Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

On 3D face reconstruction via cascaded regression in shape space*

Feng LIU, Dan ZENG, Jing LI, Qi-jun ZHAO†‡
(*College of Computer Science, Sichuan University, Chengdu 610065, China*)
†E-mail: qjzhao@scu.edu.cn

Received Apr. 16, 2017; Revision accepted Aug. 9, 2017; Crosschecked Dec. 20, 2017

Abstract: Cascaded regression has been recently applied to reconstruct 3D faces from single 2D images directly in shape space, and has achieved state-of-the-art performance. We investigate thoroughly such cascaded regression based 3D face reconstruction approaches from four perspectives that are not well been studied: (1) the impact of the number of 2D landmarks; (2) the impact of the number of 3D vertices; (3) the way of using standalone automated landmark detection methods; (4) the convergence property. To answer these questions, a simplified cascaded regression based 3D face reconstruction method is devised. This can be integrated with standalone automated landmark detection methods and reconstruct 3D face shapes that have the same pose and expression as the input face images, rather than normalized pose and expression. An effective training method is also proposed by disturbing the automatically detected landmarks. Comprehensive evaluation experiments have been carried out to compare to other 3D face reconstruction methods. The results not only deepen the understanding of cascaded regression based 3D face reconstruction approaches, but also prove the effectiveness of the proposed method.

Key words: 3D face reconstruction; Cascaded regressor; Shape space; Real-time
https://doi.org/10.1631/FITEE.1700253 **CLC number:** TP319

1 Introduction

As a fundamental problem in computer vision, reconstructing three-dimensional (3D) face shapes from two-dimensional (2D) images has recently gained increasing attention because a 3D face provides invariant features to the variations of pose, illumination, and expression. The reconstructed 3D faces are therefore useful for many real-world applications, for example, pose robust face recognition (Blanz and Vetter, 2003; Han and Jain, 2012; Hu *et al.*, 2014; Zhu XY *et al.*, 2015), 3D facial expression analysis (Chu *et al.*, 2014; Ren *et al.*, 2016), and facial animation (Cao *et al.*, 2014a; 2016). Using 3D face shapes to recognize identities is believed to be more robust and more accurate than using only 2D face images (Abiantun *et al.*, 2014). Despite its high recognition accuracy, fast acquisition of high-resolution and high-precision 3D face shapes is still difficult, especially under varying conditions or at a long distance. On the other hand, 2D face images can be much more easily captured with available cameras, and there are already a lot of 2D face image databases. Thus, there is great interest in developing efficient methods for reconstructing 3D faces from 2D face images such that the rich resources of 2D face images and facilities can be better used.

Liu *et al.* (2016) recently proposed a novel method for reconstructing 3D face shapes from single 2D images via cascaded regression in a 2D/3D

‡ Corresponding author
* Project supported by the National Key Research and Development Program of China (Nos. 2017YFB0802303 and 2016YFC0801100), the National Key Scientific Instrument and Equipment Development Projects of China (No. 2013YQ49087904), the National Natural Science Foundation of China (No. 61773270), and the Miaozi Key Project in Science and Technology Innovation Program of Sichuan Province, China (No. 2017RZ0016)
ORCID: Feng LIU, http://orcid.org/0000-0002-6625-0593

shape space. It is based on the observation that the landmarks' locations on the 2D image can be derived from the reconstructed 3D shape, and the displacement of derived landmarks from their true positions is correlated with the accuracy of the reconstructed 3D shape. This method can simultaneously locate facial landmarks and reconstruct 3D face shapes with two sets of cascaded regressors: one for updating landmarks and the other for 3D face shapes. By effectively exploring the correlation between 2D landmarks and 3D shapes, this method achieves state-of-the-art performance in both face alignment and 3D face reconstruction for face images of an arbitrary view and expression. However, some problems are still not well addressed in such shape space regression based 3D face reconstruction methods.

1. Impact of the number of 2D landmarks. Different sets of 2D landmarks are used in the face alignment and recognition literature, e.g., 68 landmarks (Sagonas *et al.*, 2013), 21 landmarks (Köstinger *et al.*, 2011), and 5 landmarks (Sun *et al.*, 2013). How will the 3D face reconstruction accuracy be affected if different numbers of 2D landmarks are used to guide the 3D face reconstruction process?

2. Impact of the number of 3D vertices. 3D face shapes can be represented by different numbers of vertices, i.e., different 3D point cloud densities and coverage. Will a sparse or narrow 3D face shape be much easier to reconstruct with a higher accuracy than a dense or wide 3D face shape? Note that a wide 3D face shape covers more areas than a narrow 3D face shape. For instance, a 3D face shape covering only eyes, eyebrows, nose, and mouth is narrow compared with a 3D face shape covering the area from the left ear to the right ear.

3. What if using standalone landmark localization methods? Although the method mentioned in Liu *et al.* (2016) can simultaneously locate 2D landmarks and reconstruct 3D shapes, it requires that the training 2D face images should be annotated with both visible and invisible landmarks. Manually marking invisible landmarks is, however, very difficult and error-prone. Is it possible to integrate standalone landmark localization methods with the 3D face reconstruction process proposed in Liu *et al.* (2016)?

4. Convergence. As an iterative approach, how many iterations would be necessary for the proposed method to achieve an acceptable performance in terms of both accuracy and efficiency? In other words, what is the convergence property of shape space regression based 3D face reconstruction methods?

We aim to investigate the shape space regression based 3D face reconstruction approach from the four aforementioned aspects. To this end, we first revise and implement the method in Liu *et al.* (2016) so that the 3D face reconstruction process can take 2D landmarks that are provided by a third party as the input, and reconstruct 3D face shapes that have the same pose and expression as the input images, rather than frontal pose and neutral expression. Fig. 1 shows the results of the method on some photos from the AFW database (Zhu and Ramanan, 2012) using the ground-truth visible 2D landmarks as the input. We then experimentally evaluate the convergence and computational complexity of the implemented 3D face reconstruction method. Afterwards, we conduct extensive experiments to assess the impact of the number of 2D landmarks and the number of 3D vertices on reconstruction accuracy. We finally attempt to integrate state-of-the-art landmark localization methods to the 3D face reconstruction process.

Fig. 1 Reconstruction results of the proposed method on face images from the AFW database (Zhu and Ramanan, 2012) with arbitrary expressions and poses

2 Related work

To solve the intrinsically ill-posed single-view 3D face reconstruction problem, different priors or

constraints have been introduced, resulting in the shape from shading (SFS) based methods and 3D morphable model (3DMM) based methods. SFS based methods (Horn and Brooks, 1989; Barron and Malik, 2012) recover 3D shapes via analyzing certain clues in the 2D texture images, with an assumption of the Lambertian reflectance and a single-point light source at infinity. While classical SFS based methods (Kemelmacher-Shlizerman and Basri, 2011; Suwajanakorn *et al.*, 2014; Li *et al.*, 2015; Zeng *et al.*, 2017) are initially designed for generic 3D shape reconstruction, their performances in recovering 3D face shapes can be further improved by using some reference 3D face models as additional constraints. These methods usually have limited accuracy because: (1) their assumed connection between 2D texture clues and 3D shape information is too weak to discriminate between different human faces; (2) they do not fully exploit the prior knowledge of 3D faces and significantly depend on the reference models; (3) they reconstruct a depth map or 2.5D shape instead of a 3D full shape since they tend to operate on a face with a narrow range of poses.

The 3D morphable model (3DMM) (Blanz and Vetter, 1999; Romdhani and Vetter, 2005; Aldrian and Smith, 2013; Zhu XY *et al.*, 2014; 2015; Bas *et al.*, 2016; Booth *et al.*, 2016), as a typical statistical 3D face model, explicitly learns the prior knowledge of 3D faces with a statistical parametric model. It represents a 3D face as a linear combination of basis 3D faces, which are obtained by applying the principal component analysis (PCA) on a set of densely aligned 3D faces. To recover the 3D face from a 2D image, the combination coefficients are estimated by minimizing the discrepancy between the input 2D face image and the one rendered from the reconstructed 3D face. These 3DMM based methods can better cope with 2D images of varying illuminations and poses. However, they are limited in individualized or detailed reconstruction because PCA conducts global modeling in essence, and they involve a time-consuming on-line optimization process to search for the optimal solution in the parameter space. Moreover, 3DMM needs an additional linear expression model to handle facial expressions, namely E-3DMM (Cao *et al.*, 2014b; Chu *et al.*, 2014; Zhu XY *et al.*, 2015). However, neither SFS-based nor 3DMM-based methods can consistently well cope with rotated or expressive face images due to invisible or deformed facial landmarks on them.

Motivated by the success of cascaded regression in 2D facial landmark localization (Xiong and de la Torre, 2013; Jourabloo and Liu, 2015; 2017; Li *et al.*, 2016), Liu *et al.* (2016) recently proposed a 2D/3D shape space regression based method for reconstructing 3D face shapes from single images of arbitrary views and expressions. The method alternately applies 2D landmark regressors and 3D shape regressors. The 2D landmark regressors estimate landmark locations by regressing over the texture features around landmarks, while the 3D shape regressors reconstruct 3D face shapes via regressing over the 2D landmarks. Unlike existing 3D face reconstruction methods, this method directly estimates 3D faces in the 3D shape space via cascaded regression, getting rid of parameterized 3D face models and assumed illumination models. As a result, it achieves state-of-the-art performance on both accuracy and efficiency of 3D face reconstruction. Fig. 2 shows example results of SFS-based, 3DMM-based, E-3DMM-based, and shape-space-regression-based methods on rotated and expressive face images. In this study, we will thoroughly assess the effectiveness of such shape space regression based 3D face reconstruction methods from various perspectives.

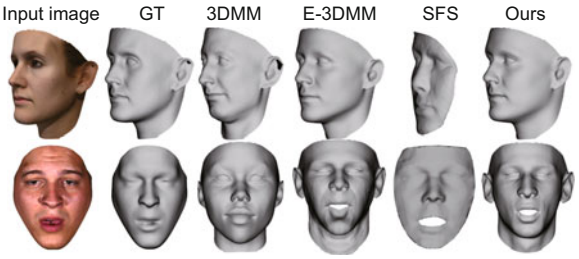

Fig. 2 Reconstruction results for images in the Basel face model (BFM) (top row) (Paysan *et al.*, 2009) and BU3DFE (bottom row) (Yin *et al.*, 2006) databases. From the left to the right columns: input images, ground-truth 3D shapes (GT), results by 3DMM (Bas *et al.*, 2016), E-3DMM (Zhu XY *et al.*, 2015), SFS (Kemelmacher-Shlizerman and Basri, 2011), and the proposed method

3 Shape space regression based approach

3.1 Overview

We denote a 3D face shape as $\boldsymbol{S} \in \mathbb{R}^{3\times n}$, which is represented by 3D locations of n vertices, and

denote $\boldsymbol{S}_{\mathrm{L}}$ as a subset of $\boldsymbol{S}$ with columns corresponding to l annotated landmarks (e.g., eye corners and nose tip). The projection of these 3D landmarks on the 2D face image $\boldsymbol{I}$ is represented by $\boldsymbol{U} \in \mathbb{R}^{2\times l}$. The relationship between 2D facial landmarks $\boldsymbol{U}$ and its corresponding 3D landmarks $\boldsymbol{S}_{\mathrm{L}}$ can be described as

$$\boldsymbol{U} = \boldsymbol{M}\boldsymbol{S}_{\mathrm{L}} = \boldsymbol{M}D_N(\boldsymbol{R}\tilde{\boldsymbol{S}}_{\mathrm{L}} + \boldsymbol{T}), \tag{1}$$

where $\tilde{\boldsymbol{S}}$ is a frontal 3D face with a neutral expression, $\{\boldsymbol{R} \in \mathbb{R}^{3\times 3}, \boldsymbol{T} \in \mathbb{R}^{3\times l}\}$, and $D_N(\cdot)$ are, respectively, rigid deformation (i.e., rotation and translation) caused by pose variations and a non-rigid deformation function caused by expression variations that occur to $\tilde{\boldsymbol{S}}$ resulting in the observed 3D face $\boldsymbol{S}$, and $\boldsymbol{M} \in \mathbb{R}^{2\times 3}$ is the camera projection matrix. Here, we employ a weak perspective projection for $\boldsymbol{M}$ as conventionally done in Zhou *et al.* (2015).

Our purpose is to reconstruct $\boldsymbol{S}$ (rather than $\tilde{\boldsymbol{S}}$) from the given 'ground-truth' visible landmarks $\boldsymbol{U}^*$ (either manually marked or automatically detected by a standalone method) for the face image $\boldsymbol{I}$. As discussed above, we achieve this by iteratively updating the initial estimate of $\boldsymbol{S}$ with a series of regressors in the 3D face shape space. These regressors calculate the adjustment to the estimated 3D face shape according to the deviation between the ground-truth landmarks and the landmarks rendered from the estimated 3D face shape. Fig. 3 shows the flowchart of the proposed method.

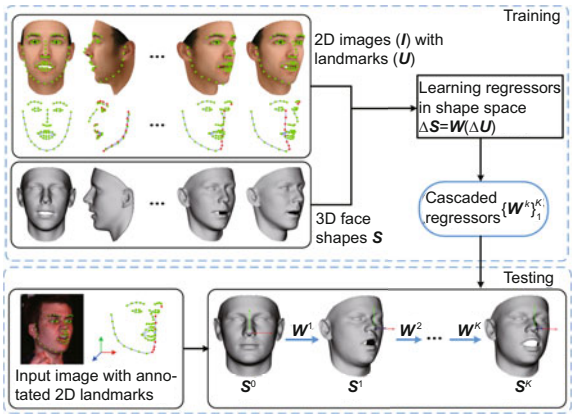

Fig. 3 Flowchart of the shape space cascaded regression based 3D face reconstruction method. Green and red points denote visible and invisible landmarks, respectively. Note that the method in this study does not require invisible landmarks' locations as the input. References to color refer to the online version of this figure

3.2 Reconstruction process

Let $\boldsymbol{U}^*$ be the 'ground-truth' landmarks (either manually annotated or automatically detected) on an input 2D image, and $\boldsymbol{S}^{k-1}$ is the currently reconstructed 3D shape after $(k-1)$ iterations. The corresponding landmarks $\boldsymbol{U}^{k-1}$ can be obtained by projecting $\boldsymbol{S}^{k-1}$ onto the image according to Eq. (1). Then the updated 3D shape $\boldsymbol{S}^k$ can be computed by

$$\boldsymbol{S}^k = \boldsymbol{S}^{k-1} + \boldsymbol{W}^k(\boldsymbol{U}^* - \boldsymbol{U}^{k-1}) + \boldsymbol{b}^k, \tag{2}$$

where $\boldsymbol{W}^k$ is the regressor in the kth iteration and $\boldsymbol{b}^k$ is a bias term (in the rest of this paper we omit the bias term for simplicity because it can be shrunk into the regressors).

3.3 Learning cascaded regressors

The K regressors $\{\boldsymbol{W}^k\}_1^K$ involved in the reconstruction process can be learned via optimizing the following objective function over the N training samples:

$$\arg\min_{\boldsymbol{W}^k} \sum_{i=1}^{N} \|(\boldsymbol{S}_i^* - \boldsymbol{S}_i^{k-1}) - \boldsymbol{W}^k(\boldsymbol{U}_i^* - \boldsymbol{U}_i^{k-1})\|^2, \tag{3}$$

where $\{\boldsymbol{S}_i^*, \boldsymbol{U}_i^*\}$ is a training sample consisting of ground-truth landmarks on the ith 2D face image and its corresponding ground-truth 3D face shape that has the same pose and expression as the face image. Mathematically, the above optimization seeks a regressor that can minimize the overall error of the entire reconstructed 3D face shapes, but not merely the error at the landmarks.

In this study, we use linear regressors $\boldsymbol{W}^k \in \mathbb{R}^{3n\times 2l}$. Then the optimization (3) can be easily solved by using least squares methods with a solution of

$$\boldsymbol{W}^k = \Delta\boldsymbol{S}^k(\Delta\boldsymbol{U}^k)^{\mathrm{T}}(\Delta\boldsymbol{U}^k(\Delta\boldsymbol{U}^k)^{\mathrm{T}})^{-1}, \tag{4}$$

where $\Delta\boldsymbol{S}^k = \boldsymbol{S}^* - \boldsymbol{S}^{k-1}$ and $\Delta\boldsymbol{U}^k = \boldsymbol{U}^* - \boldsymbol{U}^{k-1}$ are 3D shape adjustment and 2D landmark deviation, respectively. $\boldsymbol{S} \in \mathbb{R}^{3n\times N}$ and $\boldsymbol{U} \in \mathbb{R}^{2l\times N}$ denote, respectively, the ensemble of 3D face shapes and 2D landmarks of all training samples with each column corresponding to one sample. Note that we write the 3D face shape and 2D landmarks as column vectors: $\boldsymbol{S} = (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_n, y_n, z_n)^{\mathrm{T}}$ and $\boldsymbol{U} = (u_1, v_1, u_2, v_2, \ldots, u_l, v_l)^{\mathrm{T}}$. To ensure a valid solution in Eq. (4), N should be larger than $2l$ so that $\Delta\boldsymbol{U}^k(\Delta\boldsymbol{U}^k)^{\mathrm{T}}$ is invertible. Fortunately, since

the sets of used landmarks are usually sparse, this requirement can be easily satisfied in real-world applications. The whole process of learning the cascaded regressors is summarized in Algorithm 1.

Algorithm 1 3D cascaded regressor learning

Input: Training data $\{(\boldsymbol{I}_i, \boldsymbol{S}_i^*, \boldsymbol{U}_i^*) \mid i = 1, 2, \ldots, N\}$; initial shape $\boldsymbol{S}_i^0$; camera projection matrix $\boldsymbol{M}$.

1: **for** $k = 1$ to K **do**
2: Estimate 2D projection $\boldsymbol{U}_i^{k-1}$ from current 3D face $\boldsymbol{S}_i^{k-1}$ via Eq. (1);
3: Compute 2D landmark adjustment and 3D face adjustment for all samples: $\Delta\boldsymbol{U}^k = \boldsymbol{U}^* - \boldsymbol{U}^{k-1}$, $\Delta\boldsymbol{S}^k = \boldsymbol{S}^* - \boldsymbol{S}^{k-1}$;
4: Estimate $\boldsymbol{W}^k$ via Eq. (3);
5: Update 3D face $\boldsymbol{S}_i^k$ via Eq. (2);
6: **end for**

Output: Cascaded regressors $\{\boldsymbol{W}^k\}_{k=1}^K$.

4 Implementation details

4.1 Initialization

The proposed iterative method has two terms to initialize: the initial 3D face shape $\boldsymbol{S}^0$ and the camera projection matrix $\boldsymbol{M}$. Given the set of training samples, we select all the frontal faces with a neutral expression. The mean of these selected 3D face shapes is computed and used to initialize $\boldsymbol{S}^0$. Similarly, the mean of their 2D landmarks is calculated and denoted as $\boldsymbol{U}^0$. The camera projection matrix $\boldsymbol{M}$ can then be estimated by solving the following least squares fitting problem:

$$\boldsymbol{M} = \arg\min_{\boldsymbol{M}} \|\boldsymbol{U}^0 - \boldsymbol{M}\boldsymbol{S}_{\mathrm{L}}^0\|_2^2. \tag{5}$$

The obtained projection matrix $\boldsymbol{M}$ is used throughout the 3D face reconstruction process to render 2D landmarks from the reconstructed 3D face shapes.

4.2 Landmarks

Fig. 4 depicts the 68 facial landmarks ($l = 68$) considered in this study. Obviously, some of the landmarks will become invisible on the 2D face image due to self-occlusion when the face has large pose angles. These invisible landmarks are difficult to precisely annotate. Hence, we treat them as missing data and fill their corresponding entries in $\boldsymbol{U}$ with zero. In this way, these invisible landmarks will not affect the reconstruction, and thus images of arbitrary pose angles can be handled in a unified framework.

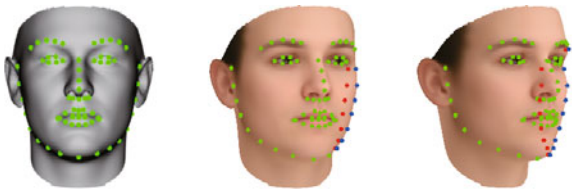

Fig. 4 Sixty-eight landmarks are used in this study. Left: landmarks annotated on a 3D face. Middle and right: corresponding landmarks annotated on its 2D images with yaw angles of 20° and 40°, respectively. Green and red points on the 2D images indicate, respectively, visible and invisibile landmarks, and blue points mark the contour instead of semantic landmarks. References to color refer to the online version of this figure

To automatically detect the visible landmarks in the testing phase, we first employ a state-of-the-art face alignment approach to automatically locate 2D landmarks positions, and then compute their visibility. Most conventional face alignment methods like the one proposed by Kazemi and Sullivan (2014) cannot detect invisible self-occluded landmarks (refer to the red points in Fig. 4). To determine the visibility of 2D landmarks projected from the reconstructed 3D face shape, given the detected 2D landmarks $\boldsymbol{U}$ on the face image and the 3D annotated landmarks $\boldsymbol{S}_{\mathrm{L}}^0$ from the initial 3D shape $\boldsymbol{S}^0$, we coarsely estimate the camera projection matrix $\boldsymbol{M}$ by Eq. (5). Suppose the 3D surface normal at landmarks in $\boldsymbol{S}^0$ is $\boldsymbol{N}$. The initial visibility $\boldsymbol{v}$ can then be measured by (Jourabloo and Liu, 2016)

$$\boldsymbol{v} = \frac{1}{2}\left(1 + \mathrm{sgn}\left(\boldsymbol{N} \cdot \left(\frac{\boldsymbol{M}_1}{\|\boldsymbol{M}_1\|} \times \frac{\boldsymbol{M}_2}{\|\boldsymbol{M}_2\|}\right)\right)\right), \tag{6}$$

where $\mathrm{sgn}(\cdot)$ is the sign function, '$\cdot$' means dot product, and '$\times$' cross-product. $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are the leftmost three elements at the first and second rows of the mapping matrix $\boldsymbol{M}$, respectively. This basically rotates the surface normal and validates whether it points toward the camera or not. Finally, to maintain consistency with the training setting, the invisible corresponding entries in $\boldsymbol{U}$ should be filled with zero.

4.3 Alignment

To simplify the camera projection model, we assume that both 3D face shapes and 2D landmarks are well aligned. Specifically, (1) all the 3D face shapes have been established as point-to-point dense registration (i.e., they have the same number of vertices, and the vertices of the same index have the same

semantic meaning), (2) all the 3D face shapes are centered at the origin of the world coordinate system, and (3) all the faces on the 2D images are also centered in the image coordinate system. With these aligned 3D & 2D face data, and as we separate face deformation from camera projection (Eq. (1)), the employed weak perspective camera projection matrix $\boldsymbol{M}$ has only one free parameter, i.e., the scaling factor or focal length, which will be estimated based on the training data.

5 Experimental results

5.1 Training data

A set of 3D face shapes and corresponding 2D face images with annotated landmarks are needed to train regressors in the proposed method. To make the trained regressors robust to pose and expression variations, samples in the training dataset should have good diversity in their poses and expressions. However, it is difficult to find (in the public domain) such datasets of 3D face shapes and corresponding annotated 2D images with various expressions/poses. Therefore, we use the Basel face model (BFM) (Paysan *et al.*, 2009) to construct synthetic 3D faces of 200 subjects (50% female), and use the expression model from Face Warehouse (Cao *et al.*, 2014b) to generate random expressions on each of the 3D faces. These expressive 3D faces are then projected onto 2D images with 55 views of 11 yaw (0°, ±15°, ±30°, ±50°, ±70°, ±90°) and 5 pitch (0°, ±15°, ±30°) rotations, resulting in a total number of 11 000 3D faces and corresponding synthetic images. Each 3D face consists of 53 215 vertices (the original BFM model has 53 490 vertices, but we discard the vertices in the tongue region). The 2D image resolution is 875×656 pixels and the inter-eye distance is about 220 pixels. The 68 landmarks on each 2D face image are recorded during the projection process (the 3D faces are densely aligned and the indices of the landmarks are known), and the invisible landmarks are marked as zero as mentioned above.

5.2 Convergence and computational complexity

We experimentally investigate the convergence of the training process of the proposed cascaded regressors. For this, we record the value of the objective function defined in Eq. (3) at each iteration during the training process. Fig. 5 shows the objective function value for 10 iterations. Clearly, the objective function value decreases substantially in the first five iterations and becomes stable after seven iterations. This demonstrates the good convergence of the proposed method. In the following experiments, we empirically set $K = 5$ as a trade-off between accuracy and efficiency.

According to our experiments on a PC with an i7-4710 CPU and 8 GB memory, the Matlab implementation of the proposed method runs at nearly 26 frames per second. This indicates that the proposed method can reconstruct 3D faces in real time.

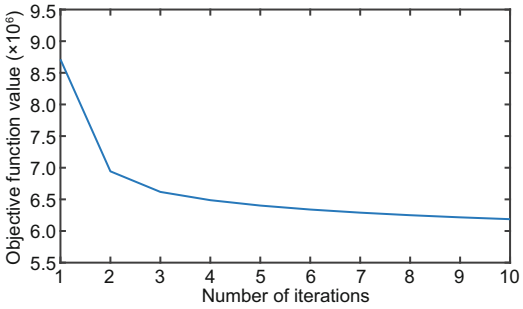

Fig. 5 Objective function values as the iteration proceeds

5.3 Reconstruction accuracy across poses on BFM

The BFM database (Paysan *et al.*, 2009) provides 10 test face subjects, each of whom has nine face images of neutral expression and different poses, including one frontal and eight yaw poses (±15°, ±30°, ±50°, ±70°). Here, the metric used to evaluate the 3D face shape reconstruction accuracy is the mean absolute error (MAE), which is defined as

$$\mathrm{MAE} = \frac{1}{N_{\mathrm{T}}} \sum_{i=1}^{N_{\mathrm{T}}} (\|\boldsymbol{S}_i^* - \hat{\boldsymbol{S}}_i\|/n), \tag{7}$$

where N_{T} is the total number of test samples, $\|\boldsymbol{S}_i^* - \hat{\boldsymbol{S}}_i\|$ is the Euclidean distance between ground-truth shape $\boldsymbol{S}_i^*$ and reconstructed 3D shape $\hat{\boldsymbol{S}}_i$ of the ith test sample. We report the MAE after Procrustes alignment.

In this experiment, we use the visible landmarks projected from ground-truth 3D face shapes as the input. The proposed method is compared with several state-of-the-art methods based on 3DMM,

including the approach proposed by Aldrian and Smith (2013), the multi-feature 3DMM framework based on contours, textured edges, specular highlights, and pixel intensity proposed by Romdhani and Vetter (2005), sparse SIFT flow 3DMM (SSF-3DMM (Zhu *et al.*, 2014)), and the edge-fitting based 3DMM approach proposed by Bas *et al.* (2016).

Table 1 shows the MAE of different methods on the BFM database with respect to different poses of face images. Obviously, the average MAE of the proposed method is lower than the counterpart of other methods. Moreover, its accuracy is quite stable across different poses. This proves the effectiveness of the proposed method in handling face images of arbitrary poses. Fig. 6 shows the reconstruction results of our method and SSF-3DMM (Zhu *et al.*, 2014) on two subjects in the BFM database.

5.4 Impact of the number of 2D landmarks

To assess how the reconstruction accuracy changes as fewer landmarks are used, we divide a face into four regions, i.e., nose, eyes, mouth, and others (Fig. 7), and use different numbers of landmarks in these regions. Note that the number of vertices in the output reconstructed 3D face shape remains unchanged. Fig. 7 shows the results, from which the following two observations can be made: (1) While using more landmarks boosts the reconstruction accuracy for all regions, the gains of different regions are not uniform; (2) When similar numbers of landmarks (e.g., seven to nine landmarks) are used, the region of eyes achieves the smallest error among the four regions, and the error of nose region is relatively high. A possible explanation is due to the varying complexity of different regions and to the different significance of different landmarks. Moreover, in the above experiment, the nose landmarks do not distribute along the nose contour, whereas the landmarks in eyes and mouth regions profile these two regions. This might be another reason why the reconstruction error of the nose region is larger than that of eyes and mouth regions. For a better evaluation of the impact of 2D landmarks, more extensive experiment is required, which will be part of our future work. In the following experiments, we will use the set of 68 landmarks (unless specified otherwise).

5.5 Impact of the number of 3D vertices

In this experiment, we study the reconstruction precision of 3D face shapes with different numbers of vertices. As we know, facial components including eyes, nose, mouth, and eye-brows are the most discriminative parts for face recognition, and thus it is required that more accurate facial component shapes can be obtained. Being aware of this, we assess the reconstruction accuracy as fewer non-facial-component vertices are used (i.e., the coverage of the 3D point cloud becomes more focused on facial components) and the number of input 2D landmarks remains unchanged (i.e., 51 landmarks located on nose, eyes, and mouth are used). Two MAEs are computed based on the whole set of 3D vertices and on the subset of facial component vertices, respectively. From the results in Fig. 8, the MAE over the whole set increases (by more than 0.5 mm) as more non-facial-component vertices are required to be reconstructed. This is because the landmarks used do not provide sufficient constraints on non-facial-component vertices. In contrast, the MAE over the facial component vertex subset is not affected by the vertices outside the facial component area. From Eq. (2), we discover that every vertex in the reconstructed 3D face shape is fully determined by the input landmarks, and different vertices are independent of each other in their reconstruction errors.

In addition, we fix the coverage of 3D point

Table 1 Mean absolute errors (MAEs) of the proposed method and four state-of-the-art methods at different poses with ground-truth landmarks

Method	MAE (mm)									Mean (mm)
	−70°	−50°	−30°	−15°	0°	15°	30°	50°	70°	
Romdhani and Vetter (2005)	2.65	2.59	2.58	2.61	2.59	2.50	2.50	2.46	2.51	2.55
Aldrian and Smith (2013)	2.64	2.60	2.58	2.64	2.56	2.49	2.50	2.54	2.63	2.58
Zhu *et al.* (2014)	3.45	2.81	3.71	4.62	4.97	4.81	3.74	2.98	3.19	3.81
Bas *et al.* (2016)	2.35	**2.26**	2.38	2.40	2.51	2.39	2.40	**2.20**	**2.26**	2.35
Ours	**2.29**	2.30	**2.35**	**2.29**	**2.31**	**2.27**	**2.36**	2.21	2.32	**2.30**

Bold numbers denote the best results

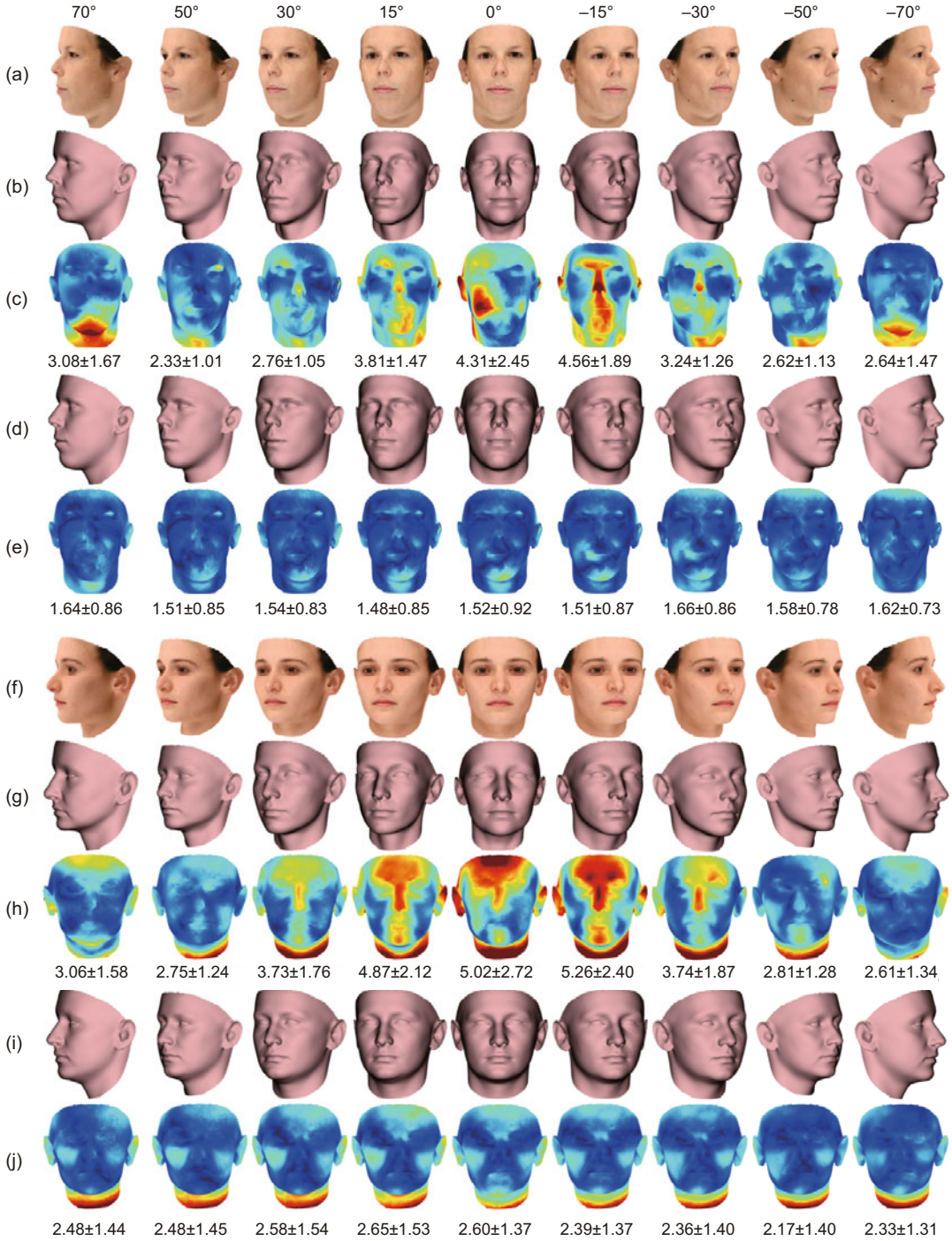

Fig. 6 Reconstruction results for two BFM samples at nine different poses: (a) and (f) are the input images; (b) and (g) are the reconstructed 3D face shapes by the SSF-3DMM method (Zhu *et al.*, 2014); (d) and (i) are those by the proposed method; (c) and (h) are the corresponding MAE error maps of the SSF-3DMM method; (e) and (j) are those by the proposed method. The colormap goes from dark blue to dark red (corresponding to an error from [0, 10]). The numbers under each of the error maps represent mean and standard deviation values (mm). References to color refer to the online version of this figure

cloud to the facial component region, and evaluate the reconstruction accuracy when different numbers of 3D vertices in that region are reconstructed (i.e., the point cloud density changes by, for example, uniform downsampling). Fig. 9 indicates that the overall reconstruction accuracy is reduced slightly

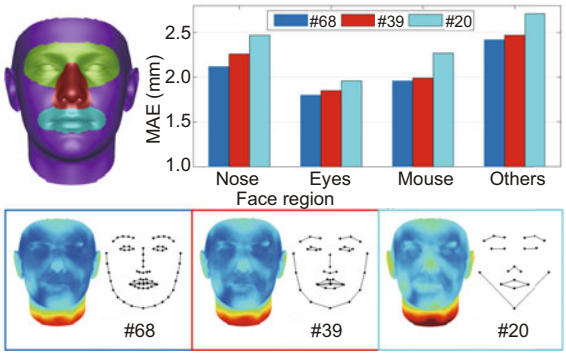

Fig. 7 MAEs of the proposed method in nose, eyes, mouse, and the other regions on the BFM test samples when different 2D landmarks are used. The bottom row shows the vertex-wise MAE maps, in which errors increase from blue to red. References to color refer to the online version of this figure

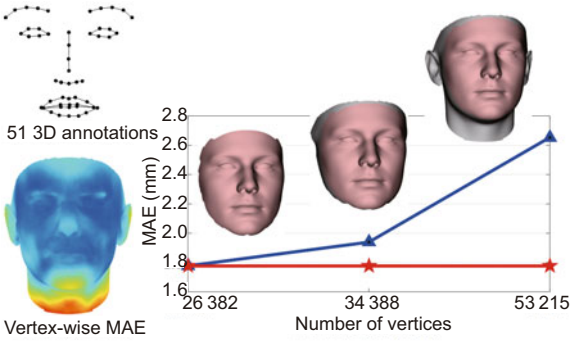

Fig. 8 MAEs of the proposed method over the whole set of 3D vertices (blue curve) and the subset of facial component vertices (red curve) on the BFM test samples as more vertices are included in the reconstructed 3D face shape and the 51 landmarks used remain unchanged. The vertex-wise MAE map shows the MAE per vertex in the 3D face (errors increase from blue to red). References to color refer to the online version of this figure

(by less than 0.001 mm) as the number of reconstructed 3D vertices decreases. This is again mainly because of the independence between different vertices as mentioned before. On the other hand, solving the optimization problem (3) is essentially to make a balance of reconstruction errors both among all training samples and among all the vertices in the reconstructed 3D face shape. Thus, different sets of vertices will theoretically result in different 'balances'. Fortunately, as long as the 2D landmarks can provide sufficient constraints on the reconstructed region of the 3D face, the point cloud density in the reconstructed 3D face region has little effect on the reconstruction accuracy (Fig. 8 shows that additional vertices outside the facial component region do

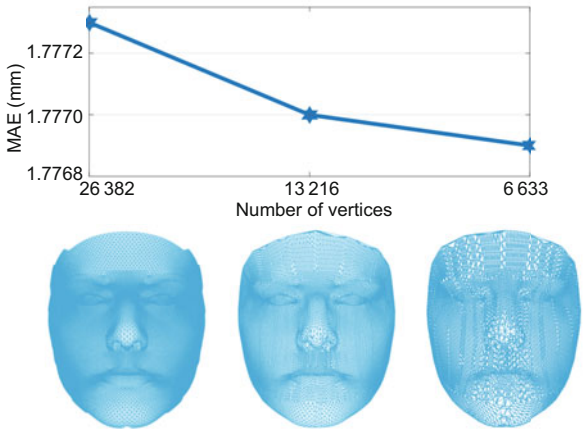

Fig. 9 MAEs of the proposed method over a fixed region of a 3D face when different numbers of vertices are used to represent that region

not change the reconstruction accuracy inside that region when facial component landmarks are used to guide the reconstruction). This is a favorite property of the proposed method, which enables people to reconstruct 3D faces of a higher resolution at the same precision without extra cost except for computational complexity (due to a higher dimensional regression output).

5.6 Using standalone landmark localization methods

In the above evaluation experiments, the 2D visible landmarks are obtained from the ground-truth 3D shapes. In this experiment we use landmarks that are automatically detected by several different methods, including SDM (Xiong and de la Torre, 2013), DLIB (Kazemi and Sullivan, 2014), TCDCN (Zhang *et al.*, 2014), and CFSS (Zhu SZ *et al.*, 2015), as the 'ground-truth' landmarks. Considering the potential errors in automatically detected landmarks, we disturb the ground-truth landmarks of training data by zero-mean Gaussian noise with a standard deviation of 25 to improve the robustness of the obtained regressors. We conduct two series of experiments: (1) training using data with ground-truth landmarks (denoted as Proposed I); (2) training using data with disturbed landmarks (denoted as Proposed II). In this experiment, the approaches of Romdhani and Vetter (2005), E-3DMM (Zhu XY *et al.*, 2015), and Bas *et al.* (2016) are selected as the baselines. We use the authors' own implementations with automatically detected landmarks. In this more challenging scenario, as shown in Table 2, our

Table 2 Mean absolute errors (MAEs) with automatically detected landmarks at different rotation angles on the BFM database

Method	MAE (mm)							Mean (mm)
	−50°	−30°	−15°	0°	15°	30°	50°	
Romdhani and Vetter (2005)	3.42	3.66	3.78	3.77	3.57	4.31	4.19	3.81
Zhu XY *et al.* (2015)	N/A	4.63	5.09	4.19	5.22	4.92	N/A	N/A
Bas *et al.* (2016)	3.20	3.19	3.09	3.30	3.36	3.36	3.84	3.33
Proposed I + SDM	4.60	3.28	3.72	3.69	3.67	3.44	4.51	3.84
Proposed I + DLIB	3.64	3.37	3.17	3.22	3.21	3.44	3.33	3.34
Proposed I + TCDCN	3.69	3.40	3.22	3.48	3.58	3.50	3.54	3.49
Proposed I + CFSS	3.34	3.48	3.27	3.39	3.22	3.41	3.52	3.38
Proposed II + SDM	**3.06**	**2.92**	3.23	3.13	3.34	3.29	3.18	3.16
Proposed II + DLIB	3.13	3.06	3.03	3.04	3.03	**3.05**	**3.02**	**3.05**
Proposed II + TCDCN	3.29	3.15	3.11	3.19	3.20	3.24	3.30	3.21
Proposed II + CFSS	3.17	3.04	**3.00**	**3.01**	**3.01**	3.08	3.26	3.08

N/A: not available. Bold numbers denote the best results

method trained with disturbed landmarks gives the best overall performance and is superior for all pose angles, especially with the DLIB face alignment method. Compared with the results obtained by using the landmarks generated from ground-truth 3D face shapes in Table 1, the accuracy by using automatically detected landmarks is worse (MAE has been increased from 2.30 mm to 3.34 mm), but can be successfully improved via disturbing the detected landmarks during training (3.05 mm).

5.7 Reconstruction accuracy across expressions on BU3DFE

The BU3DFE database (Yin *et al.*, 2006) contains 3D faces of 100 subjects displaying seven expressions of: neutral (NE), happiness (HA), disgust (DI), fear (FE), anger (AN), surprise (SU), and sadness (SA). All non-neutral expressions are acquired at four levels of intensity. We select neutral and the first level intensity of the remaining six expressions as testing sets, resulting in 700 testing samples. The reconstruction error is measured by the normalized per-vertex depth error (NPDE). NPDE is defined by the depth error at each vertex of the test sample as

$$\mathrm{NPDE}(x_j, y_j) = (|z_j^* - \hat{z}_j|)/(z_{\max}^* - z_{\min}^*), \quad (8)$$

where $z_{\max}^*$ and $z_{\min}^*$ are the maximum and minimum depth values in the ground-truth 3D face shape of the test sample respectively, and z_j^* and $\hat{z}_j$ are the ground-truth and reconstructed depth values at the jth vertex respectively. Fig. 10 shows the accuracy of the proposed method as well as

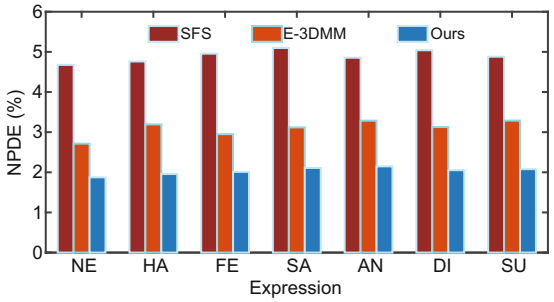

Fig. 10 Average normalized per-vertex depth errors (NPDEs) of the proposed and two counterpart methods for different expressions in the BU3DFE database. References to color refer to the online version of this figure

two counterpart methods for different expressions in the BU3DFE database. It can be seen that the proposed method achieves the lowest error for all the expressions. It successfully reduces the overall average reconstruction error from 4.89% of SFS (Kemelmacher-Shlizerman and Basri, 2011) and 3.10% of E-3DMM (Zhu XY *et al.*, 2015) to 2.03%. Fig. 11 shows the reconstruction results of our method, SFS (Kemelmacher-Shlizerman and Basri, 2011), and E-3DMM (Zhu XY *et al.*, 2015) on one subject under seven expressions.

5.8 Reconstruction accuracy for unconstrained face images on MICC

The MICC database (Bagdanov *et al.*, 2011) contains three challenging face video clips and a ground-truth 3D face model for each of 53 subjects. The videos span the range of controlled indoor to challenging unconstrained outdoor settings.

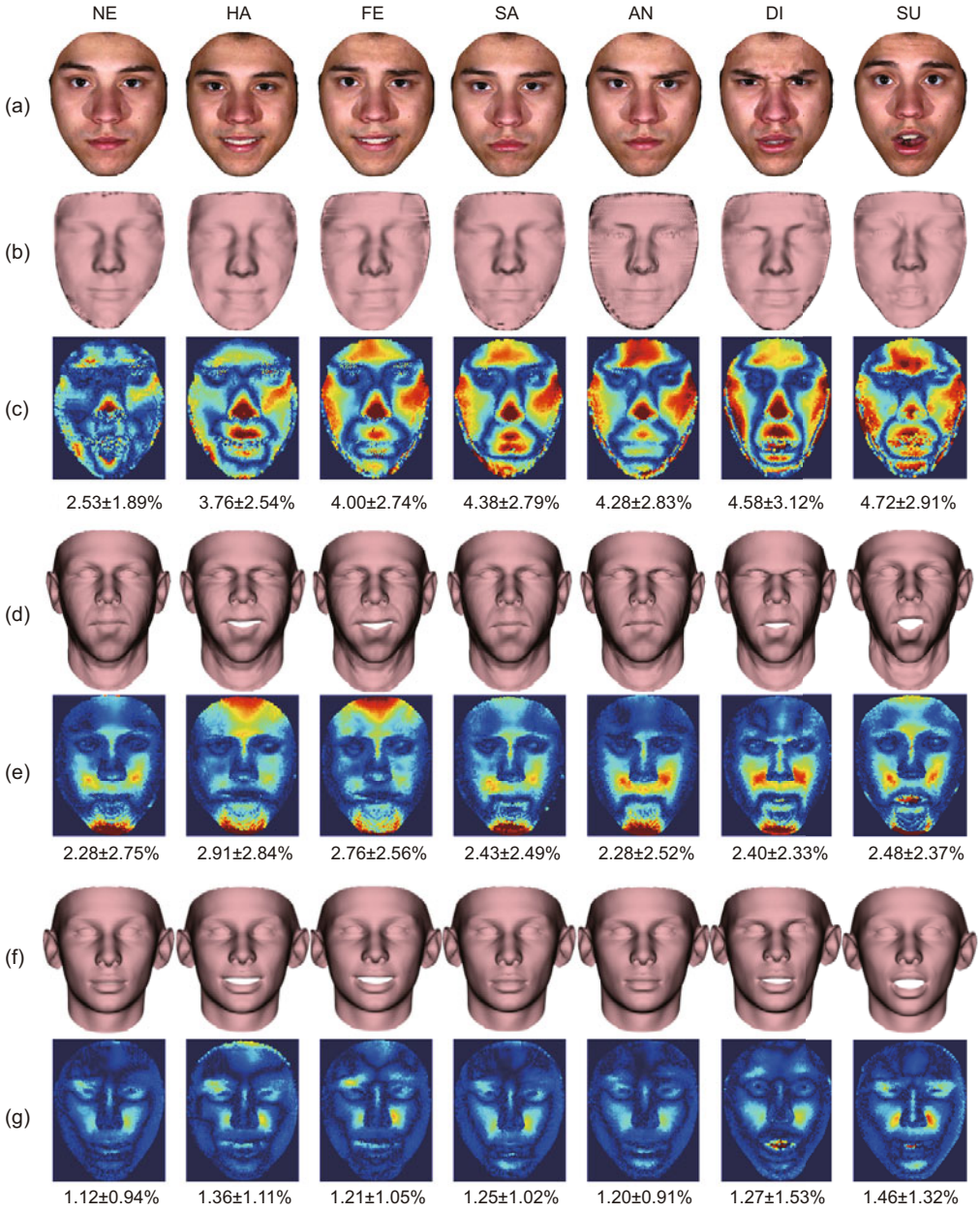

Fig. 11 Reconstruction results for the BU3DFE samples at seven different expressions: (a) the input images; (b), (d), and (f) are the reconstructed 3D face shapes by the method of SFS (Kemelmacher-Shlizerman and Basri, 2011), E-3DMM (Zhu XY *et al.*, 2015), and the proposed method, respectively; (c), (e), and (g) are the corresponding NPDE maps by the method of SFS, E-3DMM, and the proposed method, respectively. The colormap goes from dark blue to dark red (corresponding to an error from [0, 10]). The numbers under each of the error maps represent mean and standard deviation values in percent. References to color refer to the online version of this figure

The recorded outdoor videos are very challenging because of the uncontrolled lighting conditions. In this experiment, we use the outdoor videos as the input and randomly select 4 000 face images from 31 466 frames of the 53 subjects. Again, the four different automated face alignment methods are employed to detect the landmarks on the selected unconstrained face images. The cascaded regressor model is trained using data with disturbed landmarks, as introduced in Section 5.6. Fig. 12 shows the reconstruction results of the proposed method for three samples in MICC. Note that the ground-truth 3D face shapes in

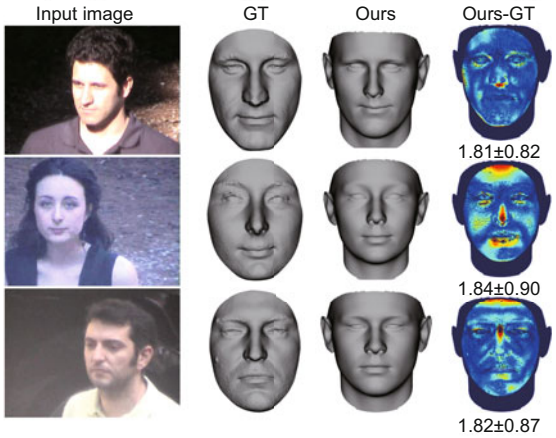

Fig. 12 Reconstruction results of the proposed method on three samples from the MICC database. From the first to the fourth columns: input images, ground-truth 3D shapes, reconstructed 3D shapes by the proposed method, and the corresponding MAE error maps, respectively

MICC have different numbers of vertices from the 3D face shapes reconstructed by the proposed method. Thus, we apply Procrustes alignment to the 3D face shapes and crop face regions around the nose tip with a radius of 95 mm. Based on the cropped face regions, we compute MAEs, and the results are shown in Table 3. Clearly, the proposed method also works for unconstrained face images.

Table 3 Mean absolute errors (MAEs) of the proposed method with landmarks automatically detected by different alignment methods on MICC

Alignment method	SDM	DLIB	TCDCN	CFSS
MAE (mm)	2.56	2.52	2.69	2.64

6 Conclusions

In this paper, we have thoroughly investigated the cascaded regression based 3D face reconstruction approach recently proposed in Liu *et al.* (2016). Our experimental results showed that: (1) more landmarks are generally helpful for accurate 3D face reconstruction, but different facial components have different gains from the increased number of landmarks; (2) the overall 3D face reconstruction accuracy will be degraded if more areas are covered by the reconstructed 3D faces while the used landmarks remain the same; (3) the reconstruction accuracy for a specific face area is not affected by the 3D point cloud density in that area or the 3D vertices outside that area as long as the input landmarks are not changed; (4) using standalone automated facial landmark detection methods together with the cascaded regression based 3D face reconstruction methods is feasible, and the reconstruction accuracy can be improved by disturbing the detected landmarks during training; (5) the cascaded regression based 3D face reconstruction methods have good convergence. In addition, the revised reconstruction method together with its training method provides a feasible alternative approach to 3D face reconstruction, for which the training data can be more easily prepared than in Liu *et al.* (2016) because invisible landmarks' locations are not required to be annotated. In the future, given the impressive accuracy and efficiency of the cascaded regression based 3D face reconstruction approach, we are going to apply it to unconstrained face recognition in real-world scenarios.

References

Abiantun, R., Prabhu, U., Savvides, M., 2014. Sparse feature extraction for pose-tolerant face recognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, **36**(10):2061-2073. https://doi.org/10.1109/TPAMI.2014.2313124

Aldrian, O., Smith, W.A.P., 2013. Inverse rendering of faces with a 3D morphable model. *IEEE Trans. Patt. Anal. Mach. Intell.*, **35**(5):1080-1093. https://doi.org/10.1109/TPAMI.2012.206

Bagdanov, A.D., Del Bimbo, A., Masi, I., 2011. The florence 2D/3D hybrid face dataset. Workshop on Human Gesture and Behavior Understanding, p.79-80. https://doi.org/10.1145/2072572.2072597

Barron, J.T., Malik, J., 2012. Shape, albedo, and illumination from a single image of an unknown object. IEEE Conf. on Computer Vision and Pattern Recognition, p.334-341. https://doi.org/10.1109/CVPR.2012.6247693

Bas, A., Smith, W.A.P., Bolkart, T., *et al.*, 2016. Fitting a 3D morphable model to edges: a comparison between hard and soft correspondences. IEEE Asian Conf. on Computer Vision, p.377-391. https://doi.org/10.1007/978-3-319-54427-4_28

Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3D faces. Proc. SIGGRAPH, p.187-194. https://doi.org/10.1145/311535.311556

Blanz, V., Vetter, T., 2003. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Patt. Anal. Mach. Intell.*, **25**(9):1063-1074. https://doi.org/10.1109/TPAMI.2003.1227983

Booth, J., Roussos, A., Zafeiriou, S., *et al.*, 2016. A 3D morphable model learnt from 10 000 faces. IEEE Conf. on Computer Vision and Pattern Recognition, p.5543-5552. https://doi.org/10.1109/CVPR.2016.598

Cao, C., Hou, Q.M., Zhou, K., 2014a. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, **33**(4):43.1-43.10. https://doi.org/10.1145/2601097.2601204

Cao, C., Weng, Y.L., Zhou, S., *et al.*, 2014b. Facewarehouse: a 3D facial expression database for visual computing.

IEEE Trans. Vis. Comput. Graph., **20**(3):413-425. https://doi.org/10.1109/TVCG.2013.249

Cao, C., Wu, H.Z., Weng, Y.L., *et al.*, 2016. Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph.*, **35**(4):126.1-126.12. https://doi.org/10.1145/2897824.2925873

Chu, B., Romdhani, S., Chen, L.M., 2014. 3D-aided face recognition robust to expression and pose variations. IEEE Conf. on Computer Vision and Pattern Recognition, p.1907-1914. https://doi.org/10.1109/CVPR.2014.245

Han, H., Jain, A.K., 2012. 3D face texture modeling from uncalibrated frontal and profile images. Int. Conf. on Biometrics: Theory, Applications and Systems, p.223-230. https://doi.org/10.1109/BTAS.2012.6374581

Horn, B.K.P., Brooks, M.J., 1989. Shape from Shading. MIT Press, Cambridge, MA, USA.

Hu, J.L., Lu, J.W., Tan, Y.P., 2014. Discriminative deep metric learning for face verification in the wild. IEEE Conf. on Computer Vision and Pattern Recognition, p.1875-1882. https://doi.org/10.1109/CVPR.2014.242

Jourabloo, A., Liu, X.M., 2015. Pose-invariant 3D face alignment. IEEE Int. Conf. on Computer Vision, p.3694-3702. https://doi.org/10.1109/ICCV.2015.421

Jourabloo, A., Liu, X.M., 2016. Large-pose face alignment via CNN-based dense 3D model fitting. IEEE Conf. on Computer Vision and Pattern Recognition, p.4188-4196. https://doi.org/10.1109/CVPR.2016.454

Jourabloo, A., Liu, X.M., 2017. Pose-invariant face alignment via CNN-based dense 3D model fitting. *Int. J. Comput. Vis.*, **4**:1-17. https://doi.org/10.1007/s11263-017-1012-z

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. IEEE Conf. on Computer Vision and Pattern Recognition, p.1867-1874. https://doi.org/10.1109/CVPR.2014.241

Kemelmacher-Shlizerman, I., Basri, R., 2011. 3D face reconstruction from a single image using a single reference face shape. *IEEE Trans. Patt. Anal. Mach. Intell.*, **33**(2):394-405. https://doi.org/10.1109/TPAMI.2010.63

Köstinger, M., Wohlhart, P., Roth, P.M., *et al.*, 2011. Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. IEEE Int. Conf. on Computer Vision Workshops, p.2144-2151. https://doi.org/10.1109/ICCVW.2011.6130513

Li, J., Long, S.Q., Zeng, D., *et al.*, 2015. Example-based 3D face reconstruction from uncalibrated frontal and profile images. IEEE Int. Conf. on Biometrics, p.193-200. https://doi.org/10.1109/ICB.2015.7139051

Li, X., Xu, Y.D., Lv, Q., *et al.*, 2016. Affine-transformation parameters regression for face alignment. *IEEE Signal Process. Lett.*, **23**(1):55-59. https://doi.org/10.1109/LSP.2015.2499778

Liu, F., Zeng, D., Zhao, Q.J., *et al.*, 2016. Joint face alignment and 3D face reconstruction. European Conf. on Computer Vision, p.545-560. https://doi.org/10.1007/978-3-319-46454-1_33

Paysan, P., Knothe, R., Amberg, B., *et al.*, 2009. A 3D face model for pose and illumination invariant face recognition. IEEE Conf. on Advanced Video and Signal-based Surveillance, p.296-301. https://doi.org/10.1109/AVSS.2009.58

Ren, J.F., Jiang, X.D., Yuan, J.S., 2016. Face and facial expressions recognition and analysis. Context Aware Human-Robot and Human-Agent Interaction, p.3-29. https://doi.org/10.1007/978-3-319-19947-4_1

Romdhani, S., Vetter, T., 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. IEEE Conf. on Computer Vision and Pattern Recognition, p.986-993. https://doi.org/10.1109/CVPR.2005.145

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., *et al.*, 2013. 300 faces in-the-wild challenge: the first facial landmark localization challenge. IEEE Int. Conf. on Computer Vision Workshops, p.397-403. https://doi.org/10.1109/ICCVW.2013.59

Sun, Y., Wang, X.G., Tang, X.O., 2013. Deep convolutional network cascade for facial point detection. IEEE Conf. on Computer Vision and Pattern Recognition, p.3476-3483. https://doi.org/10.1109/CVPR.2013.446

Suwajanakorn, S., Kemelmacher-Shlizerman, I., Seitz, S.M., 2014. Total moving face reconstruction. European Conf. on Computer Vision, p.796-812. https://doi.org/10.1007/978-3-319-10593-2_52

Xiong, X.H., de la Torre, F., 2013. Supervised descent method and its applications to face alignment. IEEE Conf. on Computer Vision and Pattern Recognition, p.532-539. https://doi.org/10.1109/CVPR.2013.75

Yin, L.J., Wei, X.Z., Sun, Y., *et al.*, 2006. A 3D facial expression database for facial behavior research. IEEE Int. Conf. on Automatic Face and Gesture Recognition, p.211-216. https://doi.org/10.1109/FGR.2006.6

Zeng, D., Zhao, Q.J., Long, S.Q., *et al.*, 2017. Examplar coherent 3D face reconstruction from forensic mugshot database. *Image Vis. Comput.*, **58**:193-203. https://doi.org/10.1016/j.imavis.2016.03.001

Zhang, Z.P., Luo, P., Chen, C.L., *et al.*, 2014. Facial landmark detection by deep multi-task learning. European Conf. on Computer Vision, p.94-108. https://doi.org/10.1007/978-3-319-10599-4_7

Zhou, X.W., Leonardos, S., Hu, X.Y., *et al.*, 2015. 3D shape estimation from 2D landmarks: a convex relaxation approach. IEEE Conf. on Computer Vision and Pattern Recognition, p.4447-4455. https://doi.org/10.1109/CVPR.2015.7299074

Zhu, S.Z., Li, C., Chen, C.L., *et al.*, 2015. Face alignment by coarse-to-fine shape searching. IEEE Conf. on Computer Vision and Pattern Recognition, p.4998-5006. https://doi.org/10.1109/CVPR.2015.7299134

Zhu, X.X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. IEEE Conf. on Computer Vision and Pattern Recognition, p.2879-2886. https://doi.org/10.1109/CVPR.2012.6248014

Zhu, X.Y., Yi, D., Lei, Z., *et al.*, 2014. Robust 3D morphable model fitting by sparse SIFT flow. IEEE Int. Conf. on Pattern Recognition, p.4044-4049. https://doi.org/10.1109/ICPR.2014.693

Zhu, X.Y., Lei, Z., Yan, J.J., *et al.*, 2015. High-fidelity pose and expression normalization for face recognition in the wild. IEEE Conf. on Computer Vision and Pattern Recognition, p.787-796. https://doi.org/10.1109/CVPR.2015.7298679