Discriminative 3D Morphable Model Fitting

Xiangyu Zhu, Junjie Yan, Dong Yi, Zhen Lei and Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Haidian District, Beijing, China.

Abstract— **This paper presents a novel discriminative method for estimating 3D shape from a single image with 3D Morphable Model (3DMM). Until now, most traditional 3DMM fitting methods depend on the analysis-by-synthesis framework which searches for the best parameters by minimizing the difference between the input image and the model appearance. They are highly sensitive to initialization and have to rely on the stochastic optimization to handle local minimum problem, which is usually a time-consuming process. To solve the problem, we find a different direction to estimate the shape parameters through learning a regressor instead of minimizing the appearance difference. Compared with the traditional analysis-by-synthesis framework, the new discriminative approach makes it possible to utilize large databases to train a robust fitting model which can reconstruct shape from image features accurately and efficiently. We compare our method with two popular 3DMM fitting algorithms on FRGC database. Experimental results show that our approach significantly outperforms the state-of-the-art in terms of efficiency, robustness and accuracy.**

I. Introduction

Since the seminal work of Blanz and Vetter [9], the 3D Morphable Model (3DMM) has been widely used to estimate 3D shape from image data, with applications ranging from relighting [39], super-resolution [25] to pose robust face recognition [10]. Given a single face image under unknown pose and illumination, the 3D Morphable Model can solve its 3D shape, texture, pose and illumination parameters simultaneously following analysis-by-synthesis framework, where Gauss-Newton optimization is applied to minimize the difference between the synthetic image and the input image.

The original 3DMM has shown its robustness to complicated pose and illumination conditions and provides a promising way to face recognition in the wild due to the explicit pose and illumination recovery [10]. However, the fitting process of 3DMM is very time-consuming and suffers from local minimum problem just as other Gauss-Newton based methods. In the last decade, many researchers have made their efforts to improve the efficiency and accuracy of 3DMM fitting algorithm. On one hand, [30] extended the Inverse Compositional Image Alignment (ICIA) to 3DMM fitting, improving the efficiency by pre-computing the derivative of the cost function. [38], [2] adopted the spherical harmonic reflectance model, making the appearance estimation completely linear. [8], [26], [1] concentrated on estimating shape with a sparse set of 2D landmarks, providing an efficient linear method for shape fitting. On the other hand, [31] presented a Multi-Features Framework (MFF) to handle the local minimum problem, where a smoother objective function is constructed by considering contours, textured edges, specular highlights and pixel intensities jointly, leading to the state-of-the-art in both robustness and accuracy. Besides 3DMM, there have proposed some algorithms which can estimate 3D information from a single image. [23] used the CCA mapping to transform the image from the pixel space to the depth space, [22] adopted the shape-from-shading framework to recover the 3D shape, [21] used the SIFT-FLOW to estimate the depth image and [11] proposed a regression method to transfer the expression from video to 3D model.

However, existing methods still have some disadvantages. For landmark based algorithms [8], [26], [1], since the only input information is the landmark positions, they heavily depend on the face alignment algorithm. Unfortunately in many cases, even though the landmarks look fine on the image, they are not accurate enough for 3D shape fitting, especially those on the eyebrow, nose and contour. Thus directly estimating 3D shape from landmarks is usually unreliable. Traditional analysis-by-synthesis 3DMM fitting algorithms [10], [31] are mainly based on explicitly modelling the image formation process and estimating shape parameters by appearance fitting. It has been shown that these methods heavily rely on the quality of initialization and have to adopt the stochastic optimization to avoid local minimum [2]. As a result, most of them are computationally expensive and always need more than one minute to fit a single face image. Besides, most training sets of 3DMM are very small (100 to 500 samples) due to the difficulty in collecting complete face scans, thus the appearance model of 3DMM is always too weak to cover the large variations of face appearance, especially in the wild. Since the central problem of analysis-by-synthesis framework is fitting the model appearance to the input image, the weak expressive ability of appearance model will lead to non-accurate results.

In this paper, we discuss 3DMM in the context of face alignment and find a new direction to overcome the problems described above. Instead of the traditional analysis-by-synthesis framework, we propose a novel discriminative 3DMM fitting algorithm based on local features and cascade regression. In section 2, we introduce the 3D Morphable Model. In section 3 we highlight our motivation by briefly revisiting face alignment algorithms. Then we propose the discriminative 3DMM fitting in section 4 and discuss some implemental details in section 5. In the experiments, we show that our algorithm outperforms existing 3DMM fitting methods in both accuracy and efficiency.

II. 3D Morphable Model

3D Morphable Model [9] is constructed from a set of 3D face scans in dense correspondence. Each scan is represented by a shape-vector $S = (x_1, y_1, z_1, ..., x_n, y_n, z_n)$ and a texture-vector $T = (r_1, g_1, b_1, ..., r_n, g_n, b_n)$, which contain the coordinate and the color of each point respectively. The points are dense enough ($n > 10000$) to directly represent a human face. PCA is applied to decorrelate texture and shape vectors respectively and a 3D face can be described as:

$$S = \overline{s} + \sum_{i=1}^{m-1} \alpha_i \cdot s_i \qquad T = \overline{t} + \sum_{i=1}^{m-1} \beta_i \cdot t_i \tag{1}$$

where $m - 1$ is the number of eigenvectors, $\overline{s}$ and $\overline{t}$ are the means of shape and texture respectively, s_i and t_i are the ith eigenvectors, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{m-1})$ and $\beta = (\beta_1, \beta_2, ..., \beta_{m-1})$ are shape and texture parameters determining S and T. With the shape and texture model, a synthetic image can be generated by projecting the 3D face onto the image plane by weak perspective projection:

$$s_{2d} = fpr(\overline{s} + \sum_{i=1}^{m-1} \alpha_i \cdot s_i + t_{3d}) \tag{2}$$

where s_{2d} is the image coordinates after projection, f is the scale parameter, p is the constant orthographic projection matrix, r is the rotation matrix and t_{3d} is the translation vector in 3D space. We represent $\gamma = (f, r, t_{3d})$ as the pose parameters of 3DMM.

In the fitting process, the 3DMM solves the parameters by minimizing the Euclidean distance between the synthetic image and the input image:

$$E = \sum_{x,y} \| I_{input}(x, y) - I_{syn}(x, y) \|^2 \tag{3}$$

where $I_{input}(x, y)$ is the input image and $I_{syn}(x, y)$ is the 3DMM synthetic image. Usually, the stochastic Gauss-Newton method is adopted to minimize the cost function.

III. Motivation

In general, 3DMM can be seen as a branch of face alignment which concentrates on fitting a face model to the input image. Recently, the face alignment in 2D has greatly advanced and can be readily applied in real applications. In this section, we discuss 3DMM in the context of face alignment, seeing how the techniques in 2D could help to fit the 3DMM.

The central problem of face alignment is minimizing the difference between the input image and the face model.

$$P = \arg\min_{P} \| Tr(I) - Tr(M(P)) \|^2 \tag{4}$$

where I is the input image, $M(P)$ is the face model with parameters P which can give rise to an observed face image and Tr is a transformation which is usually a feature extractor. Based on Equ. (4) explicitly or implicitly, the optimization can be summarized as an iterative updating process:

$$\Delta P_t = RF(I, P_{t-1}) \qquad P_t = P_{t-1} + \Delta P_t \tag{5}$$

where P_t is the model parameters in the current iteration, $F(I, P_{t-1})$ is a feature extraction function depending on the input image I and the parameters in the last iteration and R is an updater that maps the features F to the parameter update ΔP_t. Three modules can be seen from Equ. (5): the model parameter P represents how to model a human face; the feature extractor $F(I, P_{t-1})$ represents what kind of information is used for alignment, and the updater R represents how to use the information.

It is obvious that 3DMM also belongs to the face alignment framework, where P is the PCA coefficients of shape and texture model, $F(I, P_{t-1})$ is either the pixel intensity [9], landmark [1] or Multi-Features [31] and R is constructed from the Jacobian of a cost function like Equ. (3). It is promising to discuss the achievements of 2D face alignment in recent years and introduce them into 3DMM fitting. In the last decade, a number of seminal works have been proposed to find accurate and robust fitting methods in 2D alignment. We will briefly review these works on the three topics of P, F, R and extend their ideas to reinforce 3DMM fitting.

Active Appearance Model (AAM) AAM [13] which is characterized by its explicit shape and appearance model, has been widely used to match deformable objects to images in early years of face alignment. In AAM, the face shape is defined by a sparse set of landmarks and the appearance is based on the warped images on the reference frame. PCA is applied on landmark vectors and shape-free textures to construct face model just as 3DMM. In fitting process, AAM searches the best parameters that minimize the distance between the model instance and the input image by either the generative fitting [14], [24], [35] which obtains the updater from the Jacobian of the distance function or the discriminative fitting [13], [32], [15] which directly learns a regressor to map the image difference to the parameter update.

Constrained Local Model (CLM) CLM [17], [16], [33], [5], [20], [36], [7] represents an object using local image patches around landmarks. It inherits the PCA shape model from AAM, but discards the holistic appearance model and learns landmark detectors instead. During fitting process, the landmark detectors provide the response maps showing the distribution of probable landmark locations and then the shape parameters are estimated by maximizing the sum of responses of landmarks constrained by a priori. The fitting methods of CLM are also divided into generative methods [33], [20], [36] and discriminative methods [5], while the latter dramatically outstand in both accuracy and robustness.

Non Parameter Model (NPM) NPM [18], [12], [37] further removes any explicit PCA constrains on shape and directly uses landmark coordinates as shape model. Besides, robust features like HOG and SIFT are adopted to describe local spatially-coherent observations of landmarks which are proven to be more robust than image pixel and response map. Furthermore, the cascade regression [18], [6], where independent regressors are trained for each iteration, fully

TABLE I

SUMMARY OF AAM, CLM AND NPM IN THE FRAMEWORK OF FACE ALIGNMENT.

Method	Model (P)	Feature (F)	Updater (R)
AAM	PCA shape model PCA appearance model	Image Pixel	Generative or Discriminative
CLM	PCA shape model	Response Map	Generative or Discriminative
NPM	None	HOG or SIFT	All Discriminative

utilizes the flexibility of NPM and keeps the robustness at the same time, making NPM dramatically outperforms any other models and shows the state-of-the-art in face alignment [37].

Table I summarizes AAM, CLM and NPM in the framework of face alignment. In the evolution of face alignment from AAM to NPM, the holistic PCA constraints are progressively removed and more robust features are adopted to handle complicated variations. With the removal of shape constraints, discriminative methods show better performance over generative methods, especially with cascade regression.

These achievements in 2D face alignment can provide promising clues for 3DMM. Current 3DMM fitting methods are based on modelling the physical process of forming a face image and minimizing the difference between the input image and the model appearance, which obviously belongs to the generative fitting framework [32]. It has been shown that this framework heavily relies on the simulation of the image forming process and the quality of initialization [24]. Besides, the minimization is very slow and easy to get stuck in local minimum [2]. Fortunately, these problems could be solved by discriminative fitting method which directly learns a regression based fitting model from a large training set just as NPM.

However, directly applying discriminative fitting framework has some difficulties. Firstly, unlike the sparse landmarks in 2D, the shape model of 3DMM is much denser with tens of thousands of points, which makes the regression matrix too large to learn. Secondly, there is no database containing visual image and complete face scan pairs, leading to the lack of training set for discriminative methods.

In the following sections, we will adopt discriminative methods to fit 3DMM and illustrate how to preprocess existing databases to construct a training set.

IV. DISCRIMINATIVE 3DMM FITTING

In this section, we show how to use the discriminative method instead of the analysis-by-synthesis framework to estimate shape parameters.

A. Derivation of Discriminative 3DMM Fitting

As in [37], we start from the traditional fitting process. Given an image I, we want to estimate its 3D shape by minimizing the difference between the synthetic and the input image. However in many cases, the optimization will converge to a local minimum far from the global one [31]. Thus we project both the input and the synthetic images into a new space with a transformation Tr, where the cost function is smoother.

$$P = \arg\min_{P} \|f(P)\|^2 \qquad f(P) = Tr(I) - Tr(G(P)) \quad (6)$$

where P is the model parameters, $G(P)$ is the synthesis process that can generate an image from 3DMM and Tr is an unknown transformation. For simplicity, we assume $Tr(G(\cdot))$ is differentiable and use Gauss-Newton method to optimize Equ. (6).

From an initial estimate P_0, we apply Taylor expansion to $f(P)$ and minimize the cost function by equally optimizing the following function over ΔP.

$$\arg\min_{\Delta P} f(P_0 + \Delta P)^T f(P_0 + \Delta P) \quad (7a)$$

$$f(P_0 + \Delta P) = f(P_0) + J_f \Delta P \quad (7b)$$

where J_f is the Jacobian of f. Taking the derivation of Equ. (7a) over ΔP and setting it to zero, we get an update to P_0.

$$\begin{aligned} \Delta P_0 &= -(J_f^T J_f)^{-1} J_f^T f(P_0) \\ &= -(J_f^T J_f)^{-1} J_f^T (Tr(I) - Tr(G(P_0))) \end{aligned} \quad (8)$$

It is unlikely that the optimization can converge at a single iteration, thus Equ. (8) is iterated by several times.

$$\Delta P_t = -((J_f^T J_f)^{-1} J_f^T)|_{P=P_{t-1}} (Tr(I) - Tr(G(P_{t-1}))) \quad (9a)$$

$$P_t = P_{t-1} + \Delta P_t \quad (9b)$$

While in fact, we do not know the form of Tr and can not get the difference in the Tr space. Note that $(Tr(I) - Tr(G(P_{t-1})))$ depends on the image I and model parameters P_{t-1}, if we can extract features with $F(I, P_{t-1})$ that implicitly reflects the “goodness” of current fitting and learn a linear regressor A to map the features to the difference in Tr space, we can rewrite $(Tr(I) - Tr(G(P_{t-1})))$ as $AF(I, P_{t-1})$. Then Equ. (9a) becomes:

$$\begin{aligned} \Delta P_t &= -((J_f^T J_f)^{-1} J_f^T A)|_{P=P_{t-1}} F(I, P_{t-1}) \\ &= R_t F(I, P_{t-1}) \end{aligned} \quad (10)$$

Note that we merge A into the updater R_t to directly map features to parameter update. According to the Supervised Descent Method (SDM) [37], we can get a list of $R = (R_1, \ldots, R_T)$ through learning instead of numerical approximation. During the testing process, the regressor list will give a sequence of descent directions so that the P_0 will converge to the ground truth.

$$P_t = P_{t-1} + R_t F(I, P_{t-1}) \quad (11)$$

How to determine Equ. (11) is the central problem of discriminative 3DMM fitting. This function depends on the model parameters P, regressor R, and features extractor $F(I, P)$.

Even though the appearance model of 3DMM is weak, the shape model can describe most of the real-world data due to the relative small variations of face shapes [22]. Besides, the performance of CLM and NPM has shown that it is robust to directly estimate shape information without appearance fitting. Thus we discard the appearance model of 3DMM and only consider the shape PCA coefficients α and the weak perspective projection pose parameters γ in Equ. (1)(2) and let $P = \{\alpha, \gamma\}$. In the next two subsections, we provide details of the feature extractor and the regression function.

B. Feature Extraction

This section illustrates how to extract features. The HOG features around landmark positions are used as the feature extractor. We mark a set of landmarks on the 3D model following the Multi-PIE [19] 68 points mark-up, as shown in Fig. 1. In each iteration t, with pose and shape parameters $P_{t-1} = \{\alpha_{t-1}, \gamma_{t-1}\}$, the 3D shape is constructed by Equ. (1) and projected to image plane by Equ. (2). Then the HOG on the landmark positions are extracted as the features in the current iteration:

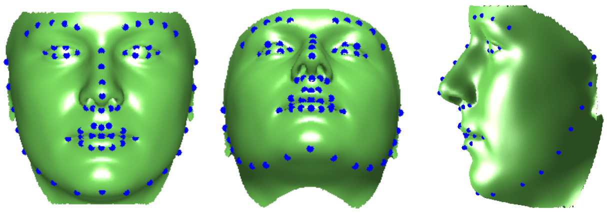


Fig. 1. The landmarks marked on the 3D face model

$$F(I,P) = HOG(I, [fpr(\bar{s} + \sum_{i=1}^{m-1} \alpha_i \cdot s_i + t_{3d})]_l) \quad (12)$$

where the $f, p, r, t_{3d}, s, \alpha$ have the same meanings as in Equ. (1)(2) and the subscript l means only the landmark points are selected.

C. Learning for Regression

In this section, we describe how to learn the regressor list $R = (R_1, \ldots, R_T)$ in Equ. (11) from a training set. Note that Equ. (11) is in fact the process of cascade regression, thus we train independent R_t for each iteration. Given a set of face images $\{I^1, \ldots, I^n\}$, their initial estimates $\{P_0^1, \ldots, P_0^n\}$ and ground truth parameters $\{P_*^1, \ldots, P_*^n\}$, we want to minimize the expected loss between the predicted update and the optimal update for all the training samples in each iteration:

$$\arg\min_{R_t} \sum_{i=1}^{n} \|(P_*^i - P_{t-1}^i) - R_t F(I^i, P_{t-1}^i)\|^2 \quad (13)$$

Writing Equ. (13) as matrix formation, we get:

$$\arg\min_{R_t} \left\| \begin{bmatrix} P_*^1 \\ \vdots \\ P_*^n \end{bmatrix}^T - \begin{bmatrix} P_{t-1}^1 \\ \vdots \\ P_{t-1}^n \end{bmatrix}^T - R_t \begin{bmatrix} F(I^1, P_{t-1}^1) \\ \vdots \\ F(I^n, P_{t-1}^n) \end{bmatrix}^T \right\|^2 \quad (14a)$$

$$= \arg\min_{R_t} \|\Delta P_t - R_t F_{t-1})\|^2 \quad (14b)$$

where ΔP_t is the optimal parameter update (ground truth minus current), and F_{t-1} is the features extracted with current parameters for each training sample. Equ. (14b) can be solved directly by linear method:

$$R_t = \Delta P_t F_{t-1}^T (F_{t-1} F_{t-1}^T + \lambda E)^{-1} \quad (15)$$

where E is the identity matrix and λ is the regularization term that avoids over fitting. Usually after each iteration, P_t will be closer to P_* than P_{t-1}, and with P_t we have a new training set and can run another iteration with Equ. (15) until coverage. In our experiments, the algorithm converges in 4 to 5 steps.

V. Data Pre-processing

To train a discriminative fitting model, we need a database with a large collection of visual images and corresponding 3D face shapes. However, unlike face alignment in 2D, the training set cannot be constructed by hand labelling because 3DMM shape model has tens of thousands of points. While using face scanners like Cyberware [34] or multiple ABW-3D [27] to collect complete face scans is so expensive and troublesome that the number of training samples is limited. The lack of training set is probably the main reason for the absence of discriminative fitting method in 3DMM.

A. Depth Image Registration

Compared with collecting complete face scans, only getting depth images is relatively easier and such work has been done in FRGC [28]. FRGC provides a large database with thousands of visual and depth image pairs in full correspondence, as shown in Fig. 2.

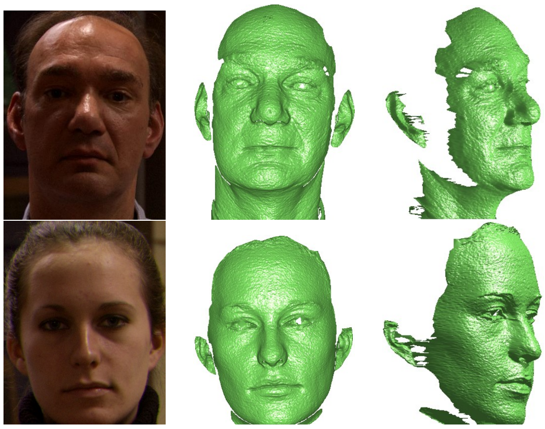


Fig. 2. Samples in FRGC, containing the visual images and corresponding depth images

However, the depth images in FRGC are not "complete", they contain much noise, holes and large missing data. Besides, for training purpose the 3D face scan for each sample should have the same semantical meaning (for example, the kth point corresponds to the left eye corner in all the samples). To make the raw data usable, registering a template to depth images for filling holes, estimating missing patches and making every scans in full correspondence are necessary.

The optimal non-rigid ICP algorithm [4] can register a template to any target surface with the same semantic. It searches for the best deformation for each point by minimizing a cost function with 3 terms: the distance term which minimizes the distances between the template points and their closest target surface points, the stiffness term which penalises the differences of the transformations of neighbouring points, and the landmark term which guides the initialization and minimizes the distances between corresponding landmarks. Although the optimal non-rigid-ICP has shown its good performance in the construction of BFM face model [27], it has difficulty in handling large missing regions, because there are no closest points for the distance term in these regions and only stiffness constraint alone will give bad results as in Fig. 3(b).

To deal with large missing patches, we fit a 3DMM to constrain the template points falling onto the missing regions. Since the distance term provides a set of correspondences between template points and target points, a 3DMM can be fitted with the target point positions by common 3D-3DMM fitting methods [3]. Note that 3DMM is controlled by PCA coefficients, the missing regions are automatically estimated. In the registration process, for the template points having no closest points on the target surface, we find their closest points on the fitted 3DMM instead. Thus every point will have a distance term constraint. Fig. 3(c) shows the results of the new method, the filling of missing regions is smooth and looks reasonable.

The registered template can be seen as an approximation of the complete face scan. Since we have known the position of every point of 3DMM, we can get the best fitted pose and shape parameters through:

$$\arg\min_{(\alpha,\gamma)} \|Rig - fr(\overline{s} + \sum_{i=1}^{m-1} \alpha_i \cdot s_i + t_{3d})\|^2 \qquad (16)$$

where Rig is the point positions of registered template, $f, r, t_{3d}, s, \alpha, \gamma$ have the same meanings as in Equ. (1)(2). Fig. 4 shows the comparison of depth image, registered template and best fitted 3DMM. We can see that even though the best fitted 3DMM loses some details because of the limited expressive ability of PCA shape model, it is close to the depth image. In the training process, the best fitted parameters will be used as the target of regression.

B. Training Data Augmentation

Considering the success of 2D face alignment in recent years, we use the landmarks detected by SDM [37] to initialize pose and shape parameters. To achieve better generalisation ability, we augment the training set by randomly disturbing the bounding box and running SDM to get multiple groups of landmarks for each sample, as shown in Fig. 5.

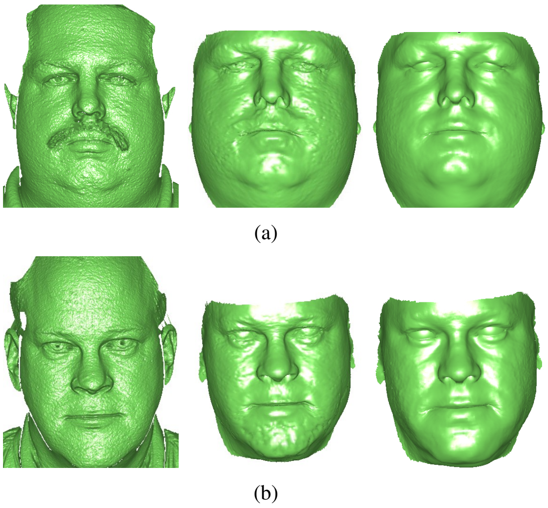


Fig. 4. 3D shapes of two subjects. For each subject, left is the depth image, middle is registered template and right is the best fitted 3DMM shape.

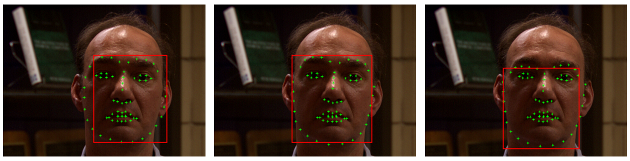


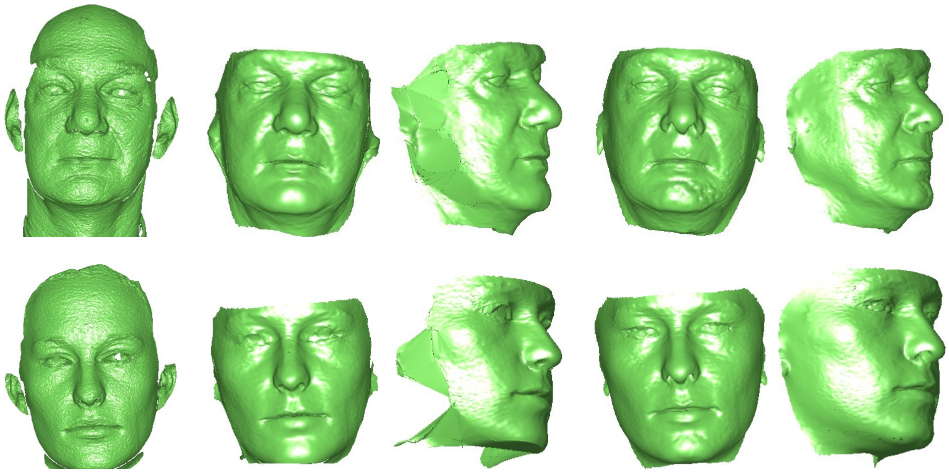
Fig. 5. The augmentation of initialization.

It can be seen that the bounding box may affect landmark detection seriously and the eyebrow and border landmarks are always not accurate enough. The non-accuracy of initialization is very common in automated 3DMM reconstruction system and the augmentation is found to be helpful to achieve the robustness to rough initialization.

VI. Experiments

We use the Basel Face Model (BFM) [27] as our 3D Morphable Model and conduct fitting experiments on the Spring2004range subset of Face Recognition Grand Challenge (FRGC) [28] database. The BFM provides a PCA shape model with 53490 vertices computed from 200 face scans. It can cover most of face shapes in the real world but cannot handle expressions. In our experiment, we eliminate the ear, neck and forehead regions because they are less important and easily occluded by hair and clothes. The processed model has 39226 vertices left. The Spring2004range has 2114 samples. Each sample consists of a frontal face image and a depth image with pixels in full correspondence. The faces with expression are discarded due to the limitation of BFM, with 1443 samples left.

For each sample in the database, the registration method in section 5.1 is used to get the target parameters and the face alignment algorithm is used to localize the landmarks for initialization. The landmarks are detected automatically using DPM face detector [40] and SDM face alignment [37], and most alignment results are accurate except for few samples (less than 10). As for error measure, we consider the depth image in FRGC as the ground truth shape of each sample. In the testing process, we first project the reconstructed 3D shape to a depth image, and the Root Mean Square Error



(a) (b) (c)

Fig. 3. (a) is the target surface, (b) is the results of optimal non-rigid ICP, (c) is the results of our registering method.

(RMSE) between the reconstructed depth image and the ground truth is measured as the fitting error. Invalid regions including holes and missing patches are ignored. Training and testing are conducted in 10-fold cross-validation without identity overlapping to measure all the 1443 samples in the database.

We compare our method with two popular 3DMM fitting algorithms. The first is the Multi-Features Framework (MF-F) [31] which considers landmark, contour, textured edge and pixel intensity jointly to fit a 3DMM. It is the state-of-the-art of traditional analysis-by-synthesis based methods. The other is the landmark based fitting method [2] which only uses landmarks to estimate 3DMM shape. It heavily depends on the accuracy of landmark but is much more efficient than MFF. Besides, the method also claims to be the state-of-the-art on a synthetic database with provided landmarks.

Considering facial component region (including eye, nose and mouth regions of BFM in [27]) is more important in most face applications and the cheek area is always occluded by hair in FRGC, we conduct two experiments by computing error on full face and facial component area respectively. Fig. 6 shows the Cumulative Error Distribution (CED) curves of the three methods and Table II shows the mean error of all 1443 samples. Clearly our method outperforms both MFF and the landmark based method. Besides, Fig. 8 shows some fitting results of landmark based method, Multi-Features Framework, and discriminative 3DMM fitting. The resulting shapes are lightened by a frontal light to highlight the difference. Note that in the first row the MFF and landmark based method both fail because the non-accurate landmarks give a bad initialization. However, our method still converges to an accurate shape, showing the robustness to rough initialization. In Fig. 7, we show some fitting results of real world images. Although without ground truth shapes, we can see the fitting results are reasonable in visual sense.

In addition to the fitting accuracy, the running time performance of our method is also promising. It takes about 12 mins to train 6500 samples and 0.9s to fit a testing image on a 3.4GHz Intel Core i7 computer, which is close to the landmark based method (0.2s) and faster than MMF which needs about 69.58s to fit an image on 3.0GHz Intel Pentium IV [29].

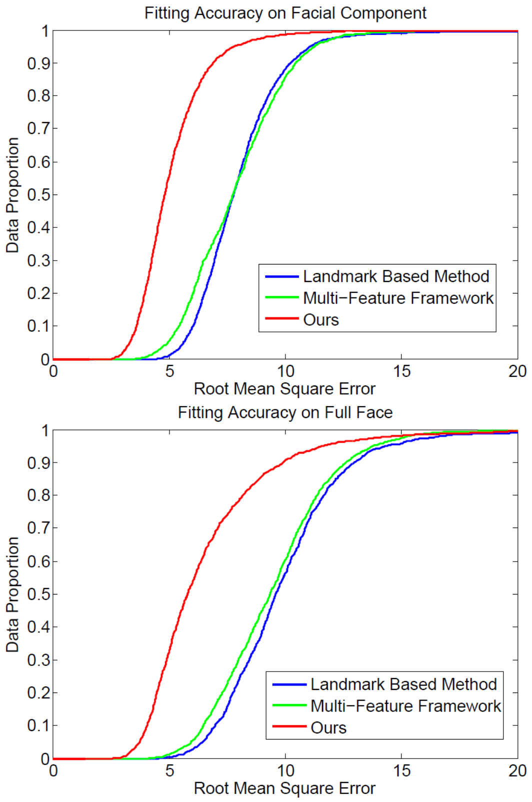


Fig. 6. First: fitting accuracy on facial component region. Second: fitting accuracy on full face.

VII. Conclusions

We have proposed a novel discriminative 3DMM fitting algorithm, where the 3D shape is estimated by directly

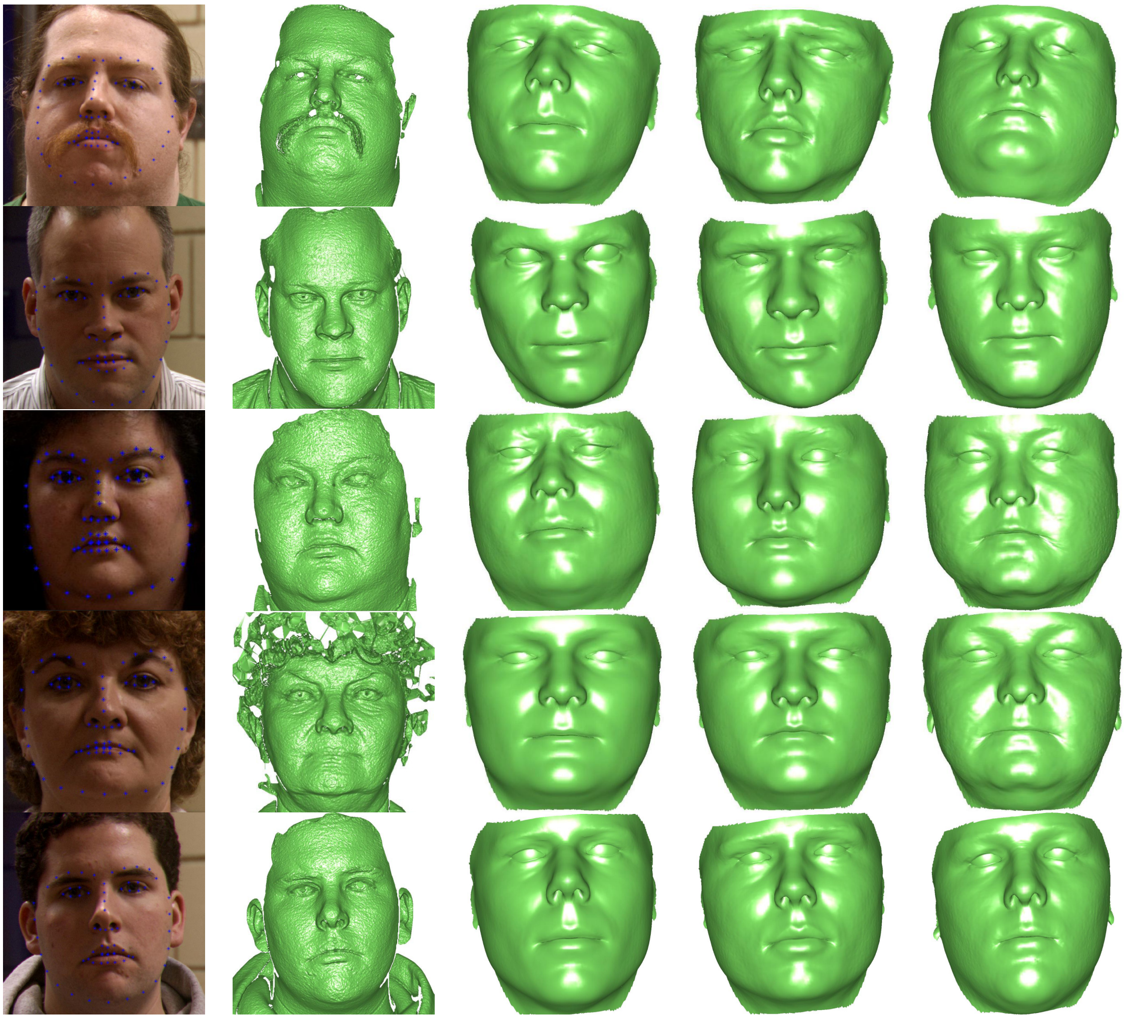


Fig. 8. Comparison of landmark based method, MFF and our method. First column: visual images and landmarks; Second column: depth images; Third column: results of landmark based method; Fourth column: results of MFF. Last column: results of ours.

TABLE II

MEAN ERROR OF LANDMARK BASED METHOD, MFF AND OUR METHOD

Region	Landmark	MultiFeatures	Ours
Facial Component	8.0837	7.8820	5.3505
Full Face	9.9956	9.5051	6.8849

regressing image features instead of minimizing a cost function with Gauss-Newton methods. The resulting method is highly accurate, robust and efficient. Experimental results on FRGC suggest that our approach significantly outperforms the popular landmark based method and the state-of-the-art Multi-Features Framework.

Despite the outstanding performance of our discriminative 3DMM fitting method, there is still large room for future improvements. For example, by training the fitting model on a database containing visual and depth image pairs in the wild, it may be able to handle more complicated scenarios.

VIII. ACKNOWLEDGMENTS

This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

REFERENCES

[1] O. Aldrian and W. A. Smith. A linear approach of 3d face shape and texture recovery using a 3d morphable model. In *Proceedings of the British Machine Vision Conference, pages*, pages 75–1, 2010.

[2] O. Aldrian and W. A. Smith. Inverse rendering of faces with a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1080–1093, 2013.

[3] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3d face recognition with a morphable model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[4] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

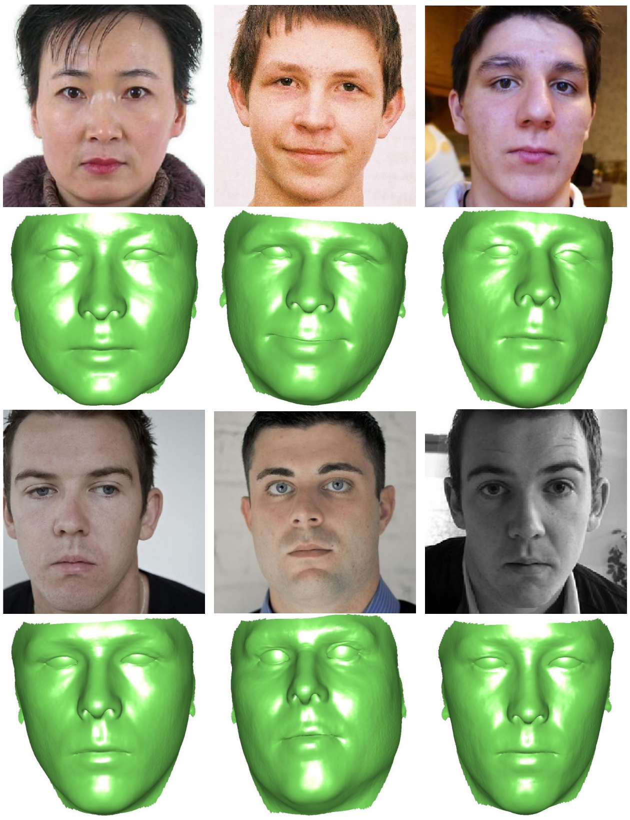


Fig. 7. Fitting results of real-world data.

[5] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1859–1866. IEEE, 2014.

[7] T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2610–2617. IEEE, 2012.

[8] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 293–300. IEEE, 2004.

[9] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[10] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, 2003.

[11] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41, 2013.

[12] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2887–2894. IEEE, 2012.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, 2001.

[14] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[15] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *Computer Vision–ECCV 2012*, pages 278–291. Springer, 2012.

[16] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[17] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 17, pages 929–938, 2006.

[18] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1078–1085. IEEE, 2010.

[19] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[20] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *Computer Vision–ECCV 2008*, pages 413–426. Springer, 2008.

[21] T. Hassner. Viewing real-world faces in 3d. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3607–3614. IEEE, 2013.

[22] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):394–405, 2011.

[23] Z. Lei, Q. Bai, R. He, and S. Z. Li. Face shape recovery from a single image using cca mapping between tensor spaces. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[24] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[25] P. Mortazavian, J. Kittler, and W. Christmas. A 3-d assisted generative model for facial texture super-resolution. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, pages 1–7. IEEE, 2009.

[26] A. Patel and W. A. Smith. 3d morphable face models revisited. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1327–1334. IEEE, 2009.

[27] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. IEEE, 2009.

[28] P. J. Phillips. Face recognition grand challenge. In *Biometric Consortium Conference*, 2004.

[29] S. Romdhani. *Face image analysis using a multiple features fitting strategy*. PhD thesis, University of Basel, 2005.

[30] S. Romdhani and T. Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 59–66. IEEE, 2003.

[31] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 986–993. IEEE, 2005.

[32] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[33] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[34] M. Tech. The bjut-3d large-scale chinese face database. Technical report, Graphics Lab, Technical Report, Beijing University of Technology, 2005.

[35] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild.

[36] Y. Wang, S. Lucey, and J. F. Cohn. Enforcing convexity for improved alignment with constrained local models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[37] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. CVPR, 2013.

[38] L. Zhang and D. Samaras. Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):351–363, 2006.

[39] L. Zhang, Y. Wang, S. Wang, D. Samaras, S. Zhang, and P. Huang. Image-driven re-targeting and relighting of facial expressions. In *Computer Graphics International 2005*, pages 11–18. IEEE, 2005.

[40] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.