Facial Landmark Detection in Uncontrolled Conditions

B. Efraty, C. Huang, S.K. Shah and I.A. Kakadiaris
Computational Biomedicine Lab
University of Houston, TX, USA

Abstract

Facial landmark detection is a fundamental step for many tasks in computer vision such as expression recognition and face alignment. In this paper, we focus on the detection of landmarks under realistic scenarios that include pose, illumination and expression challenges as well as blur and low-resolution input. In our approach, an n-point shape of point-landmarks is represented as a union of simpler polygonal sub-shapes. The core idea of our method is to find the sequence of deformation parameters simultaneously for all sub-shapes that transform each point-landmark into its target landmark location. To accomplish this task, we introduce an agglomerate of fern regressors. To optimize the convergence speed and accuracy we take advantage of search localization using component-landmark detectors, multi-scale analysis and learning of point cloud dynamics. Results from extensive experiments on facial images from several challenging publicly available databases demonstrate that our method (ACFeR) can reliably detect landmarks with accuracy comparable to commercial software and other state-of-the-art methods.

1. Introduction

Facial landmark detection is a critical step in computer vision that precedes several important tasks such as face and expression recognition, face alignment and face tracking. Despite the extensive research in this area in the past two decades, facial landmark detection is still considered to be a challenging problem. This is mainly because of the variations in the appearance of the face due to pose, illumination, and expression. These factors compromise the performance of most facial landmark detection methods.

Numerous methods have been proposed for landmark detection, most of which follow a learning-based approach using appearance or geometric constraints [18]. Among the state-of-the-art methods, approaches using Active Shape Model (ASM) and Active Appearance Model (AAM) have shown promising results under precise initializations in single- or multi-scale settings [4]. The general approach to landmark detection attempts to maximize the likelihood of landmark locations with respect to the observed image intensities while also preserving an anatomically plausible shape based on a statistical model such as the Point Distribution Model (PDM). If any two components of the objective function are in conflict, then, the final solution will converge to the trade-off point, which is often suboptimal. Dollar et al. [7] proposed a cascaded pose regression method to circumvent this problem by explicitly modeling both configuration and appearance. However, their approach suffers from sensitivity to initialization, slow convergence and is constrained to very simple shapes.

The key idea of our approach is to use a divide-and-conquer paradigm to solve a problem of complex pose regression by solving a set of simple sub-problems and combining their results. Any group of facial point-landmarks is partitioned into sets that have a small cardinality (3 points), which we refer to as sub-shapes. The main goal of the aggregated fern regression approach is to estimate the sequence of deformations for each sub-shape (initialized randomly within the search region). The robustness of the convergence is ensured by using multiple instances of sub-shapes, which creates a visual effect depicted in Fig. 1 – vertices of the sub-shapes (edges are not shown) are “attracted” to the landmark positions and, in the case of successful detection, eventually “collapse” within a small region (Fig. 1(c)). The mode of point cloud density is
computed to predict the position of the landmark and the variance determines the uncertainty of the detection (Fig. 1(d)). Our method uses very simple features derived from a grayscale image and operates in a multi-scale fashion starting from a coarse representation and gradually refining the solution. In addition, the shape and the density of point cloud formed by multiple instances of the sub-shapes is used to enhance the predictions and speedup the convergence.

The contributions of our work are: (i) a method for robust landmark detection that works accurately for a range of poses, illuminations, expressions and image defects; (ii) an algorithm for pose regression that is efficient for complex shapes; and (iii) a method that provides a reliable uncertainty measure for detected landmarks. Our method is evaluated on several challenging face databases and its performance was found to be comparable to other state-of-art methods and commercial software.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents performance evaluation and in Section 5 we summarize our findings.

2. Related Work

The existing methods for the detection of facial landmarks can be classified into two categories: generative approaches and discriminative approaches. The generative approaches attempt to fit a generative model of shape or texture to the input face image by optimizing over the multi-dimensional space of the model’s parameters. The discriminative approaches search for candidates for each of the predefined landmarks of the face using feature detection methods. They then combine results based on the topological configuration of the landmarks.

The most popular generative models are the AAM and the ASM proposed by Cootes et al. [4]. The AAM employs a statistical model for shape and texture parameters, which allows generation of new instances of facial images. The algorithm uses the texture residual between the target and estimated images to iteratively update the model’s parameters. The ASM employs only shape parameters and is guided by a local search around each point of the shape. Even though a variety of approaches have been proposed, the majority of prior work in facial landmark detection is based on variations of these two algorithms. Vetter et al. [2] developed a 3D-morphable model incorporating shape and texture which are fitted to the 2D image using optical flow. The Boosted Appearance Model proposed by Liu [16] uses a boosting-based classifier, which is used for optimization of the appearance parameters. Liang et al. [15] proposed a component-based extension to this approach, where predefined components are used to guide the shape deformation with classifiers attempting to predict direction. Lee and Kim [14] used multi-linear analysis to obtain a more robust fitting. Kanade and Gu [11] proposed a three layer generative model of landmark candidates, geometric transformation and shape constraints linked through Bayesian inference. Their framework used a large training set of 2D and 3D data. Cristinacce and Cootes [5] proposed a constrained local models method, where an AAM is used to generate a set of template regions, which are then used as feature detectors.

Concerning discriminative approaches, Rapp et al. [22] detected landmarks using Multiple Kernel Learning SVM for classification of multi-resolution patches. Valsar et al. [25] used a Support Vector Regression to guide the search for the landmarks. Their search is initialized using a face-region-based layout of the landmarks. Ding and Martinez [6] detected component-landmarks using discriminant analysis and contextual information, and then employed color information and geometrical properties to localize the points. Kozakaya et al. [13] used SIFT features to detect a special subset of points which are used to vote for the landmarks using a weighted vector concentration approach. The results of [6, 13] are accurate only for frontal face data and stable illumination. The method of Dollar et al. [7] uses cascaded regression to update all points of the simple shape simultaneously, rather than on a one-by-one basis.

3. Methodology

3.1. Overview of the Framework

The schematic layout of our framework is depicted in Fig. 2. It has four main components: (i) the shape model, which defines the partition of landmarks into simple polygonal sub-shapes; (ii) the preprocessing module, which carries out the standardization of the face region of interest (ROI) size and local intensity; (iii) the component-landmarks detector, which is used for localization of basic

Figure 2: Depiction of the proposed landmark detection framework.

Kim [14] used multi-linear analysis to obtain a more robust fitting. Kanade and Gu [11] proposed a three layer generative model of landmark candidates, geometric transformation and shape constraints linked through Bayesian inference. Their framework used a large training set of 2D and 3D data. Cristinacce and Cootes [5] proposed a constrained local models method, where an AAM is used to generate a set of template regions, which are then used as feature detectors.
component-landmarks (e.g., mouth); and (iv) the agglomerate of the cascaded fern-based regressors, which is used for precise localization of point-landmarks.

As a preprocessing step in our framework, the original image is transformed in such a way that the vertical extent of the face ROI is 300 pixels. The initial estimate of the face ROI position and scale is computed using the PittPatt face detector [21]. Additionally, we apply local illumination normalization [27] of the gray-scale image to mitigate the effects of uneven illumination. Upon the completion of this step the distribution of intensity values is roughly approximated to \( N(0, 1) \).

The agglomerate of regressors is the core component of the proposed method. While each regressor predicts a sequence of deformations for all instances of one sub-shape, their results are fused every few iterations using predicted positions for the landmarks shared by several sub-shapes.

The optimal way for initialization of multiple sub-shape instances is based on the output of the component-landmark detector. Our framework uses the algorithm proposed by Efraty et al. [9]. In that method, each component-landmark is detected independently using a cascaded classifier utilizing Bag-of-Words features. Despite the robustness of that method for various illuminations and poses, its efficiency is limited to images with moderate expressions and is unreliable if the interpupillary distance (IPD) is lower than 70 pixels. In Sec. 4, we expand our discussion regarding this module and show that its output can be ignored in order to avoid the aforementioned limitations, with only a minor penalty in performance.

### 3.2. Shape Model

Let \( \mathcal{S} \) be a set of landmarks that define the \( n \)-point shape. We propose to represent it as a union of \( T \) sub-shapes of \( h \) points, such that any landmark is a member of at least one such sub-shape. Each sub-shape is associated with a polygon in the image coordinate space and is uniquely defined by the coordinates of its vertices (Fig. 3(a)). The alternative method for the parametrization of such a polygon is chain-parametrization that encodes the lengths of line segments and angles between segments, rather than coordinates of vertices. For example, the triangle \( S \) in Fig. 3(a) may be parameterized as vector \( \hat{C}(S) = [x_1, y_1, x_2, y_2, x_3, y_3]^T \) or \( \hat{C}(S) = [x_1, y_1, \theta_1, \theta_1, \theta_2, \theta_2]^T \). For two sub-shapes \( S \) and \( S' \) the vector difference \( \Delta \hat{S} = \hat{C}(S') - \hat{C}(S) \) defines a deformation vector of \( S \) into \( S' \). The dissimilarity between two sub-shapes (or a loss function) is given by the \( L_2 \) norm:

\[
E(S, S') = \|C(S') - C(S)\|_2.
\] (1)

In Sec. 3.3 we will argue that the landmark detection task may be casted as finding the optimal deformation vectors between any instance of the sub-shape and its ground-truth position such that the dissimilarity measure is minimized.

In our framework, we consider an 8-point landmark model, which is partitioned into triangular sub-shapes \( (h = 3) \). Even though it is possible to use only one static partition, as will be presented in Sec. 4, the best results are obtained if we randomly generate a partition every few iterations. Out of all the \( \binom{6}{3} = 56 \) possible sub-shapes we use only six at a time (\( T = 6 \)) such that: (i) each landmark is a member of at least two triangles, and (ii) the most acute angle of the triangle is larger than 10° (to avoid nearly-degenerate triangles), thereby allowing the support set of every triangle to be balanced. The support set is a set of points within a certain distance from the sub-shape. We define two types of support sets for a single edge: rectangular and radial. The support set of the sub-shape is the union of support sets of its edges (Fig. 3(b)).

Only the image intensity values in the support set are used to predict the deformation parameters. Therefore, to simplify an access to their values, points within the support set are mapped into the shape coordinate system using Bookstein coordinates [8]. The Bookstein coordinate \( B(a|A) \) of the point \( a \) with respect to edge \( A \) with end points \( a_0, a_1 \) is obtained with the unique similarity transformation that maps \( a_0 \) to \((-0.5, 0)\) and \( a_1 \) to \((0.5, 0)\). Note that the rectangular and radial support sets correspond to Bookstein coordinates that satisfy \( |\tilde{x}| \leq c_1, |\tilde{y}| \leq c_2 \) and \((\tilde{x} \pm 0.5)^2 + \tilde{y}^2 \leq c_3^2\), respectively.

### 3.3. Cascaded Fern Regression

A fern classifier [19] is a special type of a decision tree classifier, which gained popularity especially because of its high computational efficiency. A fern classifier is adapted for the regression task in the same way as a decision tree classifier [3]. Dollar et al. [7] demonstrated that fern regression may be very efficient for pose regression based on sim-
Our fern regressor for sub-shape deformation prediction is an extension of the algorithm of Dollar et al. [7] and differs from it in only a few aspects, which are: (i) geometry features are used in addition to indexed features; (ii) the cardinality of the fern is updated with the propagation of the regression iterative process; and (iii) several scales of the images are used rather than using only the original scale of images.

Let the set of sub-shapes used in the training at the iteration \( n \) be \( \{ S_{m_p}^n \} \), \( n = 1, ..., N; \ t = 1, ..., T; \ p = 1, ..., P \) where \( N \) is the number of images, \( T \) is the number of sub-shapes in the shape model and \( P \) is the number of random instances per image. Each \( P \) instances \( S_{m_p}^n \) of the sub-shape are associated with a ground-truth sub-shape \( S_{m}^n \) (see Fig. 3(a)). The optimization problem for \( T \) fern regressors \( R_t^v \) is stated as follows:

\[
R_t^v = \arg\min_R \sum_{p=1}^{P} \sum_{n=1}^{N} E\left(S_{m_p}^{t+1}, S_{m}^{n}\right) \quad (2)
\]

\[
\text{s.t. :} \quad \hat{C}(S_{m_p}^{t+1}) = \hat{C}(S_{m_p}^{n}) + \hat{C}(\Delta S_{m_p}^{n}) \quad (3)
\]

\[
\Delta S_{m_p}^{n} = R\left(\{v_i (S_{m_p}^{n}), I_p\}, \{\xi_j (S_{m_p}^{n})\}\right) \quad (4)
\]

Each regressor minimizes the loss function (2) via deformations (3) and is defined as a function of features \( \{v_i (S_{m_p}^{n}), I_p\}, \{\xi_j (S_{m_p}^{n})\}\) (4), which are discussed next.

The indexed features are computed based on \( 2K \) randomly selected points within the support set, which are used to form \( K \) pairs. Among these points, \( \eta 2K \) are selected from radial support sets and the rest \( (1 - \eta)2K \) from the rectangular support set (\( \eta \in [0, 1] \)). Let \( \mathcal{P} = \{ (p_1,q_1), ..., (p_K,q_K) \} \) be the set of such pairs selected for the sub-shape \( S \). The indexed feature \( \hat{v}_i(S) \) is simply the difference of intensities in two points \( \hat{v}_i(S) = I(p_i) - I(q_i) \). The geometry features do not depend on the intensity values or on the underlying polygonal structures of the sub-shapes. The distances \( r_i \) between each pair of landmarks \( (i = 1, ..., 28) \) are used to compute 325 possible face proportion measurements \( r_i/r_j \). Then the function

\[
\hat{\xi}_{ij} = 3 \log_3 \left(\frac{r_i/r_j}{\overline{r}_i/\overline{r}_j}\right)
\]

measures the degree of change of certain proportion when compared to the same proportion measured within the mean shape. The log scaling maps ratios to symmetric range (3.1 and 0.3 are mapped to +3, 0 and -3, respectively). To convert our features to a binary format (inside the fern regression), a random threshold \( \tau \) in \([-3, +3]\) is used:

\[
v_i = H(\hat{v}_i - \tau_i), \quad \hat{\xi}_{ij} = H(\hat{\xi}_{ij} - \tau_{ij})
\]

where \( H(\cdot) \) is the Heaviside step function. Both sets of features are invariant to the similarity transform and are computationally efficient. In addition, there is no need to recompute features for every iteration. The fern regressor will use only a small fraction of features per iteration.

We take advantage of the iterative framework, which naturally progresses starting from the coarse scale optimization to the refinement stage using four different scale-space representations of the original image. According to the same principle, the parameter \( \eta \) is gradually incremented from 0.1 to 0.9 such that during the last iterations of the algorithm most of the features are extracted from the radial support regions around the landmarks. The cardinality of fern is also gradually incremented from 2 to 6, which is equivalent to increasing the number of possible deformations from 4 to 64, to accommodate fine-scale tuning.

### 3.4. Agglomerate of cascaded regressors

The agglomeration of results from multiple instances of sub-shapes are accomplished on two levels: (i) merge landmarks shared by different sub-shapes per instance, (ii) use the shape and density of point clouds from multiple instances to enhance predictions. The agglomeration operator is applied with different frequencies \( \Delta v_1, \Delta v_2 \) at the first and second levels, respectively.

Let us focus on a single landmark \( A \) (i.e., left corner of mouth) shared by \( \mathcal{F} \subset \{ 1, ..., T \} \) sub-shapes in the model. The final position of landmark \( A \) related to the instance \( p \) at iteration \( v \) in image \( n \), is obtained as the weighted average of its coordinates from sub-shapes \( S_{m_p}^n; \ t \in \mathcal{F} \). Let \( C_A(S_{m_p}^n) \) denote the coordinate of \( A \) within the sub-shape. Then, the combined coordinates of \( A \) are

\[
C_A(S_{m_p}^n) = \sum_{t \in \mathcal{F}} w_{tp}^v C_A(S_{m_p}^n).
\]

The weights are proportional to the loss function (1) reduced to one landmark

\[
E_A(S_{m_p}^n; S_a^*) = \| C_A(S_{m_p}^n) - C_A(S_a^*) \|_2,
\]

and are computed as:

\[
w_{tp}^v = \frac{\sum_{n=1}^{N} \left( E_A(S_{m_p}^n; S_{m_p}^n) - E_A(S_{m_p}^n; S_a^*) \right)}{\sum_{n=1}^{N} \left( E_A(S_{m_p}^n; S_{m_p}^n) - E_A(S_{m_p}^n; S_a^*) \right)}.
\]
\[ \Delta \mathbf{q}^* \text{ with respect to the ground truth. If we attempt to improve the prediction, but preserve the direction of } \Delta \mathbf{q}, \text{ then the new prediction will be} \]

\[ \Delta \hat{\mathbf{q}} = \Delta \mathbf{q} + \gamma \frac{\Delta \mathbf{q}}{\| \Delta \mathbf{q} \|}, \]

(7)

and

\[ \gamma^* = \arg \min_{\gamma} \| \Delta \hat{\mathbf{q}} - \Delta \mathbf{q}^* \| = \frac{\langle \Delta \mathbf{q}, \Delta \mathbf{q}^* \rangle}{\| \Delta \mathbf{q} \|} - \| \Delta \mathbf{q} \|. \]

(8)

is the optimal \( \gamma \) in Eq. (7). Let \( \phi_1, \phi_2 \) be the principal axes (eigenvectors) of \( \mathcal{A}_k \), and \( \rho_\mathbf{q} \) to be the relative density of the cloud in point \( \mathbf{q} \). We propose to model the increment \( \gamma^* \) as a function of density and the relative position of the point in the cloud using a linear model:

\[ \gamma^* = \beta_0 + \beta_1 \phi_1 \mathbf{q}^T + \beta_2 \phi_2 \mathbf{q}^T + \beta_3 \rho_\mathbf{q}. \]

(9)

The values of \( \beta_i \) are only occasionally determined during the training (i.e., \( \Delta \mathbf{v}_2 = 100 \)), to avoid over-fitting.

### 3.5. Estimating the Detection Uncertainty

One of the advantages of the proposed algorithm is that it provides a natural way to measure the uncertainty of detection per landmark based on the spatial distribution of the point cloud. Let \( \hat{f}_b(x,y) \) be a kernel density estimator for the density of points in the cloud with normal kernel and bandwidth \( h \) and let \( \mathbf{q}_1 = [x_1,y_1]^T \) and \( \mathbf{q}_2 = [x_2,y_2]^T \) be the first two local maxima points of \( \hat{f}_b(x,y) \). The minimum certainty bandwidth \( h^* \) is defined to be:

\[ h^* = \arg \min_{h > \epsilon_0} \frac{\hat{f}_b(x_1,y_1)}{\hat{f}_b(x_2,y_2)} \geq 1.25, \]

(10)

where \( \epsilon_0 \) is the minimum bandwidth of one pixel. Let \( Q(\mathbf{q}_1) \) denote the proportion of points within distance \( 2h^* \) from \( (x_1,y_1) \). Point \( \mathbf{q}_1 \) is an estimate of the landmark and its uncertainty \( U(\mathbf{q}_1) \) depends both on \( h^* \) and \( Q(\mathbf{q}_1) \):

\[ U(\mathbf{q}_1) \sim h^* \| \log (Q(\mathbf{q}_1)) + 1 \|. \]

(11)

Equations (10) and (11) provide the distribution mode along with the measurement of the distribution density at \( \mathbf{q}_1 \). The uncertainty measure may be used to select the most favorable solution if the algorithm is initialized in multiple locations (i.e., with and without component landmarks).

### 4. Experimental Results

#### 4.1. Datasets

For our experiments, we used subsets of five commonly available facial image databases: (i) CMU Multi-PIE [10], (ii) Bio-ID [1], (iii) UHDB11 [24], (iv) GBU [20] (we use only the “ugly” subset), and (v) LFW [12]. Samples of images from each database are depicted in Fig. 5. The partition of these databases to cohorts, their descriptions and assignments to either training or testing is presented in Table 1. For all following experiments, training and testing sets are disjoint w.r.t. subjects’ IDs.

| Cohort | Source database | Number of images | Number of subjects | Variation modes |
|--------|-----------------|------------------|-------------------|----------------|----------------|
| \( \mathcal{C}_1 \) | Multi-PIE | 1,170 | 125 | p.i |
| \( \mathcal{C}_2 \) | Multi-PIE | 400 | 121 | p.i |
| \( \mathcal{C}_3 \) | BioID | 104 | 10 | p.i.e |
| \( \mathcal{C}_4 \) | UHDB11 | 400 | 23 | p.i |
| \( \mathcal{C}_5 \) | LFW | 1,890 | 962 | p.i.e.o,b |
| \( \mathcal{C}_6 \) | GBU | 400 | 245 | i.e.o.b |
| \( \mathcal{C}_7 \) | combined | 500 | 403 | p.i.e.b |

Table 1: Description of the cohorts. The abbreviations in rightmost column stand for pose (p), illumination (i), expression (e), occlusions (o), blur (b). Cohort \( \mathcal{C}_1 = \mathcal{C}_{1a} \cup \mathcal{C}_{1b} \cup \mathcal{C}_{1c} \), whereas \( \mathcal{C}_7 \) is a random subset of \( \mathcal{C}_1 \).

#### 4.2. Experimental Settings

In all our experiments, we used the same internal parameters for our algorithm. The number of the sub-shape instances \( P \) is set to be 55 for training and 45 for detection. The support set sizes (see Sec. 3.2) have been set to \( c_1 = 0.7, c_2 = 0.6, c_3 = 0.25 \). The number of initial indexed features, \( K \), was set to 65 (see Sec. 3.3). These parameters were selected experimentally, while considering the trade-off between accuracy and computational cost.

For efficient training, we use a bootstrapping procedure that allows us to optimize the performance of a detector with respect to a small data set one at a time and avoid the overfitting effect by replacing the training set every 20 iterations. Specifically: (i) the training set \( \mathcal{C}_i \) of 300 images is withdrawn randomly from the training set \( \mathcal{C}_i \) at the iteration \( v \); (ii) the detection procedure of \( v - 1 \) iterations is applied to every image; (iii) the final configuration of point clouds is used as the initial configuration for iterations \( v, \ldots, v + 20 \) with \( \mathcal{C}_i \). Steps (i)–(iii) are repeated until a certain criteria is met (i.e., maximum number of iterations or minimum loss function).

#### 4.3. Experiments

In the first set of experiments, we explore the parameter space for our method (which we name ACFeR) and measure its sensitivity for various settings. In the second set of experiments, we present a comparison of our
method with other methods and a commercial software. The errors between manual landmarks and the detected landmarks are normalized using the interpupillary distance (IPD). The normalized root mean square error (NRMSE), is used as the quality measure per image. Probability $\rho = P(NRMSE \leq 0.1)$ is used as a quality measure (detection rate) per training set. The training set for most of the experiments (if not stated differently) is the union of $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$. The testing set for most of the experiments (if not stated differently) is $\mathcal{C}_3$ (see Table 1).

Analysis of ACFeR’s sensitivity for parameter settings: The parameter $T$ in ACFeR (see Sec. 3.2) controls the complexity of the shape model, determined by the number of triangles. The cumulative distribution of NRMSE with varying $T$ is depicted in Fig. 4(a). Note that, even though there is an obvious correlation between the complexity of the model and the detection accuracy, the running time for both training and testing also increases proportionally to the complexity. The next experiment (see Fig. 4(a)) compares the static partition of the sub-shapes with the dynamic partition that is updated every 20 steps. Note that there is no clear domination of one curve over another. However, the dynamic partition results in higher accuracy, while the static partition improves robustness. All the results that are depicted in Fig. 4(a) are obtained with small bootstrapping size of 100 images and 400 iterations.

The results on Fig. 4(b) demonstrate the performance of our algorithm if halted at different iterations. In general, the improvement for running more than 600 iterations is small (detection rate improves from 97.0% to 97.7%). In the next experiment, we use ACFeR with 600 iterations (as a trade-off between time and robustness) to measure the performance on different datasets, and we use ACFeR with 800 iterations to compare it to other methods on a single dataset.

Validation of ACFeR’s performance: The model of agglomerated fern regressors is trained on images from $\mathcal{C}_1$ with bootstrapping and training set $\mathcal{U}_v$ of 300 images (see Sec. 4.2). The detection is applied on images from five different face databases and the algorithm is terminated after 600 iterations. The cumulative distributions of the NRMSE are depicted in Fig. 6(a). Within the legend area of the plot we specify the detection rate $\rho$. Figure 5 depicts sample images with the detected landmarks. We found the UHDB11 database to be the most challenging for ACFeR. It contains data with illumination and pose variation (three degrees of freedom, and not only yaw rotations). As a result, the face detector is not accurate, which causes low detection rate (only 87.0%) of ACFeR.

Figure 6(b) depicts the results of ACFeR with default (“blind”) initialization along with the results of the component-landmark based initialization. Also, we combine both results using the uncertainty measure described in Sec. 3.5 by selecting only the detection with the lowest mean uncertainty per image. As mentioned in Sec. 3.1,
the component-landmark detector is efficient only for facial images with certain minimal resolution (IPD above 70 pixels), without blur effects and moderate expressions. The component-landmark based initialization results in only a slight performance increase in ACFeR (detection rate on $C_2$ increases by 0.2%, detection rate on $C_3$ increases by 0.7%). Note that the default initialization is preferable for $C_3$, where we observe about 7% drop in the detection rate when initialized with component-landmarks. Such a behavior is expected, since some of the facial images in $C_3$ have expressions or are of low resolution. Nonetheless, the results for the combination of the two methods is always better than the results for the single method (see dash-dot lines in Fig. 6(b)).

The last experiment, whose results are summarized in Table 2, performs a comparison of ACFeR (with default settings, applied for 800 iterations) to other landmark-detection methods (including commercial software) on subset of BioID database ($C_3$). The performance of ACFeR is better or comparable to other methods.

In general, the most important quality of ACFeR is robustness (measured by $\rho$). However, the accuracy (measured by $\mu_i$) of some landmarks is better for other methods (BoRMaN and STASM), which is explained by the simplicity of features used in ACFeR. Note that the results for BoRMaN and MKL are adopted from the corresponding publications [25, 22], and therefore, the comparison is not direct. Additionally, ACFeR, STASM and ASM report landmarks 5 and 6 with slight offset from BioID markers (edge of nostril wing instead of nostril) which explains the large $\mu_5$ and $\mu_6$ in Table 2.

<table>
<thead>
<tr>
<th>Landmark</th>
<th>ACFeR</th>
<th>STASM</th>
<th>ASM</th>
<th>BoRMaN</th>
<th>S.M.</th>
<th>MKL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.98</td>
<td>0.96</td>
<td>0.89</td>
<td>0.95</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>5.7%</td>
<td>4.2%</td>
<td>8.0%</td>
<td>4.4%</td>
<td>6.6%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>5.0%</td>
<td>3.6%</td>
<td>7.9%</td>
<td>5.4%</td>
<td>6.2%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>5.7%</td>
<td>4.7%</td>
<td>8.1%</td>
<td>4.3%</td>
<td>6.9%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_4$</td>
<td>5.3%</td>
<td>3.4%</td>
<td>7.7%</td>
<td>5.3%</td>
<td>6.9%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_5$</td>
<td>9.2%</td>
<td>9.5%</td>
<td>14.1%</td>
<td>3.7%</td>
<td>5.5%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_6$</td>
<td>8.8%</td>
<td>9.8%</td>
<td>14.7%</td>
<td>4.0%</td>
<td>5.3%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_7$</td>
<td>4.6%</td>
<td>6.4%</td>
<td>9.7%</td>
<td>4.4%</td>
<td>6.6%</td>
<td>--</td>
</tr>
<tr>
<td>$\mu_8$</td>
<td>4.7%</td>
<td>6.7%</td>
<td>9.7%</td>
<td>4.9%</td>
<td>7.0%</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 2: Validation of five different landmark detection algorithms on $C_3$. The first row denotes the detection rate ($\rho$), and the remaining rows denote mean normalized error ($\mu_i$) of the landmark $i$. The indexing of landmarks is depicted in Fig. 2.

Finally, the running time complexity for a single iteration

5. Conclusion

In this paper, we described a novel method for point-landmark detection using an agglomerate of fern-regressors. The proposed framework is applicable to facial images obtained under a variety of conditions. We present two operating modes for our method: (i) a component-landmark based initialization and (ii) a “blind” initialization. While
good initialization is beneficial, it is not always feasible for degraded images. In such cases a blind initialization is preferred. The method is readily applicable to any number of landmarks.

Acknowledgements: This research was funded in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL) and by the University of Houston (UH) Eckhard Pfeiffer Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, the U.S. Government, or UH.

References