https://doi.org/10.2352/ISSN.2470-1173.2019.8.IMAWM-403

Face Alignment via 3D-Assisted Features

Song Guo [a], Fei Li [a], Hajime Nada [b], Hidetsugu Uchida [b], Tomoaki Matsunami [b] and Narishige Abe [b]
[a] Fujitsu Research & Development Center Co., Ltd.; Beijing, China
[b] Fujitsu Laboratories Ltd.; Kanagawa, Japan

Abstract

We present a practical 3D-assited face alignment framework based on cascaded regression in this paper. The 3D information embedded in 2D face image is utilized to calculate two novel components to improve the performance of 2D methods in unconstrained face alignment. The two novel components for 2D image features are the projected local patch and the visibility of each landmark. First, we propose to extract the landmark related features in the projected local patches on 2D image from the corresponding 3D face model. Local patches of a fixed landmark in 3D face models for different 2D images cover the same region of face anatomically. The extracted features are more accurate for further locations regression of landmarks. Second, we propose to estimate the visibilities of 2D landmarks based on 3D face model, which are proven to be vital to address large pose face alignment problem. In this paper, we adopt Local Binary Features (LBF) to extract landmark related features in the proposed framework, and name the new method as 3D-Assisted LBF (3DALBF). An extensive evaluation on two face databases shows that 3DALBF can achieve better alignment results than the original 2D method and maintain the speed advantage of 2D method over 3D method.

Introduction

Face alignment aims to locate a set of predefined facial landmarks, such as eye corners, nose tip, mouth corners and chin center. It is an essential step for many subsequent face analysis tasks, e.g., face recognition [1], facial attributes classification [2] and 3D face reconstruction [3]. Although a large number of methods have been proposed to address this problem with various degrees of success, face alignment in unconstrained environment remains open, due to the high degree of facial appearance variations caused either by the intrinsic non-rigid of facial components, or by the change of ambient environment.

From the perspective of landmarks dimension, most existing methods can be divided into two categories: 2D face alignment and 3D face alignment. 2D face alignment, which treats face as a 2D object, aims to find the 2D locations of facial landmarks from 2D images, such as Constrained Local Model (CLM) [4, 5], Supervised Descent Method (SDM) [6] and Coarse-to-Fine Shape Searching (CFSS) [7]. In contrast, 3D face alignment, which treats face as a 3D object, aims to find the 3D locations of facial landmarks from 2D images [8], such as Pose-Invariant 3D Face Alignment (PIFA) [9], 3D Dense Face Alignment (3DDFA) [10] and 3D Face Alignment Network (3D-FAN) [11]. Generally speaking, 2D methods can achieve better results than 3D methods for faces in small to medium poses with higher speed; however, 3D methods can achieve better result than 2D methods for faces in large poses.

The reason that large pose face alignment is challenging for 2D methods is that large pose can cause self-occlusion and large shape variation in face image. Therefore, some landmarks become invisible and lose their semantic meanings. Most 2D methods assume that each landmark can be located accurately because it has distinctive visual feature. However, landmark related visual features may be incorrect in a large pose face image, which can cause failure of 2D face alignment methods. In contrast, 3D methods can estimate head pose and the visibility of each landmark with 3D facial information, which makes these methods robust to large pose variation in face image. However, most existing 3D methods convert the face alignment problem from predicting landmarks position to estimating parameters of 3D model, which are shown to be indirect and sub-optimal since smaller parameter errors are not necessarily equivalent to smaller alignment errors [12, 13]. Furthermore, 3D landmark related features are usually more complicated to extract and most 3D methods adopts deep neural network to locate the landmarks, which cause higher time cost of 3D methods.

Both 2D methods and 3D methods in face alignment have their limitations, either in accuracy or in speed, which make them not suitable for subsequent face analysis tasks in practice. Most 2D methods are incapable of estimating the visibility of 2D landmarks, assuming that all landmarks are visible. However, self-occlusion in large pose face images makes this assumption not true. Since the local features of invisible landmark are inaccurate, we attempt to extract features separately for different landmarks according to their visibilities in this paper. Furthermore, the landmark related features are usually extracted in the local patch of each landmark in traditional 2D methods, e.g. ESR [12], LBF [14], CFSS [7], etc. However, these local patches are usually untidy with complex background, causing inaccurate of the corresponding landmark related features. Therefore, we propose to extract these features in the projected local patches of visible landmarks from 3D model. For invisible landmarks, the features can be extracted in the local patches of their symmetric landmarks due to the symmetry of face shape, or set to zero to eliminate the influence of inaccurate features.

In this paper, a practical framework based on cascaded regression is proposed to deal with the problem of face alignment in unconstrained environment. A 3D face model is first constructed from a 2D face image, and then the visibility of each landmark and the 3D to 2D projection function are calculated. Landmark related features are extracted in the projected local patch in 2D image from 3D model. With the extracted features, regressors can be learnt to adjust the locations of landmarks towards their desired position. Various types of feature descriptors can be adopted to extract the landmark related feature, such as SIFT [7], HOG [7] and LBF [14]. LBF is employed in this paper, and the new method is named as 3D-Assisted LBF (3DALBF).

The main contributions of this paper can be summarized as follows:

1. More accurate features: We propose to extract the landmark related features in the projected local patches of 2D image from 3D face model. The extracted shape-indexed features are more accurate for further landmark location regression. Furthermore,

local patches of a fixed landmark in 3D face models for different 2D images cover the same region of face anatomically. Therefore, the projected local patch of one 2D image corresponds to the patch from another image anatomically. These correspondences between different images contribute to improving the accuracy of landmark related features.

2. Robust to large pose variation: We propose to estimate the visibilities of 2D landmarks based on 3D face model. Landmarks related features are then extracted separately according to their visibilities in 2D image plane.

Experimental results on 300W and ALFW2000-3D databases show that the proposed method can achieve better alignment results than original 2D method and maintain the speed advantage of 2D method over 3D method.

Related works

A large number of methods have been proposed for face alignment in the last decades. We briefly review related works in this section, including methods for 2D and 3D face alignment.

2D face alignment

Typical methods for 2D face alignment include Active Appearance Model (AAM) [15, 16], Constrained Local Model (CLM) [4, 5], Cascaded Regression (CR) [7, 12, 14] and Deep Neural Network (DNN) [17, 18, 19]. There are a number of successful methods following the cascaded regression framework, which is also adopted in our work. The basic idea of cascaded regression is to learn a series of regressors from shape-indexed features to reduce the alignment error progressively.

Cao *et al.* [12] design a two-level boosted fern regressor to progressively infer the face shape with selected pixel-difference features. Supervised descent method (SDM) [6] is proposed to predict shape increment by applying linear regression on SIFT features. Ren *et al.* [14] propose a very fast face alignment method, which learns a set of local binary features (LBF) and a linear regressor in a cascaded manner. Zhu *et al.* [7] present a novel coarse-to-fine shape searching method to locate facial landmarks stage-by-stage by using hybrid features of SIFT and BRIEF. We can see that advanced feature learning contributes to achieving higher alignment accuracy and speed [14, 20]. However, most 2D methods lack the ability to estimate the visibilities of landmarks, which are proven to have an important influence on the accuracy of exacted shape-indexed features in large pose face alignment.

3D face alignment

3D face alignment methods are introduced to address the large pose face alignment problem. 3D face model possesses natural advantages of handling a full range of head pose while still maintaining the landmark correspondences between different faces. Besides, it also provides more information for estimating the head pose and visibility of 3D facial points.

Jourabloo and Liu [9] propose to learn two regressors to predict the camera projection matrix and 3D shape parameters alternatively, following the cascaded regression framework. Besides, they also propose to estimate the visibility of 3D facial landmarks via 3D surface normal. They further present a method to fit a 3D dense shape to a face image with large poses by combining cascaded convolutional neutral network (CNN) regressors and the 3D Morphable Model (3DMM) [21]. Zhu *et al*. [10] also propose to perform 3D face alignment by fitting a 3DMM to a 2D face image via cascaded CNN with different image features and cost function. These methods transform the face alignment problem from predicting landmarks position to estimating parameters of 3D model, which are proven to be indirect and sub-optimal. Besides, complex designed feature and cascaded CNN regressors lead to higher time cost.

The proposed method

This section describes the proposed 3D-Assisted Local Binary Feature (3DALBF) method in detail. 3DALBF follows the cascaded regression scheme. In order to utilize the 3D information of face image, we need to construct a 3D face model from 2D image with landmarks and their visibilities in current stage as reference. Together with the construction of 3D model, the head pose of 3D face and the projection function from 3D model to 2D face are calculated. The visibility of each landmark is then estimated in 3D face space. The local patch of each landmark on 3D face model is projected onto the 2D image plane with corresponding projection function. Landmark related features are extracted in the projected local patch in 2D image for visible landmarks. For invisible landmarks, the features can be extracted in the projected local patches of their symmetric landmarks due to the symmetry of face shape, or set to zero to eliminate the impact of inaccurate features on learning regressors. With the features extracted for each landmark, regressors are learnt to adjust the locations of landmarks towards their desired position.

3D Face Model

We represent the 3D face model as $\mathbf{S}$, which contains the 3D locations of M vertices,

$$\mathbf{S} = \begin{pmatrix} x_1 & x_2 & \cdots & x_M \\ y_1 & y_2 & \cdots & y_M \\ z_1 & z_2 & \cdots & z_M \end{pmatrix} \tag{1}$$

3D Morphable Model (3DMM) [22] is adopted to describe $\mathbf{S}$ by a set of 3D shape bases,

$$\mathbf{S} = \mathbf{S}_0 + \sum_{n=1}^{N_{id}} c_{id}^n \mathbf{S}_{id}^n + \sum_{n=1}^{N_{\exp}} c_{\exp}^n \mathbf{S}_{\exp}^n \tag{2}$$

where $\mathbf{S}_0$ is the mean face model, $\mathbf{S}_{id}^n$ and $\mathbf{S}_{\exp}^n$ are the nth PCA basis for identity and expression. c_{id}^n and $c_{\exp}^n$ are the reconstruction coefficients for $\mathbf{S}_{id}^n$ and $\mathbf{S}_{\exp}^n$ respectively. The collection of both coefficients is denoted as the shape parameter of a 3D face model, i.e. $\mathbf{c} = (\mathbf{c}_{id}, \mathbf{c}_{\exp})$. The 3D mean model $\mathbf{S}_0$ and the identity bases $\mathbf{S}_{id}$ are from Basel Face Model [23], which contains 199 bases for describing identification variances. The expression bases $\mathbf{S}_{\exp}$ are from FaceWarehouse [24], including 29 bases for describing expression variances.

In 2D face alignment, 2D face shape $\mathbf{U}$ can be represented by the locations of L 2D landmarks, i.e.

$$\mathbf{U} = \begin{pmatrix} u_1 & u_2 & \cdots & u_L \\ v_1 & v_2 & \cdots & v_L \end{pmatrix} \tag{3}$$

A subset of L vertices of the 3D face model $\mathbf{S}$, denoted as $\mathbf{S}(:,\mathbf{d})$, corresponds to the location of 2D landmarks $\mathbf{U}$ on the image. The relationship between the 3D face model $\mathbf{S}$ and 2D shape $\mathbf{U}$ can be described as

$$\mathbf{U} = f(\mathbf{S}(:,\mathbf{d})) \tag{4}$$

where $f(\cdot)$ is the projection function from 3D model to 2D shape. $\mathbf{d}$ is a L-dim vector indicating the indexes of 3D vertexes with

semantically meaning that correspond to 2D landmarks. Using the weak perspective projection, $f(\cdot)$ can be further expanded as

$$f(\mathbf{S}) = s\mathbf{PRS} + \mathbf{t} \tag{5}$$

where s is the scale parameter, $\mathbf{P}$ is the orthographic projection matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $\mathbf{R}$ is the rotation matrix controlled by three rotation angles α, β, and γ, corresponding to pitch, yaw and roll in head pose respectively, $\mathbf{t}$ is the 2D translation parameter vector composed of t_x and t_y. All parameters of the projection function can form a vector $\mathbf{p} = (s, \alpha, \beta, \gamma, t_x, t_y)^{\mathrm{T}}$. The projection parameter $\mathbf{p}$ and shape coefficients parameter $\mathbf{c}$ guarantee the uniqueness of a 3D face model.

In the first step of 3DALBF, we need to construct a 3D face model from 2D image with landmarks and their visibilities. As mentioned before, we can identify a 3D face model with the projection parameter vector $\mathbf{p}$ and shape parameter vector $\mathbf{c}$. The construction of 3D face model can be transformed to the estimation of $\mathbf{p}$ and $\mathbf{c}$. Given 2D landmarks $\mathbf{U}$ and their visibilities $\mathbf{V}$, parameters $\mathbf{p}$ and $\mathbf{c}$ can be estimated by minimizing the following optimization function,

$$J(\mathbf{p}, \mathbf{c}) = \arg\min_{\mathbf{p},\mathbf{c}} \left\| (\mathbf{pS}(:, \mathbf{d}) - \mathbf{U}) \odot \mathbf{V} \right\|_F^2 \tag{6}$$

where $\mathbf{S}$ is expressed by Eq. (2), $\odot$ denotes the element-wise multiplication. $\mathbf{V}$ is an L-dim vector with binary elements indicating whether the landmarks are visbile (denoted as 1) or not (denoted as 0). The objective function in (6) is the difference between the locations of visible 2D landmarks and their 3D projections. Alternating optimization strategy is utilized to compute the optimal parameters $\mathbf{p}$ and $\mathbf{c}$. We initialize the 3D shape parameter $\mathbf{c}$ to 0, and estimate $\mathbf{p}$ by $\mathbf{p}^k = \arg\min_{\mathbf{p}} J(\mathbf{p}, \mathbf{c}^{k-1})$, and then estimate $\mathbf{c}$ by $\mathbf{c}^k = \arg\min_{\mathbf{c}} J(\mathbf{p}^k, \mathbf{c})$. This process continues iteratively until the changes of $\mathbf{p}$ and $\mathbf{c}$ are small enough. Both optimization problems can be efficiently solved in closed forms by least-square method. The head pose (i.e. rotation matrix $\mathbf{R}$) of 3D face can be easily estimated based on the projection parameter $\mathbf{p}$.

Landmark visibility

One of the disadvantages of traditional 2D face alignment methods is that they are not robust to large pose variation. Some landmarks are invisible because of self-occlusion in large pose, and consequently their landmark related visual features will be incorrect. However, landmark visibility estimation in 2D image is difficult, since there is little information available. On the contrary, landmark visibility can be easily computed in 3D face model.

With the calculated head pose of 3D face, we can compute the visibility of each 2D landmark by examining whether the 3D surface normal of the corresponding 3D landmark is pointing to the camera or not [9]. For each landmark, we first compute the 3D surface normal of a set of vertexes around the 3D landmark of given, and the average of these 3D normal, denoted as $\mathbf{N}$, is regarded as the surface normal of this 3D landmark. By rotating the surface normal of each landmark according to the head pose, we can compute whether the rotated surface normal is pointing toward or away from the camera, representing visible or invisible of the landmark respectively. With the 3D surface normal $\mathbf{N}_l$ of lth landmark and the head rotation matrix $\mathbf{R}$, we can compute its direction by

$$v_l = \mathbf{N}_l \cdot \mathbf{R}_{12} \tag{7}$$

where $\mathbf{R}_{12}$ is the first two rows of rotation matrix $\mathbf{R}$. If v_l is positive, the lth 2D landmark is visible, and it is invisible otherwise. The visibility of each landmark will be further utilized in extracting landmark related features in the following section.

3D-Assisted LBF

In traditional 2D face alignment methods, the landmark related visual features are extracted in the local patch of each landmark in 2D image. However, most of the 2D images contain not only people face, but also complex background, even after face detection. The local patch of each landmark may include large area of complex background, especially for the contour landmarks; therefore, the extracted features of landmarks may be inaccurate or even incorrect. And this condition will be worse in large pose face images. The inaccurate features directly deteriorate the accuracy of predicted landmarks' locations.

In this paper, the projected local patch is introduced to tackle this problem. The local patch of each 3D landmark, which is a ball centered at the landmark of interest, is defined on the surface of a 3D face model. In other words, the 3D local patch is a collection of vertices on the surface of a 3D model inside a sphere. The projected local patch of each 2D landmark can be obtained by 3D-2D projection function, which means each vertex in the local patch of 3D landmark will be projected onto the 2D image plane. The projected 2D points constitute the local patch of 2D landmark. Since there is no background in 3D model, the projected local patch will be much cleaner. The shape of projected local patch for each landmark is irregular and varying because of landmark position and 3D-2D projection function. Figure 1 shows the comparison of local patches and projected local patches in 2D face images.

Figure 1. The comparison of local patches and projected local patches in 2D face image. In each pair, the left one is a circle local patch around each landmark, the right one is the projected local patch from 3D face model, which is much cleaner than the left one. The landmark is marked with a blue point.

Furthermore, local patches of a fixed landmark in 3D face models for different 2D images cover the same region of faces anatomically except for the minor marginal difference due to shape variances (i.e. identification variances and expression variances). Therefore, the projected local patch of one 2D image corresponds to the same patch from another image anatomically. Figure 2 shows the projected local patches of some image examples from 300W database. These correspondences between different images contribute to improving the accuracy of landmark related features, which will be testified in our experiments.

After the projection, features are extracted in the projected local patches for visible landmark. For invisible landmarks, the features can be extracted in the local patches of their symmetric landmarks due to the symmetry of face shape, or set to zero to eliminate the influence of inaccurate features of invisible landmarks. The local patches in 3D model are pose invariant, so the extracted landmark related features are robust to large pose variation.

Figure 2. The projected local patches of the 17th landmark in different images have the same anatomically meaning

In this paper, we adopt 2D LBF method for feature extraction, and propose 3D-Assisted LBF (3DALBF) method for face alignment. In 2D LBF, the feature mapping function is learned by using regression random forest, whose trees are trained with the pixel-difference feature. The pixels are sampled in a local region around each landmark. Such a local region is critical to LBF, since it has been proven to be more effective to only consider candidate features in a local region instead of the global face image [14]. The local region is defined by a circle around the landmark of interested, and the radius of the region gradually shrink from early to later stage. In 3DALBF, the local region is redefined by the projected local patch from 3D face model, which is much cleaner and has anatomical correspondence across different images. The region of projected local patch depends on the radius of the ball on 3D face model, which will decrease with the cascaded regression stage. In our training, the optimal radius of the 3D ball is estimated by cross validation on a hold-out validation set at each stage.

The local feature mapping function ϕ_l^k is learned in the projected local patch via random forest, which is further used to generate local binary features for each landmark. Local features of all landmarks are then concatenated to form the 3DALBF via a global feature mapping function, i.e. $\mathbf{\Phi}^k = [\phi_1^k, \phi_2^k, \cdots, \phi_L^k]$. More specifically, a sample traverses the trees until it reaches one leaf node for each tree. The 3DALBF is a vector indicating whether a leaf node is reached or not. Supposing the total number of leaf node is Q, 3DALBF will be a Q-dimension binary vector. For each dimension in 3DALBF, its value is 1 if the sample reaches the corresponding leaf node and 0 otherwise.

After we obtain the 3DALBF, a linear regressor $\mathbf{W}^k$ at stage k is learned by minimizing the sum of alignment errors,

$$\mathbf{W}^k = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left\| \mathbf{U}_i^{gt} - \left(\mathbf{U}_i^{k-1} + \mathbf{W}\mathbf{\Phi}^k(\mathbf{I}_i, \mathbf{U}_i^{k-1}) \right) \right\| + \lambda \|\mathbf{W}\|_2^2 \quad (8)$$

where $\mathbf{I}_i$ is the ith training sample, $\mathbf{U}_i^{gt}$ is the ground truth shape of $\mathbf{I}_i$, and $\mathbf{U}_i^{k-1}$ is the estimated shape in previous stage, N is number of training samples,. A dual coordinate descent method [25] is adopted to deal with the sparse linear problem. The training procedure of 3DALBF is summarized in Algorithm 1.

Algorithm 1: The training procedure of 3DALBF

Input: Training data $\left\{\mathbf{I}_i, \mathbf{U}_i^{gt}\right\}_{i=1}^{N}$, initial landmarks with their visibility $\left\{\mathbf{U}_i^0, \mathbf{V}_i^0\right\}_{i=1}^{N}$, 3D model and bases $\left\{\mathbf{S}_0, \mathbf{S}_{id}, \mathbf{S}_{\exp}\right\}$

Output: Feature mapping functions $\left\{\mathbf{\Phi}^k\right\}_{k=1}^{K}$, cascaded regressors $\left\{\mathbf{W}^k\right\}_{k=1}^{K}$

For k = 1: K do

1: Compute the projection parameter $\mathbf{p}_i^k$ and shape parameter $\mathbf{c}_i^k$ via Eq. (6);

2: Update 3D model $\mathbf{S}_i^k$ for each sample $\mathbf{I}_i$ via Eq. (2);

3: Compute the visibility of each landmark via Eq. (7);

4: Project the local patch around each landmark on 3D face model to 2D image: $\mathbf{Patch}_{ik}^l = s^k \mathbf{P}\mathbf{R}_i^k \mathbf{S}_i^k\left(:, \mathbf{d}_{ik}^l\right) + \mathbf{t}_i^k$, $\mathbf{S}_i^k\left(:, \mathbf{d}_{ik}^l\right)$ is the local patch of landmark l in image i at stage k on 3D model, and $\mathbf{Patch}_{ik}^l$ is the corresponding projected local patch in 2D image plane;

5: Learn feature mapping function $\mathbf{\Phi}^k(\mathbf{I}_i, \mathbf{U}_i^{k-1})$ by using random forest with pixel-difference features extracted in projected local patches;

6: Learn global linear regressor $\mathbf{W}^k$ via Eq. (8);

7: Update 2D shape for each image: $\mathbf{U}_i^k = \mathbf{U}_i^{k-1} + \mathbf{W}^k\mathbf{\Phi}^k(\mathbf{I}_i, \mathbf{U}_i^{k-1})$

End for

Experiments

In this section, we evaluate the performance of 3DALBF in three databases, i.e. 300W, 300W-LP and AFLW2000-3D.

Databases

300W: 300W [26] is created from multiple databases, including AFW [27], LFPW [28], HELEN [29], IBUG [26] and XM2VTS [30]. Each image in the dataset is annotated with 68 landmarks. We adopt the same training set (3148 images) and testing set (689 images) as in [14]. The testing set is further split into three subsets: the *common subset* which consists of the test subsets of LFPW and HELEN (554 images); the *challenging subset* which consists of the IBUG dataset (135 images); the *full set* which is the summation of *common subset* and *challenging subset* (689 images).

300W-LP: The 300W-LP database [10] is a large pose extension of the 300W database, which contains 61225 samples from four databases (1786 from IBUG, 5207 from AFW, 16556 from LFPW and 37676 from HELEN).

ALFW2000-3D: AFLW dataset [15] contains 21080 in-the-wild face images, and each image is annotated with up to 21 visible landmarks and a bounding box. Zhu *et al*. [10] choose the first 2000 AFLW samples and construct a database called ALFW2000-3D for 3D face alignment. The new database contains the ground truth 3D faces and the corresponding 68 landmarks in 2D images for each sample.

Medium pose face alignment

The experiments for medium pose face alignment are conducted on 300W database, following similar protocol in [14]. Data augmentation is adopted to enlarge the training data and improve generalization ability: each training image is augmented to multiple training samples by randomly sampling the initial shape among the training set multiple times [12].

There are a few parameters in the proposed 3DALBF method: the number of stages T, the number of trees in each stage N, and the tree depth D. For comparison with the original LBF, we set the parameters as follows: $T = 5$, $N = 1224$ (i.e. 68×18), $D = 6$. We set the number of trees N to 1224 (N=1200 in LBF), which is an integer multiple of the number of landmarks in 300W database (i.e. 68). In the training process, we find that the time cost of training random forest grows exponentially with the tree depth. So we set the tree depth D to be 6 instead of 7 in the original LBF to reduce time cost, and the performance degradation is negligible in our experiments. In addition, the data augmentation number is set to be 15 in our implementation, instead of 20 in LBF.

We compare 3DALBF method with five methods that are based on cascaded regression framework. ESR [12], SMD [6], LBF [14], CFSS [7] are 2D methods, and 3DDFA [10] is a 3D method. The alignment accuracy is evaluated by the Normalized Mean Error (NME), which is the mean of the landmark distance error normalized by the inter-pupil distance.

We use the original results of the compared methods in the literature [10] for comparison. Table 1 lists the NME results of all compared methods on 300W testing sets. We can see that the proposed 3DALBF method achieve the best performance on the *common subset* and *full set*, and competitive performance on the *challenging subset*. 3DALBF achieve a significant error reduction comparing with the original LBF on all testing subset. This can be attributed to the more accurate local features extracted in the projected local patches around each landmark.

Table 1: The NME (%) of compared methods on 300W

Method	Common subset	Challenging subset	Full set
ESR	5.28	17.00	7.58
SDM	5.57	15.40	7.50
CFSS	4.73	9.98	5.76
LBF	4.95	11.98	6.32
3DDFA	6.15	10.59	7.01
3DDFA+SDM	5.53	9.56	6.31
3DALBF	3.69	10.03	4.93

Large pose face alignment

We further conduct experiment for large pose face alignment in this section. In the following experiment, we adopt 300W-LP as the training set and AFLW2000-3D as the testing set. With the ground truth 3D models, the bounding boxes enclosing all the landmarks are provided for initialization in the testing set. The parameters of 3DALBF are set to the same in the 300W experiments, except that no data augmentation is performed in the training process.

The alignment accuracy is evaluated by the NME, which is normalized by the bounding box size [9, 10]. The NME results of the compared methods are listed in Table 2. Compared with 3DDFA, 3DALBF demonstrates competitive performance in large pose face alignment task. It is worth noting that 3DALBF achieves a significant alignment error reduction of 48% comparing with the original LBF method. NME result in Table 2 testify the robustness of 3DALBF to large pose variations.

Table 2: The NME (%) of compared methods on ALFW2000-3D

Method	ALFW2000-3D
ESR	7.99
SDM	6.12
LBF	10.19
3DDFA	5.42
3DDFA+SDM	4.94
3DALBF	5.32

Time cost

We compare the time cost of our proposed 3DALBF with 2D LBF and 3DDFA on a single core CPU with a frequency of 3.20GHZ. Both the time costs of LBF and 3DDFA are the results in literature [14] and [10], and both their experiments are conducted on 3.40GHZ CPU. From Table 3, we can see that the time cost of 3DALBF is higher than LBF, but it is still much lower than 3DDFA. The increased time cost of 3DALBF over LBF focus on the reconstruction 3D face model and estimation of landmarks' visibilities, which can be further optimized in the future. The superior performance of 3DALBF in alignment accuracy and time cost make it suitable for subsequent face analysis tasks in practice. One more thing, the number of cascaded stage is set to 3 in 3DDFA, which is 5 in both LBF and 3DALBF.

Table 3: The time cost of compared methods on 300W testing set for one image

Method	Time cost (ms)
LBF	3.13
3DDFA	75.72
3DALBF	18.28

Conclusion

In this work, we have presented a practical 3D-assited face alignment framework based on cascaded regression. Different from traditional 2D methods, the landmark related features are extracted in the projected local patches from 3D face model in the novel framework. Visibilities of 2D landmarks are estimated to enhance the robustness of 2D features to large pose variations. We further employ LBF as the feature extractor and propose 3DALBF method. By incorporating 3D facial information, 3DALBF achieves better alignment results than the original 2D method and maintain the speed advantage of 2D method over 3D method. Experiments on 300W and AFLW2000-3D show the superiority of the propose method. Furthermore, it is worth exploring more 2D feature descriptors, such as SIFT and HOG, to fit in this framework.

References

[1] 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 Trans. Pattern Anal. Mach. Intell., vol. 34, no. 2, pp. 372-386, 2012.

[2] Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., "Attribute and simile classifiers for face verification", in Proceedings of the IEEE International Conference on Computer Vision, pp. 365-372, 2009.

[3] J. Roth, Y. Tong, X. Liu, "Unconstrained 3D face reconstruction", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2606-2615, 2015.

[4] A. Asthana, S. Cheng, S. Zafeiriou, M. Pantic, "Robust discriminative response map fitting with constrained local models", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3444–3451, 2013.

[5] J. Saragih, S. Lucey, J. Cohn, "Deformable model fitting by regularized landmark mean-shift", Int. J. Comput. Vis., vol. 91, no. 2, pp. 200-215, 2011.

[6] X. Xiong, F. Torre, "Supervised descent method and its application to face alignment", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 532–539, 2013.

[7] S. Zhu, C. Li, C. Loy, X. Tang, "Face alignment by coarse-to-fine shape searching", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4998-5006, 2015.

[8] S. Tulyakov, N. Sebe, "Regressing a 3d face shape from a single image", in Proceedings of the IEEE International Conference on Computer Vision, pp. 3748-3755, 2015.

[9] A. Jourabloo, X. Liu," Pose-invariant 3d face alignment", in Proceedings of the IEEE International Conference on Computer Vision, pp. 3694-3702, 2015.

[10] X. Zhu, Z. Lei, X. Liu, H. Shi, S. Li, "Face alignment across large poses: a 3d solution", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 146-155, 2016.

[11] A. Bulat, G. Tzimiropoulos. "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks)", In Proceedings of the IEEE International Conference on Computer Vision, pp.1021-1030, 2017.

[12] X. Cao, Y. Wei, F. Wen, J. Sun, "Face alignment by explicit shape regression", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2887-2894, 2012.

[13] S. Zhu, C. Li, C. Chen, X. Tang, "Unconstrained face alignment via cascaded compositional learning", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3409-3417, 2016.

[14] S. Ren, X. Cao, Y. Wei, J. Sun, "Face alignment at 3000 fps via regressing local binary features", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1685-1692, 2014.

[15] T. Cootes, G. Edwards, C. Taylor, "Active appearance models", IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681-685, 2001.

[16] G. Tzimiropoulos, M. Pantic, "Optimization problems for fast aam fitting in-the-wild", in Proceedings of the IEEE International Conference on Computer Vision, pp. 593–600, 2013.

[17] Y. Sun, X. Wang, X. Tang, "Deep convolutional network cascade for facial point detection", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3476–3483, 2013.

[18] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks", in Proceedings of the European Conference on Computer Vision, pp. 57-72, 2016

[19] Z. Zhang, P. Luo, C. Loy, X. Tang, "Learning deep representation for face alignment with auxiliary attributes", IEEE Trans. Pattern Anal. Mach. Intell. Vol. 38, no. 5, pp. 918-930, 2016

[20] V. Kazemi, J. Sullivan, "One millisecond face alignment with ensemble of regression trees", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp.1867-1874, 2014.

[21] A. Jourabloo, X. Liu, "Large-pose face alignment via cnn-based dense 3d model fitting", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4188-4196, 2016.

[22] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model", IEEE Trans. Pattern Anal. Mach. Intell. Vol. 25, no. 9, pp. 1063-1074, 2003.

[23] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, "A 3D face model for pose and illumination invariant face recognition", In Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296-301, 2009.

[24] C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, "Facewarehouse: a 3d facial expression database for visual computing", IEEE Trans. Vis. Comput. Graph., vol. 20, no. 3, pp. 413-425, 2014.

[25] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, "Liblinear: A library for large linear classification", Journal of Machine Learning Research, vol. 9, no. 9, pp. 1871-1874, 2008.

[26] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, "300 faces in-the-wild challenge: the first facial landmark localization challenge", in Proceedings of the IEEE International Conference on Computer Vision, pp. 397-403, 2013.

[27] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild", in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2879-2886, 2012.

[28] P. Belhumeur, D. Jacobs, D. Kriegman, N. Kumar, "Localizing parts of faces using a consensus of exemplars", in Proceedings of the IEEE

International Conference on Computer Vision and Pattern Recognition, pp. 545-552, 2011.

[29] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization", in Proceedings of the European Conference on Computer Vision, pages pp. 679-692, 2012.

[30] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, "Xm2vtsdb: The extended m2vts database", In Second International Conference on Audio and Video-based Biometric Person Authentication, pp.72-77, 1999.

[31] M. Kiesinger, P. Woolhat, P. Roth, H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization", in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2144-2151, 2011.

Author Biography

Song Guo *received his B.S. degree in biomedical engineering from Beijing Jiaotong University (2007) and his Ph.D. degree in signal and information processing from Beijing Jiaotong University (2018). Since then he has worked in Fujitsu Research & Development Center Co. Ltd., Beijing, China. His research interests include pattern recognition, image processing, and machine learning.*

Fei Li *received his B.S. degree in automation from Beijing University of Aeronautics and Astronautics in 2004, and his Ph.D. degree in control science and engineering from Tsinghua University in 2009. Then he joined Fujitsu Research & Development Center Co. Ltd., Beijing, China. His research interests include pattern recognition, image understanding, robot vision, and computer graphics.*

Hajime Nada *received his BS in mathematics (2007) and his MS in mathematical science (2009) from Kyoto University. Since then he has worked at Fujitsu Laboratories LTD. in Japan. He was a visiting scholar in the Department of Electrical and Computer Engineering at Rutgers University (2017-2018). His work has focused on research and development of algorithms for biometric authentication.*

Hidetsugu Uchida *received his B.S. and M.S. in design from Kyushu University (2012 and 2014 respectively) and his PhD in engineering from the University of Tokyo (2017). Since 2017, he has worked at Fujitsu Laboratories LTD. in Japan. His research interests include speech engineering and biometric authentication.*

Tomoaki Matsunami *received his Bachelor of Engineering (2009) and his Master of Information Science and Technology (2011) from the University of Tokyo. Since 2011, he has worked for Fujitsu Laboratories LTD. in Japan. His research interests include image processing and biometric authentication.*

Narishige Abe *received his B.S. in Engineering (2005) from Osaka City University and M.S. in Information Science from Osaka University (2007). Since 2007, he has worked at FUJITSU LABORATORIES LTD. He was also a visiting scholar at Stanford University (2013-2014). He received the OHM Technology Award in 2017. His research interests include image processing, machine learning, and biometric authentication algorithm.*

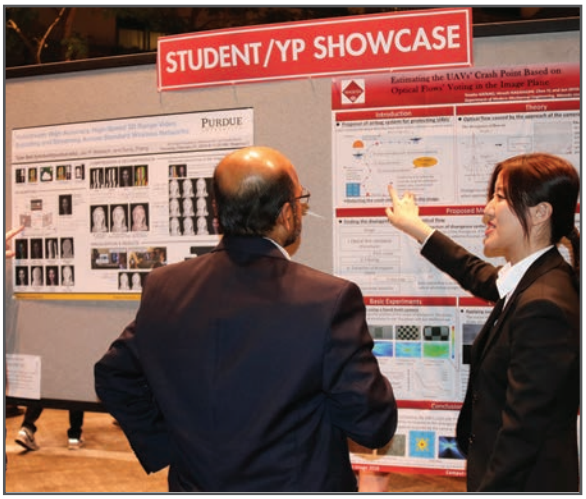
JOIN US AT THE NEXT EI!

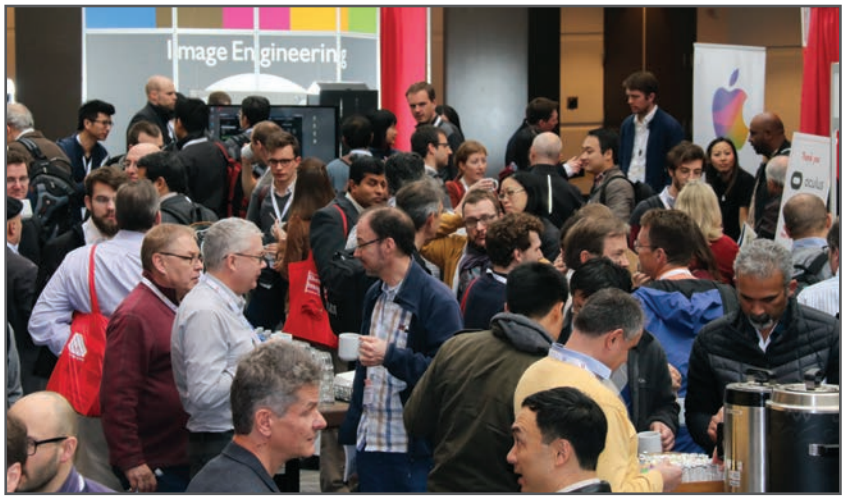
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org