

Joint Face Alignment and 3D Face Reconstruction

Feng Liu¹, Dan Zeng¹, Qijun Zhao¹ and Xiaoming Liu²

¹ College of Computer Science, Sichuan University, Chengdu, China

² Department of Computer Science and Engineering
Michigan State University, East Lansing, MI, U.S.A.

Abstract. We present an approach to simultaneously solve the two problems of face alignment and 3D face reconstruction from an input 2D face image of arbitrary poses and expressions. The proposed method iteratively and alternately applies two sets of cascaded regressors, one for updating 2D landmarks and the other for updating reconstructed pose-expression-normalized (PEN) 3D face shape. The 3D face shape and the landmarks are correlated via a 3D-to-2D mapping matrix. In each iteration, adjustment to the landmarks is firstly estimated via a landmark regressor, and this landmark adjustment is also used to estimate 3D face shape adjustment via a shape regressor. The 3D-to-2D mapping is then computed based on the adjusted 3D face shape and 2D landmarks, and it further refines the 2D landmarks. An effective algorithm is devised to learn these regressors based on a training dataset of pairing annotated 3D face shapes and 2D face images. Compared with existing methods, the proposed method can fully automatically generate PEN 3D face shapes in real time from a single 2D face image and locate both visible and invisible 2D landmarks. Extensive experiments show that the proposed method can achieve the state-of-the-art accuracy in both face alignment and 3D face reconstruction, and benefit face recognition owing to its reconstructed PEN 3D face shapes.

Keywords: Face alignment; 3D face reconstruction; cascaded regression

1 Introduction

Three-dimensional (3D) face models have recently been employed to assist pose or expression invariant face recognition [3, 14, 42], and the state-of-the-art performance has been achieved. A crucial step in these 3D face-assisted face recognition methods is to reconstruct the 3D face model from a two-dimensional (2D) face image. Besides its applications in face recognition, 3D face reconstruction is also useful in other face-related tasks, such as facial expression analysis [7, 36] and facial animation [4, 5]. While many 3D face reconstruction methods are available, they require landmarks on the face image as input, and are difficult to handle large-pose faces that have invisible landmarks due to self-occlusion.

Existing studies tackle the problems of face alignment (or facial landmark localization) and 3D face reconstruction *separately*. However, these two problems are chicken-and-egg problems. On one hand, 2D face images are projections of 3D

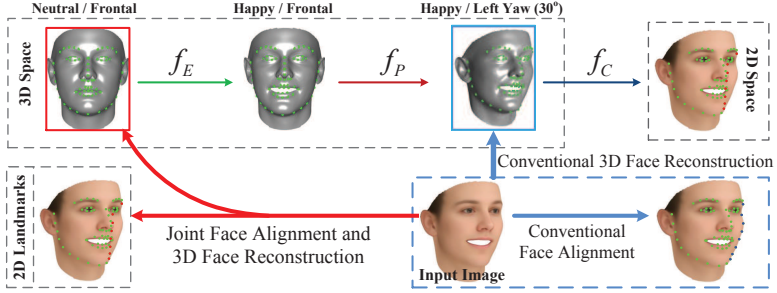


Fig. 1. We view 2D landmarks are generated from a 3D face through 3D expression (f_E) and pose (f_P) deformation, and camera projection (f_C) (top row). While conventional face alignment and 3D face reconstruction are two *separate* tasks and the latter requires the former as the input, this paper performs these two tasks *jointly*, i.e., reconstructing a pose-expression-normalized (PEN) 3D face and estimating visible/invisible landmarks (green/red points) from a 2D face image with arbitrary poses and expressions.

faces onto the 2D plane. Knowing a 3D face and a 3D-to-2D mapping function, it is easy to compute the visibility and position of 2D landmarks. On the other hand, the landmarks provide rich information about facial geometry, which is the basis of 3D face reconstruction. Figure 1 illustrates the correlation between 2D landmarks and the 3D face. That is, the visibility and position of landmarks in the projected 2D image are determined by three factors: the 3D face shape, 3D deformation due to expression and pose, and camera projection parameters. Let us denote a 3D face shape as S and its 2D landmarks as U . The formation of 2D landmarks from the 3D face can be represented by $U = f_C \circ f_P \circ f_E(S)$, where f_C is camera projection, f_P and f_E are deformation caused by pose and expression, respectively. *Given such a clear correlation between 2D landmarks U and 3D shape S , it is evident that they should ideally be solved jointly, instead of separately as in prior works - indeed this is the core of this work.*

Motivated by the aforementioned observation, this paper proposes to simultaneously solve the two problems of face alignment and 3D face shape reconstruction in one unified framework. To this end, two sets of regressors are jointly learned from a training set of pairing annotated 2D face images and 3D face shapes. These two sets of regressors are alternately applied to locate the landmarks on an input 2D image, and meanwhile reconstruct its pose-expression-normalized (PEN) 3D face shape. Note that most single image-based 3D face reconstruction algorithms aim to assist face recognition. For this purpose, we argue that reconstructing the PEN 3D shape is more useful than reconstructing the 3D shape that has the same pose and expression as the input 2D face [23, 28, 31].

The rest of this paper is organized as follows. Section 2 briefly reviews related work in the literature. Section 3 introduces in detail the proposed joint face alignment and 3D face reconstruction method. Section 4 reports experimental results. Section 5 concludes the paper.

2 Related Work

Face Alignment. Classical face alignment methods, including Active Shape Model (ASM) [9, 11] or Active Appearance Model (AAM) [8, 25], search for landmarks based on global shape models and generative texture models. Constrained Local Model (CLM) [10] also utilizes global shape models to regularize the landmark locations, but it employs discriminative local texture models. Regression based methods [6, 27, 35, 39] have been recently proposed to directly estimate landmark locations by applying cascaded regressors to an input 2D face image. These methods mostly do not consider the visibility of facial landmarks under different view angles. Consequently, their performance degrades substantially for non-frontal faces, and their detected landmarks could be ambiguous because the anatomically correct landmarks might be invisible due to self-occlusion (see Fig. 1).

A few methods focused on large-pose face alignment, which can be roughly divided into two categories: multi-view based and 3D model based. Multi-view based methods [37, 40] define different sets of landmarks as templates, one for each view range. Given an input image, they fit the multi-view templates to it and choose the best fitted one as the final result. These methods are usually complicated to apply, and can not detect invisible self-occluded landmarks. 3D model based methods, in contrast, can better handle self-occluded landmarks with the assistance of 3D face models. Their basic idea is to fit a 3D face model to the input image to recover the 3D landmark locations. Most of these methods [17, 18, 41] use 3D morphable models (3DMM) [2] – either a simplified one with a sparse set of landmarks [18, 41] or a relatively dense one [17]. They estimate the 3DMM parameters by using cascaded regressors with texture features as the input. In [18], the visibility of landmarks is explicitly computed, and the method can cope with face images of yaw angles ranging from -90° to 90° , whereas the method in [17] does not work properly for faces of yaw angles beyond 60° . In [33], Tulyakov and Sebe propose to directly estimate the 3D landmark locations via texture-feature-based regressors for faces of yaw angles upto 50° .

These existing 3D model based methods establish regressions between 2D image features and 3D landmark locations (or indirectly, 3DMM parameters). While our proposed approach is also based on 3D model, unlike existing methods, it carries out regressions both on 2D images and in the 3D space. Regressions on 2D images predict 2D landmarks, while regressions in the 3D space predict 3D landmark locations. By integrating both regressions, our proposed method can more accurately locate landmarks, and better handle self-occluded landmarks. It thus works well for images of arbitrary view angles in $[-90^\circ, 90^\circ]$.

3D Face Reconstruction. Estimating the 3D face geometry from a single 2D image is an ill-posed problem. Existing methods, such as Shape from Shading (SFS) and 3DMM, thus heavily depend on priors or constraints. SFS based methods [20, 31] usually utilize an average 3D face model as a reference, and assume the Lambertian lighting model for the 3D face surface. One limitation of SFS methods lies in its assumed connection between 2D texture clues and 3D shape, which is too weak to discriminate among different individuals.

3DMM [2, 3, 28] establishes statistical parametric models for both texture and shape, and represents a 3D face as a linear combination of basis shapes and textures. To recover the 3D face from a 2D image, 3DMM-based methods estimate the combination coefficients by minimizing the discrepancy between the input 2D face image and the one rendered from the reconstructed 3D face. They can better cope with 2D face images of varying illuminations and poses. However, they still suffer from invisible facial landmarks when the input face has large pose angles. To deal with extreme poses, Lee et al. [22], Qu et al. [26] and Liu et al. [23] propose to discard the self-occluded landmarks or treat them as missing data. All these existing 3D face reconstruction methods require landmarks as input. Consequently, they either manually mark the landmarks, or employ standalone face alignment methods to automatically locate the landmarks. Moreover, existing methods always generate 3D faces that have the same pose and expression as the input image, which may not be desired in face recognition due to the challenge of matching 3D faces with expressions [12]. In this paper, we improve 3D face reconstruction from two aspects: (i) integrating the face alignment step into the 3D face reconstruction procedure, and (ii) reconstructing PEN 3D faces, which are believed to be useful for face recognition.

3 Proposed Method

3.1 Overview

We denote an n -vertex 3D face shape of neutral expression and frontal pose as,

$$S = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \\ 1 & 1 & \cdots & 1 \end{pmatrix}, \quad (1)$$

and a subset of S with columns corresponding to l landmarks as S_L . The projections of these landmarks on the 2D face image \mathbf{I} are represented by

$$U = \begin{pmatrix} u_1 & u_2 & \cdots & u_l \\ v_1 & v_2 & \cdots & v_l \end{pmatrix} = f_C \circ f_P \circ f_E(S_L) \approx \mathbf{M} \times S_L. \quad (2)$$

Here, we use a 3D-to-2D mapping matrix \mathbf{M} to approximate the composite effect of expression and pose induced deformation and camera projection. Given an input 2D face image \mathbf{I} , our goal is to simultaneously locate its landmarks U and reconstruct its 3D face shape S . Note that, in some context, we also write the 3D face shape and the landmarks as column vectors: $\mathbf{S} = (x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_n, y_n, z_n)^T$, and $\mathbf{U} = (u_1, v_1, u_2, v_2, \cdots, u_l, v_l)^T$, where ‘ T ’ is transpose operator.

Figure 2 shows the flowchart of the proposed method. For the input 2D face image \mathbf{I} , its 3D face shape \mathbf{S} is initialized as the mean 3D shape of training faces. Its landmarks \mathbf{U} are initialized by fitting the mean landmarks of training frontal faces into the face region specified by a bounding box in \mathbf{I} via similarity transforms. \mathbf{U} and \mathbf{S} are iteratively updated by applying a series of regressors. Each iteration contains three main steps: (i) updating landmarks, (ii) updating 3D face shape, and (iii) refining landmarks.

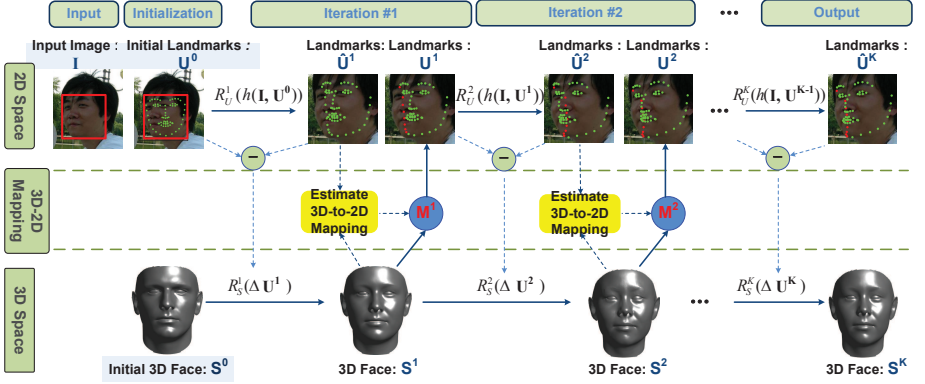


Fig. 2. Flowchart of our joint face alignment and 3D face reconstruction method.

Updating landmarks This step updates the landmarks' locations from \mathbf{U}^{k-1} to $\hat{\mathbf{U}}^k$ based on the texture features in the input 2D image. This is similar to the conventional cascaded regressor based 2D face alignment [35]. The adjustment to the landmarks' locations in k^{th} iteration, $\Delta \mathbf{U}^k$ is determined by the local texture features around \mathbf{U}^{k-1} via a regressor,

$$\Delta \mathbf{U}^k = R_U^k(h(\mathbf{I}, \mathbf{U}^{k-1})), \quad (3)$$

where $h(\mathbf{I}, \mathbf{U})$ denotes a texture feature extracted around the landmarks \mathbf{U} in the image \mathbf{I} , and R_U^k is a regression function. The landmarks can be then updated by $\hat{\mathbf{U}}^k = \mathbf{U}^{k-1} + \Delta \mathbf{U}^k$. The method for learning these landmark regressors will be introduced in Sec. 3.3.

Updating 3D face shape In this step, the above-obtained landmark location adjustment is used to estimate the adjustment of the 3D face shape. Specifically, a regression function R_S^k models the correlation between the landmark location adjustment $\Delta \mathbf{U}^k$ and the expected adjustment to the 3D shape $\Delta \mathbf{S}^k$, i.e.,

$$\Delta \mathbf{S}^k = R_S^k(\Delta \mathbf{U}^k). \quad (4)$$

The 3D shape can be then updated by $\mathbf{S}^k = \mathbf{S}^{k-1} + \Delta \mathbf{S}^k$. The method for learning these shape regressors will be given in Sec. 3.4.

Refining landmarks Once a new estimate of the 3D shape is obtained, the landmarks can be further refined accordingly. For this purpose, the 3D-to-2D mapping matrix is needed. Hence, we estimate \mathbf{M}^k based on \mathbf{S}^k and $\hat{\mathbf{U}}^k$. The refined landmarks \mathbf{U}^k can be then obtained by projecting \mathbf{S}^k onto the image via \mathbf{M}^k according to Eq. (2). During this process, the visibility of the landmarks is also re-computed. Details about this step will be given in Sec. 3.5.

3.2 Training Data Preparation

Before we provide the details about the three steps, we first introduce the training data needed for learning the landmarks and 3D shape regressors. Since the purpose of these regressors is to gradually adjust the estimated landmarks and 3D shape towards their true values, we need a sufficient number of triplet data $\{(\mathbf{I}_i, \mathbf{S}_i^*, \mathbf{U}_i^*) | i = 1, 2, \dots, N\}$, where \mathbf{S}_i^* and \mathbf{U}_i^* are, respectively, the ground truth 3D shape and landmarks for the image \mathbf{I}_i , and N is the total number of training samples. All the 3D face shapes have been established dense correspondences among their vertices; in other words, they have the same number of vertices, and vertices of the same index have the same semantic meaning. Moreover, both visible and invisible landmarks in \mathbf{I}_i have been annotated and included in \mathbf{U}_i^* . For invisible landmarks, the annotated positions should be anatomically correct positions (e.g., red points in Fig. 1).

Obviously, to make the regressors robust to expression and pose variations, the training data should contain 2D face images of varying expressions and poses. As for the 3D shape \mathbf{S}_i^* corresponding to the \mathbf{I}_i in the training data, it can either have the same expression and pose as \mathbf{I}_i , or just have neutral expression and frontal pose no matter what expression and pose \mathbf{I}_i has. In the former, the learned regressors will output 3D face shapes that have the same expression and pose as the input images; while in the latter, the learned regressors will generate neutral and frontal 3D shapes for any input images. In either case, the dense registration among all 3D shapes \mathbf{S}_i^* is needed for regressor learning. In this paper, we follow the latter for two reasons: (i) dense registration of 3D face shapes with different expressions is difficult, and (ii) the reconstructed PEN 3D shapes are preferred for being used in 3D face recognition.

It is, however, difficult to find in the public domain such data sets of 3D face shapes and corresponding annotated 2D images with various expressions/poses. Thus, we construct two sets of training data by ourselves: one based on BU3DFE [36], and the other based on LFW [16]. BU3DFE database contains 3D face scans of 56 males and 44 females, acquired in neutral plus six basic expressions (happiness, disgust, fear, angry, surprise and sadness). All basic expressions are acquired at four levels of intensity. These 3D face scans have been manually annotated with 84 landmarks (83 landmarks provided by the database and one nose tip marked by ourselves). For each of the 100 subjects, we select one scan of neutral expression as the ground truth 3D shape. For the rest six expressions, we choose the scans of the first level intensity, and project them to 2D images with recorded landmark locations. From each of the seven scans, 19 face images are generated with different poses (-90° to 90° yaw with a 10° interval). As a result, each 3D shape has 133 images of different poses and expressions. We use the method [13] to establish dense correspondence of BU3DFE neutral scans.

LFW database contains 13,233 images of 5,749 subjects. We select 150 subjects, each having at least 10 images, and use 68 landmarks on these face images that are provided by the work of [41]. From the neutral frontal image of each subject, we employ the method in [23] to reconstruct the 3D shape, which is

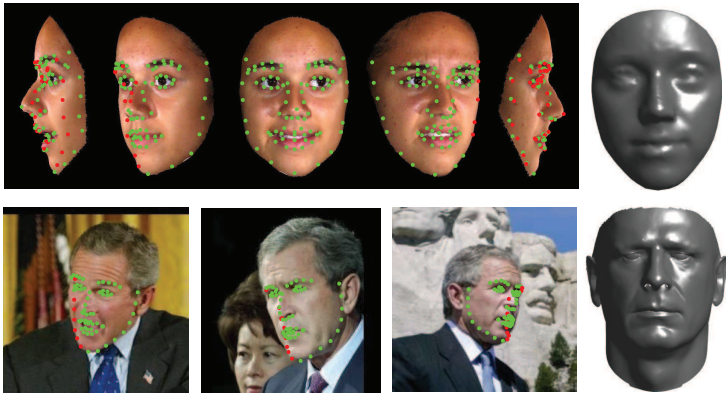


Fig. 3. Example 2D face images with annotated landmarks and corresponding neutral 3D shapes from the BU3DFE and LFW databases.

densely registered. Finally, we obtain 4,149 images of 150 subjects and their corresponding neutral 3D face shapes.

The resultant 3D shapes have $n = 9,677$ for BU3DFE and $n = 53,215$ for LFW. Figure 3 shows some example 2D face images and corresponding 3D faces in the two databases. Obviously, 3D shapes in BU3DFE consist of a sparser set of vertices, and consequently look a little bit blur in Fig. 3.

3.3 Learning Landmark Regressors

According to Eq. (3), landmark regressors estimate the adjustment to \mathbf{U}^{k-1} such that the updated landmarks \mathbf{U}^k get closer to their true positions. In the training phase, the true positions and visibility of the landmarks are given by the ground truth \mathbf{U}^* . Therefore, the objective of the landmark regressors R_U^k is to better predict the difference between \mathbf{U}^k and \mathbf{U}^* . In this paper, we employ linear regressors as the landmark regressors, and learn them by fulfilling the following optimization:

$$R_U^k = \arg \min_{R_U^k} \sum_{i=1}^N \| (\mathbf{U}_i^* - \mathbf{U}_i^k) - R_U^k(h(\mathbf{I}_i, \mathbf{U}_i^{k-1})) \|_2^2, \quad (5)$$

which has a close-form least-square solution. Note that other regression schemes, such as CNN [19], can be easily adopted in our framework.

We use 128-dim SIFT descriptors [24] as the local feature. The feature vector of h is a concatenation of the SIFT descriptors at all the l landmarks, i.e., a $128l$ -dim vector. If a landmark is invisible, no feature will be extracted, and its corresponding entries in h will be zero. It is worth mentioning that the regressors estimate the semantic positions of all landmarks including invisible landmarks.

3.4 Learning 3D Shape Regressors

The landmark adjustment $\Delta \mathbf{U}^k$ is also used as the input to the 3D shape regressor R_S^k . The objective of R_S^k is to compute an update to the initially estimated 3D shape \mathbf{S}^{k-1} in the k^{th} iteration to minimize the difference between the updated 3D shape and the ground truth. Using similar linear regressors, the 3D shape regressors can be learned by solving the following optimization via least squares:

$$R_S^k = \arg \min_{R_S^k} \sum_{i=1}^N \| (\mathbf{S}_i^* - \mathbf{S}_i^k) - R_S^k (\Delta \mathbf{U}_i^k) \|_2^2, \quad (6)$$

with its closed form solution as

$$R_S^k = \Delta \mathbf{S}^k (\Delta \mathbf{U}^k)^\top (\Delta \mathbf{U}^k (\Delta \mathbf{U}^k)^\top)^{-1}, \quad (7)$$

where $\Delta \mathbf{S}^k = \mathbf{S}^* - \mathbf{S}^k$ and $\Delta \mathbf{U}^k$ are, respectively, the 3D shape and landmark adjustment. $\mathbf{S} \in \mathbb{R}^{3n \times N}$ and $\mathbf{U} \in \mathbb{R}^{2l \times N}$ denote, respectively, the ensemble of 3D face shapes and 2D landmarks of all training samples with each column corresponding to one sample. It can be mathematically shown that N should be larger than $2l$ so that $\Delta \mathbf{U}^k (\Delta \mathbf{U}^k)^\top$ is invertible. Fortunately, since the set of used landmarks are usually sparse, this requirement is easy to be satisfied in real-world applications.

3.5 Estimating 3D-to-2D Mapping and Landmark Visibility

In order to refine the landmarks with the updated 3D face shape, we have to project the 3D shape to the 2D image with a 3D-to-2D mapping matrix. In this paper, we dynamically estimate the mapping matrix based on \mathbf{S}^k and $\hat{\mathbf{U}}^k$. As discussed earlier in Sec. 3.1, the mapping matrix is a composite effect of expression and pose induced deformation and camera projection. Here, we assume a weak perspective projection for the camera projection as in prior work [18, 38], and further assume that the expression and pose induced deformation can be approximated by a linear transform. As a result, the mapping matrix \mathbf{M}^k is represented by a 2×4 matrix, and can be estimated as a least squares solution to the following fitting problem:

$$\mathbf{M}^k = \arg \min_{\mathbf{M}^k} \| \hat{\mathbf{U}}^k - \mathbf{M}^k \times S_L^k \|_2^2. \quad (8)$$

Once a new mapping matrix is computed, the landmarks can be further refined as $\mathbf{U}^k = \mathbf{M}^k \times S_L^k$.

The visibility of the landmarks can be then computed based on the mapping matrix \mathbf{M} using the method in [18]. Suppose the average surface normal around a landmark in the 3D face shape \mathbf{S} is $\vec{\mathbf{n}}$. Its visibility \mathbf{v} can be measured by

$$\mathbf{v} = \frac{1}{2} \left(1 + \text{sgn} \left(\vec{\mathbf{n}} \cdot \left(\frac{\mathbf{M}_1}{\|\mathbf{M}_1\|} \times \frac{\mathbf{M}_2}{\|\mathbf{M}_2\|} \right) \right) \right), \quad (9)$$

where $\text{sgn}()$ is the sign function, ‘ \cdot ’ means dot product and ‘ \times ’ cross-product, and \mathbf{M}_1 and \mathbf{M}_2 are the left-most three elements at the first and second row of the mapping matrix \mathbf{M} . This basically rotates the surface normal and validates if it points toward the camera or not.

The whole process of learning the cascaded coupled landmark and 3D shape regressors is summarized in Algorithm 1.

Algorithm 1 Cascaded Coupled-Regressor Learning.

Input: Training data $\{(\mathbf{I}_i, \mathbf{S}_i^*, \mathbf{U}_i^*) | i = 1, 2, \dots, N\}$, initial shape \mathbf{S}_i^0 & landmarks \mathbf{U}_i^0 .

Output: Cascaded coupled-regressors $\{R_U^k, R_S^k\}_{k=1}^K$.

- 1: **for** $k = 1, \dots, K$ **do**
 - 2: Estimate R_U^k via Eq. (5), and compute landmark adjustment $\Delta \mathbf{U}_i^k$ via Eq. (3);
 - 3: Update landmarks $\hat{\mathbf{U}}_i^k$ for all images: $\hat{\mathbf{U}}_i^k = \mathbf{U}_i^{k-1} + \Delta \mathbf{U}_i^k$;
 - 4: Estimate R_S^k via Eq. (6), and compute shape adjustment $\Delta \mathbf{S}_i^k$ via Eq. (4);
 - 5: Update 3D face \mathbf{S}_i^k : $\mathbf{S}_i^k = \mathbf{S}_i^{k-1} + \Delta \mathbf{S}_i^k$;
 - 6: Estimate the 3D-to-2D mapping matrix \mathbf{M}_i^k via Eq. (8);
 - 7: Compute the refined landmarks \mathbf{U}_i^k via Eq. (2) and their visibility via Eq. (9).
 - 8: **end for**
-

4 Experiments

4.1 Protocols

We conduct three sets of experiments to evaluate the proposed method in 3D shape reconstruction, face alignment, and benefits to face recognition.

Datasets The training data are constructed from two public face databases: BU3DFE and LFW, as detailed in Sec. 3.2. Respectively, two different models are trained using each of the two training sets. Our test sets include BU3DFE and AFW (Annotated Faces in-the-Wild) [40]. To evaluate the 3D shape reconstruction accuracy, a 10-fold cross validation is applied to split the BU3DFE data into training and testing subsets, resulting in 11,970 training samples and 1,330 testing samples. To evaluate the face alignment accuracy, the AFW database [40] is tested using the LFW-trained model. AFW is a widely used benchmark in the face alignment literature. It contains 205 images of 468 faces with different poses within $\pm 90^\circ$. In [30], 337 of these faces have been manually annotated with face bounding boxes and 68 landmarks. We use them in our experiments.

Experiment setup During training and testing, each image is associated with a bounding box, which specifies the face region in the image. To initialize the landmarks in it, the mean of the landmarks in all neutral frontal training images is fitted to the face region via a similarity transform. In this paper, we set the number of iterations $K = 5$ (discussion of convergence issue is provided in supplemental material). SIFT descriptors are computed on 32×32 local patches

around the landmarks, and the implementation by [35] is used in our experiments.

Evaluation metrics Two metrics are used to evaluate the 3D face shape reconstruction accuracy: Mean Absolute Error (MAE) and Normalized Per-vertex Depth Error (NPDE). MAE is defined as $\text{MAE} = \frac{1}{N_T} \sum_{i=1}^{N_T} (\|\mathbf{S}_i^* - \hat{\mathbf{S}}_i\|/n)$, where N_T is the total number of testing samples, \mathbf{S}_i^* and $\hat{\mathbf{S}}_i$ are the ground truth and reconstructed 3D face shape of the i^{th} testing sample. NPDE measures the depth error at the j^{th} vertex in a testing sample as $\text{NPDE}(x_j, y_j) = (|z_j^* - \hat{z}_j|) / (z_{\max}^* - z_{\min}^*)$, where z_{\max}^* and z_{\min}^* are the maximum and minimum depth values in the ground truth 3D shape of the testing sample, and z_j^* and \hat{z}_j are the ground truth and reconstructed depth values at the j^{th} vertex.

The face alignment accuracy is measured by Normalized Mean Error (NME). It is defined as the mean of the normalized estimation error of visible landmarks for all testing samples:

$$\text{NME} = \frac{1}{N_T} \sum_{i=1}^{N_T} \left(\frac{1}{d_i} \frac{1}{N_i^v} \sum_{j=1}^l \mathbf{v}_{ij} \|(\hat{u}_{ij}, \hat{v}_{ij}) - (u_{ij}^*, v_{ij}^*)\| \right), \quad (10)$$

where d_i is the square root of the face bounding box area of the i^{th} testing sample, N_i^v is the number of visible landmarks in it, (u_{ij}^*, v_{ij}^*) and $(\hat{u}_{ij}, \hat{v}_{ij})$ are, respectively, the ground truth and estimated coordinates of its j^{th} landmark.

4.2 3D Face Reconstruction Accuracy

Reconstruction accuracy across poses Figure 4(a) shows the average MAE of our proposed method under different pose angles of the input 2D images. To give a fair comparison with the method in [23], we only compute the reconstruction error of neutral testing images, after rotating the reconstructed 3D faces to frontal view. As can be seen, the average MAE of our method is lower than that of the baseline. Moreover, as the pose angle becomes large, the error does not increase substantially. This proves the effectiveness of the proposed method in

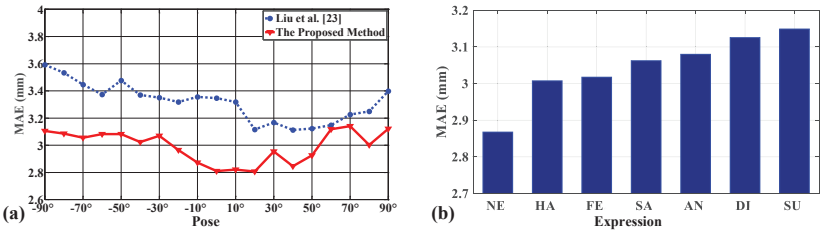


Fig. 4. MAE of the proposed method on BU3DFE (a) under different yaw angles and (b) under different expressions, i.e., neutral (NE), happy (HA), fear (FE), sad (SA), angry (AN), disgust (DI) and surprise (SU).

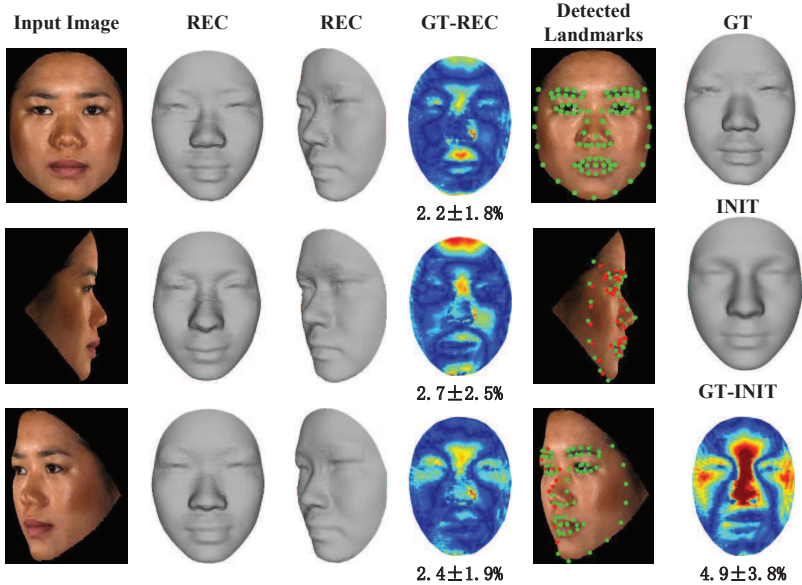


Fig. 5. Reconstruction results for a BU3DFE subject at three different pose angles. Column one are input images. Columns 2 and 3 show the reconstructed (‘REC’) 3D faces from two views. Column 4 are the NPDE between the ground truth (‘GT’) and REC 3D faces. The detected landmarks are shown in Column 5. The last column shows the GT 3D face of this subject, the initial (‘INIT’) 3D face, and the NPDE between them. NPDE increases as the color changes from blue to red. The average and the standard deviation are given below each NPDE map. Note that the same INIT 3D face is used for all input images.

handling arbitrary view face images. Figure 5 shows the reconstruction and face alignment results of one subject.

Reconstruction accuracy across expressions Figure 4(b) shows the average MAE of our proposed method across expressions. Although the error increases as expressions become intensive, the maximum increment (i.e., SU vs. NE) is below 7%. This proves the robustness of the proposed method in normalizing expressions while maintaining model individualities. Figure 6 shows the reconstruction and face alignment results of a subject under seven expressions.

Method	CDM [37]	PIFA [18]	The proposed method
NME	7.52%	5.60%	3.15%

Table 1. NME of the proposed method and two baseline methods on AFW.

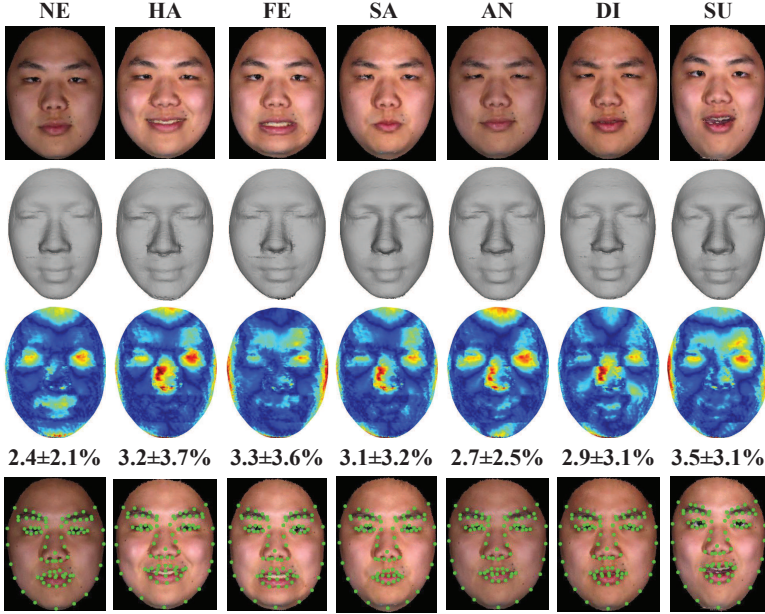


Fig. 6. Face alignment and reconstruction results for a BU3DFE subject with different expressions. Row 1 shows the input images. Row 2 shows the estimated 3D shapes, and Row 3 shows the NPDE maps with the average and standard deviation. The last row shows the detected landmarks.

4.3 Face Alignment Accuracy

As for the face alignment evaluation on AFW, we select two recent works as baseline methods: 1) CDM [37], the first method claimed to perform pose-free face alignment; 2) PIFA [18], a regression-type method that can predict the anatomically correct locations of landmarks for arbitrary view face images. We use the executable code of CDM and PIFA to compute their performance on our test set. The CDM code integrates face detection, and it successfully detects and aligns 268 out of 337 testing images. Therefore, to compare with CDM, we evaluate the NME on the 268 testing images. For PIFA and the proposed method, the face bounding boxes provided by [30] are used. One note is that the CDM detects 66 landmarks and PIFA detects 21 landmarks. For a fair comparison, we evaluate the NME on 18 landmarks that are the *intersections* of the three landmark sets. As shown in Table 1, our accuracy is better than the two baseline methods. Figure 7 shows some face alignment results.

4.4 Application to Face Recognition

While there are many recent face alignment and reconstruction work [1, 15, 21, 29, 32, 34], few work takes one step further to evaluate the contribution of alignment



Fig. 7. Detected 18 landmarks for images in AFW by the proposed method.

or reconstruction to subsequent tasks. In contrast, we quantitatively evaluate the effect of the reconstructed PEN 3D face shapes on face recognition by performing direct 3D to 3D shape matching and fuse it with conventional 2D face recognition. Specifically, we choose 70 subjects in BU3DFE to train the proposed regressors, and use the rest 30 subjects for testing. The neutral frontal face images of the testing subjects compose the gallery, and their faces under 19 poses and 7 expressions (totally 3,990 images) are the probe images. We use a commercial off-the-shelf (COTS) 2D face matcher³ as the baseline. The iterative closest points (ICP) algorithm is applied to match the reconstructed normalized 3D face shapes. It aligns the 3D shapes reconstructed from probe and gallery images, and computes the distances between them, which are then converted to similarity scores via subtracting them from the maximum distance. These scores are finally normalized to the range of $[0, 1]$, and fused with the scores of the COTS matcher (which are within $[0, 1]$ also) by using a sum rule. The recognition result for a probe is defined as the subject whose gallery sample has the highest score with it. The recognition rate is then defined as the percentage of correctly recognized subjects. Figure 8 shows the recognition rates. It can be clearly seen that the reconstructed normalized 3D face shapes do help improve the face recognition accuracy, especially for face images of large pose angles and all types of expressions. Interestingly, despite the relatively robust 2D face recognition performance w.r.t. expressions, the fusion with 3D matching still improves the performance across all expressions – a strong testimony on the discriminative capability of the expression-normalized 3D face shape.

³ <http://www.wisesoft.com.cn>

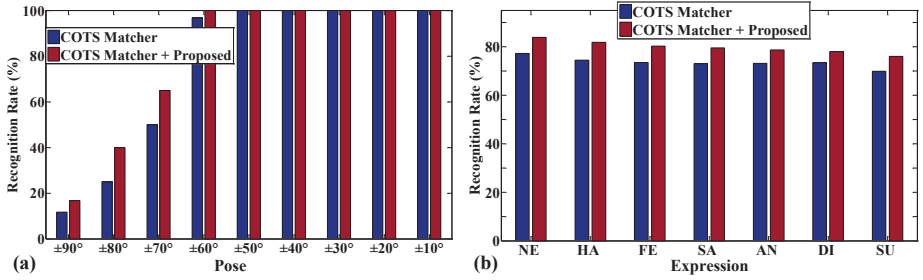


Fig. 8. Face recognition results of a COTS matcher and its fusion with proposed reconstructed 3D face based matcher under varying (a) poses and (b) expressions.

4.5 Computational Efficiency

According to our experiments on a PC with i7-4710 CPU and 8 GB memory, the Matlab implementation of the proposed method runs at ~ 26 FPS ($K = 5$ and $n = 9,677$). Hence, it can detect landmarks and reconstruct 3D face shape in real time.

5 Conclusions

In this paper, we present a novel regression based method for simultaneous face alignment and 3D face reconstruction for 2D images of arbitrary poses and expressions. It utilizes landmarks on a 2D face image as clues for reconstructing 3D shapes, and uses the reconstructed 3D shapes to refine landmarks. By alternately applying cascaded landmark regressors and 3D shape regressors, the proposed method can effectively accomplish the two tasks simultaneously in real time. Unlike existing 3D face reconstruction methods, the proposed method does not require additional face alignment methods, but can fully automatically reconstruct normalized 3D shapes from a single face image of arbitrary poses and expressions. Compared with existing face alignment methods, the proposed method can effectively handle invisible landmarks with the assistance of 3D face models. Extensive experiments with comparison to state-of-the-art methods demonstrate the effectiveness of the proposed method in both face alignment and 3D face shape reconstruction, and in facilitating face recognition as well.

6 Acknowledgment

All correspondences should be forwarded to Q. Zhao via qjzhao@scu.edu.cn. This work is supported by the National Key Scientific Instrument and Equipment Development Projects of China (No. 2013YQ49087904).

References

1. Asthana, A., Zafeiriou, S., Tzimiropoulos, G., Cheng, S., Pantic, M.: From pixels to response maps: Discriminative image filtering for face alignment in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(6), 1312–1320 (2015)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: *SIGGRAPH*. pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999)
3. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(9), 1063–1074 (2003)
4. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3D shape regression for real-time facial animation. *Transactions on Graphics (TOG)* 32(4), 41 (2013)
5. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics (TOG)* 35(4), 126 (2016)
6. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. *International Journal of Computer Vision* 107(2), 177–190 (2014)
7. Chu, B., Romdhani, S., Chen, L.: 3D-aided face recognition robust to expression and pose variations. In: *CVPR*. pp. 1907–1914. IEEE (2014)
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (6), 681–685 (2001)
9. Cootes, T.F., Lanitis, A.: Active shape models: Evaluation of a multi-resolution method for improving image search. In: *BMVC*. pp. 327–338. Citeseer (1994)
10. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. *Pattern Recognition* 41(10), 3054–3067 (2008)
11. Cristinacce, D., Cootes, T.F.: Boosted regression active shape models. In: *BMVC*. pp. 1–10 (2007)
12. Drira, H., Ben Amor, B., Srivastava, A., Daoudi, M., Slama, R.: 3D face recognition under expressions, occlusions, and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(9), 2270–2283 (2013)
13. Gong, X., Wang, G.: An automatic approach for pixel-wise correspondence between 3D faces. In: *Hybrid Information Technology*. vol. 2, pp. 198–205 (2006)
14. Han, H., Jain, A.K.: 3D face texture modeling from uncalibrated frontal and profile images. In: *BTAS*. pp. 223–230. IEEE (2012)
15. Hassner, T.: Viewing real-world faces in 3D. In: *ICCV*. pp. 3607–3614 (2013)
16. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007)
17. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3D face alignment from 2D videos in real-time. In: *FG*. IEEE (2015)
18. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: *ICCV*. pp. 3694–3702 (2015)
19. Jourabloo, A., Liu, X.: Large-pose face alignment via CNN-based dense 3D model fitting. In: *CVPR* (June 2016)
20. Kemelmacher-Shlizerman, I., Basri, R.: 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(2), 394–405 (2011)
21. Lee, D., Park, H., Yoo, C.D.: Face alignment using cascade gaussian process regression trees. In: *CVPR*. pp. 4204–4212. IEEE (2015)

22. Lee, Y.J., Lee, S.J., Park, K.R., Jo, J., Kim, J.: Single view-based 3D face reconstruction robust to self-occlusion. *EURASIP Journal on Advances in Signal Processing* 2012(1), 1–20 (2012)
23. Liu, F., Zeng, D., Li, J., Zhao, Q.: Cascaded regressor based 3D face reconstruction from a single arbitrary view image. *arXiv preprint arXiv:1509.06161* (2015)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2), 91–110 (2004)
25. Matthews, I., Baker, S.: Active appearance models revisited. *International Journal of Computer Vision* 60(2), 135–164 (2004)
26. Qu, C., Monari, E., Schuchert, T., Beyerer, J.: Fast, robust and automatic 3D face model reconstruction from videos. In: *AVSS*. pp. 113–118. *IEEE* (2014)
27. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: *CVPR*. pp. 1685–1692. *IEEE* (2014)
28. Romdhani, S., Vetter, T.: Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: *CVPR*. pp. 986–993. *IEEE* (2005)
29. Roth, J., Tong, Y., Liu, X.: Adaptive 3D face reconstruction from unconstrained photo collections. In: *CVPR* (June 2016)
30. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: *ICCVW*. pp. 397–403. *IEEE* (2013)
31. Suwajanakorn, S., Kemelmacher-Shlizerman, I., Seitz, S.M.: Total moving face reconstruction. In: *ECCV*, pp. 796–812. *Springer* (2014)
32. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: What makes tom hanks look like tom hanks. In: *ICCV*. pp. 3952–3960 (2015)
33. Tulyakov, S., Sebe, N.: Regressing a 3D face shape from a single image. In: *ICCV*. pp. 3748–3755. *IEEE* (2015)
34. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: *CVPR*. pp. 3659–3667. *IEEE* (2015)
35. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: *CVPR*. pp. 532–539. *IEEE* (2013)
36. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: *FG*. pp. 211–216. *IEEE* (2006)
37. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: *ICCV*. pp. 1944–1951. *IEEE* (2013)
38. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3D shape estimation from 2D landmarks: A convex relaxation approach. In: *CVPR*. pp. 4447–4455. *IEEE* (2015)
39. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: *CVPR*. pp. 4998–5006 (2015)
40. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: *CVPR*. pp. 2879–2886. *IEEE* (2012)
41. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3D solution. In: *CVPR* (June 2016)
42. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: *CVPR*. pp. 787–796. *IEEE* (2015)