Landmark-based 3D Face Reconstruction from an Arbitrary Number of Unconstrained Images

Wan Tian, Feng Liu, Qijun Zhao
College of Computer Science, Sichuan University, Chengdu, China
qjzhao@scu.edu.cn

Abstract—In this paper, we propose a novel method for reconstructing 3D faces from 2D images. The method is characterized in three aspects. (i) It utilizes only geometric cues in the input images, i.e., 2D facial landmarks. (ii) It works for an arbitrary number of unconstrained images, both single and multiple images. (iii) It can effectively exploit complementary information in multiple images of varying poses and expressions. The method is implemented based on cascaded regression in shape space. We have evaluated the method on three databases and observed from the experimental results that (i) the reconstruction error is reduced as more images of different poses are used, (ii) the proposed method can obtain comparable reconstruction results by using state-of-the-art automated methods to detect the 2D landmarks, and (iii) the proposed method is robust to variations in facial expressions and image qualities.

Keywords—3D face reconstruction; unconstrained images; landmark based; cascade regression;

I. INTRODUCTION

As a fundamental problem in both computer vision and computer graphics, 3D face reconstruction from 2D images has recently gained increasing attention because of its benefits to a variety of applications, such as 3D-assisted face recognition [1], [2], facial animations [3], and off-angle face alignment [4], [5], [6]. A large literature of prior work have been devoted to reconstructing 3D faces from either single or multiple 2D images. One class of techniques that can handle very general face images are single-view methods. In order to solve the single-view 3D face reconstruction problem, different priors or constraints have been introduced, including 3D Morphable Model (3DMM) and Shape from Shading (SFS). 3DMM-based methods [7] are relatively robust to varying illuminations and poses, but their reconstructed 3D faces lack personalized details that vary from one individual to another. Kemelmacher and Basri [8] designed an SFS-based single-view reconstruction approach that requires a generic 3D face shape as a reference. This approach does not fully exploit the prior knowledge of 3D faces and heavily depends on the reference models.

Numerous methods have been proposed for reconstructing 3D faces from multiple images, for example, Structure from Motion (SFM) [9], [10], stereo imaging [11] and photometric stereo [12]. These techniques, however, are not applicable to arbitrary set of unconstrained images due to lack of image correspondence, camera calibration information or controlled lighting environment. The task becomes more difficult given face images in the wild. Recently, Kemelmacher et al. [13] and Roth et al. [14], [15] developed an impressive photometric stereo-based method to produce high-quality face models from unconstrained photo collections. These methods mostly use near-frontal face images, and are thus limited in exploiting the complementary information in multi-view images. The consensus is that non-frontal, especially profile, images contain much richer geometric information, and are thus served as complements to 3D face reconstruction.

As suggested in [16], 2D geometric information (i.e., landmarks and contours), sometimes, is more ‘solid’ than photometric stereo-based clues for face reconstruction in the wild. While photometric information (including lighting, camera properties and reflectance properties) estimation is itself a ‘fragile’ process especially for unconstrained images, facial landmark detection on unconstrained images is now a mature research field [6], [17], [18]. Although facial landmarks are often utilized as constraints or for initialization for single-view (3DMM) and multi-view reconstruction (SFM) methods, the performance of these methods is limited by...
the prior models [13]. It is thus demanded to develop efficient and robust algorithms for 3D face reconstruction from unconstrained face images so that all the face images with different poses and expressions can be better used. To this end, this paper proposes a novel landmark-based 3D face reconstruction method that can cope with a set of \( p (p \geq 1) \) unconstrained images.

Motivated by state-of-the-art single-view reconstruction methods [19], [5], [20], [21], our method reconstructs 3D faces of frontal pose and neutral expression from multiple unconstrained images of a subject via cascaded regression in 2D/3D shape space. It begins with extracting 2D facial landmarks on the images. As shown in Fig. 1, for each face image, we extract a set of 68 facial landmarks using the method in [17]. With the guidance of these landmarks, it then progressively updates the estimated 3D face shape for the input subject, during which the invisible landmarks due to self-occlusion are treated as missing data and substituted by zeros. The updates are computed via a set of cascaded regressors, which are off-line learned based on a training set of pairing 3D face shapes and unconstrained face images. During the off-line training of the regressors, the optimization objective is to minimize the holistic 3D face shape errors. Meanwhile, the set of used landmarks is consistent and intrinsically constrained for all different poses and expressions. These ensure a higher reconstruction accuracy and avoid complicated online optimization.

By effectively exploring the correlation between 2D landmarks and 3D shapes, the proposed method learns shape space regression from all the input images such that every image (even profile images and images of extreme expressions) contributes to the reconstruction. To demonstrate the capabilities of the proposed approach, quantitative and qualitative experiments are performed on synthetic and in-the-wild 2D images with comparison to state-of-the-art methods. Our main contributions are summarized below.

- We propose a novel landmark-based 3D face reconstruction method that can cope with varying number of unconstrained images.
- Our method employs cascade regression to learn holistic 3D face shapes based on the consistency of all the 2D landmarks on different images. As a benefit, our approach makes better use of face images of different poses and expressions, including the profiles and extreme expressions, for more accurate reconstruction.

- We study the impact of diverse variations of images (e.g., the number of images, expressions, poses) on the reconstruction performance.

II. RELATED WORK

In this section, we review existing work from two aspects: reconstruction from texture cues or from geometric cues.

3D face reconstruction from 2D texture information

Conventional 3DMM-based methods [7], [1] establish statistical parametric models for both texture and shape, and represent a 3D face as a linear combination of basis shapes and textures. 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. However, these methods are limited in individualized or detail reconstruction because PCA conducts global modeling in essence. Although SFS-based methods [9] can yield good-looking results using the shading cues, the reconstructed facial geometry highly depends on the employed reference 3D face model. Recently, photometric stereo-based methods [13], [14], [15] have proven effective for unconstrained photo collections, but they require a sufficiently large collection of photos for reconstruction.

3D face reconstruction from 2D geometric information

A vast amount of works have used facial landmarks as a cue for 3D face reconstruction. Single-view landmark-based 3DMM methods [22], [23], [18] estimate the model parameters based on the correspondence between 2D and 3D landmarks which is, however, suffers from the problem of semantic consistence [23], [20]. SFM can be applied when facial landmarks are tracked across the face images such that correspondence is established between the images.
III. PROPOSED METHOD

A. Overview

Given an arbitrary number of unconstrained face images \( \{I_p\}_{p=1}^{P}, 1 \leq p \leq N \) of a subject \( (N \) is the maximum scale of input image sets), our goal is to reconstruct the person-specific 3D face shape of the subject with frontal pose and neutral expression. We denote a 3D face shape of neutral expression and frontal pose as \( S \in \mathbb{R}^{3 \times q} \), which is represented by 3D coordinates of its \( q \) vertices, and denote a subset of \( S \) with columns corresponding to \( l \) annotated landmarks (following the Multi-PIE [24], \( l = 68 \) in this paper) as \( S_L \). The projections of these 3D landmarks on each image \( I_i \) are represented by \( U_i \in \mathbb{R}^{2 \times l} \). The relationship between 2D facial landmarks \( U_i \) and its corresponding 3D landmarks \( S_L \) can be described as:

\[
U_i \approx f_i P_i R_i (S_L + t_i),
\]

(1)

where \( f_i \) is the scale factor, \( P_i \) is the orthographic projection matrix, \( R_i \) is the \( 3 \times 3 \) rotation matrix and \( t_i \) is the translation vector. Here, we employ weak perspective projection \( M_i \) to approximate the 3D-to-2D mapping as conventionally done in the literature [25], [23]. To fully utilize the correlation between the landmarks on all the \( p \) input images, we concatenate them to form a unified 2D facial landmark vector \( U = (U_1, U_2, \ldots, U_p, U_{p+1}, \ldots, U_N) \), where \( U_i \) are zero vectors for \( p + 1 \leq i \leq N \).

We reconstruct \( S \) from the given “ground truth” visible landmarks \( U^* \) (either manually marked or automatically detected by a stand alone method) for the unconstrained image set \( \{I_p\}_{p=1}^{P} \). For the landmarks that are invisible in an image, their coordinates are set to be a constant value (e.g., 0 in our experiments), which means missing data. As discussed above, we achieve this by iteratively updating the initial estimate of \( S \) 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. Figure 2 shows the flowchart of the proposed method.

B. The Reconstruction Process

Let \( U^* \) be the “ground truth” landmarks on the photos, and \( S^{k-1} \) the currently reconstructed 3D shape after \( k-1 \) iterations. The corresponding landmarks \( U^{k-1} \) can be obtained by projecting \( S^{k-1} \) onto the images according to Eqn. (1). The updated 3D shape \( S^k \) can be computed by

\[
S^k = S^{k-1} + W^k(U^* - U^{k-1}),
\]

(2)

where \( W^k \) is the regressor in \( k^{th} \) iteration.

C. Learning Cascaded Regressors

The \( K \) regressors \( \{W^k\}_{k=1}^{K} \) in the reconstruction process can be learned via optimizing the following objective function over the \( m \) training samples (each sample contains up to \( N \) annotated 2D images and a ground truth 3D face shape):

\[
\arg \min_{W^k} \sum_{j=1}^{m} \| (S_j^* - S_j^{k-1}) - W^k(U_j^* - U_j^{k-1}) \|_2^2,
\]

(3)

where \( \{S_j^*, U_j^*\} \) is one training sample consisting of ground truth landmarks \( U_j^* \) on the images of a subject and the subject’s ground truth 3D face shape \( S_j^* \). Mathematically, the above optimization seeks for 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 paper, we use linear regressors \( W^k \in \mathbb{R}^{3q \times (2l \times N)} \). The optimization in Eqn. (3) can be solved by using least squares methods with a solution of

\[
W^k = \Delta S^k (\Delta U^k)^T (\Delta U^k (\Delta U^k)^T)^{-1},
\]

(4)

where \( \Delta S^k = S^* - S^{k-1} \) and \( \Delta U^k = U^* - U^{k-1} \) are 3D shape adjustment and 2D landmark deviation. \( S \in \mathbb{R}^{3q \times m} \) and \( U \in \mathbb{R}^{(2l \times N) \times m} \) denote, respectively, the ensemble of 3D face shapes and 2D landmarks of all training samples with each column corresponding to one sample. Note that, here, we also write the 3D face shape \( S \) and the landmarks \( U \) as column vectors. It can be mathematically shown that, to ensure a valid solution in Eqn. (4), \( m \) should be larger than \( 2l \times N \) so that \( \Delta U^k (\Delta U^k)^T \) is invertible. Fortunately, since the set of used landmarks are usually sparse, this requirement can be easily satisfied in real-world applications.

D. Estimating 3D-to-2D Mapping

As mentioned in Sec. III-B, in order to obtain the landmarks rendered from the updated 3D face shape, we have to estimate 3D-to-2D mapping matrix of each face image in each iteration. In this paper, we dynamically estimate the mapping matrix based on \( S^k \) and \( U^k \) as suggested in the work of [5]. The mapping matrix is a composite effect of expression and pose induced deformation and camera projection. The mapping matrix \( M_i^k \) for each image \( I_i \) is represented by a \( 2 \times 4 \) matrix, and can be estimated as a least squares solution to the following fitting problem:

\[
M_i^k = \arg \min_{M_i} \| U_i^k - M_i^k S_L \|_2^2.
\]

(5)

Here, \( S_L^k \) is represented by homogeneous coordinates.

IV. IMPLEMENTATION DETAILS

A. Training Data

To train the proposed method, a set of 3D face shapes and corresponding 2D face images with annotated landmarks are needed. To make the obtained cascaded regressors robust to pose and expression variations, the 2D images should be
Table I

<table>
<thead>
<tr>
<th>Number of input images</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSSE</td>
<td>2.43</td>
<td>2.37</td>
<td>2.24</td>
<td>2.21</td>
<td>2.18</td>
<td>2.16</td>
<td>2.15</td>
<td>2.13</td>
<td>2.10</td>
</tr>
</tbody>
</table>

Table II

<table>
<thead>
<tr>
<th>Method</th>
<th>−70°</th>
<th>−50°</th>
<th>−30°</th>
<th>−15°</th>
<th>0°</th>
<th>15°</th>
<th>30°</th>
<th>50°</th>
<th>70°</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aldrian and Smith [22]</td>
<td>2.64</td>
<td>2.60</td>
<td>2.58</td>
<td>2.64</td>
<td>2.56</td>
<td>2.49</td>
<td>2.50</td>
<td>2.54</td>
<td>2.63</td>
<td>2.58</td>
</tr>
<tr>
<td>Rondhui et al. [26]</td>
<td>2.65</td>
<td>2.59</td>
<td>2.58</td>
<td>2.61</td>
<td>2.59</td>
<td>2.50</td>
<td>2.50</td>
<td>2.46</td>
<td>2.51</td>
<td>2.55</td>
</tr>
<tr>
<td>SSF-3DMM [27]</td>
<td>3.45</td>
<td>2.81</td>
<td>3.71</td>
<td>4.62</td>
<td>4.97</td>
<td>4.81</td>
<td>3.74</td>
<td>2.98</td>
<td>3.19</td>
<td>3.81</td>
</tr>
<tr>
<td>Bas et al. [28]</td>
<td>2.35</td>
<td>2.24</td>
<td>2.38</td>
<td>2.40</td>
<td>2.51</td>
<td>2.39</td>
<td>2.40</td>
<td>2.20</td>
<td>2.26</td>
<td>2.35</td>
</tr>
<tr>
<td>Liu et al. [29]</td>
<td>2.29</td>
<td>2.30</td>
<td>2.35</td>
<td>2.29</td>
<td>2.31</td>
<td>2.27</td>
<td>2.36</td>
<td>2.21</td>
<td>2.32</td>
<td>2.30</td>
</tr>
<tr>
<td>Proposed</td>
<td>2.92</td>
<td>2.30</td>
<td>2.29</td>
<td>2.18</td>
<td>2.12</td>
<td>2.16</td>
<td>2.15</td>
<td>2.27</td>
<td>2.31</td>
<td>2.23</td>
</tr>
</tbody>
</table>

We first choose MICC [30] as a part of our training data, because it includes several video sequences of varying resolutions and conditions as well as high resolution 3D face scans for each subject. The 2D image sets can be collected from the video frames. We adopt a state-of-the-art 2D face alignment method, FAN [17], to detect the landmarks on the images. However, MICC only contains 53 3D shapes. Hence, additionally, we use the Basel Face Model (BFM) [31] to construct synthetic 3D faces of 200 subjects, and use the expression model from FaceWarehouse [32] to generate random expression on each 3D face. Then these expressive 3D face shapes are projected onto 2D planes with 475 views including 19 yaw (−90° to 90° with a 10° interval), 5 pitch and 5 roll (−20° to 20° with a 10° interval) rotations. The 2D image resolution is 640 × 480 pixels. The 68 landmarks on each 2D face image are recorded during the projection procedure (note that the 3D faces are densely aligned and the indices of the landmarks in the 3D face shapes are known). We randomly choose different combinations of images for each person, resulting in sets of 1 to 20 images (i.e., N = 20 in this paper).

B. Visibility Computation

In order to determine the visibility of 2D landmarks projected from 3D face shape, given the 2D landmark U on the image and the 3D annotated landmarks $S_k^u$ from the initial 3D shape $S_k^o$, we coarsely estimate the camera projection matrix M by Eqn. (5). Suppose the 3D surface normal at a landmark in $S_k^o$ is $\hat{N}$. The visibility $v$ can be then calculated by [33]

$$v = \frac{1}{2} \left( 1 + \text{sign}(\hat{N} \cdot \frac{M_1}{\|M_1\|} \times \frac{M_2}{\|M_2\|}) \right),$$

where $\text{sign}(\cdot)$ is the sign function, ‘·’ means dot product and ‘×’ cross product, and $M_1$ and $M_2$ are the left-most three elements at the first and second row of the mapping matrix $M$. We treat the invisible landmarks on the 2D face images as missing data, and substitute their corresponding entries in $U$ with zero. This way images of arbitrary pose angles can be handled in a unified framework.

V. EXPERIMENTAL RESULTS

The proposed method has been evaluated on three databases, BFM [31], BU3DFE [34] and the Stirling/ESRC Database 1. We report the experimental results below.

A. Results on BFM Database

The BFM database [31] provides 10 test subjects, each of whom has nine face images of neutral expression and different poses (0°, +15°, +30°, +50°, +70°). With this database, we evaluate the proposed method from three aspects: (i) How does the reconstruction accuracy change as more images of different views are used? (ii) How does the proposed approach perform on diverse face angles of single images? (iii) How does the proposed method work when the facial landmarks are automatically detected?

Reconstruction Accuracy with Different Numbers of Images. In this evaluation, we consider $p \in \{1, 2, \cdots, 9\}$ image(s) as input. We randomly choose $p$ image(s) from the nine images of each subject, and compute the average reconstruction error for each $p$ in terms of Root Mean Square Error (RMSE) [35], which is defined as

$$\text{RMSE} = \frac{1}{N_T} \sum_{i=1}^{N_T} (\|S_k^o - \hat{S}_i\|/n),$$

where $N_T$ is the total number of testing samples, $S_k^o$ and $\hat{S}_i$ are the ground truth and reconstructed 3D face shape of the $i$th testing sample, $n$ is the vertices number of $S_k^o$. To compute the RMSE, the reconstructed 3D faces are first aligned to ground truth via Procrustes global alignment based on

1http://pics.stir.ac.uk/ESRC/index.htm
Table III
Reconstruction errors (RMSE) of the proposed method on the BFM database with nine images of each subject as input when the facial landmarks on the images are automatically detected by using three standalone methods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>3.06</td>
<td>2.94</td>
<td>2.48</td>
</tr>
</tbody>
</table>

68 3D landmarks as being suggested by [28], and then cropped at a radius of 95mm around the nose tip. Table I gives the results. As can be seen, the reconstruction error is reduced as more input images of different views are used. This shows that our proposed method effectively utilizes the complementary information in multi-view images for 3D face reconstruction.

Reconstruction Accuracy across Poses. In this experiment, we use single images as the input to compare with several state-of-the-art methods based on 3DMM, including the approach proposed by Aldrian and Smith [22], the multi-features 3DMM framework proposed by Romdhani et al. [26], sparse SIFT Flow 3DMM (SSF-3DMM) [27], the edge-fitting based 3DMM method by Bas et al. [28], and the cascaded regression method by Liu et al. [29].

Table II shows the RMSE of various methods on the BFM database with respect to different poses of face images. As can be seen, the average RMSE of the proposed method is obviously lower than that of the counterpart ones. This proves the effectiveness of the proposed method in tackling face images of unconstrained poses.

Reconstruction Accuracy with Automatically Detected Landmarks. The above experiments use ground truth landmarks. In this experiment, we use stand-alone methods to automatically detect facial landmarks, and evaluate the accuracy of our proposed method with these automatically detected landmarks. Specifically, three state-of-the-art landmark localization methods, i.e., DLIB [36], FAN [17], and SDM [37], are used. Table III shows the result when nine images of each subject are used as input. Obviously, the error increases compared with the result in Table I where ground truth landmarks are used. Yet, as automated landmark detectors improve, the proposed method benefits from the more accurately detected landmarks, and its reconstruction error decreases accordingly.

B. Results on BU3DFE Database

BU3DFE [34] contains 3D face scans of 56 males and 44 females, with a neutral and six basic expressions (happiness, disgust, fear, anger, surprise and sadness). All basic expressions are attained at four levels of intensity. Higher levels mean more intensive expressions.

Reconstruction Accuracy Under Different Facial Expressions. For each of the subjects in BU3DFE, we form four image sets, each of which contains one level of the six basic expression images of frontal pose. Hence, in this experiment, the number of input images is 6. The result is presented in Table IV, from which we can clearly see the robustness of the proposed method to expression variations.

<table>
<thead>
<tr>
<th>Level of expressions</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>2.38</td>
<td>2.35</td>
<td>2.33</td>
<td>2.34</td>
</tr>
</tbody>
</table>

C. Results on Stirling/ESRC Database

In this experiment, a test set from the Stirling/ESRC database\(^3\) has 2,000 2D images, including 656 high-quality and 1,344 low-quality images, of 134 subjects. We evaluate the error of reconstructed face shape from every single image, and from all the images of one subject as well. We chose a state-of-the-art face alignment method in [17] to automatically detect the facial landmarks on the input 2D face images. For some low-quality photos, the method failed, and we manually annotate the 68 landmarks on them. Figure 3 shows the reconstructed 3D face shapes.

\(^3\)https://www.face2vm.org/ig2018/
of four subjects. To further demonstrate the effectiveness of our proposed method, Figure 4 shows two additional reconstruction results on the LFW database [38].

The average error of our method is $2.65 \pm 0.67$ for high-quality images and $2.87 \pm 0.81$ for low-quality images. When all the images of a subject are used for reconstruction, on average, our proposed method achieves a RMSE of 2.26 with a standard deviation of 0.72. These results show that our method is relatively robust to variations in poses, illuminations, expressions, and even image quality (assuming that the facial landmarks are successfully detected). The decreased reconstruction errors when more images are used prove that our proposed method effectively utilizes the complementary information in multiple images for 3D face reconstruction. The quantitative reconstruction errors for different subjects are plotted in Figure 5.

VI. CONCLUSION

We have presented a method that can reconstruct 3D faces from a set of $p$ ($p \geq 1$) unconstrained images. Given the ground truth 2D landmarks on the images, the method starts from an initial 3D face shape, and iteratively computes the required adjustment to the 3D face shape via regression over the deviation between the ground truth landmarks and the landmarks obtained by projecting the reconstructed 3D face shape onto 2D plane. To utilize all the input images, we concatenate the landmark deviations on all of them as the input of the regressors. Evaluation results of our proposed method on three databases demonstrate its effectiveness in exploiting complementary information in multiple images for accurate 3D face reconstruction, its flexibility in reconstructing 3D faces from an arbitrary number of images, either single or multiple images, and its robustness to variations in poses, expressions and image qualities.

ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Program of China (Nos. 2017YFB0802300 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).

REFERENCES

