient shape prediction model, whose predictions are used to generate coarse initial shapes for each sample. At this stage, a fusion subspace partition method is added to enhance the pose robustness of the algorithm. This method can divide the samples with similar posture into the same subset. On this basis, we train CRM for each subset, which can significantly improve the pose robustness of the algorithm.

The two-stage cascade regression model is proposed. In the first stage, multiple CRMs are used to predict the salient shapes. In the second stage, based on the coarse initial shapes generated from the above results, we still use the CRM to achieve more accurate alignment. Experimental results show that the two-stage model can effectively improve the accuracy of alignment and has pose robustness.

2 TRADITIONAL CASCADE REGRESSION MODEL

Assuming that a face shape \( s = [x_{1}, y_{1}, ..., x_{L}, y_{L}] \) is composed of \( L \) landmarks. Given a face image \( I \), starting with the initial shape \( s'_{0} \), the face shape is updated by predicting the shape residuals between the ground truth shapes and the current shapes in each iteration. The update process of the \( t \)-th iteration is

\[
s'_{t} = s'_{t-1} + W_t \Phi'(I, s'_{t-1})
\]

where \( \Phi' \) is the feature mapping function, and \( W_t \) is the linear regression matrix that maps texture features to shape residuals.

From Equation (1), we can see that the texture features depend both on the face image \( I \) and the face shape of the previous iteration \( s'_{t-1} \). Specifically, it extracts the local features from the surrounding area of each landmarks of the face shape, and combines all the local features into the global feature, which is called shape-index feature [20]. In this way, the feature mapping function of the \( t \)-th local feature is denoted as \( \phi_{j,t} \), and the one of the shape-index feature in the \( t \)-th iteration is denoted as \( \Phi_{t} = [\phi_{1,t}, ..., \phi_{j,t}, ..., \phi_{N,t}] \). Considering the speed and performance, we use local binary features (LBF) [21] as the local features.

Local binary features are learnt by random forest. Assuming that the random forest in each iteration contains \( G \) decision trees. During the training process, the training samples are divided into \( G \) parts, which are, respectively, traversed in the corresponding decision tree. The split strategy of the decision tree depends on the pixel difference feature [20]. We randomly extracted 500-pixel difference features from local region around each landmark. The feature that gives rise to maximum variance reduction is selected as the split threshold. When one of the samples falls into a certain leaf node of the decision tree, the code at that leaf node is \( 1 \), and the code at the other leaf nodes in the decision tree is \( 0 \). Combining the coding results of the leaf nodes of all decision trees in the random forest, we can get a very sparse binary vector, which is called local binary feature. Figure 1 shows the encoding process of local binary features and shape-index features.

After extracting the features, there should be a linear regression matrix to map the features to the shape residuals. The process of learning the linear regression matrix in the \( t \)-th iteration can be expressed as the minimized objective function:

\[
\min_{W_t} \sum_{i=1}^{N} \| \Delta s'_{t} - W_t \psi_{i,t} \|_{2}^{2} + \lambda \| W_t \|_{2}^{2}
\]

where \( N \) is the total number of samples, \( \Delta s'_{t} = s'_{t} - s'_{t-1} \) and \( \psi_{i,t} = \Phi'(I, s'_{t-1}) \) means the shape residual and the shape-index.
TWO-STAGE CASCADE REGRESSION

Salient shape

feature of the \(i\)th sample. Since the local binary feature is pretty high dimensional, there is a L2 regularisation term for \(W\), where \(\lambda\) is the regularisation coefficient. We use a dual coordinate descent method [24] to calculate \(W\) in our implementation.

Due to the difference in posture and expression of face shapes, it is difficult to map texture features to shape residuals with only a single regressor. The global supervised descent method [22] proposes to divide the optimisation space into multiple domains of homogeneous descent (DHD), and each DHD learns the regressor separately. This optimisation problem (we omit the index \(i\) in this section for brevity):

\[
\min_{U_k, W_k} \sum_{k=1}^{K} \sum_{i \in U_k} \| \Delta y_i - W_k \psi_i \|^2_2
\]

s.t. \(\Delta y_k W_k \psi_i > 0\)

where \(K\) is the total number of the subsets, \(U_k\) is the \(k\)th sample subset. Equation (4) ensure that \(W_k \psi_i\) is strictly monotonic in the vicinity of the optimal solution. When the condition is met, \(W_k\) could be a general regressor for the \(k\)-th subset. Since Equation (3) is an NP-hard problem [22], proposed some constraint conditions to approximately solve the problem:

\[
\Delta t_k \Delta y_k > 0, \forall \Delta t_k \in U_k
\]

\[
\Psi_k \psi_i > 0, \forall \Delta t_k \in U_k
\]

where \(\Delta t_k\) and \(\Psi_k\) are matrices composed of \(\Delta t\) and \(\psi\) in each row.

If we partition the subsets with Equations (5) and (6) directly, the number of subsets will increase exponentially according to the dimensions of \(\Delta t\) and \(\psi\). Therefore, Xiong and Torre [22] proposes that, firstly use principal component analysis (PCA) to extract subspaces of \(\Delta t\) and \(\psi\) in the training set, then according to the first 2D subspace of \(\Delta t\) and the first-dimensional subspace of \(\psi\), divide the samples into \(2^{d+1}\) subsets. The PCA-based method is only suitable for features which follow the normal distribution, but face shapes would not follow the condition. Therefore, it is difficult to achieve ideal partition result. In addition, the ground truth shapes are unknowable during the test, so the PCA-based method could not apply to the alignment of static face images.

3 TWO-STAGE CASCADE REGRESSION MODEL

In this section, we will introduce the two-stage cascaded regression model, which aims to achieve more accurate alignment results by generating coarse aligned initial shapes. The first stage is used to predict salient shapes, and the other stage is used to further predict full shapes. A 3D fitting method is used between the two stages to generate a coarse full shape through fitting and projection. In addition, to improve the robustness of the algorithm across poses, we add a partition method based on fusion subspace to the first stage, dividing the samples into multiple subsets, and training the feature mapping functions and regressors separately for each subset. Figure 2 shows the main work of our proposed method.

3.1 Salient shape

In 300-W dataset [25], each face has 68-point annotation, of which the first 17 points mark the face contour, and the other 51 points mark the eyebrows, eyes, nose and mouth, respectively. Although the shape of the human face is complicated, only a few landmarks play a key role in face alignment. [20] proposed that only a few salient landmarks (such as the centre of the eyes, the corners of the mouth) can be reliably characterized by their image appearances, while most of the other landmarks (such as the face contour) need to rely on the constraints of the whole shape. Therefore, in this paper, we call all the landmarks of a sample as full shape \(s\), and the set of salient landmarks of a sample as salient shape \(\tilde{s}\).

Many face alignment methods based on deep convolutional networks [9, 11] only predicted five salient landmarks, including the centre of the eyes, the tip of the nose, and the corner of the mouth. This salient shape can greatly reduce the computational cost and achieve more accurate alignment result. When we follow the 5-point shape in [11] as the salient shape, and use the 3D fitting method in Section 3.3 to generate a 68-point face shape, the result is shown in Figure 3(a). It can be seen that the full shape generated from the 5-point shape can achieve rough alignment of the eyebrows, eyes, nose and mouth, and can also approximately perform the posture of the face. However, since there are no landmarks at the contour part in 5-point shape to constrain the full shape, all the contour landmarks of the full shape perform poor alignment result.

To solve this problem, we enrich the salient shape from 5 points to 14 points. The 14-point shape contains four points at the corners of the eyes, three points at the nose and wings, four points at the edge of the lips and three points at the contour of face. The full shape generated by the new salient shape is shown in Figure 3(b). It can be apparently seen that the alignment result is more accurate, and the landmarks at the contour of the face can also be roughly aligned, which will reduce the difficulty of

FIGURE 1 The encoding process of shape-index features and local binary features

[Diagram showing the encoding process of shape-index features and local binary features]
further accurate alignment. In Section 4.1, we will use experimental data to prove that the 14-point shape can generate the best initial shape for the full shape alignment.

3.2 Partition method based on fusion subspace

Xiong et al. [22] proposed a method of divide samples into multiple subsets according to the PCA subspace. The constraint conditions in Equations (5) and (6) can be understood as if the $i$-th and $g$-th samples have the same sign in each dimension of the PCA subspace, they will belong to the same subset. This strategy is as follows:

$$\text{sign}(\psi_{i,j}) = \text{sign}(\psi_{g,j}), \forall i, g \in U_{i,j}, j = 1:m$$  (7)

$$\text{sign}(\Delta s_{i,j}) = \text{sign}(\Delta s_{g,j}), \forall i, g \in U_{i,j}, j = 1:n$$  (8)

They can be seen that this partition method separates the shape-index features and the shape residuals, ignoring the correlation between them. Inspired by [26], we propose a method of projecting the shape-index features to the fusion subspace that has a strong correlation with the shape residuals, and dividing the samples according to the subspace.

We treat the shape-index features and the shape residuals as two different features, re-denote the shape-index feature $\psi$ as $x \in \mathbb{R}^m$, and the shape residual $\Delta s$ as $y \in \mathbb{R}^n$. The sets of these two features of all samples in the training set can be denoted respectively as $X_{\times N \times m} = [x_1, \ldots, x_i, \ldots, x_N]$, $Y_{\times N \times n} = [y_1, \ldots, y_i, \ldots, y_N]$, where $N$ is the total number of training set samples, $m$ and $n$ are the dimensions of the two features. Assuming that there are two projection matrix $P$ and $Q$, where

$$P_{\times m \times r} = [p_1, \ldots, p_j, \ldots, p_r]^T, p_j \in \mathbb{R}^m$$  (9)

$$Q_{\times n \times r} = [q_1, \ldots, q_j, \ldots, q_r]^T, q_j \in \mathbb{R}^n$$  (10)

they project the matrices $X$ and $Y$ into subspaces with same dimension respectively, and ensure that (1) there is maximum correlation between two subspaces; (2) the samples belonging to the same subset have the same sign (positive and negative) on the same dimension of the two subspaces. At this time, even if $Y$ (shape residual) is unknown during the test, we can only use the projection subspace of the matrix $X$ to partition. This subspace combines the characteristics of the two features, we call it the fusion subspace.

If the projections of the matrices $X$, $Y$ along the projection direction $p_j$, $q_j$ are denoted as $\zeta_j = Xp_j = [u_{1,j}, \ldots, u_{i,j}, \ldots, u_{N,j}]^T$, $w_j = Yq_j = [v_{1,j}, \ldots, v_{i,j}, \ldots, v_{N,j}]^T$, the problem of using the fusion subspace to divide the samples can be expressed as:

$$\min_{P,Q} \sum_{j=1}^{r} \|Yq_j - Xp_j\|_2 = \min_{P,Q} \sum_{j=1}^{r} \|w_j - \zeta_j\|_2$$  (11)

$$\text{s.t.} \sum_{j=1}^{r} \sum_{i=1}^{N} \text{sign}(v_{i,j}u_{i,j}) = rN$$  (12)
Equation (11) means to minimise the difference between the projections of the matrices \( \mathbf{X} \) and \( \mathbf{Y} \), and Equation (12) ensures that the signs of the projections of the two matrices are consistent.

When the vector \( z_j \) and \( w_j \) are normalised, which means the mean value of \(-w_j \) is 0, where \( i = 1, ..., N \), and the standard deviation is \( 1/N \), Equation (11) can be simplified as Equation (13):

\[
\min_{\mathbf{P}, \mathbf{Q}} \sum_{j=1}^{r} \left\| w_j - z_j \right\|_2^2 = \max_{\mathbf{P}, \mathbf{Q}} \sum_{j=1}^{r} w_j^\top z_j
\]  

(13)

according to the analysis in [26], after normalisation,

\[
\frac{1}{N} \sum_{i=1}^{N} \text{sign}(v_{ij}w_{ij}) \approx w_j^\top z_j
\]  

(14)

which means when \( w_j^\top z_j \) takes the maximum value 1, the constraint condition Equation (12) will be satisfied. Therefore, when the condition of standardisation is satisfied, Equations (11) and (12) will hold.

When extracting shape-index feature matrix \( \mathbf{X} \), we choose PHOG [27] as feature descriptor, which is less disturbed by noise and rotation than HOG and doesn’t need training like LBF. After extracting \( \mathbf{X} \) and \( \mathbf{Y} \), we remove means and divide standard deviation for normalisation. Then we use canonical correlation analysis (CCA) to calculate the two projection directions \( \mathbf{p}_1, \mathbf{p}_2 \) that make the two normalised matrices have maximum correlation, and continue to look for another pair of projection directions that can achieve the maximum correlation but has no correlation with the previous one. Repeat \( r \) times until the projection matrix \( \mathbf{P}_{\mathbf{X}, \mathbf{Y}} \) and \( \mathbf{Q}_{\mathbf{X}, \mathbf{Y}} \) are completed. Finally, we only use \( \mathbf{P} \) to project the fusion subspace and divide all samples into subsets according to Equations (7) and (8). This process is shown in Algorithm 1.

\section*{3.3 3D fitting method}

In this paper, we propose a 3D fitting method to generate initial shape for the full shape alignment. Specifically, we use the first-stage CRM to predict the salient shape firstly, then fit the 3D salient shape to the face image according to the aligned shape, and retain the affine transformation parameters. At last we perform the same affine transformation on the 3D full shape and project it on the 2D plane, obtain a 2D full shape that roughly fitting the face image.

3D Morphable Model (3DMM) [28] is a widely used 3D face model. In this paper, we use 3DMM provided by the BMF dataset [29] for 3D fitting. The model contains a neutral face shape composed of a dense set of 3D points, and some feature matrices that represent shape, expression and texture information, which obtained by principal component analysis of 3D face shapes sampled from the real world. By modifying the weight parameters corresponding to these feature matrices, the model can be fitted as different face shapes.

We only use the neutral face shape of the 3DMM in our 3D fitting method. In addition to this, according to the points required for face alignment, we extract a set of 14 points and the other set of 68 points from the dense 3D point set, denoted as 3D salient shape \( \tilde{s}_{3d} \in \mathbb{R}^{14 \times 3} \) and 3D full shape \( s_{3d} \in \mathbb{R}^{68 \times 3} \). In order to fit \( s_{3d} \) to the face image, we project \( \tilde{s}_{3d} \) onto the 2D plane by weak perspective projection [30]:

\[
\tilde{s}_{2dp} = f \mathbf{P}_{\mathbf{X}} \mathbf{R}(\alpha, \beta, \gamma)(\tilde{s}_{3d} + t_{3d})
\]  

(15)

where \( \tilde{s}_{2dp} \in \mathbb{R}^{14 \times 2} \) is the projection of \( \tilde{s}_{3d} \) onto the 2D plane, \( f \) is scale factor, \( \mathbf{P}_{\mathbf{X}} \) is the orthographic projection matrix \( \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \), \( \mathbf{R}(\alpha, \beta, \gamma) \) is the 3X3 rotation matrix composed of pitch angle \( \alpha \), yaw angle \( \beta \) and roll angle \( \gamma \), \( t_{3d} \) is the translation vector.

We denote the aligned 2D salient shape which outputs from first-stage CRM as \( \tilde{s}_{2d} \in \mathbb{R}^{14 \times 2} \). By minimising the Euclidean distance between \( \tilde{s}_{2dp} \) and \( \tilde{s}_{2d} \), we can get the appropriate values of the parameters \( f, \mathbf{R} \) and \( t_{3d} \) that make \( \tilde{s}_{2dp} \) fit the face image:

\[
\arg \min_{f, \mathbf{R}, t_{3d}} \left\| \tilde{s}_{2d} - \tilde{s}_{2dp} \right\|
\]  

(16)

After calculating the parameters through the least-squares method, we replace \( \tilde{s}_{3d} \) in the Equation (16) with the 3D full shape \( s_{3d} \in \mathbb{R}^{68 \times 3} \), so that we can get the projection of the \( s_{3d} \) fitted to the image plane, which is denoted as \( s_{2dp} \in \mathbb{R}^{68 \times 2} \).

The first column in Figure 4 shows the projection \( s_{2dp} \) obtained through the above method. Note that the contour landmarks on one side have a large deviation. This is because when the pose changes, the contour landmark which marked on the 3D model will be occluded by the cheek, but the real contour landmarks should be located at the boundary of the cheek [31]. Puts forward an assumption for this phenomenon: when the pose changes, if the contour landmark is visible, it will not move; otherwise, it will move in a direction parallel to the visible boundary point. Based on this assumption, we establish the point subsets for 16 contour landmarks (excluding the bottom contour landmark), containing all points parallel to themselves. Using the yaw angle obtained in the above fitting process, it can
FIGURE 4 Process of the contour points movement. (a) Original projected shape. (b) Fitted 3D model. The yellow lines represent the parallel line of the contour points. (c) Shape after movement

ALGORITHM 2 Generate initial shapes by 3D fitting method

**Input:** 3D salient shapes \( \tilde{s}_{3d} \), 3D full shapes \( s_{3d} \), aligned 2D salient shapes \( s_{2d} \), parallel point subsets of contour landmarks \( X_i = \{ (x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots \} \), \( l = 1, ..., 8, 10, 17 \).

**Output:** fitted full shapes \( s_{2d} \).

1. Calculate the projection \( s_{2d}^p \) according to Eq. 15.
2. Optimize Eq. 16 using the least square method, get the affine transformation parameters \( f, \alpha, \beta, \gamma \) and \( t_{3d} \).
3. Perform the same affine and projection transformation:
   \[ s_{2d}^p = fP \cdot R(\alpha, \beta, \gamma)(s_{3d}^p + t_{3d}). \]
4. If \( \beta > 0 \) then
   5. Face to the left
   6. Find the points with minimum \( x \)-coordinate in \( X_i, l = 1, ..., 8 \), replace the left contour points \( \{ (x_{l1}, y_{l1}), (x_{l2}, y_{l2}), \ldots \} \) in \( s_{2d}^p \), output the result after replacement.
   7. Else
   8. Face to the right
   9. Find the points with maximum \( x \)-coordinate in \( X_i, l = 10, ..., 17 \), replace the right contour points \( \{ (x_{r1}, y_{r1}), (x_{r2}, y_{r2}), \ldots \} \) in \( s_{2d}^p \), output the result after replacement.

be judged whether the face is to the left or the right. When the face is to the left, the eight left contour landmarks search for the point with the minimum \( x \)-coordinate in the corresponding point subset, respectively, and these points will be the new contour landmarks; when the face is to the right, the eight right contour landmarks search for the point with maximum \( x \)-coordinate. The complete 3D fitting process is shown in Algorithm 2.

3.4 Summary of algorithm

This section is a summary of 2-SCRM. The method in this paper includes three parts: alignment of salient shapes, generate initial shapes and alignment of full shapes. Algorithm 3 shows the training flow of this algorithm, which corresponds to Figure 2. The testing flow is similar to training. First, extract the PHOG-based shape-index features, project the feature matrix into the fusion subspace with the projection matrix \( P \) obtained in the training phase, and divide the samples into different subsets according to the subspace; then predict the salient shapes according to the corresponding feature mapping function and the regressor, and generate initial full shapes with 3D fitting method; finally, based on the initial shapes, predict the full shapes with the trained feature mapping function and the regressor.

ALGORITHM 3 Training process of 2-SCRM

**Input:** face images and corresponding ground truth shapes in the training set \( \{(I_i, \tilde{s}_i)\}, i = 1 \ldots N \), number of iterations \( T \), number of subsets \( K \).

**Output:** projection matrix of fusion subspace \( P \), feature mapping functions and regressors for salient shapes \( \{ (\Phi_k, W_k) \}, k = 1 \ldots K \), feature mapping function and regressor for full shapes \( (\Phi, W) \).

1. get subsets \( \{U_1, U_2, \ldots, U_K\} \), projection matrix \( P \leftarrow \text{PartitionMethod}(\{(I_i, \tilde{s}_i)\}) \) using Algorithm 1.
2. for \( t = 1 \rightarrow T \)
3. for \( k = 1 \rightarrow K \)
4. Train a feature mapping model based on random forest \( \Phi_k^T \) according to \( \{ (I_g, \tilde{s}_g, s^0_g) \}, \forall g \in U_k \), calculate \( W^T_k \) according to Eq. 2.
5. update \( s^0_g \) according to Eq. 1.
6. \( s^0_i \leftarrow \text{FittingMethod}(s^T_i) \) using Algorithm 2
7. for \( t = 1 \rightarrow T \)
8. Train a feature mapping model based on random forest \( \Phi^T \) according to \( \{ (I_i, \tilde{s}_i, s^0_i) \} \), calculate \( W^T \) according to Eq. 2.
9. update \( s_i \) according to Eq. 1.

4 EXPERIMENTS AND EVALUATION

The two-stage cascade regression model includes salient shape alignment based on partition method, initial shape generation by 3D fitting and full shape alignment. In this section, we will introduce some details of the experiments, evaluate the above three parts, and compare out method with state-of-art methods.

Datasets. The samples used in this paper come from two widely used face alignment datasets: HELEN dataset [32] and 300-W dataset [25]. These data sets contain unconstrained face images with varying pose, occlusion, and illumination, which are challenging.

HELEN dataset [32]: The images in this dataset come from the Internet and contain 2330 high-resolution facial images, of which 2000 are training samples and 330 are test samples. The original dataset labels 194 landmarks for each sample. In our experiment, we use 68 landmarks provided by Sagonas, et al. [25].

300-W dataset [25]: This dataset contains the existing LFPW [33], AFW [34], HELEN [32] and XM2VTS [35], as well as the newly added IBUG dataset, which is more challenging. The
facial image in 300-W is labeled with 68 landmarks. Refer to the partition method as [20, 21], we combine the training samples of HELEN, the training samples of LFPW and the entire AFW into the training set, a total of 3148 images; merge the test samples of HELEN and LFPW into the common subset, a total of 554 images; set 135 images from IBUG as the challenge subset; two subsets constitute the full set, a total of 689 images.

Implementation details. In the training phase, we use similar data augmentation as in [20, 21]. All samples are flipped horizontally, and then 10 face shapes are randomly selected from other training samples as initial shapes. The total number of samples increased by 20 times. This sample augmentation method can also improve the robustness of the model across poses. In the test phase, the first-stage model still takes the mean face shape as the initial shape, while the second-stage model takes the 3D fitting results as the initial shape.

In the experiment of this paper, the number of decision trees in each random forest $G = 10$, the depth of trees $D = 5$, and the number of iterations $T = 7$. Because the shape space of the second-stage model is smaller than that of the first stage, the two models adopt different ranges when selecting pixel difference features. As the number of iterations increases, the radius of the first stage decreases from 0.4 to 0.08, and the radius of the second stage decreases from 0.3 to 0.06 (normalised by the face bounding boxes).

Evaluation criteria. We use the same metric as most face alignment algorithms, called normalised mean error (NME). The specific calculation process is:

$$NME = \frac{1}{N \times L} \sum_{i=1}^{N} \frac{\|s_i - \hat{s}_i\|^2}{d_{ipd}} \tag{17}$$

where $N$ is the total number of samples; $L$ is the number of landmarks of face shape; $s_i$ and $\hat{s}_i$ is the predicted shape and ground truth shape of the $i^{th}$ sample; $d_{ipd}$ is the Euclidean distance between the eyes centre of the $i^{th}$ sample.

4.1 Evaluation of different salient shape

Before we begin, it is necessary to determine which landmarks are used as the salient shape. Based on the 5-point shape of eye centre, nose tip and mouth corner, some landmarks are added to the eyes, nose, mouth and contour, respectively, and six salient shapes are selected, as shown in Figure 5. In order to select the most suitable one from these shapes, we evaluate it from two aspects: (1) based on each salient shape training one-stage CRM, the NME of each model on the test samples is denoted as the alignment error; (2) using the alignment results obtained in (1), the full shape is generated according to the 3D fitting method in Section 3.3. The NME between the generated full shape and the ground truth shape is denoted as fitting error. Figure 5 shows the alignment and fitting errors of the six salient shapes on the Helen dataset according to the above criteria. At the same time, in order to prove the advantages of the salient shape, the alignment and fitting errors according to the full shape are also marked in the figure.

According to the alignment error curve and fitting error curve in Figure 5, compared with 5-point shape and 14-point shape, the 8-point shape and 16-point shape increase 3-point and 2-point, respectively, in the contour part, and the alignment error is also increased, which indicates that the contour shape cannot be well characterized by the image appearances. However, the fitting error of 8-point shape is much less than that of 5-point shape, which means that adding a small number of contour landmarks can constrain the whole shape and reduce the fitting error. Compared with their previous shape, the 8-point, 10-point, 12-point and 14-point shapes have, respectively, added 3 contour landmarks, changed two binocular centre landmarks to four binocular corner landmarks, added two nose landmarks and two lip landmarks. With the increase of landmarks, the fitting error is also gradually reduced, which proves that adding landmarks of eyes, nose and mouth can get better fitting results. However, considering the computational complexity, we do not continue to try to add more landmarks.

From Figure 5, it can also be seen that the alignment error of the full shape (68-point shape) is higher than that of all salient shapes, indicating that a large number of non-critical landmarks will make alignment more difficult. Among the six selected salient shapes, the alignment and fitting errors of the 14-point shape are minimal, which are 27.23% and 12.44% lower than those of 5-point shape. It proves that the salient shape can not only improve the alignment accuracy, but also prevent the poor fitting of the 3D model caused by the mutual constraint of too many landmarks. In the follow-up experiments, the 14-point shape was used as salient shape.

4.2 Evaluation of partition method

In order to prove the effectiveness of the proposed partition method based on fusion subspace, we will evaluate it from two
Due to the manifold distribution of face images, effective feature subspace should also satisfy this characteristic. According to the method in Section 3.2, the shape-index features of 300-W dataset are extracted and projected into the fusion subspace. At this time, the distribution of some face images in the first two dimensions of the fusion subspace is shown in Figure 6. It is not difficult to see that along the horizontal and vertical direction of the figure, the yaw and roll angle of the faces gradually change. The pose rotation angle of the faces distributed in the centre of the image is small, mainly the front face; while the pose rotation angle of the faces distributed around is gradually increased according to the change trend of the horizontal and vertical directions. It proves that the fusion subspace satisfies the manifold distribution characteristics of human face, and can be used to partition the faces with various pose into different subsets.

In order to evaluate the influence of the number of sample subsets $K$ on the alignment results of the first-stage CRM, we calculate the NME of the output from the model with various $K$ values on different datasets. The specific results are shown in Figure 7. Where $K = 1$ means training a single regression model without partition. It can be seen that on the Helen dataset, the NME decreases when $K = 1$ to 4, specifically from 4.46% to 4.32%, which proves that the partition method can effectively reduce the alignment error. However, when $K = 8$ and 16, the NME increases again, and when $K = 16$, it even exceeds the case of no partition. The reasons may be as follows: first, with the increase of the number of subsets, the number of training samples in each subset decreases, resulting in the decline of robustness of the model; second, most of the pose rotation angles in Helen dataset are small, when the number of subsets is too large, it will produce confusing partition result. On 300-W challenge subset, which has more faces with different poses, the proposed partition method can significantly reduce the alignment error when $K = 1$ to 8, specifically from 13.83% to 11.48%. When $K = 16$, the NME rises slightly, and the reason is similar to that of Helen dataset. According to the experimental data, when our algorithm compares with others, the values of $K$ on Helen and 300-W dataset are taken as 4 and 8, respectively.

In order to prove that the fusion subspace-based partition method has more advantages than the PCA subspace-based in [22], we compare the alignment error of salient shape on 300-W challenge subset when $K = 2, 4$ and 8, as shown in Figure 8. The position of the dotted line in the figure is the alignment error when the $K = 1$. It can be seen that the two partition methods can both reduce the alignment error, but the partition method based on fusion subspace is more effective. Since the method in [22] cannot be applied to static faces, only the shape-index feature is used in this experiment to realise the PCA subspace-based partition method. This method ignores the correlation between the shape-index feature and the shape residual, so the partition result is confusing; while proposed method combines two features, it can divide the samples more accurate and achieve lower alignment error with each value of $K$. 

FIGURE 6  Face image distribution map in 2D subspace

FIGURE 7  The NME of salient shape alignment on different datasets with different values of $K$. (a) compared on Helen dataset; (b) compared on 300-W challenge subset
4.3 | Evaluation of two-stage cascade regression model

To prove the two-stage cascade regression model is better than the single one, we compare the full shape alignment results of the original LBF and our method with different $K$ values on the Helen and 300-W dataset.

Because some of the parameters used in our experiments are different from those in [21], there are certain differences between the experimental data in Table 1 and the data published in the original paper. We call the implemented algorithm as LBF Implement. It can be seen from Table 1 that on all datasets, the performance of the two-stage cascaded regression model without partition ($K = 1$) is better than that of unimproved LBF, which proves that the two-stage model is better than the one-stage model. After adding the fusion subspace-based partition method, as the value of $K$ increases, the alignment error on the Helen dataset and 300-W challenge set gradually decreases, and the NMEs are reduced to the lowest when $K = 4$ and $K = 8$. Compared with the original LBF, they were reduced by 4% and 23.55%, respectively. Due to the experimental results in Section 4.2, the range of $K$ is only taken to 4 and 8. On 300-W common subset, there is little difference among the NMEs of the two-stage cascaded regression model with different $K$ values. This is mainly due to the fact that the 300-W common subset is mainly composed of frontal face images, which are densely distributed in the centre of the full test set and close to the boundary of the partition. Therefore, when using the proposed partition method, some frontal faces may be divided into inappropriate subsets. Nevertheless, the alignment error of the 300-W full set will decrease with the increase of $K$ value, which is 11.24% lower than the original algorithm.

Figure 9 shows the NME of each iteration of the second-stage CRM on 300-W dataset with different $K$ values. The results of the 7th iteration are consistent with the data in Table 1. It can be seen that the accurate initial shape can make the second-stage model have an advantage in the first iteration, and can keep this advantage until the last iteration, resulting in better alignment.

![Figure 8](image1.png) **Figure 8** The NME of salient shape alignment on 300-W dataset with different partition strategy

![Figure 9](image2.png) **Figure 9** (a) The NME of the full shape alignment on HELEN dataset with different $K$ values; (b) the NME of the full shape alignment on 300-W challenge subset with different $K$ values

| Method        | HELEN | 300-W   |
|---------------|-------|---------|
|               | Common Subset | Challenge Subset | Full Set |
| LBF implement | 4.84  | 4.67    | 14.44 | 6.58 |
| 2-SCRM ($K = 1$) | 4.76  | 4.57    | 13.23 | 6.24 |
| 2-SCRM ($K = 2$) | 4.72  | 4.60    | 12.05 | 6.06 |
| 2-SCRM ($K = 4$) | 4.62  | 4.59    | 11.15 | 5.87 |
| 2-SCRM ($K = 8$) | —     | 4.58    | 10.94 | 5.82 |
TABLE 2  The NME of the state-of-art methods on HELEN and 300-W datasets (%)

| Method  | HELEN Common Subset | Challenge Subset | Full Set |
|---------|---------------------|------------------|---------|
| ESR [20] | 5.68 17.00 | 7.58 |
| SDM [19] | 5.50 15.40 | 7.50 |
| CFAN [10] | 5.53 9.98 | 6.32 |
| LBF [21] | 4.95 11.98 | 6.32 |
| CFSS [23] | 4.63 9.58 | 5.76 |
| 3DDFA [13] | 6.15 10.59 | 7.01 |
| PIFA-S [14] | 5.43 9.88 | 6.30 |
| MCO [18] | 5.02 12.30 | 6.36 |
| 2-SCRM | 4.62 10.94 | 5.82 |

4.4  | Compare with state-of-art methods

In order to compare with the state-of-art methods, we take the NMEs, cumulative errors distribution (CED) curves and specific alignment results of each method on different datasets as the evaluation criteria.

Table 2 shows the NME of different methods on Helen and 300-W datasets. These data are from the original paper of the method. Among them, ESR, SDM, LBF, CFSS and MCO are methods based on cascade regression, CFAN, 3DDFA and PIFA-S are based on deep learning.

From the data in Table 2, it can be seen that the proposed method has better performance among all the comparison methods on Helen dataset and 300-W common subset, and even better than the three methods based on deep learning. The reason may be that CFAN only changes the feature mapping function to deep autoencoder networks, which is still a one-stage cascade regression in essence; 3DDFA and PIFA-S are based on 3DMM, and the 3D face alignment results are not dominant under the evaluation criteria of 2D face alignment. This also proves that the alignment process of our model from coarse to fine improves the accuracy of the method. On the 300-W challenge subset, the proposed method ranks second in the cascade regression-based methods, second only to CFSS. At the same time, compared with the two methods based on deep learning, our method can produce competitive results. CFSS selects the candidate shapes by calculating the probability matrix, and fuses the updating results of multiple candidate shapes to generate new shapes, which improves the alignment accuracy at the cost of large amount of calculation; 3DDFA and PIFA-S both use convolutional neural network to optimise 3DMM parameters to achieve 3D face alignment, which has strong robustness to pose and occlusion changes. On 300-W full set, our method is second only to CFSS, and better than other methods based on cascade regression and two methods based on deep learning. Therefore, our method has a good competitive advantage compared with the state-of-art methods.

Figure 10 shows the CED curves of the proposed method, SDM, LBF and CFSS on 300-W full test set. The NME of implemented SDM is 7.04%, which is better than 7.50% published in the paper. For implemented LBF, the NME is 6.59%, which is slightly worse than the original 6.32%. This is because we reduce the number of decision trees from 1200 to 680, in order to prevent excessive memory consumption caused by the high dimension of features. We use the code on GitHub provided by the original author to implement CFSS, and the NME is 6.00%, it can be seen that our method outperforms the other methods.

Figure 11 shows the alignment results of our method, SDM, LBF and CFSS on some samples of the 300-W full test set. The first row of the figure is ground truth, and the second to fifth rows are, respectively, the alignment results of our method, CFSS, LBF and SDM. In the first to fourth column, our method can obtain results that are very similar to the ground truth shape, while the other three algorithms have visible errors; in the fifth to eighth column, the results of our method have some errors, while SDM and LBF are basically failed to align. The results of CFSS in the fifth and sixth columns are better than ours, but in the seventh and eighth column where the faces have large pitch angles, it gets worse results than ours. Figure 12 shows more results of our method. According to these results, it can be seen that the proposed method has robustness across poses.

5  | CONCLUSION

In this paper, we propose a new two-stage model for face alignment. Each stage is based on the cascade regression model. The first stage is used to predict the salient shape, and the second stage is used to further predict the full shape. We propose a new samless partition method based on fusion subspace. Our partition method divides all training samples into several subsets, and then each subset is used to train different cascaded regression sub-models, which together constitute the first stage of the two-stage model. Since the correlation between the two
features is considered in our method, similar samples can be better divided into the same subset, which improves the pose robustness. Based on the aligned salient shape, the full shape is roughly aligned by our 3D fitting method, and then used as the initial shape of the second stage. These initial shapes are already very close to the true shapes, resulting in a effective improvement in subsequent alignment work. The experimental results further prove that the fusion subspace partition method can improve the robustness of the method for pose changes, and the framework of two-stage cascade regression solves the limitation of initial shape to a certain extent. In our 3D fitting method, only a few landmarks are used to extract the pose parameters, while the expression parameters are ignored, which result in the generated initial shapes being less robust to expression changes. In future work, we will consider using depth network to predict the parameters in the 3D fitting process to obtain more accurate results.

ACKNOWLEDGMENTS

This research is supported by Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety (Grant No. 2020ZDSYSKFKT03-2), Postgraduate Research &
Practice Innovation Program of Jiangsu Province (Grant No. KYCX20_1959) and National Natural Science Foundation of China (Grant No. 71971031).

ORCID
Jiahui Zhang https://orcid.org/0000-0001-8860-958X
Lan Di https://orcid.org/0000-0002-7071-4899
Jiachen Liang https://orcid.org/0000-0003-3511-3374

REFERENCES
1. Gundogdu, B., Bianco, M.J.: Collaborative similarity metric learning for face recognition in the wild. IET Image Process. 14(9), 1733–1739 (2020)
2. Barman, A., Dutta, P.: Influence of shape and texture features on facial expression recognition. IET Image Process. 13(8), 1349–1363 (2019)
3. Xu, X., et al.: Weighted similarity and distance metric learning for unconstrained face verification with 3D frontalisation. IET Image Process. 13(2), 399–408 (2019)
4. Cootes, T.F., et al.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23(6), 681–685 (2001)
5. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: Proceedings of the 11th International Conference on Computer Vision, pp. 1–9. IEEE, Piscataway (2007)
6. Lee, H.S., Kim, D.: Tensor-based AAM with continuous variation estimation. Application to variation-robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31(6), 1102–1116 (2008)
7. Dollár, P., et al.: Cascaded pose regression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1078–1085. IEEE, Piscataway (2010)
8. Sun, Y., et al.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483. IEEE, Piscataway (2013)
9. Zhang, J., et al.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: 13th European Conference on Computer Vision, pp. 1–16. Springer, Cham (2014)
10. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23(10), 1499–1503 (2016)
11. Liu, Y., et al.: Dense face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1619–1628. IEEE, Piscataway (2017)
12. Zhu, X., et al.: Face alignment in full pose range: A 3d total solution. IEEE Trans. Pattern Anal. Mach. Intell. 41(1), 78–92 (2017)
13. Jourabloo, A., et al.: Pose-invariant face alignment with a single cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3200–3209. IEEE, Piscataway (2017)
14. Wu, W., et al.: Look at boundary: A boundary-aware face alignment algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2138. IEEE, Piscataway (2018)
15. Zhao, Y., et al.: Mobilefacenet: Transferring deep hidden representation for face alignment. Pattern Recognit. 100, 107114 (2020)
16. Zhu, S., et al.: Unconstrained face alignment via cascaded compositional learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3409–3417. IEEE, Piscataway (2016)
17. Liu, C., et al.: Face alignment via multi-regressors collaborative optimization. IEEE Access 7, 4101–4112 (2018)
18. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539. IEEE, Piscataway (2013)
19. Cao, X., et al.: Face alignment by explicit shape regression. Int. J. Comput. Vis. 107(2), 177–190 (2014)
20. Ren, S., et al.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692. IEEE, Piscataway (2014)
21. Xiong, X., De la Torre, F.: Global supervised descent method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2664–2673. IEEE, Piscataway (2015)
22. Zhu, S., et al.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4998–5006. IEEE, Piscataway (2015)
23. Fan, R.E., et al.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9(Aug), 1871–1874 (2008)
24. Sagonas, C., et al.: A semi-automatic methodology for facial landmark annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 896–903. IEEE, Piscataway (2013)
25. Zhuang, Y., et al.: Sign-correlation partition based on global supervised descent method for face alignment. In: Asian Conference on Computer Vision, pp. 281–295. Springer, Cham (2017)
26. Bai, Y., et al.: A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 3305–3308. IEEE, Piscataway (2009)
27. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. 25(9), 1063–1074 (2003)
28. Paysan, P., et al.: A 3D face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. IEEE, Piscataway (2009)
29. Bruckstein, A.M., et al.: Optimum fiducials under weak perspective projection. Int. J. Comput. Vis. 35(3), 223–244 (1999)
30. Zhu, X., et al.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 787–796. IEEE, Piscataway (2015)
31. Le, V., et al.: Interactive facial feature localization. In: European Conference on Computer Vision, pp. 679–692. Springer, Berlin Heidelberg (2012)
32. Bellumeur, P.N., et al.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. 35(12), 2930–2940 (2013)
33. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE, Piscataway (2012)
34. Messer, K., et al.: Xm2vtsdb: The extended m2vts database. In: Second International Conference on Audio and Video-Based Biometric Person Authentication, Vol. 964, pp. 965–966 Springer-Verlag, Berlin, Heidelberg (1999)

How to cite this article: Zhang J, Di L, Liang J. Face alignment based on fusion subspace and 3D fitting. IET Image Process. 2021;15:16–27.
https://doi.org/10.1049/ipr2.12002