Shape Augmented Regression for 3D Face Alignment

Chao Gou[1,3(✉)], Yue Wu[2], Fei-Yue Wang[1,3], and Qiang Ji[2]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{gouchao2012,feiyue.wang}@ia.ac.cn
[2] ECSE, Rensselaer Polytechnic Institute, Troy, USA
{wuy9,jiq}@rpi.edu
[3] Qingdao Academy of Intelligent Industries, Qingdao, China

Abstract. 2D face alignment has been an active topic and is becoming mature for real applications. However, when large head pose exists, 2D annotated points lose geometric correspondence with respect to actual 3D location. In addition, local appearance varies more dramatically when subjects are with large pose or under various illuminations. 3D face alignment from 2D images is a promising solution to tackle this problem. 3D face alignment aims to estimate the 3D face shape which is consistent across all poses. In this paper, we propose a novel 3D face alignment method. This method consists of two steps. First, we perform 2D landmark detection based on the shape augmented regression. Second, we estimate the 3D shape using the detected 2D landmarks and 3D deformable model. Experimental results on benchmark database demonstrate its preferable performances.

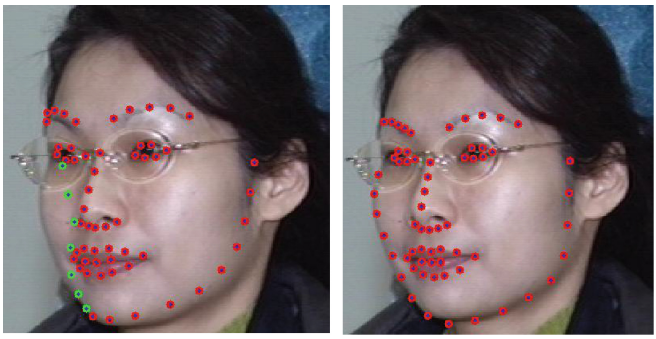
Keywords: Shape augmented regression · 3D face alignment

1 Introduction

Face alignment aims to estimate the locations of semantic facial landmarks such as eye corners, mouth corners and nose tip in a given image. 2D facial landmark detection has been an important research topic due to its wide applications such as facial action unit recognition [1], face recognition [2], head pose estimation [3] and 3D face reconstruction [4]. Recently, cascade regression framework has shown good performances for 2D facial landmark detection [5–7]. It begins with an initial guess about the facial landmark locations and it iteratively updates the landmark locations based on the local appearance features. Different regression models for each cascade level are applied to map the local appearance features to shape updates. Cascade regression framework is promising because iteratively updating through a supervised scheme is more efficient than solving an optimization problem for each image.

One limitation of 2D face alignment is that it can not capture the actual 3D shape correspondence especially if the face is with large poses. As shown in Fig. 1,

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part II, LNCS 9914, pp. 604–615, 2016.
DOI: 10.1007/978-3-319-48881-3_42



(a) (b)

Fig. 1. 3D face alginment are more consistent than 2D face alignment across different poses. Landmarks on cheek occluded by head pose are marked green for better view. (a) 3D face alignment. (b) 2D face alginment. (Color figure online)

different from 2D face alignment which estimates the 2D landmark locations in the image plane, 3D face alignment aims to estimate the 3D landmark locations corresponding to the real 3D information of face. Recently, 3D face alignment from a 2D image is becoming an active research topic due to its robustness to pose and strong representational power. There are two major paradigms for 3D face alignment: one is first 2D landmark detection followed by fitting a 3D face model to estimate a 3D face shape, and another is directly estimating 3D deformable parameters and 3D shape based on discriminate features. It is worth nothing that, direct estimating the 3D shape in one step needs large number of 3D annotation data for training to cover the various 3D texture and shape while the two-step based methods need a few 3D data to train the 3D deformable shape model. Hence, in this paper, we follow the first paradigm. We detect 2D landmark based on cascade regression framework first, followed by fitting a off-line trained 3D deformable model to estimate 3D shape. Different from 2D face alignment task, the shape information is more important for 3D face alignment because it can capture the actual correspondence of 3D shape. For 2D landmark detection, we incorporate the shape and appearance information in a cascade regression framework. Then we combine the detected 2D landmarks and 3D morphable model to estimate the 3D shape.

In the rest of this paper, we first review the related works of 2D and 3D face alignment in Sect. 2. Our proposed approach is described in Sect. 3. Section 4 reports the experimental results with discussions. Section 5 concludes the paper.

2 Related Work

Face alignment can be classified into 2D face alignment and 3D face alignment. The goal of 2D face alignment is predicting locations of semantic facial landmarks in a given image with limited head pose. 3D face alignment is an extension of 2D

face alignment and estimate 3D facial landmarks w.r.t a pre-defined coordinate system (eg. camera coordinate system).

In particular, 2D face alignment methods can be classified into holistic methods [8–10], Constrained Local Model (CLM) [11–13] and regression based [5,6]. Holistic method learns models that can capture the global appearance and face shape information. It focuses on designing algorithms that minimize the difference between the current estimate of appearance and ground truth. CLM learns a set of local appearance models and a global shape models. For inference, it estimates each landmark locations based on local searching region features and global shape constraints. Regression based methods estimate the landmark locations or displacements through local appearance features using the off-line trained regression models. Cascade regression framework has been successfully applied to facial landmark detection and achieves state-of-the-art performance recently [5]. In this paper, we also utilize the cascade regression framework. Different from conventional cascade regression framework that the regression parameters are constant for each iteration, we propose shape augmented regression to adjust the parameters iteratively based on the current estimated shape and corresponding local appearance features.

Many works are done on 3D face shape estimation from a single image [7,14–20]. The related works can be classified into two types: (I)two-step based methods that perform 2D landmark detection first followed by fitting 3D model to estimate the 3D shape, (II)one-step based methods that directly estimate the 3D shape based on discriminative shape invariant features. For the two-step based methods, Gu and Kanade [15] align 3D morphable model to a single image based on local patches related to a set of sparse 3D points. Cao *et al.* [17] propose to recover face pose and facial expression by fitting a user-specific blendshape model for landmark detection in 2D video frames. In [16] and [21], the authors propose to estimate the locations of landmarks and related visibility. They then recover the 3D shape by fitting a part-based 3D model. For the one-step based methods, Tulyakov and Sebe [7] estimate the 3D shape from a single image based on cascade regression framework using the shape invariant features. Jourabloo and Liu [18] present a cascaded coupled-regressor to jointly update the projection matrix and 3D deformable parameters for 3D landmark locations based on local appearance features. In [19], they further extend it to combine the 3DMM and cascaded CNN regressor for 3D face shape estimation. Zhu *et al.* [20] consider a dense 3DMM and the projection matrix as a representation of 2D face image. They propose to use CNN as the regressor in the cascaded framework to learn the mapping between the 2D face image and 3DMM with projection matrix.

3 Approach

Our overall framework is illustrated in Fig. 2. We perform two steps for 3D face alignment. The 2D landmarks are detected first, followed by 3D morphable fitting with off-line trained deformable model. Then we can estimate the 3D

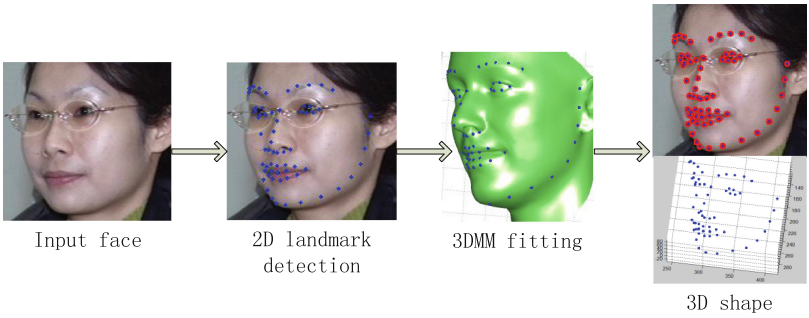


Fig. 2. Overall framework of our proposed method. It performs 2D landmark detection based on shape augmented regression method. It then fits the 3D demorphable model to estimate the 3D shape.

Algorithm 1. General cascaded regression framework.

Input: Facial landmark locations $\mathbf{x}^0$ are initialized by mean face.
Do cascade regression:
for t=1,2,...,T **do**
Update the key point locations $\mathbf{x}^t$ given the current key point locations $\mathbf{x}^{t-1}$ and image $\mathbf{I}$.
$f_t : \mathbf{I}, \mathbf{x}^{t-1} \rightarrow \Delta\mathbf{x}^t$
$\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t$
end for
Output:
Landmark locations $\mathbf{x}^T$.

facial shape. In the following, we firstly describe the general cascaded regression framework for 2D landmark detection. Then we discuss the shape augmented method for 2D landmark detection. Finally we introduce the method that fits the 3D deformable model based on 2D landmark for 3D shape estimation.

3.1 General Cascaded Framework

General cascaded framework approximately solves the optimization problem by learning several sequential regressors based on the local appearance. The Supervised Decent Method (SDM) [5] is one popular cascade framework as shown in Algorithm 1. The facial landmark locations are denoted as $\mathbf{x}^t = \{x_1^t, x_2^t, ..., x_D^t\}$, where D denotes the number of landmarks and t denotes the iteration in cascaded regression framework. At iteration t, given the image $\mathbf{I}$, it uses the linear regression function f_t to map the high dimension features (eg. SIFT [22]) around the landmarks to the updates $\Delta\mathbf{x}^t$ of landmark locations. $\mathbf{x}^0$ is usually initialized as the mean face.

3.2 Shape Augmented Cascaded Regression

For 3D face alignment, it is critical to capture the shape correspondence when 3D face is projected onto image plane. To incorporate the shape information, we utilize the shape augmented regression method [6] to adjust the regression parameters iteratively according to the current estimated shape and related local appearance. The overall framework is shown in Algorithm 2. In this paper, given the image $\mathbf{I}$, 2D face alignment objective function can be formulated as Eq. 1:

$$f(\mathbf{x}) = \frac{1}{2}||\Phi(\mathbf{x}, \mathbf{I}) - \Phi(\mathbf{x}^*, \mathbf{I})||^2 + \frac{1}{2}||\Psi(\mathbf{x}) - \Psi(\mathbf{x}^*)||^2, \tag{1}$$

where $\mathbf{x}$ are the landmark locations, $\mathbf{x}^*$ are the ground truth locations, $\Phi(\mathbf{x}, \mathbf{I})$ are the local SIFT features around the current landmark locations, and $\Psi(\mathbf{x})$ are the shape features which are the difference among pairs of landmarks. Hence, landmark locations can be estimated by solving the optimization problem $\tilde{\mathbf{x}} = arg\min_{\mathbf{x}} f(\mathbf{x})$. We further apply a second order Taylor expansion on Eq. 1:

$$f(\mathbf{x}) = f(\mathbf{x}^0 + \Delta\mathbf{x}) \approx f(\mathbf{x}^0) + J_f(\mathbf{x}^0)^T \Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T H_f(\mathbf{x}^0)\Delta\mathbf{x}, \tag{2}$$

where $J_f(\mathbf{x}^0)$ and $H_f(\mathbf{x}^0)$ are the Jacobian and Hessian matrices of function $f(\cdot)$ evaluated at the current location $\mathbf{x}^0$, respectively. After taking the derivation of $f(\mathbf{x})$ in Eq. 2 and set it to zero, we can get the update for landmark locations as shown in Eq. 3.

$$\begin{aligned}
\Delta\mathbf{x} =& - H_f(\mathbf{x}^0)^{-1} J_f(\mathbf{x}^0) \\
=& - H_f(\mathbf{x}^0)^{-1}[J_\Phi(\mathbf{x}^0)(\Phi(\mathbf{x}^0, \mathbf{I}) - \Phi(\mathbf{x}^*, \mathbf{I})) + J_\Psi(\mathbf{x}^0)(\Psi(\mathbf{x}^0) - \Psi(\mathbf{x}^*))] \\
=& - H_f(\mathbf{x}^0)^{-1} J_\Phi(\mathbf{x}^0)\Phi(\mathbf{x}^0, \mathbf{I}) - H_f(\mathbf{x}^0)^{-1} J_\Psi(\mathbf{x}^0)\Psi(\mathbf{x}^0) \\
& + H_f(\mathbf{x}^0)^{-1}(J_\Phi(\mathbf{x}^0)\Phi(\mathbf{x}^*, \mathbf{I}) + J_\Psi(\mathbf{x}^0)\Psi(\mathbf{x}^*))
\end{aligned} \tag{3}$$

It is computationally expensive to estimate Hessian and its inverse. In addition, the ground truth landmark locations $\mathbf{x}^*$ are unknown but fixed as constant during inference. Similar to SDM, we introduce the related parameters as below:

$$\begin{aligned}
\mathbf{P} &= -H_f(\mathbf{x}^0)^{-1} J_\Phi(\mathbf{x}^0) \\
\mathbf{Q} &= -H_f(\mathbf{x}^0)^{-1} J_\Psi(\mathbf{x}^0) \\
\mathbf{b} &= H_f(\mathbf{x}^0)^{-1}(J_\Phi(\mathbf{x}^0)\Phi(\mathbf{x}^*, \mathbf{I}) + J_\Psi(\mathbf{x}^0)\Psi(\mathbf{x}^*))
\end{aligned} \tag{4}$$

At iteration t for cascade regression, we can rewrite Eq. 3 as Eq. 5 to estimate the updates of landmark locations:

$$\Delta\mathbf{x}^t = \mathbf{P}^t\Phi(\mathbf{x}^{t-1}, \mathbf{I}) + \mathbf{Q}^t\Psi(\mathbf{x}^{t-1}) + \mathbf{b}^t \tag{5}$$

Hence, we need to learn the parameters in Eq. 4 for cascade regression. Given the i-th face image $\mathbf{I}_i$ with estimated landmark locations $\mathbf{x}_i^{t-1}$, the local appearance features $\Phi(\mathbf{x}^{t-1}, \mathbf{I})$ and shape features $\Psi(\mathbf{x}^{t-1})$ of i-th image can be calculated. For iteration t, the updates $\Delta\mathbf{x}_i^t$ of landmark locations can be acquired by

Algorithm 2. Shape augmented regression framework.

Input: Facial landmark locations $\mathbf{x}^0$ are initialized by mean face.
Do cascade regression:
for t=1,2,...,T **do**
Given the current key point locations $\mathbf{x}^{t-1}$ and image $\mathbf{I}$, estimate the update of landmarks through Eq. 5.
$$\Delta\mathbf{x}^t = \mathbf{P}^t\Phi(\mathbf{x}^{t-1}, \mathbf{I}) + \mathbf{Q}^t\Psi(\mathbf{x}^{t-1}) + \mathbf{b}^t$$
Update the key point locations $\mathbf{x}^t$
$$\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t$$
end for
Output:
Landmark locations $\mathbf{x}^T$.

subtracting the current locations $\mathbf{x}_i^{t-1}$ from the ground truth locations $\mathbf{x}_i^*$. The initialization of landmark locations are mean face denoted by $\mathbf{x}_i^0$. The learning of $\mathbf{P}^t$, $\mathbf{Q}^t$ and bias $\mathbf{b}^t$ can be formulated as a standard least-squares formulation with closed form solution:

$$\mathbf{P}^{t^*}, \mathbf{Q}^{t^*}, \mathbf{b}^{t^*} = arg \min_{\mathbf{P}^t, \mathbf{Q}^t, \mathbf{b}^t} \sum_{i=1}^{K} \parallel \Delta\mathbf{x}_i^t - \mathbf{P}^t\Phi(I_i, \mathbf{x}_i^{t-1}) - \mathbf{Q}^t\Psi(I_i, \mathbf{x}_i^{t-1}) - \mathbf{b}^t \parallel^2 \tag{6}$$

where K is the number of training samples.

For testing, given the face image $\mathbf{I}$ and current key point locations $\mathbf{x}^{t-1}$ at iteration t, we can estimate the update locations $\Delta\mathbf{x}^t$ by learned parameters $\mathbf{P}^t$, $\mathbf{Q}^t$ and bias $\mathbf{b}^t$. Then the landmark locations can be acquired through $\mathbf{x}^t = \mathbf{x}^{t-1} + \Delta\mathbf{x}^t$.

3.3 3D Morphable Model Fitting

Given the detected 2D landmark locations on the testing image and a 3D morphable model [14], we can recover 3D face by estimating the 3D pose and nonrigid deformation via the fitting process. 3DMM is defined as a shape model with dense mesh. In particular, it can be simplified by the 3D vertex locations(landmark points) of the related dense mesh. Hence, it can describe 3D face nonrigid shape variations with mean 3D shape and PCA space linearly as below:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Bq} \tag{7}$$

Here, $\mathbf{s}$ is the 3D shape of N landmarks in head coordinate system denoted by $\mathbf{s} = \{x_1, y_1, z_1, ..., x_N, y_N, z_N\}$, $\bar{\mathbf{s}}$ is the mean 3D shape, $\mathbf{B}$ represents the learned orthonormal bases, and $\mathbf{q}$ denotes the nonrigid shape variation parameters. We learn the mean 3D shape $\bar{\mathbf{s}}$ and PCA bases $\mathbf{B}$ of 3D model in Eq. 7 from the annotations provided in [7].

Assuming 3D face is projected onto the image plane with weak perspective projection, the k-th landmark location in image plane can be calculated by:

$$\begin{bmatrix} u_k \\ v_k \end{bmatrix} = \mathbf{Ms} + \mathbf{t} = \begin{bmatrix} \frac{1}{\bar{z}_c} f s_x \mathbf{r}_1 \\ \frac{1}{\bar{z}_c} f s_y \mathbf{r}_2 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} + \mathbf{t}, \tag{8}$$

where $\bar{z}_c$ is the mean depth in camera coordinate, f is the focal length, s_x and s_y are the sampling frequency in rows and columns (also known as scaling factors), $\mathbf{r}_i = (r_{i1}, r_{i2}, r_{i3})$ is the i-th row of 3 by 3 rotation matrix $\mathbf{R}$ which is encoded by three head poses(pitch α, yaw β, roll γ), and $\mathbf{t} = (t_1, t_2)^T$ is the 2D translation vector. Since f, s_x and s_y are intrinsic parameters and $\bar{z}_c$ is constant, we set $\frac{1}{\bar{z}_c} f s_x = \lambda_1$ and $\frac{1}{\bar{z}_c} f s_y = \lambda_2$ as two unknown parameters for simplicity. As a result, we can rewrite Eq. 8 as below:

$$\begin{bmatrix} u_k \\ v_k \end{bmatrix} = \mathbf{x}_{2d}(\mathbf{p}) = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{R}_{2\times 3}(\bar{\mathbf{s}}_k + \mathbf{B}_k \mathbf{q}) + \mathbf{t} \tag{9}$$

where $\mathbf{x}_{2d}$ is landmark location in image plane, $\mathbf{R}_{2\times 3} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}$ is the first two rows of rotation matrix $\mathbf{R}$ and $\mathbf{p} = \{\lambda_1, \lambda_2, \alpha, \beta, \gamma, t_1, t_2, \mathbf{q}\}$ denotes the parameters of the model.

Given the pre-trained 3D model and 2D landmark locations $\mathbf{x}$, we can estimate the model parameters by minimizing the misalignment error of projected locations $\mathbf{x}_{2d}$ and detected 2D landmark locations $\mathbf{x}$ for all landmark points:

$$\begin{aligned} \mathbf{p} &= arg\min_{\mathbf{p}} \sum_{k=1}^{K} \parallel \mathbf{x}_k - \mathbf{x}_{2d,k} \parallel^2 \\ &= arg\min_{\mathbf{p}} \sum_{k=1}^{K} \parallel \mathbf{x}_k - (\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{R}_{2\times 3}(\bar{\mathbf{s}}_k + \mathbf{B}_k \mathbf{q}) + \mathbf{t}) \parallel^2 \end{aligned} \tag{10}$$

where $\mathbf{p} = \{\mathbf{h}, \mathbf{q}\} = \{\lambda_1, \lambda_2, \alpha, \beta, \gamma, t_1, t_2, \mathbf{q}\}$ denotes the model parameters. To solve this optimization problem, we alternatively update the transformation parameters $\mathbf{h} = \{\lambda_1, \lambda_2, , \alpha, \beta, \gamma, t_1, t_2\}$ and deformable parameters $\mathbf{q}$ until it converges. We first initialize the deformable parameters $\mathbf{q}$ as zeros and we can get the 3D locations in head coordinate systems. Then we solve the linear optimization problem of Eq. 10 to get the parameters of $\mathbf{h}$. After estimating $\mathbf{h}$, we feed it into Eq. 10 and estimate the deformable parameters $\mathbf{q}$. We repeat until the max update of pose parameters (α, β, γ) is less than 0.1 in degree.

After estimating parameters $\mathbf{p}$, we can calculate the 3D shape as follows (according to the definition of the 3DFAW challenge and [16,21]):

$$\mathbf{x}_{3d} = \lambda \mathbf{Rs} + \mathbf{T} = \lambda \mathbf{R}(\bar{\mathbf{s}} + \mathbf{Bq}) + \mathbf{T} \tag{11}$$

where $\mathbf{x}_{3d}$ denotes the 3D landmark locations, λ is the scale factor, $\mathbf{R}$ is a 3 by 3 rotation matrix, $\mathbf{T} = (t_1, t_2, t_3)^T$ is the 3D translation vector. In this paper,

we approximate the scale factor by $\lambda = \frac{\lambda_1 + \lambda_2}{2}$, t_1, t_2 are the same as Eq. 8, t_3 is set to zero and $\mathbf{r}_3$ is the cross product of $\mathbf{r}_1$ and $\mathbf{r}_2$. After calculating the 3D shape, we normalize the depth to zero mean.

4 Experiments

In this section, we first describe the implementation details. Then, we conduct experiments and comparisons.

4.1 Database

The experimental dataset is from ECCV2016 workshop on 3D face alignment in the wild (3DFAW) challenge. It consists of MultiPIE [23], BU-4DFE [24], BP4D-Spontaneous [25] and some image frames of videos collected from web. The landmark annotations consist 23606 images of 66 3D points and the depth information is recovered using a model-based structure from motion technique [16,21]. It is divided into four sub-datasets: 13671 images for training, 4725 images for validation, 4912 for testing and 298 images for extra training.

The facial images are normalized to 200 pixels in width. Mean face shape are calculated on all normalized facial images. Similar to [6] during training, we generate multiple initial face shapes by rotating, scaling and shifting the mean face shape to improve the robustness. For cascade regression, the number of iteration is 4. We use the detected 2D landmarks as the estimated locations of 3D shape.

4.2 Evaluation Criteria

For fair comparisons, we use the widely accepted evaluation matrices named Ground Truth Error (GTE) and Cross View Ground Truth Consistency Error (CVGTCE). GTE is the average point-to-point Euclidean error normalized by the outer corners of the eyes (inter-ocular) formulated as below:

$$GTE(\mathbf{x}_{gt}, \mathbf{x}_{pre}) = \frac{1}{K} \sum_{k=1}^{K} \frac{||\mathbf{x}_{gt}^k - \mathbf{x}_{pre}^k||_2}{d_i} \tag{12}$$

where K is the number of points, $\mathbf{x}_{gt}$ is the ground truth 3D shape, $\mathbf{x}_{pre}$ is the prediction and d_i is the inter-ocular distance for the i-th image. CVGTCE is used to evaluate cross-view consistency of the predicted landmarks from 3D model. It is computed as below:

$$CVGTCE(\mathbf{x}_{gt}, \mathbf{s}, \mathbf{p}) = \frac{1}{K} \sum_{k=1}^{K} \frac{||\mathbf{x}_{gt}^k - (f\mathbf{P}\mathbf{s}^k + \mathbf{t})||_2}{d_i} \tag{13}$$

where $\tilde{\mathbf{x}}_{pre} = f\mathbf{P}\mathbf{s} + \mathbf{t}$ is the predicted 3D shape for another view of the subject, $\mathbf{p} = \{f, \mathbf{P}, \mathbf{t}\}$ is the model parameters, f is the scale factor and $\mathbf{P}$ is the rotation matrix.

4.3 Experimental Results

We first train shape augmented cascade regression models for 2D landmark detection on training dataset. We conduct comparisons with SDM [5] which follows the same procedure and retrain on the same database. After detecting the 2D landmark, we use the same 3D deformable model fitting process to estimate the 3D shape. The 3D face shape detection experimental results on validation dataset are shown in Table 1. Some qualitative results are shown in Fig. 3 and inaccurate results are shown in Fig. 4. From Table 1, incorporating the shape information help improve the results especially for the 3D face alignment. As we discussed before, 3D face alignment is more consistent with respect to large pose. Shape information between pair of points is very important when 3D shape are projected onto image plane. As shown in Fig. 4 for the inaccurate detection results, it fails when the subjects are with extreme head pose and the appearance is very ambiguous. For our proposed method, it is important to predict 2D landmarks which are used to estimate the 3D model parameters.

Table 1. 3D landmark detection comparison on validation dataset

Mehtod	GTE(%)
SDM [5]+3DMM	6.34
Ours	**5.90**

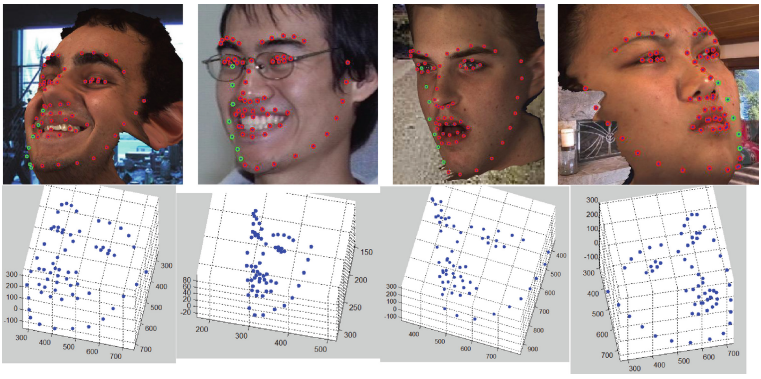


Fig. 3. Qualitative 3D face alginment results on validation dataset. Landmarks on cheek occluded by head pose are marked green for better view. (Color figure online)

We further train our model on training, extra training and validation dataset and test on testing dataset. Experimental results are shown in Table 2 and Fig. 5. The performance of our method are preferable on challenging testing dataset.

Fig. 4. Inaccurate detection results on validation dataset. The calcuated GTE for (a), (b), (c) are 16.18, 23.05, 13.87, respectively.

Table 2. 3D landmark detection results on testing dataset

CVGTCE(%)	GTE(%)
6.21	4.95

Fig. 5. Qualitative 3D face alginment resutls on testing data. Landmarks on cheek occluded by head pose are marked green for better view. (Color figure online)

As shown in Table 2, we can achieve 4.95 of GTE on all the testing dataset. In addition, we also get 6.21 of CVGTCE which demonstrates that our method is cross-view consistent for 3D face alignment. As shown in Fig. 5, our method is robust to illuminations and can achieve reasonable detection results when subjects are with extreme head pose.

5 Conclusions

In this paper, we propose to firstly estimate the location of a set of landmarks based on shape augmented regression framework, then recover the 3D face shape by fitting a 3D morphable model. By incorporating the shape and local appearance information in the cascade regression framework, our proposed method can

capture the geometric correspondence between pair of points for 3D face alignment. An alternative optimizing method for estimating 3D morphable model parameters is adopted to estimate the 3D shape including the depth information. Experimental results on large scale of testing dataset validate the robustness and effectiveness of proposed method.

The appearance of occluded landmarks is not reliable for the prediction of location. Future work will focus on inferring the visibility of landmarks which can be used to weight the related appearance. In addition, we will iteratively update the 2D landmark location and corresponding 3D morphable model parameters during a unified cascade regression framework based on the local appearance.

Acknowledgments. This work was completed when the first author visited Rensselaer Polytechnic Institute (RPI), supported by a scholarship from University of Chinese Academy of Sciences (UCAS). The authors would like to acknowledge support from UCAS and RPI. This work was also supported in part by National Science Foundation under the grant 1145152 and by the National Natural Science Foundation of China under Grant 61304200 and 61533019.

References

1. Wu, Y., Ji, Q.: Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
2. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Toward a practical face recognition system: robust alignment and illumination by sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(2), 372–386 (2012)
3. Narayanan, A., Kaimal, R.M., Bijlani, K.: Estimation of driver head yaw angle using a generic geometric model. IEEE Trans. Intell. Transp. Syst. **PP**(99), 1–15 (2016)
4. Roth, J., Tong, Y., Liu, X.: Adaptive 3D face reconstruction from unconstrained photo collections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
5. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539. IEEE (2013)
6. Wu, Y., Ji, Q.: Shape augmented regression method for face alignment. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 26–32 (2015)
7. Tulyakov, S., Sebe, N.: Regressing a 3D face shape from a single image. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 3748–3755. IEEE (2015)
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 681–685 (2001)
9. Matthews, I., Baker, S.: Active appearance models revisited. Int. J. Comput. Vis. **60**(2), 135–164 (2004)
10. Lucey, S., Navarathna, R., Ashraf, A.B., Sridharan, S.: Fourier lucas-kanade algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1383–1396 (2013)

11. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1034–1041. IEEE (2009)
12. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. **1**(2), 3 (2006)
13. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1862–1874 (2015)
14. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999)
15. Gu, L., Kanade, T.: 3D alignment of face in a single image. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 1, pp. 1305–1312. IEEE (2006)
16. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3D face alignment from 2D videos in real-time. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–8. IEEE (2015)
17. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3D shape regression for real-time facial animation. ACM Trans. Graph. (TOG) **32**(4), 41 (2013)
18. Jourabloo, A., Liu, X.: Pose-invariant 3D face alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3694–3702 (2015)
19. Jourabloo, A., Liu, X.: Large-pose face alignment via CNN-based dense 3D model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
20. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2016)
21. Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3D face alignment from 2D video for real-time use. Image and Vision Computing (2016)
22. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
23. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vis. Comput. **28**(5), 807–813 (2010)
24. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2008, pp. 1–6. IEEE (2008)
25. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. Image Vis. Comput. **32**(10), 692–706 (2014)