Face Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Model

Xiang Yu, *Member, IEEE*, Junzhou Huang, *Member, IEEE*,
Shaoting Zhang, *Senior Member, IEEE*, and Dimitris N. Metaxas, *Fellow, IEEE*

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. In initialization stage, we propose a group sparse optimized mixture model to automatically select the most salient facial landmarks. By introducing 3D face shape model, we apply procrustes analysis to provide pose-aware landmark initialization. In landmark localization stage, 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 simultaneously handles face detection, pose-robust landmark localization and tracking in real time. Extensive experiments are conducted on both laboratory environmental databases and face-in-the-wild databases. The results reveal that our approach consistently outperforms state-of-the-art methods for face alignment and tracking.

Index Terms—Face landmark localization, face tracking, deformable shape model, part based model

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 [1], [2], facial expression analysis [3], [4], [5], 3D face modeling [2], [6], video editing [7], etc. All the applications require accurate landmark positions. However, due to complicated background, lighting conditions and particularly occlusion and pose variations, accurate landmark localization remains challenging in practice.

Many face alignment algorithms rely on the facial area detection results, which is the first and key step in landmark localization. For example, in CLM [8] and SDM [9], their frameworks apply the famous face detector Viola and Jones [10]. Though general face detection algorithms such as Viola and Jones face detector show good performance in general cases, it cannot handle the conditions with pose variation or partial occlusion. Therefore, the more state-of-the-art face detection algorithms are proposed to overcome the extreme conditions [11], [12]. Even provided with good region of interests under extreme conditions, improper landmark initialization still leads to misalignment. For example, Active Appearance Models (AAMs) [13] are very sensitive to initial positions, because complex appearance with illumination and noise may result in local minima. Like the more state-of-the-art face detection algorithms above, a robust initialization which simultaneously detects face under challenging conditions and provides facial key points alleviates the complexity of face alignment.

Pose variation and partial occlusion of faces are two main causes of missing facial appearance. Pose variation leads to the self-occlusion by viewing from certain viewpoints and partial occlusion directly obstacle the facial appearance. The missing appearance provides no evidence for the algorithms to evaluate whether the landmarks (especially the landmarks in the missing region) are in proper positions. Regarding occlusion, Saragih et al. [8] introduced a robust error function to control the unseen landmarks' variation. Yang et al. [14] introduced a sparse shape error term to compensate the occluded landmarks' deviation. Roh et al. proposed a supervised RANSAC method [15] to detect non-occluded facial landmarks and use the non-occluded points for alignment. The method shows better accuracy and stability than a Bayesian inference solution based on tangent shape approximation [16]. However, RANSAC achieves high accuracy at the price of low efficiency. Yu et al. [17] proposed a consensus over multiple occlusion-specific regressors to overcome the missing appearance problem. Burgos-Artizzu et al. [18] proposed a block-wise occlusion prior learned from training set to guide the cascaded pose regression. Ghiasi and Fowlkes [19] considered a hierarchical deformable part model (DPM) to detect occlusion at part level and propagate such occlusion information for the holistic landmark alignment.

To alleviate pose variation problem, Cootes and Taylor [16] imported mixture model for representing shape

- *X. Yu and D.N. Metaxas are with the Department of Computer Science, Rutgers University, NJ, USA, 08854. E-mail: {xiangyu, dnm}@cs.rutgers.edu.*
- *J. Huang is with the Department of Computer Science and Engineering, University of Texas at Arlington. E-mail: jzhuang@uta.edu.*
- *S. Zhang is with the Department of Computer Science, University of North Carolina at Charlotte.*

Manuscript received 29 Apr. 2014; revised 30 Oct. 2015; accepted 17 Nov. 2015. Date of publication 16 Dec. 2015; date of current version 10 Oct. 2016. Recommended for acceptance by D. Ramanan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2509999

variation. Zhou et al. [20] also provided a Bayesian mixture model for multi-view face alignment. Although multi-view face shape models partially solve the pose variation problem, they cannot cover unlimited possibilities of view changes. Therefore, 3D shape models [21] are 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 [22] are based on face detectors, which makes the problem recursive. A better choice is to train fast and accurate facial anchor point detectors.

Previous face alignment methods have achieved wide success. The parametric models, e.g., Active Shape Models (ASMs) [23], AAMs [13], [24], establish shape representation and an either holistic object appearance or series of local patches' appearance by conducting Principal Component Analysis (PCA) on a set of training data. Such methods assume the testing faces lie in the linear subspace of the training faces. Most of them are sensitive to the initial positions. While in searching the optimal solutions, the inverse of hessian matrix for AAMs is time-consuming. Constrained Local Models (CLMs) [8], [25], [26], largely speed up the optimization because the local update can be achieved by efficient EM algorithm and the global update is close-form. The discriminative approaches attempt to learn a mapping from image features to parameters' increment or landmarks' displacement. Cristinacce and Cootes [27] proposed to fit ASMs by learning a linear regression between the increment of motion parameters and the appearance differences. Further improvements investigated directly mapping the image features variation to landmark displacement [9], [28]. These methods attempt to numerically approximate the Jacobian from locations to parameters. But due to data completeness and noise, there is no guarantee the approximation is practically equivalent to the real Jacobian.

In this paper, we propose a novel framework to deal with all the above-mentioned problems. The initialized anchor points are selected by a group sparse learning strategy. Modified from Zhu and Ramanan's work [29], 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. The sparse landmarks reduce the risk of error propagation from dense misaligned landmarks. Then a bi-stage cascaded deformable shape model is presented to achieve global optimum. The first step is globally searching all the landmarks' positions by the holistic shape constraint. The second step is locally refining each components alignment by each component shape constraint.

A preliminary method is proposed in [30]. The difference between [30] and this paper is mainly summarized into three aspects. 1) We extensively analyze the effectiveness of each module proposed in our method, i.e., we evaluate the Optimized Part Mixtures (OPM) versus Tree Structure Part Model (TSPM) [29] by using the two modules separately as our initializer and applying the bi-stage cascaded deformable shape model to fit the landmarks. Also we investigate the component-wise active contour by checking the performance of the proposed framework with and without the module. These additional discussions are very important for justifying the parts of our proposed method. 2) We add two more state-of-the-art comparing methods [31] and [9], which are the regression based methods and have shown better performance so far. Rivera and Martinez [31] proposed to build kernel regression (KR) relationship between the feature space and coordinate space, which is a non-linear strategy for alignment. Xiong and De la Torre [9] introduced multi-step linear regression method to approach the non-linear face manifold. These comparisons further confirmed the advantages of the proposed approach. 3) In the sparse learning step, we provide an efficient and more concrete algorithm to solve the sparse max-margin learning problem. The initial work [30] does not introduce the detail of solving those learning problems. Considering the completeness and reproducibility of our work, in this paper, we illustrate all the learning process sequentially to form a clear workflow for implementation.

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-stage cascaded deformable shape model is able to deal with large pose variation and capture subtle shape variations from classical parametric shape models. **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.

The remainder of the paper is organized as follows. Section 2 introduces the fundamentals of facial feature detection by reviewing some relevant papers. The proposed approach is described in Sections 3 and 4. Our experimental results are evaluated quantitatively and qualitatively in Section 5 , and the conclusions are drawn in the final section.

2 Related Work

With respect to face detection, Viola and Jones [10] proposed a widely used boost framework. It is fast and effective for most frontal faces. Although it is able to handle non-frontal faces by assembling side-view training data, it provides no pose information for proper facial landmark initialization. A number of improvements and modifications are made based on the Viola and Jones detector. Some focused on extracting more effective features [32] and others concentrated on classification learning methods [33]. Nonetheless, several detection based landmark localization approaches are proposed to directly provide the landmark positions without knowing the face region. Sivic et al. [34] used mixture of tree structure to optimize the landmark positions of the whole face. Karlinsky and and Ullman [35] exhibited face component detector learning to ensemble the facial component detectors and parse the facial attributes. Uricar et al. [36] proposed a seven anchor point detector based on Deformable Part Models [37] which depicts an object with several parts and the connection in between each other part. They adopted structure-output SVMs to further localize the landmarks 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 [29] proposed another framework based on mixture of part model. Different such mixtures can handle different view-point faces. 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. ASMs [21], [23], [27], [38] constrains the pre-defined landmark vector $s = [x_1, x_2, \ldots, x_N]$, which is concatenated by N landmark coordinates $x_i, i = 1, \ldots, N$, in the linear subspace spanned by the eigen vectors from training data. The objective function is illustrated as Eq. (1):

$$\arg\min_{u \in \mathcal{R}^m} \|s - (\bar{s} + \mathcal{Q}u)\|_2^2, \tag{1}$$

where s is the objective shape, $\bar{s}$ is the mean shape, $\mathcal{Q}$ is the matrix formed by shape eigenvectors and u is the coefficients to be pursued. Each landmark uses pixel value or gradient feature to search the optimal position in neighborhood. Then the global linear subspace constrain is applied to realign all the landmarks. The ASMs and their improvements assume the shape s can be linearly represented by training data and they consider no consistent appearance feature in the optimization, e.g., only exhaustive local search in gradient or pixel feature. AAMs [13], [24], [39] improve the ASMs by considering the minimization of holistic facial appearance error:

$$\arg\min_{c,\mathcal{P}} \|I(W(\bar{s}, \mathcal{P})) - \mathcal{U}c\|_2^2. \tag{2}$$

$\mathcal{P}$ is the parameters for similarity transformation W. U is the appearance matrix formed by concatenating appearance eigenvectors, which is learned from PCA over holistic face appearance features. c is the coefficients and I is the facial image. AAMs make the same assumption for shapes as ASMs. In addition, they assume the facial appearance also lies in the linear subspace spanned by eigen appearance vectors from PCA in the same way of shapes. Thus objective function Eq. (2) aims to find optimal parameters, p and c to reconstruct the facial appearance. These methods are expected to achieve more precise results. However, during the optimization, the calculation of Jacobian of parameters is very expensive and it must be updated for each iteration. Practically such methods are sensitive to initial landmark positions since holistic facial appearance may contain large number of local minimum. In order to further overcome the fallacies, Cristinacce and Cootes firstly proposed a more effective framework CLMs [25]. The appearance model is similar to that used in AAMs. But the appearance is not holistic anymore. They extract local patch around each landmark and combine all the local patches to indicate a face, in a feature template manner

$$\arg\max p(\mathcal{P}|s, I) \propto p(\mathcal{P}|I)p(s|\mathcal{P}, I) \tag{3}$$

$$= p(\mathcal{P}) \prod_{i=1}^{N} p(s_i|\mathcal{P}, I). \tag{4}$$

Instead of minimizing the reconstruction error of appearance, CLM aims to maximize the overall alignment probability by matching current local patches with pretrained templates and assuming each patch alignment is independent given the model. In Eq. (3), it maximizes the probability of model parameter $\mathcal{P}$ given current landmark positions s and facial image I. A Bayesian inference is derived in Eq. (3). In Eq. (4), the posteriori of each landmark alignment is conditionally independent. In Cristinacce and Cootes work [25], the posteriori is estimated as a Gibbs-Boltzmann distribution. While in Saragih et al.'s work [8], it is assumed a mixture of Gaussian and then an Expectation Maximization (EM) approach is adopted to search the optimum. Another improvement is using Random Forests to vote for the best position for each point [40]. Recently a fast AAM algorithm was presented for real time alignment [41], and an ensemble of AAM [42] was proposed to jointly register landmarks for image sequence. The combination of a part model and CLM [30] was proposed to alleviate pose variations, while other CLM frameworks focused on local patch expert learning [43].

Nonparametric shape regression is another way for shape registration. Belhumeur et al. [44] proposed a data-driven method that employed RANSAC to robustly fit exemplar landmark configurations drawn from a database to a set of local landmark detections. Similar methods [45], [46] either considered temporal feature similarity for joint face alignment or used graph matching to enhance the landmark localization. Cristinacce and Cootes [27] introduced a boosted regression predictor which learns the relationship between the local neighborhood appearance and the displacement from the true feature location under the ASM framework. Following the same way, several discriminative methods are introduced. Gradient Boosting [47] models the nonlinearities in relationship of the feature motion and the coordinates displacement effectively. Valstar et al. [48] used a combination of Support Vector Regression and Markov Random Fields to largely reduce the search time. Martinez et al. [49] proposed local evidence aggregation for regression based alignment. Rivera and Martinez [31] investigated kernel regression to map from image features directly to landmark location and received good performance especially for low-resolution images. Cootes et al. [40] recently presented Random Forest voting based shape fitting method in improving the locating accuracy. Cao et al. [28] presented a holistic regression method to directly infer the whole landmark set from facial images. Dantone et al. [50] introduced conditional regression forests to treat faces with different poses separately. Dollar et al. [51] proposed cascaded pose regression to approximate 2D pose of objects. Burgos-Artizzu et al. [18] continued using the cascaded pose regression and apply it to predict the occlusion likelihood of landmarks. Xiong and De la Torre [9] proposed a supervised descent learning to establish such relationship and achieved high efficiency. Besides the mainstream, Liang et al. [52] trained directional classifiers to discriminatively search facial components.

Those algorithms are among state-of-the-art performance. However, most of them lack the flexibility in representing pose-variate cases. They are not designed for handling partial occlusion because many of them need to compute the feature motion and partial occlusion provides no feature at all. Since they need to build the direct

relationship between coordinate displacement and feature motion, the training data must be prepared with experience and the volume of training faces is large. Nevertheless, since one step of regression usually cannot guarantee convergence of landmark search, e.g., in Xiong and De la Torre's work [9], they proposed several same steps of such search procedure, it is difficult to determine the number of iterations in practice. 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 can handle partial face occlusion by explicitly detecting the occlusion from normal facial parts. Also it achieves real time performance due to the group sparse selection and cascaded two-stage deformation strategy. In addition, merely several hundreds of facial images are needed to provide model prior.

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 follow a pictorial structure [29], [37] to organize the landmarks. Section 3.1 serves as preliminary background introduced in [29]. Based on that, we aim to simplify the dense structure of TSPM. Observing that not all dense landmarks are needed in order to localize the facial area, we introduced group sparse constraint over all the landmarks in Section 3.2. To obtain the coefficients of each landmark, a max-margin learning framework is proposed in Section 3.3. In the meanwhile, we further introduced an iterative updating algorithm to efficiently solve the group-sparse constraint problem.

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 [29].

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

$$f_i(I, \mathbf{s}) = \sum_{j \in V_i} q_i(I, s_j) + \sum_{(j,k) \in E_i} g_i(s_j, s_k). \tag{5}$$

The first term in Eq. (5) is a local patch appearance evaluation function $q_i : I \times \mathcal{S}_i \rightarrow \mathbb{R}, i \in (1, N)$, defined as,

$$q_i(I, s_j) = \left\langle \mathbf{w}_j^{iq}, \Phi_j^{iq}(I, s_j) \right\rangle \tag{6}$$

indicating how likely a landmark is in an aligned position. The second term is the shape deformation cost $g_i : \mathcal{S}_j \times \mathcal{S}_k \rightarrow \mathbb{R}, (j, k) \in E$, defined as,

$$g_i(s_j, s_k) = \left\langle \mathbf{w}_{jk}^{ig}, \Phi_{jk}^{ig}(s_j, s_k) \right\rangle \tag{7}$$

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 [29]. Further, we formulate the two evaluation functions in a uniform way to obtain a more compact representation

$$f_i(I, \mathbf{s}) = \langle \tilde{\mathbf{w}}_i, \tilde{\Phi}_i \rangle, \tag{8}$$

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 Eq. (9):

$$\mathbf{s}^* = \underset{\mathbf{s} \in \mathcal{S}, i \in (1,M)}{\arg\max} \; f_i(I, \mathbf{s}). \tag{9}$$

3.2 Group Sparse Learning for Landmark Selection

Facial landmarks are usually defined manually or human-selected without any consistent rules. Evidence is that the annotation among different face datasets is largely different, e.g., LFPW [44] database has 29 points, LFW [53] database has seven points, while AR [54] database has 22 labels. 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. Although one can manually select the most common landmark points for a new facial structure, we intend to automatically select those landmarks by learning from training data to well represent facial structures and the number of landmarks should meet real-time requirement for inference.

The goal of sparse selection is to robustly initialize the landmark positions more than accurately localize the landmark positions. In fact, the landmarks which are more salient than others, i.e., the corner points of eyes and mouth, in some sense would be easier to detect. However, considering the harder landmarks to localize as TSPM [29] does, it is far from robust when severe pose variation or occlusion is present. Removing the vague points decreases the false alarms. In our framework, the learning based selection serving as initial detection and boosting the localization accuracy is placed on the latter two-stage deformable shape fitting, not the detection step. Based on the above observations, we intend to build an optimization algorithm to simplify the dense structure.

Visually salient key points have a higher probability to be selected as the optimized structure. Since the saliency significance is different among the original TSPM's landmarks, selecting the most salient landmarks is feasible. Technically,

as shown in Fig. 2, each landmark is denoted as a patch which is centered at the landmark with certain square size. Each such patch is with a weight matrix of the same size. On the landmark level, the weight matrices should be sparse; on the matrices' element level, the weight elements are all either non-zeros or near zeros. The sparse constraint is on the group level while inside each group the weights are not necessarily sparse. Such property is well characterized as group sparsity [55].

The sparse group constraint is defined as in Eq. (10). assume a partition $\cup_{j=1}^{m} G_j$ of β, which is rearranging the elements inside the vector β and grouping the neighboring elements as group G_j. The m groups are disjoint $G_1, G_2, \ldots, G_m : G_i \cap G_j = \varnothing$ when $i \neq j$. In this way, the coefficient vector becomes $\beta = [\beta_{G_1}, \beta_{G_2}, \ldots, \beta_{G_m}]$. We expect inside each group, only a subset $F \subset \{1, \ldots, m\}$ of those groups are non-zero elements while those non-zero elements are small

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n} \|X\beta - y\|_2^2 + \sum_{j=1}^{m} \|\beta_{G_j}\|_2 \right\}. \tag{10}$$

Notice that the constraint on β is $L_{2,1}$ norm. At the inter-group level $G_j, j = 1, \ldots, m$, the constraint is L_1 regularized while at the intra-group level the coefficients are L_2 regularized which is considered dense but small. The $L_{2,1}$ constraint is considered the group sparse property.

3.3 Max-Margin Learning for Landmark Parameters

In our learning process, we collect positive samples from MultiPIE database [56], 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 $\tilde{\mathbf{w}}_i$. For each landmark, we know that $\tilde{\mathbf{w}}_i = [\mathbf{w}_j^{ig}, \mathbf{w}_{jk}^{ig}]$, which is the weight vector consists of unary weights $\mathbf{w}_j^{ig}$ and pair-wise weights $\mathbf{w}_{jk}^{ig}$. The pair-wise weights are set according to the tree structure edge set E. $\mathbf{w}_{jk}^{ig}$ includes all the weights that are connected to node j. Similarly, in node k, there is such edge weight $\mathbf{w}_{kj}^{ig}$. They are not necessarily equivalent since parent node and child node may take each other in different importance. For simplicity, we denote $\tilde{\mathbf{w}}_i$ as $\tilde{\mathbf{w}}$ in the following notations. Based on Eq. (5), considering the group sparse constraint from Section 3.2, we establish a max-margin framework in Eq. (11):

$$\begin{aligned} \arg\min_{\tilde{\mathbf{w}}, \varepsilon \geq 0} & \left(\sum_{n \in \mathcal{C}} \varepsilon_n + \lambda_1 \|\tilde{\mathbf{w}}\|_2^2 + \lambda_2 \sum_{t=1}^{m} \|\tilde{\mathbf{w}}_t\|_2 \right) \\ & s.t. \forall n \in \mathcal{C}_+, \left\langle \tilde{\mathbf{w}}, \tilde{\Phi}(I_n, \mathbf{s}_n) \right\rangle \geq 1 - \varepsilon_n \\ & \forall n \in \mathcal{C}_-, \forall \mathbf{s}, \left\langle \tilde{\mathbf{w}}, \tilde{\Phi}(I_n, \mathbf{s}) \right\rangle \leq -1 + \varepsilon_n, \end{aligned} \tag{11}$$

where $\tilde{\mathbf{w}} = [\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_m]$. We omit the pose angle i here because the optimization is a unified framework to all the pose angles. Then the holistic weight vector $\tilde{w}$ is constructed by $\tilde{\mathbf{w}}_t$, each of which is a rearranged weight vector at part t combining both the unary weights and pair-wise weights. $\tilde{\Phi}$ is a feature descriptor, i.e., hog feature all through the optimization. The positive features are extracted over positive samples with ground truth and the negative features are extracted with arbitrary configurations.

To solve the problem, a group sparse optimization method is used. We refer the readers to [55] for details of algorithms. From the objective function 11, we know that:

$$\varepsilon_n \geq 1 - y_n \tilde{\mathbf{w}}^T \tilde{\Phi}. \tag{12}$$

Minimizing objective function 11 pushes ε_n to $1 - y_n \tilde{\mathbf{w}}^T \tilde{\Phi}$, where y_n is the class label of node n. For simplicity, we denote $\tilde{\mathbf{w}}$ as W and $\tilde{\Phi}$ as Φ. We define the search process as:

$$S_{i+1} = W_{i+1} + \beta_i (W_{i+1} - W_i), \tag{13}$$

where sequence $\{W_i\}$ are the approximate solutions and $\{S_i\}$ are the search points. Learning rate β_i is a properly chosen coefficient. To compute W_{i+1}, an approximating model was proposed in [55] as Eq. (14). The loss function in Eq. (14) is defined in Eq. (15). Y is a matrix variable of the loss function $F_{L,W}(Y)$. In Eq. (14), W_{i+1} is achieved by minimize the loss function $F_{L,W}(Y)$ over Y

$$\begin{aligned} F_{L,W}(Y) = & [loss(W) + \langle loss'(W), Y - W \rangle] \\ & + \lambda_2 \sum_{t=1}^{m} \|Y_t\|_2 + \frac{L}{2} \|Y - W\|_2^2, \end{aligned} \tag{14}$$

$$loss(W) = 1 - yW^T \Phi + \lambda_1 \|W\|_2^2. \tag{15}$$

Combining the above two equations, W_{i+1} is derived by minimizing the approximating model $F_{L,W}(Y)$ as shown in Eq. (16):

$$\begin{aligned} W_{i+1} &= \arg\min_{Y} F_{L_i, S_i}(Y) \\ &= \arg\min_{Y} \frac{1}{2} \|Y - S_i\|_2^2 + C \sum_{t=1}^{m} \|Y_t\|_2. \end{aligned} \tag{16}$$

Further we notice that for each group in m groups, the optimization is independent, which is denoted as following:

$$W_{i+1,t} = \arg\min_{Y_t} \left(\frac{1}{2} \|Y_t - S_{i,t}\|_2^2 + C \|Y_t\|_2 \right). \tag{17}$$

The above objective function is convex and smooth, which can be solved by gradient descent methods.

Then we can select the most salient $\tilde{\mathbf{w}}_i$ to form a new tree. Those nodes with small weights are eliminated. Because the tree structure is changed, we have to re-train our weights. Training is achieved by solving the traditional max-margin problem:

$$\begin{aligned} \arg\min_{\tilde{\mathbf{w}}, \varepsilon \geq 0} & \left(\sum_{n \in \mathcal{C}} \varepsilon_n + \lambda_1 \|\tilde{\mathbf{w}}\|_2^2 \right) \\ & s.t. \forall n \in \mathcal{C}_+, \left\langle \tilde{\mathbf{w}}, \tilde{\Phi}(I_n, \mathbf{s}_n) \right\rangle \geq 1 - \varepsilon_n \\ & \forall n \in \mathcal{C}_-, \forall \mathbf{s}, \left\langle \tilde{\mathbf{w}}, \tilde{\Phi}(I_n, \mathbf{s}) \right\rangle \leq -1 + \varepsilon_n, \end{aligned} \tag{18}$$

which is a classic quadratic programming problem. We use the dual coordinate descent method proposed in [29] to obtain the optimized weights.

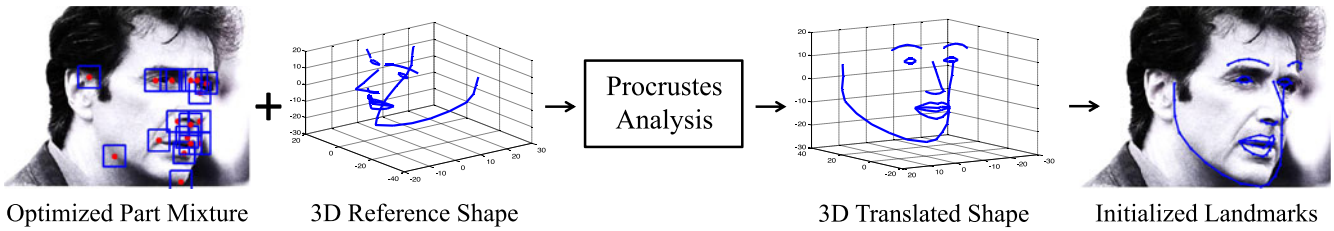


Fig. 1. Pose-free facial landmark initialization using Procrustes analysis on 3D reference shape and detected optimized part mixture.

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. The 3D model is trained off line based on a 3D labeled face shape dataset [57] which serves as a prior all through our method. As head is a near-rigid object in 3D space, the 3D to 2D mapping is unique. The process is illustrated in Fig. 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 $\mathbf{s} = [s_1, \ldots, s_N]$, 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. [23], 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, $\mathcal{Q} = [Q_1, \ldots, Q_k]$ contains the k shape basis, $u \in \mathbb{R}^k$ is the coefficient vector.

The general Point Distribution Model (PDM) introduced by Cootes and Taylor [16], takes global transformation into consideration. Considering rigid transformation in 3D space, scaling, rotation and translation are the only three 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 Eq. (19):

$$s_j = aR(\bar{s}_j + \mathcal{Q}u_j) + T. \tag{19}$$

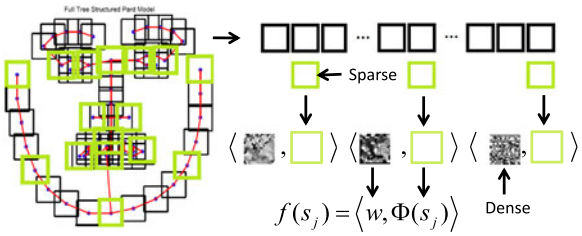


Fig. 2. The group sparse structure illustration. The most salient boxes denoted as green comparing to all the boxes are sparse. Each box is considered a group. At the group level, the selection is sparse. While inside each box, the corresponding coefficient matrix denoted as the gray patch is dense because inside the area of each box, all the pixels contribute to the score $f(s_j)$ calculation.

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 $v = [v_1, \ldots, v_N]$ 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

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} p\big(\mathbf{s}|\{v_i = 1\}_1^N, I\big) \tag{20}$$

$$\propto \arg\max_{\mathbf{s}} p(\mathbf{s}) p\big(\{v_i = 1\}_{i=1}^n|\mathbf{s}, I\big) \tag{21}$$

$$= \arg\max_{\mathcal{P}} p(\mathcal{P}) \prod_{i=1}^{n} p(v_i = 1|s_i, I). \tag{22}$$

Bayesian rule allows Eq. (20) being derived to Eq. (21). From Eq. (21) to Eq. (22), 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 in Eq. (23):

$$p(v_i = 1|s_i, I) = \frac{1}{1 + exp\{\vartheta\varphi + b\}}, \tag{23}$$

which has shown its effectiveness in [26]. φ is the feature descriptor of landmark patch i, ϑ and b are the regressor weights trained from collected positive and negative samples.

The parameter $\mathcal{P}$ are set from the PDM model which applies PCA to a set of registered shapes. The distance in PCA subspace is measured by Mahalanobis distance, which is a kernel l_2-norm measurement. Thus, we assume that the prior conforms to Gaussian distribution

$$p(\mathcal{P}) \propto \mathcal{N}(\mu; \Lambda), \Lambda = diag\{[\lambda_1, \ldots, \lambda_k]\}, \tag{24}$$

where λ_i is the ith eigenvalue corresponding to the ith 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

$$p(v_i = 1|s_i, I) = \sum_{y_i \in \Psi_i} \pi_{y_i} \mathcal{N}(y_i, \sigma_i \mathbf{I}), \tag{25}$$

where $\pi_{y_i} = p(v_i = 1|y_i, I)$. By Bayesian rule, $p(y_i|v_i, s_i, I) = \frac{p(v_i=1, y_i|s_i, I)}{p(v_i=1|s_i, I)}$, we obtain

$$p(y_i|v_i, s_i, I) = \beta_{y_i} = \frac{\pi_{y_i} \mathcal{N}(y_i, \sigma_i \mathbf{I})}{\sum_{z_i \in \Psi_i} \pi_{z_i} \mathcal{N}(z_i, \sigma_i \mathbf{I})}. \tag{26}$$

An Expectation Maximization approach is raised to solve the problem of Eq. (22). Assuming all the landmarks' candidate distribution has the same deviation σ, we derive the objective function in Eq. (27):

$$\arg\min_{\mathcal{P}, s_i} \left(||\mathcal{P} - \mu||^2_{\Lambda^{-1}} + \sum_{i=1}^{n} \sum_{y_i \in \Psi_i} \frac{\beta_{y_i}}{\sigma^2} ||s_i - y_i||^2 \right). \tag{27}$$

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, we obtain the updating function of parameter $\mathcal{P}$

$$\Delta\mathcal{P} = \left(\sigma^2 \Lambda^{-1} + J^T J\right)^{-1} \left[J^T U - \sigma^2 \Lambda^{-1} (\mathcal{P} - \mu)\right]. \tag{28}$$

In Eq. (28), $U = [U_1, \ldots, U_N], U_i = \sum_{y_i \in \Psi_i} \beta_{y_i} y_i - s_i$. 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 Eq. (22), we obtain a refined objective function as Eq. (29):

$$\arg\max_{\mathcal{P}} p(\mathcal{P}) \prod_{i=1}^{n} p(v_i = 1|s_i, I) \prod_{i=1}^{n} exp(-\eta e_i). \tag{29}$$

η is a regularization term. We take the linear combination of the three constraints as shown in Eq. (30):

$$e_i = \gamma \begin{bmatrix} ds = [\Delta x \Delta y] \\ ds^2 = [x'' y''] \\ \nabla I \\ exp(-d) + log(1 + d) \end{bmatrix} = \gamma \mathbf{\Gamma} \mathbf{s}, \tag{30}$$

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 $\mathbf{s}$

$$\Delta\mathcal{P} = \left(\sigma^2 \Lambda^{-1} + J^T J\right)^{-1} \cdot \left[J^T \left(U + \frac{1}{2} \eta\gamma \mathbf{\Gamma}\right) - \sigma^2 \Lambda^{-1} (\mathcal{P} - \mu)\right]. \tag{31}$$

Similarly we give out the overall rule for parameter update in Eq. (31), 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. Our bi-stage fitting procedure is summarized in Algorithm 1.

Algorithm 1. Two-Stage Deformable Shape Fitting with Optimized Part Mixtures

Require: facial image I.
Ensure: optimized $\mathcal{P}$.
1: **Initialization:** given trained $\tilde{\mathbf{w}}$, run landmark detection (9) to get $\mathbf{s}_0$.
2: run Procrust process on $\mathbf{s}_0$ and 3D shape model $\mathbf{s}_{3d}$ to obtain initial $\mathcal{P}$.
3: **repeat**
4: Local Patch Mean-shift: run $\mathcal{P}$ updating function (27), $\mathcal{P} \leftarrow \mathcal{P} + \Delta\mathcal{P}$
5: Component-wise Active Contour: run $\mathcal{P}$ updating function (30), $\mathcal{P} \leftarrow \mathcal{P} + \Delta\mathcal{P}$
6: **until** $\mathcal{P}$ converges

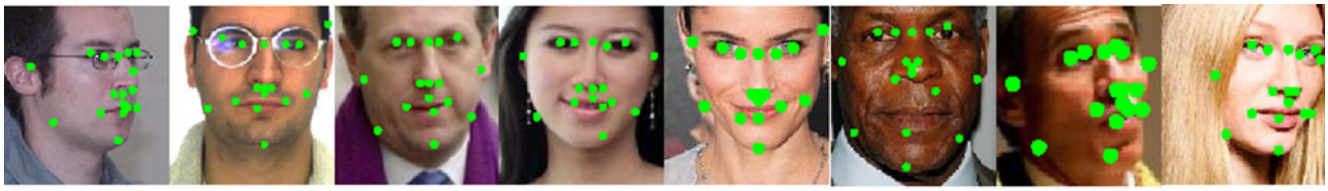
5 Experiments

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

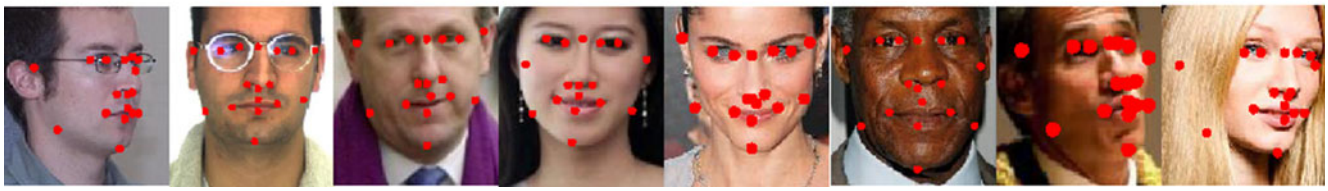
MultiPIE [56] contains images of 337 people with different poses, illumination and expressions. We collected 1,300 images from it, which include 13 different poses and each pose contains 100 images from different people. The training of optimized part mixtures is based on this database.

Images in AR [54] are frontal with different facial expressions, illumination and occlusion. We take 509 images of 126 people with different facial expressions to raise the experiment.

LFPW [44], [58], LFW [53] and AFW [29] 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 selected 12,007 of 13,233 images which have valid annotations. For AFW, we collected 205 testing images. iBug [58] is a recently published



(a) TSPM selected anchor points in green dots to match the 17 point setup of OPM



(b) OPM detected anchor points in red dots

Fig. 3. Visual comparison of converted TSPM with OPM. The converted TSPM is the manual selected 17 point setup which matches the 17 point setup in OPM. The results are evaluated on MultiPIE, AR, LFW, LFPW and AFW. The first column is from MultiPIE. The second column is from AR. The third and fourth columns are from LFW. The fifth and sixth columns are from LFPW and the last two columns are from AFW. (a) Result of converted TSPM in green dots as anchor points. (b) Result of OPM in red dots as anchor points.

even more challenging dataset in which the head pose variation and occlusion are the extreme conditions. The test dataset consists of 135 images with labeled ground truth.

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 [29] algorithm. We then conduct the near-frontal face alignment comparison with Multi-view ASMs [38], CLM [8], Oxford landmark detector [34], TSPM, Kernel Regression [31], SDM [9] and RCPR [18]. The databases are AR and near-frontal images from MultiPIE. In evaluating the pose robustness of our method, we introduced a more recent and challenging dataset iBug [58]. However, since not all the comparing methods can provided proper results on this dataset, i.e. ASM performs pool on the dataset, we do not present the comparison over all the method on this dataset as shown in Table 4. 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 [59] and compare it with CLM and Multi-ASM algorithms.

Quantitatively, the alignment error is measured by normalizing the absolute pixel error over the square root of face size, reflected by the rectangle hull of aligned landmarks. We uniformly apply the face size other than inter-ocular distance as the normalization measure because there are many cases, in which not both eyes are visible.

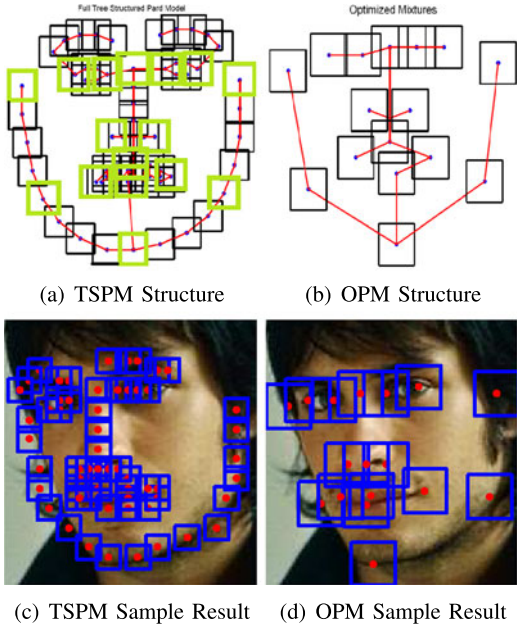
5.1 Optimized Part Mixtures versus Tree Structure Part Model

Zhu and Ramanan [29] 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 AR and LFPW datasets. As the code provided by the authors is based on Matlab, we compare the running time on the same Matlab platform.

Fig. 4 visualizes the TSPM dense model and our optimized mixture model. The TSPM consists of 68 points surrounded with 68 black bounding boxes. Each point is a node of the node set V. The neighboring points are connected in red lines. Each line is an element of the line set E. The graph T = (V, E) consists of the node set V and the edge set E. The center is located at the center of the nose. It is expanded in an undirected and noncyclic way. Thus, the graph is a tree structure. In Fig. 4b, the structure is a simplified one from Fig. 4a. All the blue dots and surrounded black bounding boxes are the ones selected from the dense Tree Structure Part Model, also denoted as the green rectangles in Fig. 4a. We could see that the OPM maintains a part of the original structure of TSPM but neglects the intermediate nodes in between. Though with fewer nodes, the OPM depicts a face completely.

From the TSPM structure, how much portion that the OPM should preserve is also an interesting point to investigate. In implementation, each landmark patch is related with several filters. Each filter corresponds to a weight vector for one view-point. Several view points at the same landmark may share a common filter. Different filters at the same landmark deal with different view-points. For all the filters, we firstly calculate the euclidean norm of the weight vectors and plot the curve as shown at the top of Fig. 5.

From Fig. 5, we notice that for each of the peaks, the width of the peaks is significant, which means that the filters in the neighborhood all have significant or insignificant weights. Meanwhile, the filters with indices less than 50 and larger than 100 are more significant weighted than the filters from 50 to 100. It again verifies that the importance of different filters is different across all the landmarks. By setting threshold varying from 0.03 to 0.09, the weight vector distribution changes shown at the bottom of Fig. 5. When threshold is low, most of the patches are selected. When threshold is high, only the most salient patches (shown in brighter blocks) are selected. We could visually find the most salient patches at the bottom of Fig. 5 matches the simplified structure shown in Fig. 4b.



(a) TSPM Structure (b) OPM Structure

(c) TSPM Sample Result (d) OPM Sample Result

Fig. 4. Facial landmark models of TSPM and Optimized Part Mixtures. (a) TSPM landmark model with 68 red dots as landmark positions and blue rectangles as local patches. (b) The Optimized Part Mixture model with only 17 red-dot landmarks and blue rectangles as local patches.

An interesting and important argument is whether the group sparse selection is beneficial. From the visual saliency map in Fig. 5, we manually selected the most salient points, i.e., four eye corner points, one nose tip point, one eyebrow center point, two mouth corner points, one upper lip and one lower lip points, in total 10-point setup as the baseline model. In the meanwhile, we provided another baseline by randomly selecting 17 points out of the dense 68 points, which has the same number of points layout as our proposed method. By independently train the baselines and our proposed 17-point model, we test the three methods on MultiPIE. As shown in Fig. 6, the proposed optimized part mixture performs more than 10 percent proportion gap over the random and manual selection, which verifies the effectiveness of the group sparse selection.

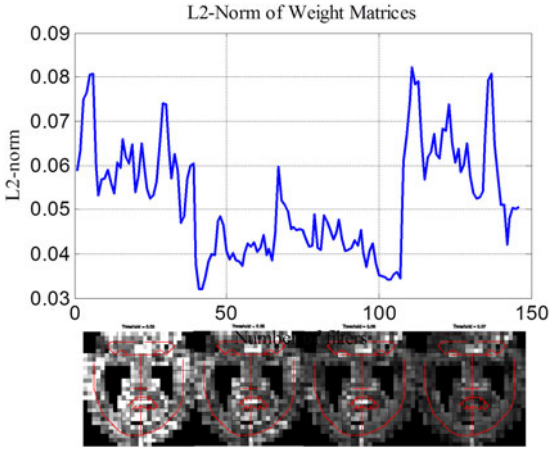


Fig. 5. The visualization of weight vector norms and the gray scale patch image to show the weight distributions at various norm thresholds. The top part is the plot of weight vector norm of each filter. The bottom part are the gray scale patch images under norm threshold 0.03, 0.05, 0.06 and 0.07.

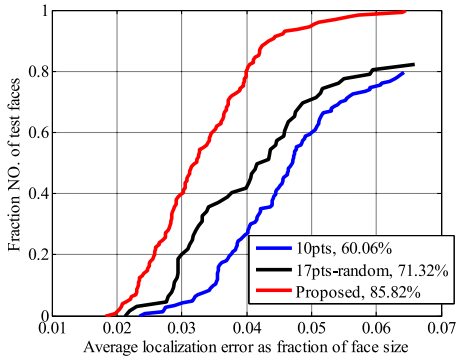


Fig. 6. Cumulative error distribution curves on MultiPIE compared with 17-point random selection method and 10-point baseline method. The proportion reported in the legend is under the relative error 0.05.

We present more visual comparisons between the two algorithms across different databases as shown in Fig. 3. In Fig. 3, the result of TSPM (68 points) is manually selected to match with the 17 points' result of OPM, in which the definition of the 17 points in both TSPM and OPM is the same. We could see from the results that the proposed OPM provides more reasonable initial anchor points than TSPM, e.g., the red dots in Fig. 3 distributes with less deviation and are more consistent than the green dots.

Quantitatively, in Table 1, we observe that TSPM performs slightly better than the optimized mixture model on AR database, of which the gap under 5 percent relative error is only 2.81 percent. But the running time for proposed method is two times less. The performance gap is because the images in AR database are near-frontal with harmonic illumination conditions. The assumption of large pose variation and partial occlusion may not hold at this database. If without pose variation or occlusion, the landmark's detection error is expected to be small. In this situation, the more landmarks of detection evidence, the better of the detection result.

In contrast, the proposed method on LFPW is marginally better than TSPM and running time is 4 times less. The case for LFPW is that pose variation or partial occlusion exists. Each single landmark's detection error is expected to be enlarged. If the landmarks are dense, the error of each landmark influences its neighborhood more than the sparse structure, which is the case of Table 1 on LFPW. Our main purpose designing the group sparse structure on top of the original TSPM is to simplify the dense structure and speed up the process. In well-constrained environments, the TSPM performs better as shown in Table 1 AR database. While in unconstrained environments, the OPM shows

TABLE 1
Percentage of Images Less than Given Relative Error Level of TSPM and the Proposed Optimized Mixtures on AR and LFPW Datasets and Average Running Time per Image

		$<5\%$	$<10\%$	$<15\%$	time(s)
AR	TSPM	**69.4%**	**97.0%**	**99.3%**	14.03
	OPM	57.0%	85.4%	96.2%	**5.81**
LFPW	TSPM	71.8%	95.2%	97.7%	8.23
	OPM	**80.1%**	**96.1%**	**98.5%**	**2.25**

TABLE 2
Proportion of Image Volume Less than Given Relative Error Level on LFPW and AFW Comparing with TSPM-Convert, the Proposed Method and the Proposed Method without Component-Wise Active Contour (No Snake)

		$<5\%$	$<10\%$	$<15\%$
LFPW	TSPM-convert	71.8%	95.2%	97.7%
	proposed	**81.1%**	**96.1%**	**98.4%**
	No Snake	79.2%	94.3%	96.7%
AFW	TSPM-convert	60.4%	92.5%	97.9%
	proposed	**71.4%**	**95.8%**	**99.7%**
	No Snake	66.7%	93.7%	98.2%

some advantage on accuracy over TSPM to overcome the error propagation. We cannot guarantee the OPM as an initialization can always provide sufficiently good result. Otherwise, there is no need for the following two-stage deformable fitting algorithm. On the other side, low-quality initialization such as gross failures obviously results in wrong localization. Thus, it is necessary for the OPM to provide reasonably accurate initialization. As indicating in Eq. (11), our goal is to minimize the localization error margin ϵ_n, which suggests that we require the detection to be as good as possible. From our experimental results, the initialization from OPM provides good enough accuracy such that the latter process achieves competitive while sometimes better results than other state-of-the-art methods. Therefore, our OPM effectively handles challenging situations in practice, and is an important module to improve the overall accuracy and efficiency.

5.2 Algorithm Component Analysis

In our framework, we have introduced Optimized Part Mixtures to simplify TSPM's dense structure and initialize the landmarks. The two-step cascaded deformable shape fitting consists of local patch mean-shift and component-wise active contours. We investigate all modules of the framework to reveal the effectiveness of each component.

Since the TSPM's annotation is different from the OPM's, we manually select the landmarks from TSPM to match the landmark setup in our optimized part model for fair comparison, which denoted as TSPM-convert. For active contour, we directly compare the performance of our framework with and without such refinement. The experiment is conducted on LFPW and AFW wild face databases. We evaluate on the proportion of image volume when relative error is under 5, 10 or 15 percent. Quantitative results are shown in Table 2.

Using converted TSPM as initialization, we see the accuracy drops significantly from the proposed one. Thus applying group sparse selection of anchor points is a novel and key step in our framework. Comparing to the result without active contour, the proposed method is consistently and marginally better, which reveals Snake's effectiveness in improving performance.

5.3 Evaluation on Pose Robustness

In our work, a major task is to align the face shape under severe pose variation. We adopt the MultiPIE, selected images with large pose variation from LFPW (LFPW-P) and images with large pose variation from iBug (iBug-P) for the evaluation.

TABLE 3
The Success Rate of the Detection, the Proportion of Successfully Detected Images over the Database Volume on MultiPIE, LFPW-P and iBug-P

succ. rate	Viola-Jones	TSPM	Proposed
MultiPIE	0.54	1.0	**1.0**
LFPW-P	0.92	0.91	**0.94**
iBug-P	0.88	0.69	**0.91**

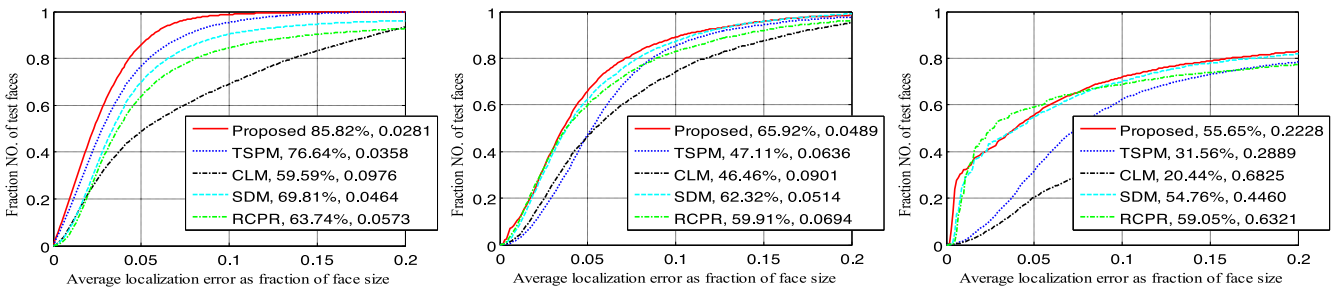
For LFPW-P, 112 images out of 215 test images are manually selected with significantly large head pose variation. The annotation is from 300-W and all the comparing methods are tested on the sub-dataset. Though the landmark setup for the comparing methods is different, the proportion of images well aligned against the normalized alignment error is a fair measure for all the methods.

For iBug-P, 81 images out of 135 test images are manually selected for testing. Most of the 135 test images are with large pose variation. However, since some of the images contain multiple faces, in which some comparing methods only predict one face shape, there is a mismatch to evaluate the alignment accuracy. Thus, we remove the images in this situation for fair comparisons.

The evaluation is conducted on the success rate of all the methods on the three databases and the curve of proportion of database volume versus the normalized error. The success rate is shown in Table 3.

In MultiPIE, the TSPM and the proposed method achieves 100 percent detection rate on the test dataset while Viola-Jones is far less. It is because the detector almost all failed on the pose variations larger than 60 degree in MultiPIE. Even if the multi-view face detector succeeds in the large pose variation, those localization methods cannot facilitate to the large pose situations because their shape fitting schemes may fail in searching such large pose space. To validate this assessment, we equally provide all the compared methods the same face bounding boxes and evaluate their localization accuracy in Fig. 7. The advantageous performance may result from that the TSPM and the proposed method simultaneously detect the key landmarks and the face region utilizing a multi-view deformable part model. The multi-view DPM breaks the entire pose search space into discrete subspaces. The propagation of part detection results enhance the overall success rate.

A more quantitative comparison is on the cumulative error function over the relative error as shown in Fig. 7. The proposed method performs consistently better across the three large pose variation datasets. On MultiPIE, our method achieves 10 percent better than other state-of-the-art methods while on LFPW-P and iBug-P, the gap shrinks because the the faces in these two datasets are more challenging in head poses and all types of occlusion. Even so, our method manages to maintain high performance. Note that we equally provide all the compared methods with the same face bounding boxes. The difference shows only the capability of deforming the initial shapes to handle all kinds of pose variations. The results from Fig. 7 indicate that the proposed method is advantageous in dealing with pose variation.

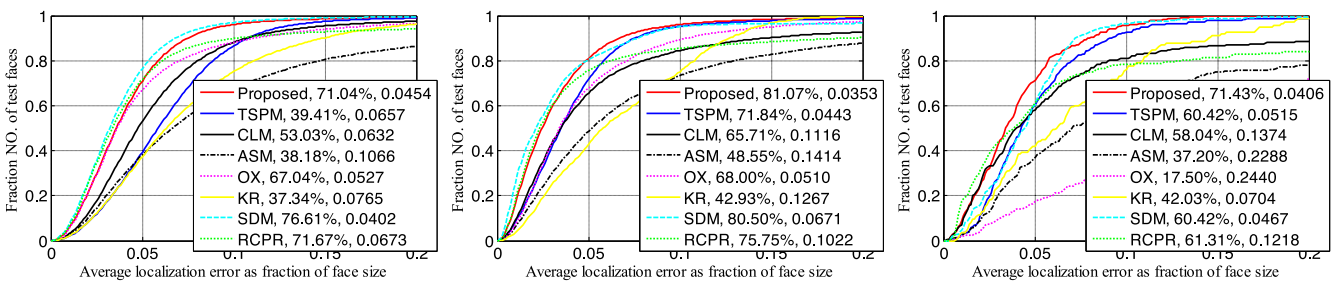


(a) MultiPIE (b) LFPW-P (c) iBug-P

Fig. 7. Cumulative error distribution curves for landmark localization on large pose variation databases. (a) Error distribution tested on MultiPIE. (b) Error distribution tested on LFPW-P. (c) Error distribution tested on iBug-P.

TABLE 4
Proportion of Image Less than Given Relative Error Level on AR, MultiPIE, LFW, LFPW and AFW Comparing with Oxford Detector (Ox), ASM, Kernel Regression (KR), CLM, TSPM, RCPR and SDM

Method	AR			MultiPIE			LFW			LFPW			AFW		
	≤5%	≤10%	≤15%	≤5%	≤10%	≤15%	≤5%	≤10%	≤15%	≤5%	≤10%	≤15%	≤5%	≤10%	≤15%
Ox	0.72	0.97	0.99	0.48	0.71	0.80	0.67	0.88	0.94	0.68	0.89	0.95	0.18	0.33	0.54
ASM	0.59	0.79	0.87	0.29	0.53	0.68	0.38	0.67	0.80	0.48	0.73	0.83	0.37	0.62	0.75
KR	0.60	0.91	0.98	0.29	0.64	0.91	0.37	0.75	0.90	0.43	0.75	0.96	0.42	0.76	0.91
CLM	0.71	0.95	0.99	0.44	0.63	0.76	0.53	0.88	0.95	0.66	0.85	0.90	0.58	0.81	0.87
TSPM	0.69	0.97	0.99	0.77	0.95	0.99	0.39	0.87	0.97	0.72	0.95	0.97	0.60	0.93	0.98
RCPR	0.73	0.89	0.92	0.66	0.87	0.93	0.71	0.89	0.93	0.76	0.85	0.88	0.61	0.79	0.81
SDM	**0.82**	**0.98**	**0.99**	0.72	0.93	0.97	**0.76**	**0.96**	**0.99**	0.80	0.95	0.96	0.60	0.96	0.98
Proposed	0.81	0.97	0.99	**0.85**	**0.98**	**0.99**	0.71	0.96	0.99	**0.81**	**0.96**	**0.98**	**0.71**	**0.96**	**0.99**



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

Fig. 8. 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).

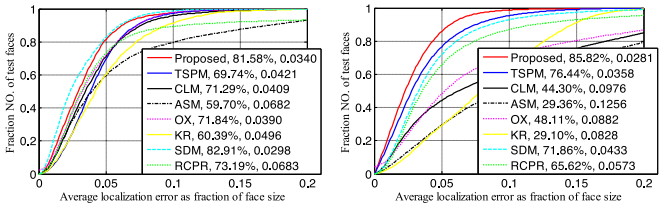
5.4 Comparison with Previous Work

We compare our approach (optimized mixtures with cascaded deformable shape model) with the following methods. (1) Multi-view ASMs [38], (2) Constrained local model [8], (3) Oxford facial landmark detector (OX) [34], (4) tree structure part model [29], (5) Kernel Regression [31], (6) Supervised Descent Method (SDM) [9] and (7) Robust Cascaded Pose Regression (RCPR) [18]. For wild faces, TSPM, RCPR and SDM 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 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 degree, in which case all landmarks are visible. For the relative error (Fig. 9a), our proposed method achieves top performance except 1.4 percent gap below SDM. In Fig. 9b, the proposed method shows superior performance with significant margin to other methods.

Quantitatively, we evaluate all the algorithms on percentage of database volume against the normalized error shown in Table 4. In AR and MultiPIE, our method stands in the top two performance with only 0.01 alignment gap at error 0.05 compared to SDM. While compared to all other methods, the advantage in alignment accuracy is significant.

Further investigation is focused on the performance of all the methods on LFW, LFPW and AFW. Fig. 8 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. Although giving advantage to those methods,



(a) AR (b) Near-frontal MultiPIE

Fig. 9. 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 percent of the pupil distance. (b) Error distribution tested on near-frontal MultiPIE database. The percentage is the ratio of error less than 5 percent of ground truth face size.

the proposed method achieves 71.0 percent of total face volume within relative error 5 percent on LFW, 81.1 percent fraction on LFPW and 71.4 percent on AFW, which consistently retains the localization accuracy in a very high level. Quantitative percentage results in Table 4 also supports the conclusion. One may notice that there is a small gap between SDM and our proposed method on LFW. One reason is that images in LFW are with less pose variations comparing to the other two wild face databases. The other thing is that SDM is trained based on LFW and MultiPIE. The smaller training and testing sample distribution gap within the same database may also result in the advantage.

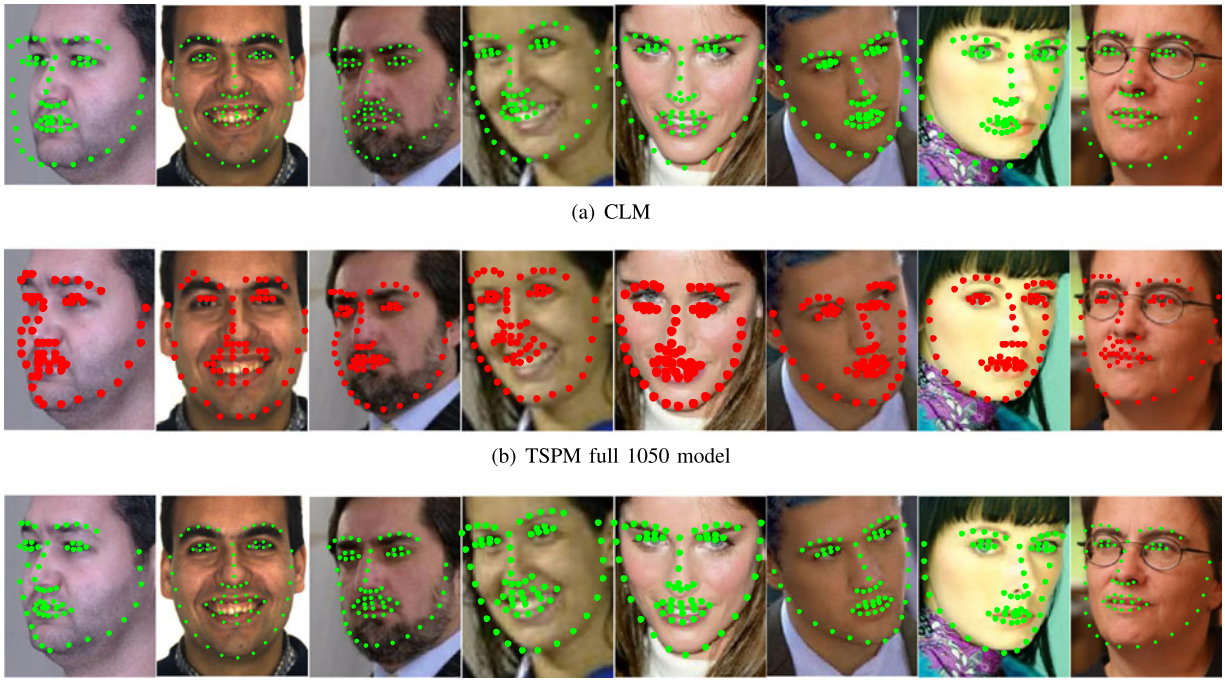
From visualization point of view, we present some localization results of TSPM (full independent 1050 part model), CLM (implemented by the ourselves) and our proposed method in Fig. 10. The CLM results pertain most of the landmarks in good positions. But for pose variation and subtle local variance, the output is not promising. TSPM can handle different kinds of head poses due to its multi-view model. However, its local shape constraint is too strong such that the holistic face shape is not precise. In contrast, the proposed method attempts to strike balance between the global shape constraint and the local shape constraint. Inside the bi-step procedure, the first step emphasizes on the global shape constraining while the second step aims at refining each facial component locally. By alternatively applying the two steps fitting, our method achieves state-of-the-art performance.

5.5 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 next frames.

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 5 shows that our method outperforms the other two methods with distinct margin. Visualization from Fig. 11 convinces our conclusion that the error by proposed method is consistently smaller than the other two methods.

Computational complexity. Our algorithm consists of three parts. (1) Optimized part mixtures. Restricting the part model tree structure, a dynamic programming and distance



(a) CLM

(b) TSPM full 1050 model

(c) proposed

Fig. 10. Visual comparison of CLM, TSPM with full 1,050 independent part model and our proposed method evaluated on MultiPIE, AR, LFW, LFPW and AFW databases. The first column is a test sample from MultiPIE. The second column is from AR database. The third and fourth columns are from LFW database. The fifth and sixth columns are with LFPW images and the last two columns are from AFW dataset. (a) Localization result by CLM. (b) Localization result by Tree Structure Part Model with full independent 1,050 parts model which achieves the highest accuracy among all its models. (c) Localization result from proposed method.

TABLE 5
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**

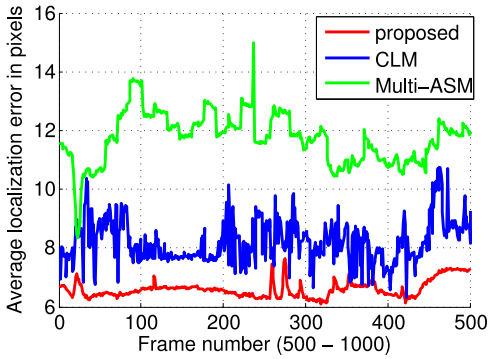


Fig. 11. Average landmark tracking error in pixels of talking face video from frame 500 to frame 1,000.

transform strategy [37] is used in pursuing Eq. (9). It achieves $O(Nh)$ running time, where N is the number of landmarks and h is grid size defined in distance transform bounded by image size. (2) Local patch mean-shift. Assuming response map size ρ, the running time is $O(N\rho)$. (3) Component-wise active contour. The component number is a constant. For each component, we should evaluate ds, ds^2 and $exp(-d) + log(1+d)$. Each takes $O(N)$. With k iterations, the running time is $O(Nk)$. Overall, our algorithm achieves $O(N(h+\rho+k)) \approx O(Nh)$, which is because in practice, $k \leq 3$ and $h \gg \rho$. Comparing TSPM running time $O(Nh)$, CLM with $O(N\rho)$ and ASM with $O(N\rho)$ (assuming the same patch size ρ for CLM and ASM), our algorithm is at the same running time level of those real-time methods. Typically, our N is a quarter of that in TSPM, which leads to the result that our method is at least two times faster than TSPM as shown in Table 1.

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.

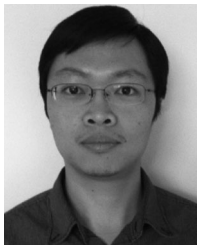
Acknowledgments

This paper is based upon work supported in part by National Science Foundation: Information and Intelligent Systems (IIS)-1064965, IIS-1451292, IIS-1423056, Industrial Innovation and Partnerships (IIP)-1069258, Division of Graduate Education (DGE)-0549115, Civial Mechanical and Manufacturing Innovation (CMMI)-1434401 and Computer, Network Systems (CNS)-1405985 and National Aeronautics and Space Administration (NCC) 9-58. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of the U.S. Government. J. Huang is the corresponding author.

References

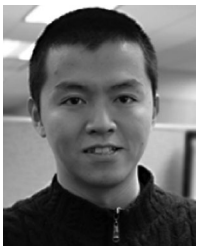
[1] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, pp. 399–458, Dec. 2003.
[2] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.
[3] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.
[4] I. Mpiperis, S. Malassiotis, and M. Strintzis, "Bilinear elastically deformable models with application to 3d face and facial expression recognition," in *Proc. IEEE 8th Int. Conf. Autom. Face Gesture Recog.*, 2008, pp. 1–8.
[5] V. Bettadapura, "Face expression recognition and analysis: The state of the art," in *Arxiv*, 2012.
[6] I. Kemelmacher-Shlizerman and R. Basri, "3D face reconstruction from a single image using a single reference face shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 394–405, Feb. 2011.
[7] F. Yang, L. Bourdev, E. Shechtman, J. Wang, and D. Metaxas, "Facial expression editing in video using a temporally-smooth factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 861–868.
[8] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, pp. 200–215, 2011.
[9] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 532–539.
[10] M. Jones and P. Viola, "Fast multi-view face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2003.
[11] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," Microsoft Res., Redmond, WA, USA, Tech. Rep., 2010.
[12] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, "Efficient boosted exemplar-based face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1843–1850.
[13] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proc. 5th Eur. Conf. Comput. Vis.*, 1998, pp. 484–498 .
[14] F. Yang, J. Huang, and D. Metaxas, "Sparse shape registration for occluded facial feature localization," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshops*, 2011, pp. 272–277.
[15] M. Roh, T. Oguri, and T. Kanade, "Face alignment robust to occlusion," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2011, pp. 239–244.
[16] T. Cootes and C. Taylor, "A mixture model for representing shape variation," in *Proc. Brit. Mach. Vis. Conf.*, 1997.
[17] X. Yu, Z. Lin, J. Brandt, and D. Metaxas, "Consensus of regression for occlusion-robust facial feature localization," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 105–118.
[18] X. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1513–1520.
[19] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1899–1906.
[20] Y. Zhou, W. Zhang, X. Tang, and H. Shum, "A Bayesian mixture model for multi-view face alignment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 741–746.
[21] 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 *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–7.

[22] E. M. Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.
[23] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models—their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[24] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, 2004.
[25] D. Cristinacce and T. Cootes, "Automatic feature localization with constrained local models," *Pattern Recog.*, vol. 41, pp. 3054–3067, 2008.
[26] Y. Wang, S. Lucey, and J. Cohn, "Enforcing convexity for improved alignment with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[27] D. Cristinacce and T. Cootes, "Boosted regression active shape models," in *Proc. Brit. Mach. Vis. Conf.*, 2007, pp. 880–889.
[28] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2887–2894.
[29] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2879–2886.
[30] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1944–1951.
[31] S. Rivera and A. Martinez, "Learning deformable shape manifolds," *Pattern Recog.*, vol. 45, no. 4, pp. 1792–1801, Apr. 2012.
[32] Z. Kalal, J. Matas, and K. Mikolajczyk, "Weighted sampling for large-scale boosting," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 42.1–42.10.
[33] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2009.
[34] M. Everingham, J. Sivic, and A. Zisserman, "'Hello! my name is... buffy'—automatic naming of characters in tv video," in *Proc. 17th Brit. Mach. Vis. Conf.*, 2006.
[35] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 326–339.
[36] M. Uricar, V. Franc, and V. Hlavac, "Detector of facial landmarks learned by the structured output SVM," in *Proc. VISAPP*, 2012.
[37] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, pp. 55–79, 2003.
[38] Y. Huang, Q. Liu, and D. Metaxas, "A component based deformable model for generalized face alignment," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[39] X. Liu, "Generic face alignment using boosted appearance model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[40] T. Cootes, M. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using random forest regression voting," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 278–291.
[41] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 593–600.
[42] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey, "Rank minimization across appearance and shape for AAM ensemble fitting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 577–584.
[43] A. Asthana, S. Cheng, S. Zafeiriou, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3444–3451.
[44] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 545–552.
[45] B. Smith and L. Zhang, "Joint face alignment with non-parametric shape models," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 43–56.
[46] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1025–1032.
[47] H. Gao, H. Ekenel, and R. Stiefelhagen, "Face alignment using a ranking model based on regression trees," in *Proc. Brit. Mach. Vis. Conf.*, 2012.
[48] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2729–2736.
[49] B. Martinez, M. Valstar, X. Binefa, and M. Pantic, "Local evidence aggregation for regression-based facial point detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1149–1163, May 2013.
[50] M. Dantone, J. Gall, G. Fanelli, and L. Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2578–2585.
[51] P. Dollar, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1078–1085.
[52] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 72–85.
[53] G. Huang, M. Ramesh, T. Berg, and E. Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.
[54] A. Martinez and R. Benavente, "The AR face database," CVC Tech. Rep. 24, 1998.
[55] J. Liu, S. Ji, and J. Ye. (2009). *SLEP: Sparse Learning with Efficient Projections*. Arizona State Univ., Tempe, AZ, USA [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP
[56] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, pp. 807–813, 2010.
[57] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recog.*, 2006, pp. 211–216.
[58] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2013, pp. 397–403.
[59] [Online]. Available: http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html



Xiang Yu received the BE degree from the Electronic Engineering Department, Tsinghua University, in 2007, the MS degree from the Electronic Engineering Department, Tsinghua University, Beijing, China, in 2010, and the PhD degree from the Computer Science Department, Rutgers University on September 2015. He is a senior associate researcher at NEC Laboratories America, Media Analytics Department. His research interests include computer vision and machine learning, especially on face alignment, tracking, object pose analysis, extended facial, and biometric analysis. He is a member of the IEEE.

Junzhou Huang received the BE degree from Huazhong University of Science and Technology, Wuhan, China, in 1996, the MS degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2003, and the PhD degree from Rutgers University, New Brunswick, New Jersey, in 2011. He is an assistant professor in the Computer Science and Engineering Department, University of Texas at Arlington. His research interests include biomedical imaging, machine learning and computer vision, with focus on the development of sparse modeling, imaging, and learning for large scale inverse problems. He is a member of the IEEE.



Shaoting Zhang received the BE degree from Zhejiang University in 2005, the MS degree from Shanghai Jiao Tong University in 2007, and the PhD degree in computer science from Rutgers in January 2012. He is an assistant professor in the Department of Computer Science, University of North Carolina at Charlotte. Before joining UNC Charlotte, he was a faculty member in the Department of Computer Science, Rutgers-New Brunswick (research assistant professor, 2012-2013). His research is on the interface of medical imaging informatics, large-scale visual understanding, and machine learning. He is a senior member of the IEEE.

Dimitris N. Metaxas received the BE degree from the National Technical University of Athens Greece in 1986, the MS degree from the University of Maryland in 1988, and the PhD degree from the University of Toronto in 1992. He is a professor in the Computer Science Department, Rutgers University. He is directing the Computational Biomedicine Imaging and Modeling Center (CBIM). He has been conducting research toward the development of formal methods upon which computer vision, computer graphics, and medical imaging can advance synergistically. He is a fellow of the IEEE.