Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model

Xiang Yu* Junzhou Huang† Shaoting Zhang§ Wang Yan* Dimitris N. Metaxas*
*Rutgers University †Univ. of Texas at Arlington §Univ. of North Carolina at Charlotte
xiangyu, wy109, dnm@cs.rutgers.edu jzhuang@uta.edu szhang16@uncc.edu *

Abstract

This paper addresses the problem of facial landmark localization and tracking from a single camera. We present a two-stage cascaded deformable shape model to effectively and efficiently localize facial landmarks with large head pose variations. For face detection, we propose a group sparse learning method to automatically select the most salient facial landmarks. By introducing 3D face shape model, we use procrustes analysis to achieve pose-free facial landmark initialization. For deformation, the first step uses mean-shift local search with constrained local model to rapidly approach the global optimum. The second step uses component-wise active contours to discriminatively refine the subtle shape variation. Our framework can simultaneously handle face detection, pose-free landmark localization and tracking in real time. Extensive experiments are conducted on both laboratory environmental face databases and face-in-the-wild databases. All results demonstrate that our approach has certain advantages over state-of-the-art methods in handling pose variations[1].

1. Introduction

Facial landmark localization and tracking have been studied for many years in computer vision. Landmark localization addresses the problem of matching a group of predefined 2D landmarks to a given facial image. Landmark tracking is to continuously capture the predefined landmarks in a facial image sequence. Such tasks are prerequisite for many applications, such as face recognition, facial expression analysis, 3D face modeling, video editing, etc. All the applications require accurate landmark positions. However, due to complicated background, lighting conditions and particularly pose variations, accurate landmark localization remains challenging in practice.

Initialization is the first and key step in landmark localization and tracking. Many alignment algorithms heavily rely on the initialization. Some of them are gradient descent based methods and may encounter the local optimum. For example, Active Appearance Model (AAM) [5] is very sensitive to initial positions because complex appearance with illumination and noise may result in local minima. In addition, even before initialization, most alignment algorithms [10, 11, 18, 26, 31] require to locate face region from face detectors [19]. Though they are widely used in many facial applications, they may lack flexibility of handling large facial pose variations.

Although multi-view face shape models [6, 33, 34] partially solve the pose variation problem, they cannot cover unlimited possibilities of view changes. Therefore, 3D shape model [15, 29] is proposed to handle continuous view change. There are two possible ways to explicitly project 3D shape onto 2D images. One way is to use facial anchor points, *e.g.* eye corners and mouth corners, mapping from 3D shape; The other is to leverage the view information from head pose estimators. Since most pose estimators [4, 24] are based on face detectors, which makes the problem recursive, a better choice is to train fast and accurate facial anchor point detectors.

For aligning landmarks, traditional parametric methods, *e.g.* Active Shape Model (ASM) [7, 9], AAM [5, 23], have achieved success for their wide applicability. However, they are sensitive to initial shapes. Holistic algorithms can not handle subtle shape variations [7]. By exhaustive local search, Constrained Local Model (CLM) [8, 26, 30] is expected to pull the landmarks out of local minima. Assuming so, auxiliary local discriminative search may further approach the global optimum.

In this paper, we propose a unified framework to handle all the above-mentioned problems. A group sparse learning method is proposed to automatically select the optimized anchor points. Then a two-level cascaded deformable shape model is presented to search global optimal positions. Starting from Zhu and Ramanan's work [35], we set up weights for each landmark patch in the part mixture model indicat-

*This work was partially supported by grants from National Science Foundation (IIS-1064965, IIS-1065013, CNS-1059281, CNS-1059218, IIS-0964597, IIS-0964385) and National Space Biomedical Research Institute through NASA NCC 9-58.

[1]The code is available at http://www.research.rutgers.edu/~xiangyu/face_align.html

ing the likelihood of choosing these parts. By regularizing the weights to be group sparse, maximizing the margin over positive and negative training samples generates effective weights to simplify the mixtures of parts. With initialized landmarks, we firstly perform mean-shift search on pre-trained response map for each landmark with CLM, pulling the landmarks into the convergence basin globally. Then component-wise active contour model is used to refine each component of face, *e.g.* eyebrows, eyes, etc. Exhaustive local search inside the convergence basin with global optimum is expected to approach the optimal solution.

Our framework primarily leads to the following **contributions**. **1)** The proposed optimized mixtures and two-step cascaded deformable shape model achieve real-time performance in facial landmark tracking. **2)** The proposed two-step cascaded deformable shape model enhances the flexibility to capture subtle shape variations from classical parametric shape models by integrating component-wise active contours. **3)** Extensive experiments have been conducted to demonstrate that our pose-free landmark fitting framework consistently achieves more significant results comparing to state-of-the-art methods on not only laboratory environmental face databases but also face-in-the-wild databases.

2. Related Work

For face detection, Viola and Jones [19] proposed a widely used framework. It is fast and effective for most near-frontal faces, but lacks flexibility dealing with large pose variations. Sivic et al. [11] used mixture of tree structure to estimate landmarks. Uricar et al. [27] proposed a seven-anchor point detector based on deformable part models (DPM) [12] and structure-output SVMs which achieve fast speed and high accuracy. However, when the detection error occurs, seven points are not sufficient to provide steady initial landmarks. Zhu and Ramanan [35] proposed another framework based on mixture of part model. However, the size of parts pool in their model is large, which impedes the potential for real-time landmark tracking.

Parametric models have been widely used in face alignment. Active Shape Model (ASM) [7, 9] and Active Appearance Model (AAM) [5, 23] have achieved good performance in face alignment. But it is difficult to represent face shapes merely using linear shape combination or appearance subspace in extremely varying views . Constrained Local Model (CLM) [8, 26, 30], another successful deformable fitting model, performs exhaustive local search and optimizes the overall likelihood of the landmarks' alignment. To alleviate the varying view problem, multi-view shape models [6, 34] were proposed either by local search to estimate the head pose or by incrementally combining models from different views.

Nonparametric shape regression is another way for shape registration. Cristinacce and Cootes [9] introduced boosted regression for individual landmarks. Valstar et al. [28] combined the boosted regressor with graph model. Liang et al. [20] trained directional classifiers to discriminatively search facial components. Pose or shape fern regressors [3, 10] was proposed to handle different shape variations. Rivera and Martinez [25] proposed kernel regression to deal with low-resolution images. Xiong and De la Torre [31] introduced supervised descent method in approximating the regression matrix mapping from features to locations. But those methods either lack flexibility in representing pose-variate cases or require large amount of training faces. In contrast, our framework simultaneously tackles face detection and landmark initialization using proposed optimized anchor point detectors. The framework deals with arbitrary head pose conditions by introducing 3D shape model. It achieves real time performance due to the group sparse selection and cascaded two-stage deformation strategy.

3. Robust Initialization via Optimized Part Mixtures

Before shape alignment or landmark tracking, robust initialization promotes the performance and prevents the fitting process from falling into local minima. We introduce a pictorial structure [13] to organize the landmarks. In order to achieve real-time performance for tracking, we propose a group sparse learning based method to automatically select the landmarks and reorganize them into a new tree structure part mixture, which dramatically decreases the number of landmarks and still preserves the detection effectiveness. A max-margin method is used to learn the weights for the landmark detector.

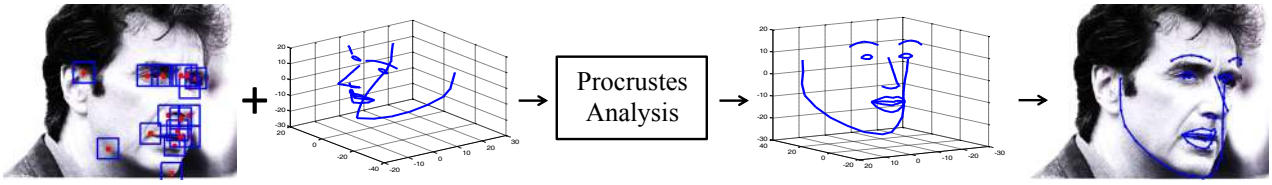
3.1. Mixtures of Part Model

Every facial landmark with predefined patch neighborhood is a part. Same landmark in different viewpoints may be different parts. As a consequence, the landmarks of a face are a mixture of those parts. We define the shared pool of parts as V. The connection between two parts forms an edge in E. In connecting the landmarks, specific tree structures are superior to general complete graphical models for not only the simplicity of representation but also the efficiency in inference [13, 35].

For each viewpoint i, we define a tree $T_i = (V_i, E_i), i \in \{1, 2, ..., M\}$. Given a facial image $I^{H \times W}$, the j^{th} landmark position $s_j = (x_j, y_j) \in \mathcal{S}_j \subset \{1, ..., H\} \times \{1, ..., W\}, j \in \{1, 2, ..., N\}$. The measuring of a landmark configuration $\mathbf{s} = (s_1, ..., s_N)$ is defined by a scoring function $f : I \times \mathcal{S} \rightarrow \mathbb{R}, \mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_N\}$.

$$f_i(I, \mathbf{s}) = \sum_{j \in V_i} q_i(I, s_j) + \sum_{(j,k) \in E_i} g_i(s_j, s_k) \qquad (1)$$

The first term in Equation 1 is a local patch appearance evaluation function $q_i : I \times \mathcal{S}_i \rightarrow \mathbb{R}, i \in (1, N)$, defined as



Optimized Part Mixture | 3D Reference Shape | 3D Translated Shape | Initialized Landmarks

Figure 1. Pose-free facial landmark initialization using Procrustes analysis on 3D reference shape and detected optimized part mixture.

$q_i(I, s_j) = \left\langle \mathbf{w}_j^{iq}, \Phi_j^{iq}(I, s_j) \right\rangle$, indicating how likely a landmark is in an aligned position. The second term is the shape deformation cost $g_i : \mathcal{S}_j \times \mathcal{S}_k \to \mathbb{R}, (j,k) \in E$, defined as $g_i(s_j, s_k) = \left\langle \mathbf{w}_{jk}^{ig}, \Phi_{jk}^{ig}(s_j, s_k) \right\rangle$, balancing the relative positions of neighboring landmarks. $\mathbf{w}_j^{iq}$ is the weight vector convolving the feature descriptor of patch j, $\Phi_j^{iq}(I, s_j)$. $\mathbf{w}_{jk}^{ig}$ are the weights controlling the shape displacement function defined as $\Phi_{jk}^{ig}(s_j, s_k) = (dx, dy, dx^2, dy^2)$, where $(dx, dy) = s_k - s_j$. Such quadratic deformation cost controls the model with only four parameters and has shown its effectiveness in face alignment [35]. Further, we formulate the two evaluation functions in a uniform way to obtain a more compact representation $f_i(I, \mathbf{s}) = \left\langle \tilde{\mathbf{w}}_i, \tilde{\Phi}_i \right\rangle$, where $\tilde{\mathbf{w}}_i = [\mathbf{w}_j^{iq}, \mathbf{w}_{jk}^{ig}]$ and $\tilde{\Phi}_i = [\Phi_j^{iq}(I, s_j), \Phi_{jk}^{ig}(s_j, s_k)]$ for each viewpoint i.

Given an image I, for each possible configuration of landmark positions, we evaluate the score of each configuration in each viewpoint. The largest score potentially provides the most likely localization of the landmarks. Thus the landmark positions can be obtained by maximizing Equation 2.

$$\mathbf{s}^* = \underset{\mathbf{s} \in \mathcal{S}, i \in (1,M)}{\arg\max} \; f_i(I, \mathbf{s}) \tag{2}$$

3.2. Group Sparse Learning for Landmark Selection

Facial landmarks are usually defined manually without any consistent rules. Evidence is that the annotation among different face datasets is largely different. However, we observe that there are some common points defined by those different datasets, such as eye corners, eyebrow corners, mouth corners, upper lip and lower lip points, etc. We intend to automatically select those landmarks which well represent facial structure while the number of landmarks meets real-time requirement for inference.

The chosen landmarks are sparse because only several key positions are needed to depict the component. Considering the location and rough shape of a mouth, two corner points, upper and lower lip points are sufficient to locate the mouth. However, within each landmark patch, every pixel contributes to the feature descriptor. We should not expect pixels for appearance are sparse. Consequently, such property is well characterized by group sparsity [17, 21].

3.3. Max-Margin Learning for Landmark Parameters

In our learning process, we collect positive samples from MultiPIE database [14], which contains annotations and viewpoint information, denoted as $\mathcal{C}_+$. Negative samples are collected from arbitrary natural scenes but without faces, denoted as $\mathcal{C}_-$. The overall training set is $\mathcal{C} = \mathcal{C}_+ \cup \mathcal{C}_-$. For each viewpoint i, we need to train the weights $\tilde{\mathbf{w}}_i$. Hence we denote $\tilde{\mathbf{w}}_i$ as $\tilde{\mathbf{w}}$ in the following notations. Based on Equation 1, considering the group sparse constraint from section 3.2, we establish a max-margin framework in Equation 3.

$$\begin{aligned} \underset{\tilde{\mathbf{w}}, \varepsilon \geq 0}{\arg\min} & \left(\sum_{n \in \mathcal{C}} \varepsilon_n + \lambda_1 \left\| \tilde{\mathbf{w}} \right\|_2^2 + \lambda_2 \sum_{t=1}^{m} \left\| \tilde{\mathbf{w}}_t \right\|_2 \right) \\ & s.t. \forall n \in \mathcal{C}_+, \left\langle \tilde{\mathbf{w}}, \tilde{\Phi}(I_n, \mathbf{s}_n) \right\rangle \geq 1 - \varepsilon_n \\ & \forall n \in \mathcal{C}_-, \forall \mathbf{s}, \left\langle \tilde{\mathbf{w}}, \tilde{\Phi}(I_n, \mathbf{s}) \right\rangle \leq -1 + \varepsilon_n \end{aligned} \tag{3}$$

where $\tilde{\mathbf{w}} = [\tilde{\mathbf{w}}_1, ..., \tilde{\mathbf{w}}_t]$, $\tilde{\mathbf{w}}_t$ is a portion of the reorganized form of $\tilde{\mathbf{w}}$, each of which stands for the regularized weights within one landmark patch. $\tilde{\Phi}$ for negative samples are extracted with arbitrary configurations. To solve the problem, a group sparse optimization method is used. Please refer to [21] for further details of algorithms.

4. Two-step Cascaded Deformable Model

With initial anchor points detection, we use general Procrustes analysis to project our 3D shape model onto the facial image. As head is a near-rigid object in 3D space, the 3D to 2D mapping is unique. The process is illustrated in Figure 1. In this section, we firstly formulate the problem into parametric forms. Assuming the aligning of neighborhood landmarks conditionally independent, we apply Bayesian inference to build a probabilistic model. Further assuming the response map of each landmark patch mixture of Gaussian, we propose a two-step cascaded deformable shape model to refine the locations of landmarks.

4.1. Problem Formulation

In section 3.1, we have defined the landmarks as vector $\mathbf{s} = [s_1, ..., s_N]$, each landmark s_j is formed by concatenating the x and y coordinates. Let I denote the image

potentially containing faces. The task is to infer $\mathbf{s}$ from I. Proposed by Coots et al. [7], ASM represents face shapes by a mean shape and a linear combination of k selected shape basis, $\mathbf{s} = \bar{\mathbf{s}} + \mathcal{Q}u$, where $\bar{\mathbf{s}}$ is the mean shape vector, $\mathcal{Q} = [Q_1, ..., Q_k]$ contains the k shape basis, $u \in \mathbb{R}^k$ is the coefficient vector.

The general Point Distribution Model (PDM) takes global transformation into consideration. For rigid transformation in 3D space, scaling, rotation and translation are the only 3 deterministic factors. Considering local deformation, the ASM shape basis is able to depict it as long as the training set contains enough variate shapes and the number of basis k is large enough. Hence we establish the relationship between any two points in 3D space in Equation 4.

$$s_j = aR(\bar{s}_j + \mathcal{Q}u_j) + T \tag{4}$$

s_j is one of the defined landmarks, R is a rotation matrix, a is a scaling factor and T is the shift vector. The PDM provides us a way to depict arbitrary shape from a mean shape by deforming the parameter $\mathcal{P} = \{a, R, u, T\}$. The problem is to find such parameter $\mathcal{P}$ to map the 3D reference shape to a fitted shape which best depicts the faces in an image.

4.2. The Two-step Cascaded Model

We introduce a random variable vector $v = [v_1, ..., v_N]$ to indicate the likelihood of alignment, $v = 1$ means landmarks are well aligned and $v = 0$ means not. In this way, maximizing $p(\mathbf{s}|v = 1, I)$ demonstrates the aim that we are pursuing.

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} p(\mathbf{s}|\{v_i = 1\}_1^N, I) \tag{5}$$

$$\propto \arg\max_{\mathbf{s}} p(\mathbf{s}) p(\{v_i = 1\}_{i=1}^n|\mathbf{s}, I) \tag{6}$$

$$= \arg\max_{\mathcal{P}} p(\mathcal{P}) \prod_{i=1}^{n} p(v_i = 1|s_i, I) \tag{7}$$

Bayesian rule allows Equation 5 being derived to Equation 6. From Equation 6 to Equation 7, we assume that the degree of landmark i's alignment is independent to other landmarks' alignment given current landmarks' positions and the image. Since $\mathbf{s}$ is uniquely determined by parameter $\mathcal{P}$ given 3D shape model, $p(\mathcal{P}) = p(\mathbf{s})$.

We build a logistic regressor to represent the likelihood, $p(v_i = 1|s_i, I) = \frac{1}{1+exp\{\vartheta\varphi+b\}}$, which has shown its effectiveness in [30, 32]. φ is the feature descriptor of landmark patch i, ϑ and b are the regressor weights trained from collected positive and negative samples. We assume that the prior conforms to Gaussian distribution, $p(\mathcal{P}) \propto \mathcal{N}(\mu; \Lambda), \Lambda = diag\{[\lambda_1, ..., \lambda_k]\}$, where λ_i is the i^{th} eigenvalue corresponding to the i^{th} shape basis in $\mathcal{Q}$ from the nonrigid PCA approach, μ is the mean parameter vector respectively.

Step 1: local patch mean-shift. Given a near-optimal landmark s_i, we intend to search its neighborhood to get the optimal alignment likelihood. Naturally the possible optimal candidates y_i form a region Ψ_i. We assume y_i conforms to Gaussian distribution $\mathcal{N}(s_i, \sigma_i\mathbf{I})$. Hence, the alignment likelihood is modeled as a mixture of Gaussian of the candidates y_i, $p(v_i = 1|s_i, I) = \sum_{y_i \in \Psi_i} \pi_{y_i} \mathcal{N}(y_i, \sigma_i\mathbf{I})$, where $\pi_{y_i} = p(v_i = 1|y_i, I)$.

An Expectation Maximization (EM) approach is raised to solve the problem of Equation 7, which follows the solution presented in [26]. Assuming all the landmarks' candidates distribution have the same deviation σ, the objective function is shown in Equation 8.

$$\arg\min_{\mathcal{P}, s_i} \left(||\mathcal{P} - \mu||^2_{\Lambda^{-1}} + \sum_{i=1}^{n} \sum_{y_i \in \Psi_i} \frac{\beta_{y_i}}{\sigma^2} ||s_i - y_i||^2 \right) \tag{8}$$

where $\beta_{y_i} = p(y_i|v_i, s_i, I)$. Taking the first order approximation $\mathbf{s} = \mathbf{s}^* + J\Delta\mathcal{P}$, $J = \frac{\partial \mathbf{s}}{\partial \mathcal{P}}$ the Jacobian of shape points, the updating function of parameter $\mathcal{P}$ has a close form in Equation 9.

$$\Delta\mathcal{P} = \left(\sigma^2\Lambda^{-1} + J^T J\right)^{-1} \left[J^T U - \sigma^2\Lambda^{-1}(\mathcal{P} - \mu)\right] \tag{9}$$

where $U = [U_1, ..., U_N], U_i = \sum_{y_i \in \Psi_i} \beta_{y_i} y_i - s_i$. Actually U is the mean-shift vector on response map Ψ. By iteratively updating the mean-shift vectors on each local patch response map, the parameter $\mathcal{P}$ is updated until converging to the global optimum.

Step 2: component-wise active contour. Local patch mean-shift performance relies heavily on the response map. We found in some cases merely mean-shift strategy cannot find the correct positions. Possibly the global constrain of $\mathcal{P}$ after mean-shift does not guarantee fitting each component exactly. But the result of mean-shift is expected to fall in the convergence basin of the global minima. We aim to take external force constrain to push the landmarks in each component aligning to its global minimum. It is component-wise because there is seldom such general external force for all the landmarks. By adding shape constrain similar as $\Phi^{ig}_{jk}(s_j, s_k) = (dx, dy, dx^2, dy^2)$ defined in section 3.1, we expect to preserve the structure of shape.

For each landmark, we evaluate its alignment by another measurement $exp(-\eta e_i)$. e_i is positive energy item including shape constrain, appearance constrain and external force constrain. Combining with objective function Equation 7, we obtain a refined objective function as Equation 10.

$$\arg\max_{\mathcal{P}} p(\mathcal{P}) \prod_{i=1}^{n} p(v_i = 1|s_i, I) \prod_{i=1}^{n} exp(-\eta e_i) \tag{10}$$

η is a regularization term. We take the linear combination of the three constraints as shown in Equation 11.

$$e_i = \gamma \begin{bmatrix} \Delta s = [\Delta x \Delta y] \\ \Delta s^2 = [x'' y''] \\ \nabla I \\ exp(-d) + log(1+d) \end{bmatrix} = \gamma \Gamma \mathbf{s} \qquad (11)$$

where γ is the linear combination coefficients and d is a distance measure. We choose the Mahananobis distance of pixel value as d, which is the distance between the value of current landmark's pixel and the average value of face skin pixels. We notice that ∇I is the function of I and $\mathbf{s}$ while d is the function of I and $\mathbf{s}$ too. Once I is known, they are just the function of $\mathbf{s}$.

$$\Delta \mathcal{P} = \left(\sigma^2 \Lambda^{-1} + J^T J\right)^{-1} \cdot \left[J^T (U + \frac{1}{2}\eta\gamma\Gamma) - \sigma^2 \Lambda^{-1} (\mathcal{P} - \mu)\right] \qquad (12)$$

Similarly we give out the overall rule for parameter update in Equation 12, which can be achieved by gradient descent method. The reason not merging the two steps together is because in step 1, some patches' mean-shift may deviate due to low quality of response map before global shape constraint. If we directly raise the component-wise active contour on the deviated landmarks, the error may propagate. But if step 1's result is regularized by global shape constraint, the deviation is mediated and step 2 finds the convergence point with fewer iterations.

5. Experiments

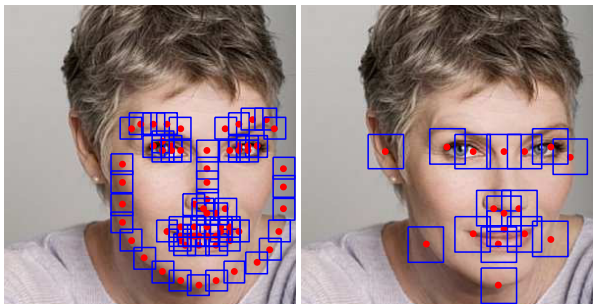
To evaluate our method, we introduce five main face databases used in our experiments, i.e. MultiPIE, AR, LFPW, LFW and AFW. They are collected either under specific experimental conditions or under natural conditions. All of them present challenges in different aspects.

MultiPIE [14] contains images of 337 people with different poses, illumination and expressions. We collected 1300 images from it, which include 13 different poses and each pose contains 100 images from different people. The training of optimized part mixtures is based on this database. Images in AR [22] are frontal with different facial expressions, illumination and occlusion. We take 509 images of 126 people with different facial expressions to conduct the experiment.

LFPW [2], LFW [16] and AFW [35] are image databases collected in wild conditions. The images contain large variations in pose, illumination, expression and occlusion. For LFPW, we collected 801 training images and 222 testing images. For LFW, we used 12007 images which have valid annotations. For newly published AFW, we can only access 205 testing images.

As each of them has different number of annotation landmarks, when evaluating different algorithms on the same database, we use the landmarks from database annotation which are common in all the algorithms. We firstly verify the group sparse learning selection based landmark detectors by comparing to the Tree Structure Part Model (TSPM) [35] algorithm. We then raise the near-frontal face alignment comparison with Multi-view ASMs [18], CLM [26], Oxford landmark detector [11] and TSPM. The databases are AR and near-frontal images from MultiPIE. Based on LFPW, LFW and AFW, we compare the algorithms on the unconstrained cases. In addition, our method is potentially capable of tracking facial landmarks because of its fast update between two consecutive frames. We test it on talking face video [1] and compare it with CLM and Multi-ASM algorithms.

Quantitatively, the alignment error is measured by the distance from ground truth normalized by the distance of two centers of eyes for frontal face databases. For those non-frontal databases, in which case not all two pupils are visible, we normalize the error by the square root of face size, reflected by the rectangle hull of aligned landmarks.



(a) TSPM (b) Optimized Mixtures

Figure 2. Facial landmark models of TSPM and Optimized Mixtures. (a) TSPM landmark model with 68 red dots as landmark positions and blue rectangles as local patches. (b) The Optimized Mixture model with only 17 red-dot landmarks and blue rectangles as local patches.

5.1. Optimized Mixtures vs. Tree Structure Part Model

Zhu and Ramanan [35] proposed a tree structure part model to simultaneously detect face and localize landmarks. The landmarks in their model are densely distributed. We propose a group sparse learning method to select the most representative landmarks. We conduct the comparison of the average localization error on AFW and LFPW datasets. As the code provided by the authors is based on Matlab, we compare the running time on the same Matlab platform.

Figure 2 visualizes the TSPM dense model and our optimized mixture model. The optimized model attempts to capture the most significant anchor points while omitting the intermediate landmarks, which reduces the risk of error propagation from misaligned landmarks. Quantitatively, in Table 1, the proposed optimized mixture method outper-

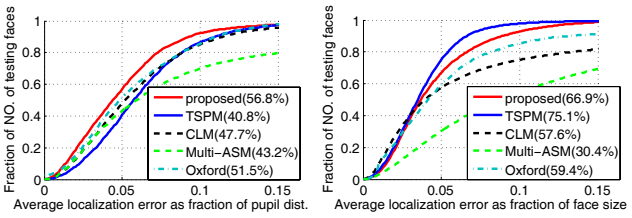
Table 1. Percentage of images less than given relative error level of TSPM and the proposed optimized mixtures on AFW and LFPW datasets and average running time per image.

		$< 5\%$	$< 10\%$	$< 15\%$	time(s)
AFW	TSPM	61.3%	88.9%	92.6%	14.03
	proposed	**68.9%**	**95.6%**	**98.4%**	**5.81**
LFPW	TSPM	72.8%	87.8%	91.2%	8.23
	proposed	**81.1%**	**96.1%**	**98.5%**	**2.25**

forms TSPM in both AFW and LFPW datasets. The running time is about 3 to 4 times less than TSPM method. One reason is that the number of simplified model landmarks is less, hence with less possibility of misalignment. Another reason is that the sparse structure interferes the misaligned landmark's error from being passed to the neighborhood landmarks. The accuracy of initialization is not critical in our algorithm. But the running time is a key factor in the whole framework. Since TSPM is claimed possible to be real-time [35], we expect to push the proposed method real-time by certain implementing techniques other than matlab.

5.2. Comparison with Previous Work

We compare our approach (optimized mixtures with cascaded deformable shape model) with the following methods. (1) Multi-view ASMs [18], (2) Constrained local model (CLM) [26], (3) Oxford facial landmark detector [11], (4) tree structure part model (TSPM) [35]. TSPM and CLM are two of the state-of-the-art methods in face alignment. Especially for wild faces, TSPM has reported superior performance over many other state-of-the-art methods. For non-frontal comparison, we hard code ground truth face rectangle to Multi-ASMs, CLM and Oxford as face detection results because in those cases such methods may fail to locate



(a) AR (b) Near-frontal MultiPIE

Figure 3. Cumulative error distribution curves for landmark localization on near-frontal images. (a) Error distribution tested on near-frontal AR database. The numbers in legend are the percentage of testing faces that have average error below 5% of the pupil distance. (b) Error distribution tested on near-frontal MultiPIE database. The percentage is the ratio of error less than 5% of ground truth face size. TSPM is a little marginally better than proposed method on MultiPIE. It is because TSPM is trained on MultiPIE database which potentially causes over-fitting. But in AR, the proposed method outperforms all the other methods including TSPM.

faces merely using Viola-Jones detector.

We firstly evaluate performance on frontal and near-frontal faces in AR and MultiPIE database. For MultiPIE, we select the near-frontal portion of all the pose-variant images. The near-frontal is defined as faces with yaw angle varying from $-45°$ to $45°$, in which case all landmarks are visible. For the relative error (Figure 3(a)), our proposed method outperforms other methods with 5% improvement compared to the second one at relative error level 0.05. In Figure 3(b), the TSPM method achieves the best performance. However, the proposed method still shows competitive performance as compared to the other methods. The reason is that TSPM is trained on MultiPIE database. Potentially it may be over-fitting to the particular database. Later experimental results verified the assumption.

Table 2. Mean Average Pixel Error (MAPE) on AR and Near-frontal MultiPIE datasets in pixels.

	CLM	ASM	Oxford	TSPM	proposed
AR	9.50	15.63	9.04	9.72	**7.87**
MultiPIE	19.65	23.51	19.59	**6.38**	7.34

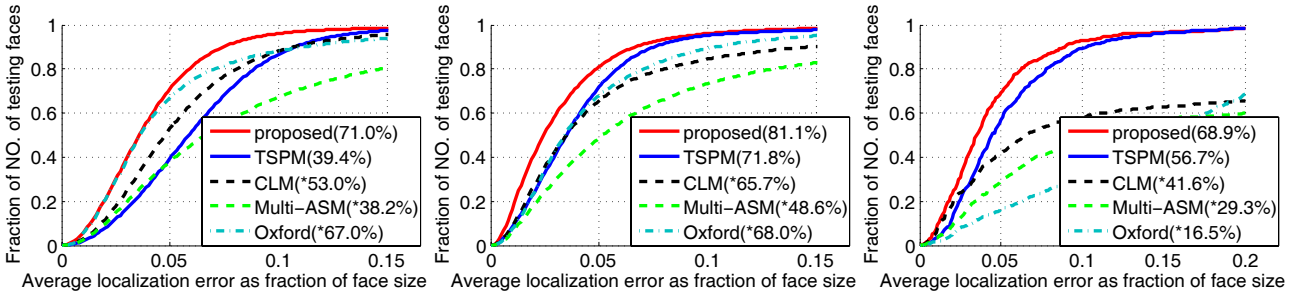
From the absolute pixel error point of view, we evaluate all the algorithms on mean average pixel error (MAPE) measure. Our method achieves 7.87 pixels MAPE on AR database and 7.34 pixels MAPE on near-frontal MultiPIE database as shown in Table 2. Though TSPM leads the accuracy in MultiPIE, our method controls the absolute error in a very low level which consistently outperforms the state-of-the-art methods.

Further investigation is conducted about the performance of all the methods on LFW, LFPW and AFW. Figure 4 shows that our method consistently outperforms other methods with a significant margin. For fair comparison, we provide ideal face bounding boxes for compared methods , CLM, Multi-ASM and Oxford, as they may fail to detect faces in side-view face images. Star sign is shown in Figure 4 for those manually labeled methods. Although giving advantage to those methods, the proposed method achieves 71.0% of total face volume within relative error 5% on LFW, 81.1% fraction on LFPW and 68.9% on AFW, which consistently retains the localization accuracy in a very high level. Absolute pixel error results in Table 3 also supports the conclusion.

Table 3. Mean Average Pixel Error (MAPE) on LFW, LFPW and AFW datasets in pixels.

	CLM	ASM	Oxford	TSPM	proposed
LFW	5.08	8.53	4.23	5.26	**3.64**
LFPW	11.36	17.33	10.21	9.26	**7.37**
AFW	19.32	20.22	27.44	11.09	**9.13**

We notice that the proposed method is better than T-SPM with a large marginal gap (at least 9%) all through the three datasets. It shows that TSPM may be over-fitting



(a) LFW (b) LFPW (c) AFW

Figure 4. Cumulative error distribution curves for landmark localization on face-in-the-wild databases. (a) Error distribution tested on Life Face in the Wild (LFW) dataset. (b) Error distribution tested on Labeled Face Parts in the Wild (LFPW). (c) Error distribution tested on Annotated Face in the Wild (AFW).

to MultiPIE environmental conditions. Moreover, in Figure 4 (c), CLM, Multi-ASM and Oxford achieves abnormally poor performance which indicates that those models can not accommodate to the extremely bad face-in-the-wild conditions.

5.3. Evaluation on Talking Face Video

We claim that the proposed method (optimized mixtures with cascaded deformable shape model) has potential to track videos and image sequences. The reason is that in our model, initialization is simplified from TSPM which is claimed real-time detection performance and the two-step cascaded strategy is based on mean-shift and component-wise active contour. We can directly use information from past frames as the initialization for the following frames.

Table 4. Percentage of talking face image frames less than given relative error level and Mean Average Pixel Error (MAPE) in pixels.

Relative error	$< 5\%$	$< 10\%$	$< 15\%$	MAPE
Multi-ASM	38.07%	73.72%	95.67%	12.22
CLM	73.16%	98.01%	99.80%	8.59
proposed	**79.19%**	**99.70%**	**99.98%**	**7.31**

Since TSPM is a detection based method without any plug-in of tracking strategy, we only compare the results on talking face video with CLM and Multi-ASM, which are able to raise video tracking. The relative error is defined as the fraction of average localization error over pupil distance. Table 4 shows that our method outperforms the other two methods with distinct margin. Visualization from Figure 5 convinces our conclusion that the error by proposed method is consistently smaller than the other two methods.

6. Conclusion

We present a two-stage cascaded deformable shape fitting method for face landmark localization and tracking. By introducing 3D shape model with optimized mixtures of parts, we achieve pose-free landmark initialization. Extensive experiments demonstrate the advantage of our method in aligning wild faces with large pose variation. It also outperforms CLM and Multi-ASM in face landmark tracking. Future work may further investigate local discriminative search and its efficiency.

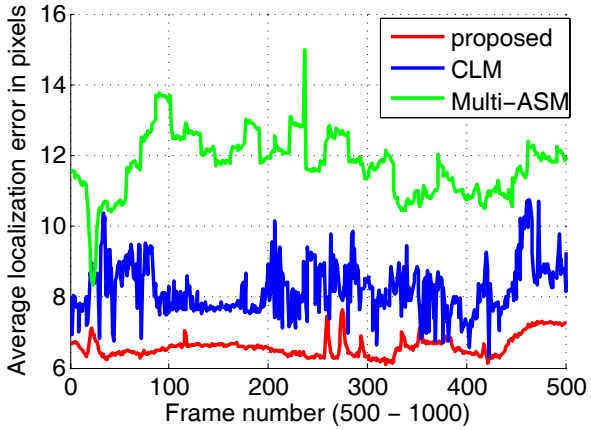


Figure 5. Average landmark tracking error in pixels of talking face video from frame 500 to frame 1000.

References

[1] http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html. 5

[2] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011. 5

[3] X. Cao, Y. Wei, F. Wen, and J.Sun. Face alignment by explicit shape regression. In *CVPR*, 2012. 2

[4] E. M. Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. on PAMI*, 2009. 1

[5] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998. 1, 2

[6] T. Cootes and C. Taylor. A mixture model for representing shape variation. In *BMVC*, 1997. 1, 2

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995. 1, 2, 4

Figure 6. Selected results from Talking Face video, AR, MultiPIE, LFPW, LFW and AFW databases. The first two columns are from MultiPIE, AR, and Talking Face video. The second two columns are from LFPW. The third two columns are from LFW. The last two columns are from AFW.

[8] D. Cristinacce and T. Cootes. Automatic feature localization with constrained local models. *PR*, 2007. 1, 2

[9] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, 2007. 1, 2

[10] P. Dollar, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085, 2010. 1, 2

[11] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy"-automatic naming of characters in tv video. In *BMVC*, 2006. 1, 2, 5, 6

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on PAMI*, 2009. 2

[13] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61, 2003. 2

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 2010. 3, 5

[15] L. Gu and T. Kanade. 3d alignment of face in a single image. In *CVPR*, 2006. 1

[16] G. Huang, M. Ramesh, T. Berg, and E. Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Technical Report*, 2007. 5

[17] J. Huang, X. huang, and D. Metaxas. Learning with dynamic group sparsity. In *ICCV*, 2009. 3

[18] Y. Huang, Q. Liu, and D. Metaxas. A component based deformable model for generalized face alignment. In *ICCV*, 2007. 1, 5, 6

[19] M. Jones and P. Viola. Fast multi-view face detection. In *CVPR*, 2003. 1, 2

[20] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *ECCV*, 2008. 2

[21] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. 3

[22] A. Martinez and R. Benavente. The ar face database. In *CVC Tech. Report number 24*, 1998. 5

[23] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004. 1, 2

[24] D. Metaxas and S. Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 2013. 1

[25] S. Rivera and A. Martinez. Learning deformable shape manifolds. *Pattern Recognition*, 2012. 2

[26] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 2010. 1, 2, 4, 5, 6

[27] M. Uricar, V. Franc, and V. Hlavac. Detector of facial landmarks learned by the strctured output svm. In *VISAPP*, 2012. 2

[28] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010. 2

[29] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, and D. Metaxas. The best of both worlds: Combining 3d deformable models with active shape models. In *ICCV*, 2007. 1

[30] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *CVPR*, pages 1–8, 2008. 1, 2, 4

[31] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 1, 2

[32] X. Yu, F. Yang, J. Huang, and D. Metaxas. Explicit occlusion detection based deformable fitting for facial landmark localization. In *AFGR*, 2013. 4

[33] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. Metaxas, and X. Zhou. Sparse shape composition: A new framework for shape prior modeling. In *CVPR*, 2011. 1

[34] Y. Zhou, W. Zhang, X. Tang, and H. Shum. A bayesian mixture model for multi-view face alignment. In *CVPR*, 2005. 1, 2

[35] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *CVPR*, 2012. 1, 2, 3, 5, 6