Deep Face Feature for Face Alignment and Reconstruction

Boyi Jiang, Juyong Zhang, Bailin Deng, Yudong Guo and Ligang Liu

arXiv:1708.02721v1 [cs.CV] 9 Aug 2017

***Abstract*—In this paper, we propose a novel face feature extraction method based on deep learning. Using synthesized multi-view face images, we train a deep face feature (DFF) extractor based on the correlation between projections of a face point on images from different views. A feature vector can be extracted for each pixel of the face image based on the trained DFF model, and it is more effective than general purpose feature descriptors for face-related tasks such as alignment, matching, and reconstruction. Based on the DFF, we develop an effective face alignment method with single or multiple face images as input, which iteratively updates landmarks, pose and 3D shape. Experiments demonstrate that our method can achieve state-of-the-art results for face alignment with a single image, and the alignment can be further improved with multi-view face images.**

***Index Terms*—Feature Learning, Face Alignment, MVS**

I. Introduction

Face alignment from images have been an active research topic since 1990s. Improving the accuracy of face alignment accuracy is beneficial for many computer vision tasks related to facial analysis, because it is used as a prerequisite for these tasks, e.g., 3D face reconstruction [17] and face recognition [28]. Although AAM-based approaches [27] and regression-based approaches [6], [21], [32], [37] work well for face images with small poses, they usually can not handle profile face images as they do not consider the visibility of landmarks.

In recent years, several methods introduced the 3D Morphable Model (3DMM) [2] for face alignment and achieved better results [15], [38], [17]. This is because we can easily compute the visibility and position of 2D landmarks with the help of 3D face shape. These methods used a 3D face model and can handle challenging cases with large pose variation. However, the reconstruction accuracy of such methods is often insufficient. Existing approaches usually use only one face image to reconstruct the 3D face shape, which is a highly ill-posed problem that can lead to unsatisfactory results. Besides, 3DMM shape and expression parameters do not regress well from single face image texture information because of their highly non-linear relationship. The reconstruction error can further introduce bias in the visibility and location of the landmarks, which finally influences the face alignment accuracy.

Boyi Jiang, Juyong Zhang(Corresponding author), Yudong Guo, Ligang Liu are with School of Mathematical Sciences, University of Science and Technology of China. E-mail: jby1993@mail.ustc.edu.cn, juyong@ustc.edu.cn, gyd2011@mail.ustc.edu.cn, lgliu@ustc.edu.cn.

Bailin Deng is with School of Computer Science and Informatics, Cardiff University. E-mail: DengB3@cardiff.ac.uk.

Inspired by multi-view stereo (MVS) reconstruction [31], [30], we would like to utilize multi-view face images to improve 3D face reconstruction and face alignment. However, if we directly apply MVS to reconstruct face shapes with SIFT [18], the reconstructed point cloud may contain holes due to poor correspondence of the local features computed from SIFT. To solve this problem, we specifically design a feature for multi-view face images, which can accurately represent the same anatomical 3D face points across face images with large posture variation. With such a feature descriptor, we can perform co-alignment and reconstruction with an arbitrary number of multi-view face images, while maintaining robustness even with only one face image. We train our feature extractor and cascaded regressor using a large number of multi-view face images with registered ground truth 3DMM faces. However, it is not easy to obtain real face images with ground truth 3D shapes, especially profile view face images. To solve this problem, we synthesize large-scale multi-view face images for deep face feature learning and cascaded regression learning. In summary, the main contributions of this work include:

- For multi-view face images, we propose a CNN-based deep face image feature extractor, which outperforms general feature descriptors such as SIFT.
- Based on our new feature extractor, we propose a simple yet effective cascaded regression-based approach for joint face alignment and reconstruction for an arbitrary number of multi-view face images.

II. Related Works

Classical Face Alignment. Classical face alignment methods, including Active Shape Model (ASM) [26], [9] and Active Appearance Model (AAM) [7], [24], [27], simulate the image generation process and perform face alignment by minimizing the difference between the model appearance and the input image. These methods can achieve accurate reconstruction results, but require a large number of face models with detailed and accurate point-wise correspondence, as well as expensive parameter fitting. Constrained Local Model (CLM) [1], [8] employs discriminative local texture models to regularize the landmark locations. The CLM algorithm is more robust than the AAM search method, which relies on the image reconstruction error to update the model parameters. Recently, regression based methods [4], [6] have been proposed to directly estimate landmark locations from the discriminative features around landmarks. Most regression based algorithms do not consider the visibility of facial landmarks under different view angles. As a result, their performance can degrade substantially for profile view images.

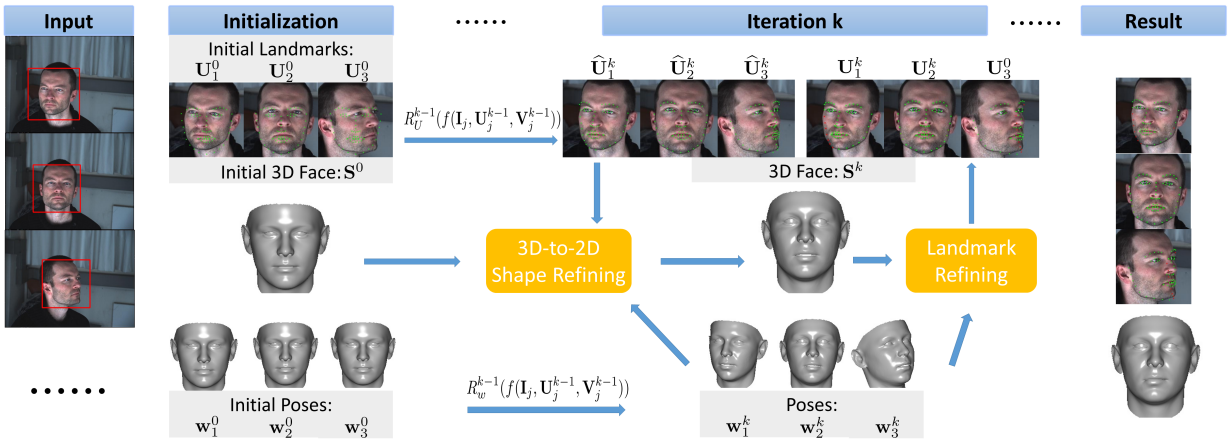

Figure 1: An overview of our algorithm pipeline. The inputs are multi-view face images. The DFF features are first extracted. During each regression process, the landmarks $\mathbf{U}_j^k$ and pose $\mathbf{w}_j^k$ are updated, and the 3D shape is refined with updated landmarks and pose. The landmark and its visibility would be revised according to the updated 3D shape. The iterative process will terminate until convergence.

Face Alignment via Deep Learning. Various deep learning approaches have been applied to face alignment and achieved remarkable results. The methods of [25], [34], [36] use multi-stage CNN to regress sparse face image landmark locations. To boost the performance of face alignment, Zhang *et al.* [35] combine face detection, face alignment, and other tasks into the training of CNN. Alignment of faces with large pose variation is a very challenging problem, because each face might have a different set of visible landmarks. Early works for large-pose face alignment rely on MVS based methods [33], [39]. These methods use different landmark templates for different views, which can cause high computation costs. Recently, 3D model based techniques [12], [14], [15], [38] have been proposed to address the problem of alignment accuracy for challenging inputs, e.g. those with non-frontal face poses, low image quality, and occlusion, etc. These techniques utilize a 3D morphable model to handle self-occluded landmarks and large-pose landmark detection. Jourabloo *et al.* [15] integrate 2D landmark estimation into the 3D face model fitting process, use cascaded CNN to replace simple regressor, and are able to detect 34 landmarks under all view angles. Zhu *et al.* [38] reduce the CNN regression complexity by estimating a Projected Normalized Coordinate Code map, which can detect 68 anatomical landmarks with visibility judgement.

3D Face Reconstruction. 3DMM establishes statistical linear parametric models for both texture and shapes of human faces, where a 3D face is represented using coefficients for its shape and texture basis. To recover the face from a 2D image, 3DMM-based methods [2], [22], [3] estimate the shape and texture coefficients by maximizing the similarity between the input 2D face image and the projected 3D face. However, such methods are not robust enough to handle facial landmarks under large pose variation. Multi-view stereo (MVS) [31], [30] is a classical reconstruction method that requires dense correspondence between neighboring images to achieve satisfactory results. When such methods are applied on multi-view face images, the reconstructed face point cloud might contain holes due to insufficient detected matched points.

III. Our Method

In the following, we propose a face alignment method for an arbitrary number of multi-view face images. Suppose we have N different faces, each with several images from different view angles. We represent each 3D face shape using a triangle mesh with n vertices with the same connectivity. Then its shape can be determined from the vertex positions, which can be written as a $3 \times n$ matrix

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

where each column vector corresponds to the 3D coordinates of a vertex. We denote the i-th face shape as $\mathbf{S}_i$ and its j-th view image as $\mathbf{I}_{ij}$. In addition, we assume the mapping from a 3D face shape $\mathbf{S}$ to a 2D image $\mathbf{I}$ is a weak perspective projection, represented using a camera parameter vector $\mathbf{w} = (s, \alpha, \beta, \gamma, t_x, t_y)$. Here, (t_x, t_y) represent the translation on the image plane, s the scaling factor, α the pitch angle, β the yaw angle, and γ the roll angle. The three angles can determine a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$. Then the 2D projection $P(\mathbf{S})$ of $\mathbf{S}$ can be written as:

$$P(\mathbf{S}) = s\ \mathbf{P}_{\mathrm{r}}\ \mathbf{R}\ \mathbf{S} + \mathbf{t}_{2\mathrm{d}}, \tag{2}$$

where $\mathbf{P}_{\mathrm{r}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and $\mathbf{t}_{2\mathrm{d}} = \begin{pmatrix} t_x \\ t_y \end{pmatrix}$.

Our feature extractor and cascaded regressor are trained using a large set of data, consisting of ground truth face shapes and their multi-view images with corresponding camera parameters. Details on the construction of such training data are provided in Sec. III-B.

The key idea of our method is to utilize the correspondence between multi-view face images to fit a unique dense 3DMM

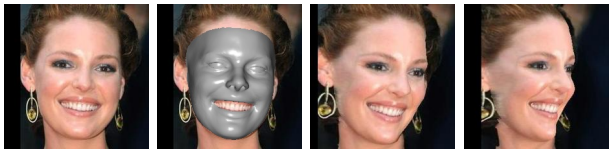

Figure 2: Examples of our construct data. From left to right: a frontal face image; fitted 3DMM face model; two constructed profile face images.

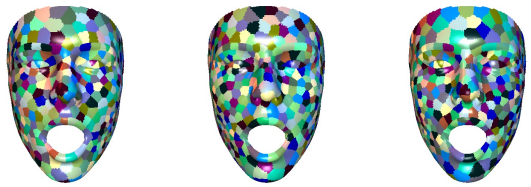

Figure 3: Examples of random segment face mesh into 500 different patches.Each segmentation represents a classification problem.

shape $\mathbf{S}$ and further refine the face landmarks for each view. We use a convolutional neural network (CNN) to train a deep face feature (DFF) extractor which can establish the dense correspondence more accurately than general feature detectors such as SIFT. The training of the DFF extractor will be discussed in Sec. III-C.

In the landmark regression framework, the 2D landmark coordinates and the camera parameters for each image are iteratively estimated through two set of cascade of linear regressors. With the shared 3D shape $\mathbf{S}$, we can further refine the locations of 2D landmarks for each view, and utilize person-specific 3D surface normals to infer visibility of landmarks easily. Fig. 1 shows the overall process of the proposed method.

A. 3D Morphable Model

We use the 3D Morphable Model (3DMM) [2] to represent the 3D face shape $\mathbf{S}$ with 53K vertices:

$$\begin{aligned} \widehat{\mathbf{S}} &= \mathbf{S}_0 + \mathbf{S}_{\text{id}}\mathbf{p}_{\text{id}} + \mathbf{S}_{\text{exp}}\mathbf{p}_{\text{exp}}, \\ \widehat{\mathbf{T}} &= \mathbf{T}_0 + \mathbf{T}_{\text{al}}\mathbf{p}_{\text{al}}, \end{aligned} \tag{3}$$

where $\widehat{\mathbf{S}} \in \mathbb{R}^{3n}$ stacks the vertex coordinates of $\mathbf{S}$, and $\widehat{\mathbf{T}} \in \mathbb{R}^{3n}$ represent the albedo values for the vertices. Here $\mathbf{S}_{\text{id}} \in \mathbf{R}^{3n\times 80}$, $\mathbf{S}_{\text{exp}} \in \mathbf{R}^{3n\times 79}$, and $\mathbf{T}_{\text{al}} \in \mathbf{R}^{3n\times 80}$ denote the basis of identity, expression, and texture respectively, while $\mathbf{p}_{\text{id}}$, $\mathbf{p}_{\text{exp}}$, $\mathbf{p}_{\text{al}}$ denote their linear combination coefficients. $\mathbf{S}_0 = \overline{\mathbf{S}}_{\text{id}} + \overline{\mathbf{S}}_{\text{exp}} \in \mathbf{R}^{3n}$ is the mean face shape, with $\overline{\mathbf{S}}_{\text{id}}$ and $\overline{\mathbf{S}}_{\text{exp}}$ being the mean identity and expression, respectively. $\mathbf{T}_0 \in \mathbf{R}^{3n}$ denotes the mean albedo values. We denote $\mathbf{p} = (\mathbf{p}_{\text{id}}^T, \mathbf{p}_{\text{exp}}^T)^T$ as the collection of shape parameters of the 3D face. We use the Basel Face Model (BFM) [19] as the identity bases, and use the FaceWarehouse [5] as the expression bases.

We also select L landmarks from the 3D face shape $\mathbf{S}$, and project them onto the 2D image. We represent the 2D coordinates of the projected landmarks using a matrix $\mathbf{U} \in \mathbb{R}^{2\times L}$. Additionally, we indicate the visibility of these landmarks using a vector $\mathbf{V} = \begin{pmatrix} v_1 & v_2 & \cdots & v_L \end{pmatrix}^T$, with each value v_i being either 0 (invisible) or 1 (visible). The relationship between the 3D shape $\mathbf{S}$ and the 2D landmarks $\mathbf{U}$ is given in Eq. (2) with projection parameter $\mathbf{w}$.

We can now represent a set of different view face images $\mathbf{I}_{ij}$ as the projections of a shared 3D face shape $\mathbf{S}_i$, using the camera parameters $\mathbf{w}_{ij}$ for the j-th face image and shared face shape parameter $\mathbf{p}_i$. In this way, the alignment problem is reduced to the estimation of these parameters.

B. Training Data Construction

As mentioned before, we need a large set of training data $\{(\mathbf{S}_i, (\mathbf{I}_{ij}, \mathbf{w}_{ij})_{j=1,2,\cdots,m_i})_{i=1,2,\cdots,N}\}$, where $\mathbf{S}_i$ is the i-th ground truth face shape, m_i is the number of its images, and $\mathbf{I}_{ij}, \mathbf{w}_{ij}$ are its j-th image and the corresponding camera parameters. All the face shapes $\mathbf{S}_i$ share the same index for their corresponding vertices. From $\mathbf{S}_i$ and $\mathbf{w}_{ij}$, we can easily compute the ground truth landmarks location $\mathbf{U}_{ij}$ and visibility $\mathbf{V}_{ij}$, by projecting the pre-selected landmarks from face shape $\mathbf{S}_i$ to the image and checking the visibility in 3D. However, there is no public large-scale multi-view face images with labeled landmarks and associated ground truth 3D models. Therefore, we construct a set of multi-view images with registered 3D face shapes as training data.

Specifically, we select 4308 frontal face images from the 300W [23] and Multi-PIE [11] datasets. We follow the method of [13] to fit a 3DMM face to each image, producing the face shape, albedo, and the camera parameters. We use second order Spherical Harmonics [20] to model the illumination, and utilize inverse render technology to increase our fitting accuracy per pixel. Afterwards, we follow the profile face image construction method in [38] to construct about 13 additional images from different views, and compute their corresponding camera parameters. In this way, we obtain $N = 4308$ synthetic ground truth face shapes, and 73412 face images with camera parameters. Fig. 2 shows an example of our fitting result and generated multi-view images.

C. Deep Face Feature Training

We propose a deep learning approach to extract a per-pixel feature from a face image, to distinguish pixels from different anatomical regions of the face. Instead of using patch-based features, we develop an end-to-end feature extraction method. This is because the former approach requires tricky and significant preprocessing of training data, including projection of every pre-selected landmarks under different poses and scales; moreover, patch-based features cannot utilize information from the whole image and cannot be easily extended to more landmarks.

Given a face image $\mathbf{I}$, its deep face feature (DFF) descriptor $\mathbf{f}$ should meet two criteria:

1) $\mathbf{f}$ depends on the location of the pixel on the human face, such that for two pixels from the same anatomical region, their feature vectors should be nearly identical.
2) $\mathbf{f}$ should be smooth, while $\|\mathbf{f}(p) - \mathbf{f}(q)\|$ is small when p and q are the projections of nearby human face points, and large for faraway points.

Inspired by the dense human body correspondence computation in [29], we formulate this feature extraction problem

Figure 4: Examples of a segmentation face mesh projecting back into 3 different view face images. Each pixel of same patch has same label.

as a multi-classification problem. We randomly generate 100 uniform segmentations of the 3DMM model, each consisting of 500 patches. To make feature more discriminative for different face regions, we run the above-mentioned random segmentation method independently on eyebrow, eye, mouth and the remaining face region, while ensuring the total number of patches is 500. Fig. 3 shows some segmentation results. Given a segmented 3DMM model and the camera parameters, we project all visible patches to the image plane, where each projected patch represents a class, and the pixels within that path have the same label. Now all the images can be segmented to different labeled patches. Fig. 4 shows some patched images. And we use convolutional Neural Network (CNN) to regress per pixel classification problem.

Since we train the DFF extractor for all 100 classification problems, the DFF extractor should be shared by different classification problem. The DFF extractor part weight is denoted as $\mathbf{Q}$. We use the Softmax loss function for each classification problem, with feature transfer $\mathbf{Q}_i$ for i-th problem. We formulate the DFF extractor learning as minimizing a combination of loss functions of all classification problems:

$$\mathbf{Q}_i^*, \mathbf{Q}^* = \arg\min_{\mathbf{Q}_i, \mathbf{Q}} \sum_i^N \text{softmax}(\mathbf{Q}_i, \mathbf{Q}).$$

After training, we take the optimized DFF extractor with weights $\mathbf{Q}^*$ as output.

Our DFF extractor is extended from AlexNet [16] architecture by adding some de-convolution layer and concatenating with shallow feature map. The input of the DFF extractor is a 224×224 gray-scale image and the output is a 64-dimensional feature for each pixel.

D. Cascaded Regression Algorithm

As shown in Fig. 1, the input of our method is a set of multi-view images $(\mathbf{I}_{ij})_{j\in 1,2,\cdots m_i}$. The face shape $\mathbf{S}_i$ is initialized by the mean shape $\mathbf{S}_0$, and the landmarks $\mathbf{U}_{ij}$ and visibility $\mathbf{V}_{ij}$ are initialized by projecting $\mathbf{S}_0$ to corresponding image with mean camera parameters $\mathbf{w}_0$.

In the k-th iteration, we extract DFF $f(\mathbf{I}_{ij}, \mathbf{U}_{ij}^k, \mathbf{V}_{ij}^k)$ from each image I_{ij} based on current landmark locations $\mathbf{U}_{ij}^k$ and visibility $\mathbf{V}_{ij}^k$. The feature vector f is a concatenation of the DFF descriptors around $\mathbf{U}_{ij}$ on $\mathbf{I}_{ij}$, which means that f is a $64L$-dimensional vector. If a landmark is invisible, its corresponding entries in f will be set to be zero. Then, we use f to regress $\mathbf{U}_{ij}$ and $\mathbf{w}_{ij}$ respectively. Based on the updated $(\mathbf{U}_{ij}, \mathbf{w}_{ij})_{j\in m_i}$, we iteratively optimize face shape $\mathbf{S}_i$, and then refine $(\mathbf{U}_{ij}, \mathbf{V}_{ij})_{j\in m_i}$ until convergence.

a) Landmark Updating.: In the k-th iteration, we use linear regressor R_U^k to update $\mathbf{U}_{ij}^k$ to $\widehat{\mathbf{U}}_{ij}^{k+1}$ via the following form:

$$\widehat{\mathbf{U}}_{ij}^{k+1} = \mathbf{U}_{ij}^k + R_U^k f(\mathbf{I}_{ij}, \mathbf{U}_{ij}^k, \mathbf{V}_{ij}^k), \tag{4}$$

and R_U^k is learned by minimizing the energy function:

$$R_U^k = \arg\min_{R_U^k} \sum_{i=1}^N \sum_{j=1}^{M_i} \|(\mathbf{U}_{ij}^* - \mathbf{U}_{ij}^k) - R_U^k f_{ij}\|_2^2 + \lambda_1 \|R_U^k\|_F^2, \tag{5}$$

where $\mathbf{U}_{ij}^*$ is the ground truth landmark locations on image $\mathbf{I}_{ij}$. This optimization problem has a close-form least-square solution.

b) Pose Updating.: We update the camera parameters $\mathbf{w}_{ij}$ based on $f(\mathbf{I}_{ij}, \mathbf{U}_{ij}^k, \mathbf{V}_{ij}^k)$ in the following form:

$$\mathbf{w}_{ij}^{k+1} = \mathbf{w}_{ij}^k + R_w^k f(\mathbf{I}_{ij}, \mathbf{U}_{ij}^k, \mathbf{V}_{ij}^k), \tag{6}$$

and R_w^k is learned by minimizing the energy function:

$$R_w^k = \arg\min_{R_w^k} \sum_{i=1}^N \sum_{j=1}^{M_i} \|(\mathbf{w}_{ij}^* - \mathbf{w}_{ij}^k) - R_w^k f_{ij}\|_2^2 + \lambda_2 \|R_w^k\|_F^2, \tag{7}$$

where $\mathbf{w}_{ij}^*$ is the ground truth camera parameters of shape $\mathbf{S}_i$ on image $\mathbf{I}_{ij}$.

c) Refining 3D Shape.: After $(\widehat{\mathbf{U}}_{ij}^{k+1})_{j\in 1,2,\cdots m_i}$ and $(\mathbf{w}_{ij}^{k+1}))_{j\in 1,2,\cdots m_i}$ have been updated, we can refine the face shape from $\mathbf{S}_i^k$ to $\mathbf{S}_i^{k+1}$ based on the new camera parameters $\mathbf{w}_{ij}^{k+1}$ and landmark coordinates $\widehat{\mathbf{U}}_{ij}^{k+1}$. The face shape is refined by minimizing an objective function:

$$E(\mathbf{p}) = \omega_{\text{lan}} E_{\text{lan}}(\mathbf{p}) + \omega_{\text{reg}} E_{\text{reg}}(\mathbf{p}), \tag{8}$$

where $\mathbf{p} = (\mathbf{p}_{\text{id}}^T, \mathbf{p}_{\text{exp}}^T)^T$ are the 3DMM model shape parameters, and $\omega_{\text{lan}} = 1.0, \omega_{\text{reg}} = 5.0$. The landmark fitting term E_{lan} and the regularization term E_{reg} are defined as:

$$\mathbf{S}_i^k = \mathbf{S}_0 + \mathbf{S}_{\text{id}} \mathbf{p}_{\text{id}}^k + \mathbf{S}_{\text{exp}} \mathbf{p}_{\text{exp}}^k,$$

$$E_{\text{lan}} = \frac{1}{m_i} \sum_{j=1}^{m_i} \|\widehat{\mathbf{U}}_{ij}^{k+1} - (s\mathbf{R}_{ij} \mathbf{S}_i^k(\ell) + \mathbf{t}_{ij})\|_2^2,$$

$$E_{\text{reg}} = \sum_{i=1}^{80} (\frac{\mathbf{p}_{\text{id},i}^k}{\sigma_{\text{id},i}})^2 + \sum_{i=1}^{79} (\frac{\mathbf{p}_{\text{exp},i}^k}{\sigma_{\text{exp},i}})^2.$$

where ℓ represents the vector of the L landmark vertex indices, and $\mathbf{S}_i^k(\ell)$ represent the coordinates of these vertices. $\mathbf{R}_{ij}$ is the first 2 rows of a 3×3 rotation matrix deduced from $\mathbf{w}_{ij}^k$, and $\mathbf{t}_{ij}$ is the translation vector deduced from $\mathbf{w}_{ij}^k$. $\widehat{\mathbf{U}}_{ij}^{k+1}$ is a $2L$-dimensional vector, and $s\mathbf{R}_{ij}\mathbf{S}_i^k(\ell) + \mathbf{t}_{ij}$ is also reshaped to the same dimension. By solving the optimization problem (8), we can update $\mathbf{S}_i^k$ to $\mathbf{S_i}^{k+1}$.

d) Refining Landmarks.: After updating $\mathbf{S}_i^{k+1}$ and all $(\mathbf{w}_{ij}^{k+1}))_{j\in 1,2,\cdots m_i}$, we can further refine $(\widehat{\mathbf{U}}_{ij}^{k+1})_{j\in 1,2,\cdots,m_i}$ to $(\mathbf{U}_{ij}^{k+1})_{j\in 1,2,\cdots m_i}$ based on Eq. (2). This refinement step constrains the 2D landmarks locations to stay in the projection space of a 3D face shape, and thus can eliminate singular landmarks distribution. The visibility $\mathbf{V}_{ij}^{k+1}$ of the landmarks $\mathbf{U}_{ij}^{k+1}$ can be determined easily on the GPU, by rendering the triangles of the 3D face mesh and the landmark vertices to the OpenGL depth buffer, with the OpenGL projection matrix determined from the corresponding camera parameters.

IV. Experiments

In this section, we conduct experiments to verify the effectiveness of the proposed approach in three aspects. We first evaluate the trained DFF against SIFT on multi-view face images. Then we compare the performance between our method and other existing approaches for large-pose face alignment in the wild. Finally, we show our alignment results with multiple inputs.

A. The Peforamance of DFF

Since SIFT is widely used for feature matching and landmark regressions, we select two multi-view face images with large difference between their view angles, to perform feature matching with SIFT and DFF. For SIFT matching, we perform SIFT points extraction on two images respectively, and match the two point sets. For each descriptor in first set, we find the closet descriptor in second set. The pair is valid only if the ratio of the distance to the second best matching point and the distance to the best matching point is greater than 1.3, this value is fine-tuned from the setting of D. Lowe[18] . For our DFF matching, we first choose one face image and use SIFT to obtain some keypoints, which guarantees that the candidate keypoints are the same with SIFT. For each descriptor in the first set, we find the closet descriptor on each pixel of the second image, and the pair is valid only if the ratio of the distance between the pair and the norm of the first descriptor is less than 0.38. The number 0.38 in DFF matching is used for excluding interference of points that are not within the face part. The threshold is mainly applied to check the effectiveness of pair whether it contains point outside the face region. Increasing the value would lead more incorrect pairs caused by candidate points outside face region.

Fig. 5 compares the matching results between SIFT and our DFF, using two images from significantly different views. The result from SIFT contains a large number of mismatched pairs, while DFF produces more consistent and accurate matching.

In Fig. 6, we visualize the smoothness of the correspondence from DFF, on two images with almost 90° difference in the yaw angle. We color the face area of the first image with Projected Normalized Coordinate Code (PNCC) [38]. For each colored pixel in the first image, we find from the face area of the second image a corresponding pixel with the closest feature vector, and assign it the same color. In Fig. 6, despite some bad matching at the boundary of the face area, the color on the second image varies smoothly, indicating the smoothness of the DFF-based correspondence.

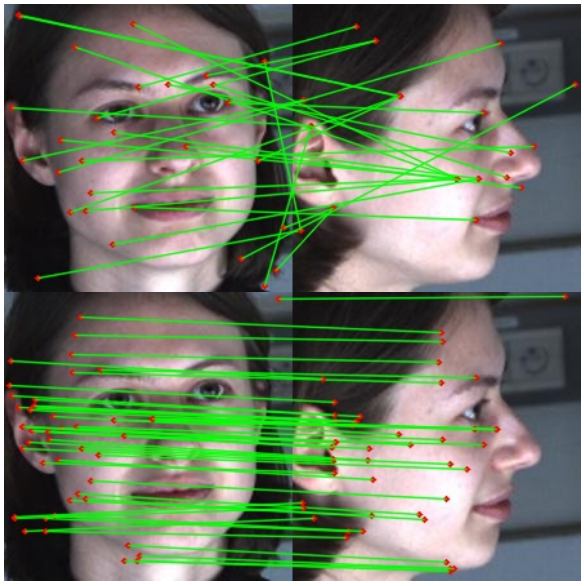

Figure 5: Example of matching results with SIFT (top) and our DFF descriptor (bottom).

Figure 6: Color-coded visualization for the smoothness of DFF-based correspondence. Each pair of corresponding pixels are assigned the same color.

In Fig. 7, we train our model using SIFT and DFF respectively, and compare their alignment accuracy. It can be seen that the model with DFF produces much more accurate results.

B. Large-pose Face Alignment

We also test the performance of our face alignment approach using single face images in the wild. We test our algorithm on the AFLW dataset[1] and the AFLW2000-3D dataset[2]. The

[1]https://lrs.icg.tugraz.at/research/aflw/

[2]http://www.cbsr.ia.ac.cn/users/xiangyuzhu/projects/3DDFA/main.htm

Figure 7: Alignment result comparison. The top shows the results trained with SIFT, and the bottom shows the results trained with DFF.

AFLW dataset consists of face images with 21 visible ground truth landmarks, while the AFLW2000-3D dataset consists of fitted 3D faces of the first 2000 AFLW samples which can be used for 3D face alignment evaluation. In our experiments, the accuracy of alignment is measured by the Normalized Mean Error (NME), which is the average of visible landmark error normalized by the bounding box [14]. For AFLW, we use the bounding boxes provided by AFLW. For AFLW2000-3D, we use the face box that hulls of all the 68 ground truth landmarks, just as [38] did. To make the comparisons fair, we train the model using the 300W-LP dataset instead of our constructed data.

For testing on the AFLW dataset, we use 22608 in-the-wild faces with large pose variations (yaw from $-90°$ to $90°$). To show the Robustness of our method for large yaw angles, we divide the testing set into 3 subsets according to their absolute yaw angles: $[0°, 30°]$,$[30°, 60°]$ and $[60°, 90°]$, with 13398, 5635 and 3575 samples respectively. Tab. I shows the comparison between our approach and existing face alignment methods on large pose data using AFLW database. The results of other methods are gathered from [38].

Tab. II shows the comparison between our approach and existing face alignment methods on large pose data using AFLW2000-3D database. We follow the same experimental setting in [38] on AFLW2000-3D. Our method significantly outperforms RCPR [4], ESR [6] and SDM [32], especially for samples within $[60°, 90°]$ yaw angles. Our method achieves similar performance with 3DDFA+SDM for $[0, 60°]$ yaw angles, and much better results for $[60°, 90°]$ yaw angles. Fig. 8 shows some alignment results using our method on challenging samples.

Table I: The NME(%) on AFLW Dataset(21 pts)

Method	[0,30]	[30,60]	[60,90]	Mean	Std
RCPR(300W-LP)	5.43	6.58	11.53	7.85	3.24
ESR(300W-LP)	5.66	7.12	11.94	8.24	3.29
SDM(300W-LP)	4.75	5.55	9.34	6.55	2.45
3DDFA	5.00	5.06	6.74	5.60	0.99
3DDFA+SDM	4.75	4.83	**6.38**	5.32	**0.92**
Our method(SIFT)	5.65	6.23	9.24	7.04	1.93
Our method(DFF)	**4.12**	**4.59**	6.62	**5.11**	1.33

Table II: The NME(%) on AFLW2000-3D Dataset(68 pts)

Method	[0,30]	[30,60]	[60,90]	Mean	Std
RCPR(300W-LP)	4.26	5.96	13.18	7.80	4.74
ESR(300W-LP)	4.60	6.7	12.67	7.99	4.19
SDM(300W-LP)	3.67	4.94	9.76	6.12	3.21
3DDFA	3.78	4.54	7.93	5.42	2.21
3DDFA+SDM	3.43	**4.24**	7.17	4.94	1.97
Our method(SIFT)	5.35	6.75	8.23	6.78	1.44
Our method(DFF)	**3.35**	4.34	**5.89**	**4.53**	**1.28**

C. Multi-view Input Refinement

The DFF feature is trained using synthetic multi-view face images, which is very effective for alignment of large-pose faces. To show the benefits of multi-view inputs against single-view inputs with our cascaded regression algorithm, we select one large-pose face image from a set of multi-view images to perform face alignment with our method, and compare it with the result using multi-view inputs.

We test the algorithm with the stereo face database [10]. The database has in total 1600 images from 100 faces, where each face has 16 different view images captured by 2 cameras. In most cases, single input and multi-view input produce similar accurate results. But for several challenging examples, multi-view inputs can produce more accurate alignment results as shown in Fig. 9. This improvement is due to the correlation constraint between the multi-view images as introduced in Eq. (8).

D. Computation Time

In our experiments, it takes about 25 ms to extract DFF for one image on a GTX 960 GPU, and about 16ms for each iteration on a 4.00 GHz CPU. All the results of our method in this paper are generated using three iterations. As we only need to extract DFF once for each image, the total running time does not exceed 75ms. This computation time is comparable with the method in [38], while our method achieves more accurate alignment results.

V. Conclusions

In this paper, we proposed a novel method for face alignment from single-view or multi-view face images. Utilizing the correlation between image from different views, we devised a deep learning based approach to train a deep face feature extractor, which shows great performance compared with general feature descriptors such as SIFT. Besides, the proposed regression algorithm iteratively updates the landmarks, pose and 3D shape with the multi-view constraints. As far as we know, this is the first work to do alignment for multi-view images.

References

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

[2] 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.

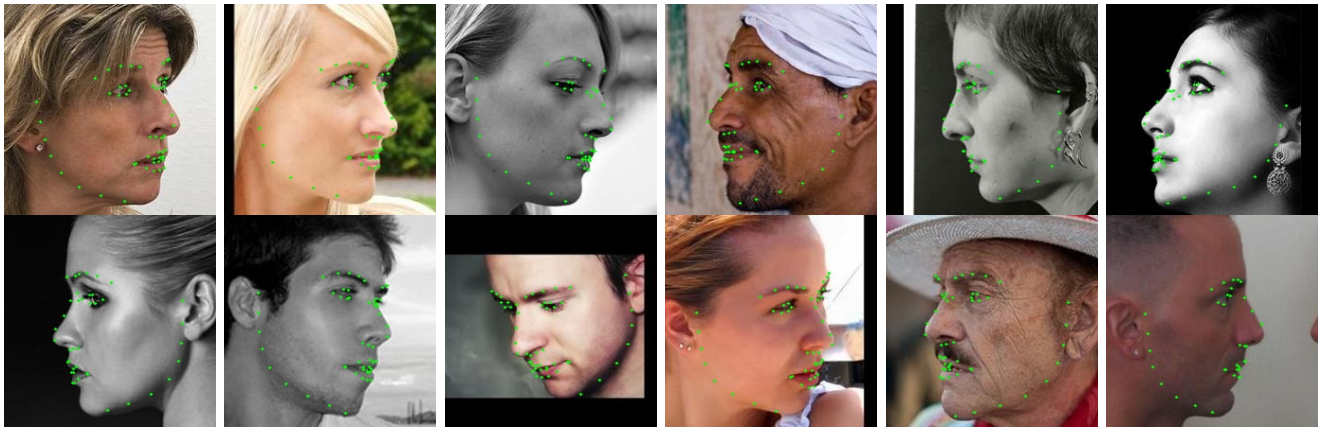

Figure 8: Examples of face alignment on large poses from the AFLW2000-3D database.

Figure 9: Comparison of results by our method with a single view input (top) and with 8 multi-view images as input (bottom). Using multiple input images leads to more accurate landmark positions and visibilities.

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003.

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[5] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

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

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

[9] D. Cristinacce and T. F. Cootes. Boosted regression active shape models. In *BMVC*, volume 2, pages 880–889, 2007.

[10] R. Fransens, C. Strecha, and L. Van Gool. Parametric stereo for multi-pose face recognition and 3d-face modeling. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 109–124. Springer, 2005.

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

[12] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[13] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu. 3d face reconstruction with geometry details from a single image. *arXiv preprint arXiv:1702.05619*, 2017.

[14] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015.

[15] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[19] 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.

[20] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500. ACM, 2001.

[21] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

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

[23] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[24] 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.

[25] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[26] C. TECootes and A. Lanitis. Active shape models: Evaluation of a multi-resolution method for improving image search. In *Proc. British Machine Vision Conference*, pages 327–338. Citeseer, 1994.

[27] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE international conference on computer vision*, pages 593–600, 2013.

[28] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):372–386, 2012.

[29] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1544–1553, 2016.

[30] C. Wu. Towards linear-time incremental structure from motion. In *3DTV-Conference, 2013 International Conference on*, pages 127–134. IEEE, 2013.

[31] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3057–3064. IEEE, 2011.

[32] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[33] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[34] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[35] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[36] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[37] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[38] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016.

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