This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

PIEFA: Personalized Incremental and Ensemble Face Alignment

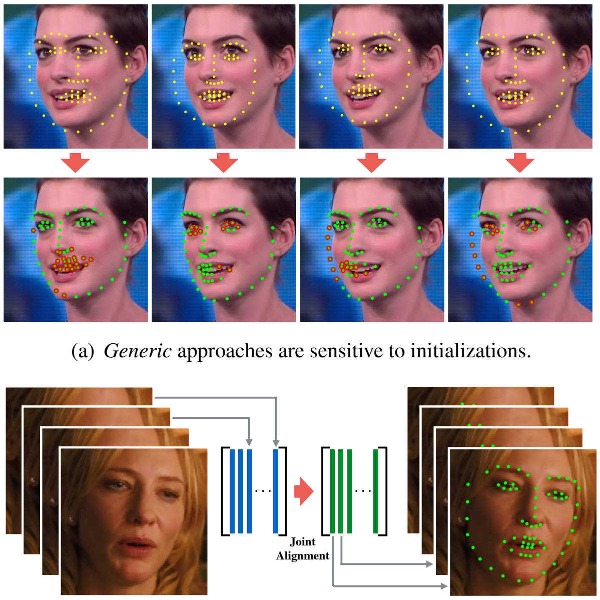
Xi Peng* Shaoting Zhang† Yang Yu* Dimitris N. Metaxas*

*Rutgers University
Piscataway, NJ, 08854
xpeng.nb,yyu,dnm@cs.rutgers.edu

†The University of North Carolina at Charlotte
Charlotte, NC, 28223
szhang16@uncc.edu

Abstract

Face alignment, especially on real-time or large-scale sequential images, is a challenging task with broad applications. Both generic and joint alignment approaches have been proposed with varying degrees of success. However, many generic methods are heavily sensitive to initializations and usually rely on offline-trained static models, which limit their performance on sequential images with extensive variations. On the other hand, joint methods are restricted to offline applications, since they require all frames to conduct batch alignment. To address these limitations, we propose to exploit incremental learning for personalized ensemble alignment. We sample multiple initial shapes to achieve image congealing within one frame, which enables us to incrementally conduct ensemble alignment by group-sparse regularized rank minimization. At the same time, personalized modeling is obtained by subspace adaptation under the same incremental framework, while correction strategy is used to alleviate model drifting. Experimental results on multiple controlled and in-the-wild databases demonstrate the superior performance of our approach compared with state-of-the-arts in terms of fitting accuracy and efficiency.



(a) *Generic* approaches are sensitive to initializations.

(b) *Joint* approaches are restricted to offline batch alignment.

Figure 1: Limitations of existing methods. Yellow points: different initial shapes. Green points: well-aligned landmarks. Red points: mis-aligned landmarks.

1. Introduction

Recently, analysing image sequences in large-scale and unconstrained conditions attracts increasing interest in computer vision community [40]. In the context of face-related topics, sequential face alignment, *i.e.*, fitting facial landmarks on consecutive frames, is a crucial task with a wide range of applications, such as Face Verification [31], Facial Action Unit (FAU) analysis [35] and Human-Computer Interaction (HCI) [25]. It is challenging due to the extensive rigid and non-rigid variations of human faces [15], as well as unconstrained imaging conditions such as illumination changes and occlusions.

It has been shown that either *generic face alignment* [6, 23, 28, 30, 33, 38], aligning each frame independently in a tracking-by-detection manner by various facial landmark detectors [10, 22], or *joint face alignment* [7, 8, 27, 36], aligning all frames simultaneously in a batch manner, can be employed for sequential face alignment.

Generic face alignment starts the fitting process from an initial shape, *e.g.*, a mean face [6, 33] or the result of the last frame [1, 28]), and deform the shape constrained by facial deformable models (FDMs) to minimize the reconstruction residual by either gradient descend optimization [28, 30] or cascade/boosted regression [6, 33]. They have shown great success on single image with respect to the efficiency, *e.g., face alignment at 3000 FPS* [23], and unconstrained scenarios, *e.g., face alignment in the wild* [33]. However, they have significant limitations when applied to dynamic streams with extensive variations: **(1)** They lack the capability to capture the personalized information and imaging continuity in consecutive frames, due to their reliance on offline-trained static FDMs and thus the difficulty of incorporating any motion information. **(2)** Many of them are heavily sensitive to initializations, as illustrated in Figure 1(a), since both gradient descend and regression-based methods may be trapped in local optima when starting from poor initial guess.

Joint face alignment, on the other hand, take the advantage of the shape and appearance consistency to simultaneously minimize fitting errors for all frames [7, 8, 27, 36]. They are more robust to illumination changes and partial occlusions than generic methods [8]. However, they still have limitations in two aspects. **(1)** Most of them can only handle offline tasks as they require all frames to conduct batch alignment, as illustrated in 1(b), which severely impedes their applications on either real-time or large-scale tasks. **(2)** Some of them attempt to achieve personalized modeling without correction, which may inevitably result in drifting during the model update.

In this paper, we propose personalized incremental ensemble alignment to address aforementioned issues. Instead of the single initialization, we incorporate motion information to sample multiple initial shapes and conduct generic alignment in parallel at each frame. The image congealing is then achieved within one frame, which enables the ensemble alignment to be performed in an incremental manner by constrained robust decomposition. At the same time, personalized modeling is achieved by subspace adaptation under the same incremental framework, while correction strategy is used to alleviate model drifting. To sum up, our main contributions are the follows:

- To the best of our knowledge, this is the first work to address initialization-sensitivity issue of generic methods by ensemble initialization with motion models.
- The proposed incremental framework is radically different from existing joint alignment with respect to the congealing manner, *i.e.*, intra- *v.s.* inter-frame.
- The proposed group-sparse regularized rank minimization is well designed for incremental framework, and guarantees robust personalized modeling on the fly.

By conducting extensive experiments on multiple public and unconstrained databases, we show that our approach has significant accuracy improvement compared with state-of-the-arts, while constant computational cost w.r.t. both CPU time and memory usage is guaranteed. These merits make our approach very suitable for real-time and large-scale applications.

2. Related Work

We briefly reviews both *generic* and *joint* approaches in this section. Based on the different FDMs employed, existing face alignment approaches can be categorized as methods based on either *holistic* models, *e.g.*, *active appearance models* (AAMs) [9], or *part-based* models, *e.g.*, *constrained local models* (CLMs) [28].

Among *generic face alignment* approaches, part-based FDMs combined with regression-based fitting strategies attract intensive interest. For instance, Cao *et al.* [6] achieved *explicit shape regression* (ESR) by combining shape-indexed feature selection and multi-layer boosted regression. Xiong *et al.* [33] proposed *supervised descent method* (SDM) for fast optimization by concatenating SIFT features and applying cascade non-linear regression. Although they have shown great success on single image [23], however, the static FDMs and initialization-sensitivity issue severely limited their performance on streaming data with extensive variations.

Multiple efforts were devoted to address these limitations. For instance, Asthana *et al.* [2] improved SDM by updating cascade regressors in parallel for *incremental face alignment* (IFA). Yan *et al.* [34] proposed to rank and *combine multiple hypotheses* (CMH) in a structural SVM framework to address the initialization-sensitivity issue. Zhu *et al.* [39] proposed to search the best initial shape in a coarse-to-fine manner. However, the temporal constraints, w.r.t. personalized shape, appearance and motion cues in sequential inputs, are hardly investigated in these approaches.

To this end, *joint face alignment* approaches, which take the advantage of consistency constraints to minimize fitting errors for all frames, are mainly applied. For instance, Zhao *et al.* [36] proposed to regularize the holistic texture by enforcing all frames to lie in a low-rank subspace. The drawback of this method is that it did not incorporate any face prior, which may result in arbitrary deformations during joint optimization. To address this problem, Cheng *et al.* [7] proposed to use anchor shapes to penalize arbitrary deformations; while Sagonas *et al.* [27] proposed to employ a clean face subspace trained offline to restrict optimization directions. The most prominent limitation of these joint methods is that they can hardly handle real-time or large-scale applications since they lack the capability to incrementally utilize consecutive information.

More recently, Zhang *et al.* [37] proposed to use dictionary learning to achieve sparse representations for rigid object tracking [32]. However, it is nontrivial to apply dictionary learning for sequential face alignment as face may undergo extensive non-rigid deformations. Moreover, It remains a challenging task to simultaneously address the initialization-sensitivity issue and adapt FDMs for person-specific modeling in a unified framework.

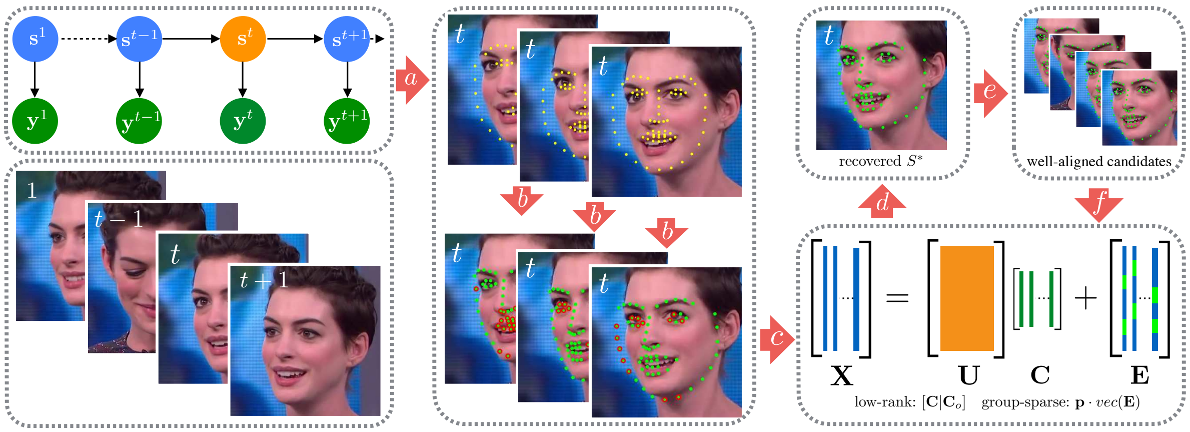


Figure 2: Overview of our approach: (a) Ensemble Initialization (3.1), (b) generic face alignment in parallel, (c) Constrained Decompostion (3.3), (d) Fitting Recovery (3.4), (e) fitting evaluation and (f) Personlized Adaptation (3.5).

3. Proposed Approach

In this paper, we propose a novel approach for sequential face alignment. To address the sensitivity issue, we incorporate motion model to sample multiple initial shapes at each frame, *i.e.*, *Ensemble Initialization*. Then we employ off-the-shelf generic approach to conduct batch face alignment in parallel. The alignment results are re-organized in a part-based manner, *i.e.*, *Part-based Representation*. By conducting low-rank and group-sparse optimization, *i.e., Constrained Decomposition*, we can recover the best fitting S^* from the part-based representation, *i.e.*, *Fitting Recovery*. Finally, personalized modeling is achieved by robust subspace adaptation, *i.e.*, *Personalized Adaptation*. Please refer to Figure 2 for an overview of our approach.

3.1. Ensemble Initialization

Initialization is the first and key step in landmark localization. It is easy to get a landmark correctly aligned if it is initialized closely to the ground-truth. This fact motivates us to incorporate Bayesian motion models [11] to sample multiple initial shapes for ensemble initialization.

Let $\mathbf{s}$ denote the latent state, *i.e.*, the scale, rotation, translation and deformation of initial shapes, $\mathbf{y}$ denote the observation, *i.e.*, fitting results, we can sample an ensemble of particles, *i.e.*, initial shapes, at time t from the prediction:

$$p(\mathbf{s}^t|\mathbf{y}^{1:t-1}) = \int q(\mathbf{s}^t|\mathbf{s}^{t-1})p(\mathbf{s}^{t-1}|\mathbf{y}^{1:t-1})d\mathbf{s}^{t-1}, \quad (1)$$

where $q(\mathbf{s}^t|\mathbf{s}^{t-1})$ is the state transition probability, and the integration can be approximated by efficient *Markov chain Monte Carlo* (MCMC) sampling [21]. The posterior state distribution is then updated at time t by:

$$p(\mathbf{s}^t|\mathbf{y}^{1:t}) \propto p(\mathbf{y}^t|\mathbf{s}^t)p(\mathbf{s}^t|\mathbf{y}^{1:t-1}), \quad (2)$$

where $p(\mathbf{y}^t|\mathbf{s}^t)$ is the observation model, which is the key component that evaluates the goodness of the corresponding initial shape. We model it using group-sparse fitting errors and introduce the details in Section 3.4.

This motion model guarantees that more initial shapes with higher weights are sampled near the optimum, which can effectively overcome the sensitivity issue. More importantly, the ensemble makes it possible to conduct joint alignment in an incremental manner since the congealing can be achieved within the same frame.

3.2. Part-based Representation

Once K initial shapes are sampled, we can employ an off-the-shelf generic face alignment approach, *e.g.,* ESR [6] and SDM [33], to obtain a batch of erroneous fittings $\{S_1, \ldots, S_K\}$. It is worth noting that the efficiency is guaranteed since generic approaches are highly efficient [23] and we can conduct batch alignments in parallel.

To conduct batch alignment, former approaches [7, 27, 36] usually use holistic FDMs to parameterize the shape and appearance separately and bridge the two by image warping [9]. Apart from the very time-consuming warping operations, this representation is susceptible to occlusions and illumination changes due to the limitations of holistic FDMs.

We propose a new part-based representation to jointly depict the shape and appearance:

$$\mathbf{A} = [(\mathbf{x}_1 - \bar{\mathbf{x}})^T \ \mathbf{f}(\mathbf{x}_1)^T \ldots (\mathbf{x}_L - \bar{\mathbf{x}})^T \ \mathbf{f}(\mathbf{x}_L)^T]^T,$$

where $(\mathbf{x}_1 - \bar{\mathbf{x}}) \in \mathbb{R}^2$ are centralized landmark coordinates, $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d$ are feature vector extracted from the image patch centered at $\mathbf{x}$. This part-based representation is extremely fast to compute. The direct concatenation of the landmark coordinates and feature vectors can greatly facilitate the constrained decomposition in the next section.

3.3. Constrained Decomposition

The next goal is to recover the best fitting S^* from $\{S^1, \dots, S^K\}$. We propose a constrained decomposition based on the following facts and observations. **(a)** Each of $\{S^1, \dots, S^K\}$ is aligned to the same face but with fitting errors. **(b)** With respect to the k^{th} shape, only a small number of its landmarks are misaligned. **(c)** With respect to the l^{th} landmark, only a small number of shapes are misaligned.

Low-Rank Representation Constraint. Let $\mathbf{U} \in \mathbb{R}^{N\times M}$ denote an orthogonal subspace learned from annotated training images, $\mathbf{X} = [\mathbf{A}^1, \dots, \mathbf{A}^K] \in \mathbb{R}^{N\times K}$ denote the batch observation matrix. Based on the observation **(a)** we have the following low-rank constraint:

$$\arg\min_{\mathbf{C},\mathbf{E}} \mathbf{rank}(\mathbf{C}), \text{ s. t. } \mathbf{X} = \mathbf{U}\mathbf{C} + \mathbf{E}, \tag{3}$$

where $\mathbf{C} \in \mathbb{R}^{M\times K}$ is the encoding matrix, $\mathbf{E} \in \mathbb{R}^{N\times K}$ is the error matrix. In the ideal case, $\mathbf{rank}(\mathbf{C}) = 1$, since all columns represent the same face. However, the correct fitting is not unique, *e.g.*, profile landmarks remain well-aligned even when they move a little along the face contour. Therefore, we seek for rank minimization for robustness.

In experiments, we find that only using encodings of aligned shapes in current frame may cause the recovered S^* to deform arbitrarily in certain cases. To address this problem, we incorporate prior knowledge for temporal consistency in the low-rank constraint. That is, we minimize $\mathbf{rank}([\mathbf{C}|\mathbf{C}_o])$ instead of $\mathbf{rank}(\mathbf{C})$, where $\mathbf{C}_o \in \mathbb{R}^{M\times K_o}$ are the encodings of well-aligned candidates from tracked frames. We set $K_o = K/10$ in our experiments.

Group-Sparse Error Constraint. Owing to the special-designed part-based representation, the error matrix $\mathbf{E}$ in Equation 3 has the group structure:

$$\begin{aligned} \mathbf{E} &= \begin{bmatrix} \epsilon_1^1 & \cdots & \epsilon_1^K \\ \vdots & \ddots & \vdots \\ \epsilon_L^1 & \cdots & \epsilon_L^K \end{bmatrix}, \\ vec(\mathbf{E}) &= \left[\epsilon_1^1, \dots, \epsilon_1^K, \dots, \epsilon_L^1, \dots, \epsilon_L^K\right], \end{aligned} \tag{4}$$

where $\epsilon_l^k \in \mathbb{R}^{2+d}$ is the fitting errors of the l^{th} landmark in the k^{th} shape. $vec(\cdot)$ performs block-wise vectorization.

According to observation **(b)-(c)**, the nonzero entries of $\mathbf{E}$ should be sparse with respect to both columns and rows, which is equivalent to the group-sparse constraint:

$$\arg\min_{\mathbf{C},\mathbf{E}} \|\mathbf{p} \cdot vec(\mathbf{E})\|_{2,0}, \text{ s. t. } \mathbf{X} = \mathbf{U}\mathbf{C} + \mathbf{E}, \tag{5}$$

where $\mathbf{p} = \begin{bmatrix} \mathbf{I}_{2\times 2} \otimes \rho & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d\times d} \end{bmatrix}$ balances the error contributions between the shape and appearance. ρ is the mean ratio from feature vectors to centralized landmark coordinates.

Robust Decomposition. Given the constraints in Equation 3 and 5, we can achieve the object function for robust decomposition:

$$\begin{aligned} \arg\min_{\mathbf{C}_n,\mathbf{E}_v,\mathbf{C},\mathbf{E}} \quad & \|\mathbf{Z}\|_F^2 + \lambda_1 \mathbf{rank}(\mathbf{C}_n) + \lambda_2 \|\mathbf{E}_v\|_{2,0} \\ \text{subject to} \quad & \mathbf{Z} = \mathbf{X} - \mathbf{U}\mathbf{C} - \mathbf{E}, \\ & \mathbf{C}_n = [\mathbf{C}|\mathbf{C}_o], \; \mathbf{E}_v = \mathbf{p} \cdot vec(\mathbf{E}). \end{aligned} \tag{6}$$

where $\lambda_1, \; \lambda_2$ are non-negative parameters to balance contributions between the two constraints. We present an efficient solution for the optimization in Section 4.

3.4. Fitting Recovery

The error matrix $\mathbf{E}$ guarantees robust decomposition against outliers such as illumination changes and partial occlusions [24]. More importantly, we can recover a well-aligned S^* from $\{S^1, \dots, S^K\}$ by investigating its group-sparse structure.

Equation 4 indicates that ϵ_l^k measures the fitting errors of S^k at the l^{th} landmark. Therefore, each row of $\mathbf{E}$ models the errors distribution of all aligned shapes with respect to the same indexed landmark. Let $S^* = \{x_1^*, \dots, x_L^*\}$, $\mathbf{x}_l^*$ can be recovered by using the intra-row ℓ_2-norm of $\mathbf{E}$ to weight the same indexed landmark of all aligned shapes:

$$\mathbf{x}_l^* = \frac{1}{\mathbf{q}_l} \Sigma_{k=1}^K e^{-\|\epsilon_l^k\|_2} \mathbf{x}_l^k, \text{ where } \mathbf{q}_l = \Sigma_{k=1}^K e^{-\|\epsilon_l^k\|_2}. \tag{7}$$

Besides the row structure, we also investigate the column structure of $\mathbf{E}$ to present the observation model $p(\mathbf{y}^t|\mathbf{s}^t)$ of Equation 2. Considering the fact that the k^{th} column of $\mathbf{E}$ measures the overall fitting errors of S^k, we can use inter-column ℓ_2-norm of $\mathbf{E}$ to represent the observation model:

$$p(\mathbf{y}^{t,k}|\mathbf{s}^t) = \frac{e^{-\mathbf{r}_k}}{\Sigma_{k=1}^K e^{-\mathbf{r}_k}}, \text{ where } \mathbf{r}_k = \Sigma_{l=1}^L \|\epsilon_l^k\|_2. \tag{8}$$

We compute $p(\mathbf{y}^{t,k}|\mathbf{s}^t)$ for each aligned shape at frame t, and apply Equation 1 to predict the latent state for ensemble initialization in frame $t + 1$.

3.5. Personalized Adaptation

The offline trained $\mathbf{U}$ has limited representation power to capture extensive online variations especially in wild conditions, which motivates us to incrementally update $\mathbf{U}$ for personalized modeling.

Given S^* recovered, we first extract the part-based representation $X^* \in \mathbb{R}^N$, and then compute $C^* \in \mathbb{R}^M$ and $E^* \in \mathbb{R}^N$ by robust decomposition:

$$\arg\min_{C^*,E^*} \|\mathbf{p} \cdot vec(E^*)\|_{2,0}, \text{ s.t. } X^* = \mathbf{U}C^* + E^*, \tag{9}$$

which can be efficiently solved using the same ALM optimization in Section 4 by introducing the augmented Lagrangian $\mathcal{L}^* = \|\mathbf{p} \cdot vec(E^*)\|_{2,0} + \mathbf{Y}^T(X^* - \mathbf{U}C^* - E^*) +$

$\frac{\mu}{2}\|X^* - \mathbf{U}C^* - E^*\|_F^2$, where $\mathbf{Y}$ and μ are Lagrange multiplier and penalty parameter.

To efficiently update $\mathbf{U}$, we adopt the concept of incremental subspace adaptation on Grassmannian [17], which in our case is a Riemannian manifold of all subspaces of $\mathbb{R}^N$ with fixed dimension M. The key step is to specify the gradient along the geodesic of Grassmannian:

$$\frac{d\mathcal{L}^*}{d\mathbf{U}} = \chi_\pi \left[\mathbf{Y}C^{*T} - \mu(X^* - \mathbf{U}_\pi C^* - E^*)C^{*T}\right], \quad (10)$$

where χ_π and $\mathbf{U}_\pi$ are the $|\pi|$ columns of an identity matrix and $\mathbf{U}$, respectively [13]. Due to the space limitation, we directly present the final result of incremental update step:

$$\Delta\mathbf{U} = \left[(cos(\psi) - 1)\frac{\mathbf{U}C^*}{\|C^*\|} - sin(\psi)\frac{\Omega}{\|\Omega\|}\right]\frac{C^{*T}}{\|C^*\|}, \quad (11)$$

where $\|\Omega\| = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)(Y^* + \mu^* h^*)$, $\psi = \eta\|\Omega\|\|C^*\|$ and η is the gradient step.

The incremental subspace adaption is well suited for our intra-frame ensemble alignment framework. It is highly efficient and takes only $\mathcal{O}(M^2)$ operations. Moreover, unlike former personalized approach that blindly updates the subspace without effective correction strategy [29], we can utilize $\mathbf{E}$ and E^* to obtain a robust criterion for erroneous detection, which can significantly alleviate model drifting.

We use $\|\mathbf{E}\|_2 = \Sigma_{k=1}^K \Sigma_{l=1}^L \|\epsilon_l^k\|_2$ to model the prior confidence to recover S^* since it measures the overall fitting errors of all aligned shapes. Similarly, $\|E^*\|_2 = \Sigma_{l=1}^L \|\epsilon_l^*\|_2$ measures the posterior fitting errors given S^* recovered. Therefore we can achieve a robust criterion to distinguish well and erroneous fittings:

$$max(\|\mathbf{E}\|_2, K\|E^*\|_2) < \tau, \quad (12)$$

where K is the number of sampled initial shapes and threshold τ can be computed from training images. S^* that satisfy Equation 12 are well-aligned candidates. We use them to compose $\mathbf{C}_o$ for consistency constraint and update $\mathbf{U}$ for drift-free personalized modeling.

4. ALM Optimization

Directly minimizing $rank(\cdot)$ and $\ell_{2,0}$-norm in Equation 6 is NP-hard [24]. Therefore, we reformulate the optimization with relaxed ℓ_*-norm and $\ell_{2,1}$-norm, respectively:

$$\begin{aligned} \underset{\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}}{\arg\min} \quad & \|\mathbf{C}_n\|_* + \lambda\|\mathbf{E}_v\|_{2,1} \\ \text{subject to} \quad & \mathbf{X} = \mathbf{U}\mathbf{C} + \mathbf{E}, \\ & \mathbf{C}_n = [\mathbf{C}|\mathbf{C}_o]\,,\ \mathbf{E}_v = \mathbf{p} \cdot vec(\mathbf{E}). \end{aligned} \quad (13)$$

The intermediate variables $\mathbf{C}_n$ and $\mathbf{E}_v$ allows us to efficiently solve the Equation 6 with *Augmented Lagrange Multiplier* (ALM) method [20].

Algorithm 1 Alternating Optimization of Equation 13

```
Input: $\mathbf{X}$, $\mathbf{U}$, $\mathbf{C}_o$, $\mathbf{p}$, $\lambda$, $\gamma$
Output: $\mathbf{C}_n$, $\mathbf{E}_v$, $\mathbf{C}$, $\mathbf{E}$
Initialize: $\mathbf{C} = \mathbf{0}$, $\mathbf{C}_n = [\mathbf{C}|\mathbf{C}_o]$, $\mathbf{E} = \mathbf{0}$,
$\mathbf{E}_v = \mathbf{p} \cdot vec(\mathbf{E})$, $\mathbf{Y}_{1-3} = \mathbf{0}$, $\mu_{1-3} = 0$.
while not converged do
    $\mathbf{C}_n \leftarrow \underset{\mathbf{C}_n}{\arg\min}\, \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
        $\Rightarrow \mathbf{C}_n^* = \mathcal{J}_{\frac{1}{\mu_2}}\Big[\,[\mathbf{C}|\mathbf{C}_o] + \frac{1}{\mu_2}\mathbf{Y}_2\Big]$,
    $\mathbf{E}_v \leftarrow \underset{\mathbf{E}_v}{\arg\min}\, \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
        $\Rightarrow \mathbf{E}_v^* = \mathcal{L}_{\frac{\lambda}{\mu_3}}\Big[\mathbf{p} \cdot vec(\mathbf{E}) + \frac{1}{\mu_3}\mathbf{Y_3}\Big]$,
    $\mathbf{C} \leftarrow \underset{\mathbf{C}}{\arg\min}\, \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
        $\Rightarrow \mathbf{C}^* = \Lambda_1\Big[\mathbf{M} + \frac{1}{\mu_1}(\mathbf{U}^T\mathbf{Y}_1 - [\mathbf{Y}_2]_{1:K})\Big]$,
        $where\ \Lambda_1 = (1 + \frac{\mu_2}{\mu_1})^{-1}\mathbf{I}$,
            $\mathbf{M} = \mathbf{U}^T(\mathbf{X} - \mathbf{E}) + \frac{\mu_2}{\mu_1}\,[\mathbf{C}_n]_{1:K}$ ,
    $\mathbf{E} \leftarrow \underset{\mathbf{E}}{\arg\min}\, \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
        $\Rightarrow \mathbf{E}^* = \Lambda_2\Big[\mathbf{W} + \frac{1}{\mu_1}\left(\mathbf{U}^T\mathbf{Y}_1 - vec^{-1}(\mathbf{Y}_3)\right)\Big]$,
        $where\ \Lambda_2 = (1 + \frac{\mu_3}{\mu_1})^{-1}\mathbf{I}$,
            $\mathbf{W} = \mathbf{X} - \mathbf{U}\mathbf{C} + \frac{\mu_3}{\mu_1}\left[vec^{-1}(\mathbf{p}^{-1}\mathbf{E}_v)\right]$ ,
    $\mathbf{Y}_1 \leftarrow \mathbf{Y}_1 + \mu_1(\mathbf{X} - \mathbf{U}\mathbf{C} - \mathbf{E})$,
    $\mathbf{Y}_2 \leftarrow \mathbf{Y}_2 + \mu_2([\mathbf{C}|\mathbf{C}_o] - \mathbf{C}_n)$,
    $\mathbf{Y}_3 \leftarrow \mathbf{Y}_3 + \mu_3(vec(\mathbf{E}) - \mathbf{E}_v)$,
    $\mu_1 \leftarrow \gamma\mu_1,\ \mu_2 \leftarrow \gamma\mu_2,\ \mu_3 \leftarrow \gamma\mu_3$.
end while
```

Let $\mathbf{Y}_{1-3}$ and μ_{1-3} denote the Lagrange multipliers and non-negative penalty parameters respectively, the augmented Lagrangian can be written in a scaled form:

$$\begin{aligned} \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3}) &= \|\mathbf{C}_n\|_* + \lambda\|\mathbf{E}_v\|_{2,1} \\ &+ \mathbf{Y}_1^T(\mathbf{X} - \mathbf{U}\mathbf{C} - \mathbf{E}) + \frac{\mu_1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{C} - \mathbf{E}\|_F^2 \\ &+ \mathbf{Y}_2^T([\mathbf{C}|\mathbf{C}_o] - \mathbf{C}_n) + \frac{\mu_2}{2}\|\,[\mathbf{C}|\mathbf{C}_o] - \mathbf{C}_n\|_F^2 \\ &+ \mathbf{Y}_3^T(\mathbf{p} \cdot vec(\mathbf{E}) - \mathbf{E}_v) + \frac{\mu_3}{2}\|\mathbf{p} \cdot vec(\mathbf{E}) - \mathbf{E}_v\|_F^2. \end{aligned} \quad (14)$$

The augmented Lagrangian function can be optimized by solving each of the variables alternately until converges. The complex objective is broken into a sequence of least-squares problems with efficient closed-form solutions. We briefly describe the alternating optimization in Algorithm 1, where the soft shrinkage operators are defined as:

$$\begin{cases} \mathcal{J}_\tau(x) = U\left[sgn(x)\left[|x| - \tau\right]_+\right]V^T, \\ \mathcal{L}_\tau(\mathbf{x}) = \mathbf{x}\left[1 - \dfrac{\tau}{\|\mathbf{x}\|_2}\right]_+. \end{cases} \quad (15)$$

The alternative iterations in Algorithm 1 converge very fast with quadratic rate [3]. The computational bottleneck

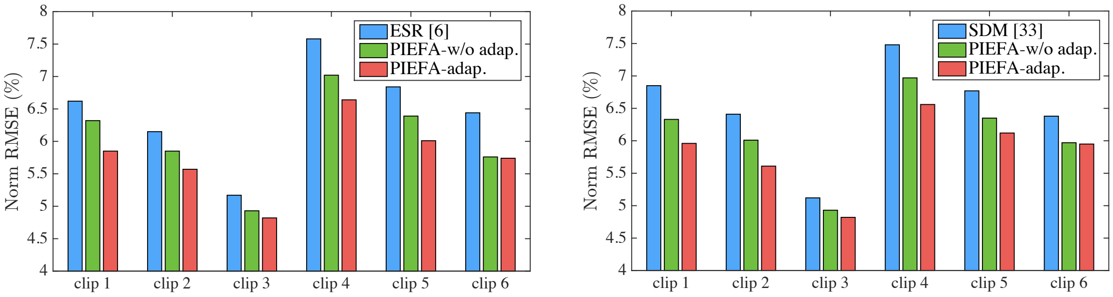


Figure 3: Comparions with *generic* alignment approaches on *Face Movie* dataset.

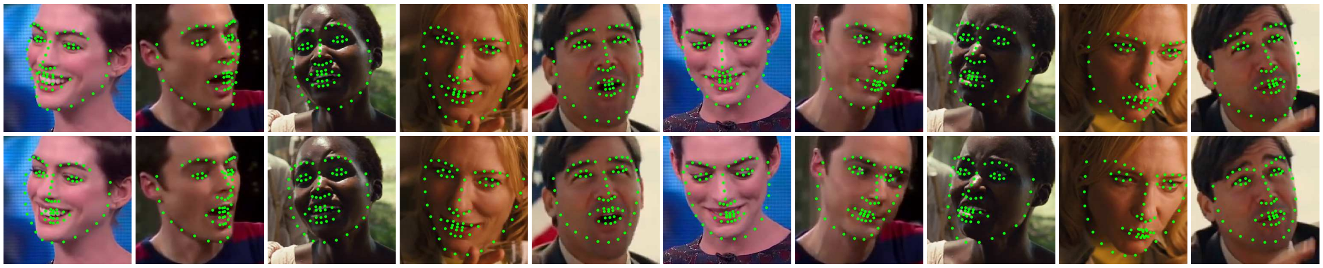


Figure 4: Examples on *Face Movie* dataset: first row, PIEFA-adap.; second row, ESR (column 1-5) and SDM (column 6-10). PIEFA-adap. has substantial fitting accuracy improvement in wild conditions.

is the *singular value decomposition* (SVD) to update $\mathbf{C}_n^*$ in Step 5, which needs $\mathcal{O}(M^2K)$ operations. Other steps consist of simple linear algebra with an average of $\mathcal{O}(NK)$ operations. Therefore, the total computational complexity is $\mathcal{O}\left((M^2K + NK)\epsilon^{-0.5}\right)$, where $\mathcal{O}(\epsilon^{-0.5})$ is the iteration number [37]. It is worth mentioning that the computation complexity can be further cut down by employing more efficient SVD algorithm [5], given the fact that $[\mathbf{C}|\mathbf{C}_o]$ is low rank. We leave it as our future work.

5. Experiments

In this section, we first introduce the implementation details and experimental settings. Than, we compare our approach with both *generic* and *joint* alignment methods on three video datasets: *Face Movie dataset*, *Talking Face dataset* and *YouTube Celebrities dataset*.

5.1. Implementation Details

To train the representation subspace $\mathbf{U}$, we construct a training set which consists of:

- *Multi-PIE dataset* [14] contains images of 337 subjects under 15 view points, 7 expressions and a range of illumination changes recorded in experimental environment. We collect 1,300 images from this dataset.
- *LFPW dataset* [4] is recorded in wild conditions, showing extensive variations in both subject and imaging conditions. Only 1,035 out of 1,400 images are successfully downloaded due to some broken links.
- *Helen dataset* [19] is also collected in unconstrained conditions. We include all the 2,000 training and 330 testing images in our training set.

All training images are annotated with 68-point scheme defined in *300-W challenge* [26]. We first perform procrustes analysis [28] based on a mean shape to remove any rigid 2D transformation among all training images. The interocular distance is set to 50 pixels. Then, SIFT feature [16] is extracted around each landmark for part-based representation as it is robust to illumination and scale variations. Finally, $\mathbf{U}$ is trained by performing PCA and preserving 80% variations on the normalized training set.

For ensemble initialization, we define the latent state $\mathbf{s} \in \mathbb{R}^4$ to control scale, rotation and 2D translation of the initial shape. We change the particle number K from 10 to 200, and test the average fitting error and time cost. The results indicate that we can empirically set $K = 30$, as it provides a good trade-off between the fitting accuracy and efficiency. Moreover, to address the particles degeneration issue, we employ resampling technique [11] to initialize particles in every 50 frames.

To set the threshold τ for robust subspace adaptation, we construct $\{\mathcal{S}_+^*, \mathcal{S}_-^*\}$ from the training set. $\{\mathcal{S}_+^*\}$ contains one annotated shape while $\{\mathcal{S}_-^*\}$ contains ten perturbed

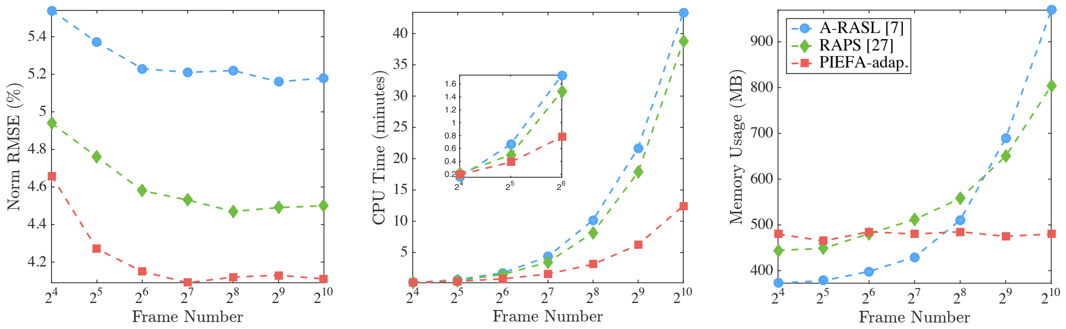


Figure 5: Comparions with *joint* alignment approaches on *Talking Face* dataset.

shapes of each image to simulate well and erroneous fittings, respectively. A perturbed shape is generated by randomly shifting a subset ($< 20\%$) of all landmarks away from the annotation with a Gaussian distributions. Given $\{\mathcal{S}^*_+, \mathcal{S}^*_-\}$, robust decomposition is performed using Equation 12 to get the corresponding error sets $\{\mathcal{E}^*_+, \mathcal{E}^*_-\}$. τ is set to best separate $\{\mathcal{E}^*_+\}$ and $\{\mathcal{E}^*_-\}$. We use Normalized Root Mean Square Error (Norm RMSE) [26] for fitting evaluation in all experiments.

5.2. Comparison on *Face Movie* Dataset

We conduct comparison between the proposed method and *generic alignment* approaches on *Face Movie* dataset. This in-the-wild dataset consists of movie clips that present challenges in different aspects, such as violent head movement, drastic expression variations and dynamic lighting changes. We collected 6 clips and manually labeled 2150 frames for evaluation. Two state-of-the-art generic methods are employed for the comparison: **(1)** ESR [6], and **(2)** SDM [33]. We use the same training set to train their static FDMs offline. To best evaluate the performance, we tested our approach with four different settings:

- Use ESR for alignment with personalized adaptation.
- Use ESR for alignment w/o personalized adaptation.
- Use SDM for alignment with personalized adaptation.
- Use SDM for alignment w/o personalized adaptation.

We report the average Norm RMSE of different approaches in Figure 3 and show some fitting comparisons in Figure 4. The results show that both PIEFA-w/o adap. and PIEFA-adap. outperform ESR and SDM with a substantial margin on all clips. The superior performance of our approach is most obvious especially for landmarks around mouth and face contour when extensive pose variations and expression changes exist, *e.g.*, clip 4. In these cases, the single initial shape used in ESR and SDM is usually far away from the ground-truth, which inevitably results in local optimum and unsatisfactory. Our approach, on the other hand, takes advantage of both motion cues and person-specific information for multiple initialization, and can significantly improve the fitting accuracy in challenging conditions. More specifically, it has average 16.4% and 15.1% accuracy improvement compared to SDM and ESR, respectively. This result highlights the validity of the propose ensemble initialization and constraint decomposition to address the poor initialization issue.

The results also show that the proposed person-specific modeling can also significantly improve fitting accuracy, which demonstrates the validity of the proposed incremental subspace adaption. We also notice that person-specific modeling has less fitting accuracy improvement in clip 6, which contains a large number of blurring frames, than in other clips. Since the personalized adaption is severely impeded by a large E^* recovered in this case.

5.3. Comparison on *Talking Face* Dataset

We compare our approach with *joint alignment* approaches on *Talking Face* dataset [12]. This dataset contains 5 consecutive clips of totally 5000 frames recorded in controlled environment. We convert the original landmark annotations to the standard 68-point scheme [26] for evaluation consistency. We implement two joint alignment approaches: **(1)** A-RASL [7], and **(2)** RAPS [27]. For fairly comparison, we train the clean face subspace for RAPS on the training set, and use SDM to provide initial fittings and anchor shapes for RAPS and A-RASL, respectively.

For each of the 5 clips, we record the experimental results as the number of frames are increasing from 16 to 1000. The average Norm RMSE, CPU time and memory usage are reported in Figure 5. We have three observations. **(1)** For all the three methods, the average fitting errors decrease as the frame number increases, which makes sense since more personalized information is involved in image congealing [27]. The ensemble initialization and person-specific modeling make our approach have the best performance in general w.r.t. both converge speed and finial accuracy. **(2)** The CPU time costs of both A-RASL and RAPS

Table 1: Comparisons of average Norm RMSE with *state-of-the-arts* on *YouTube Celebrities* dataset.

ESR [6]	SDM [33]	RAPS [27]	A-RASL [7]	RLMS [28]	RLB [23]	IFA [2]	PIEFA-adap.
5.61%	5.85%	5.44%	8.63%	7.27%	5.37%	6.79%	**4.92%**

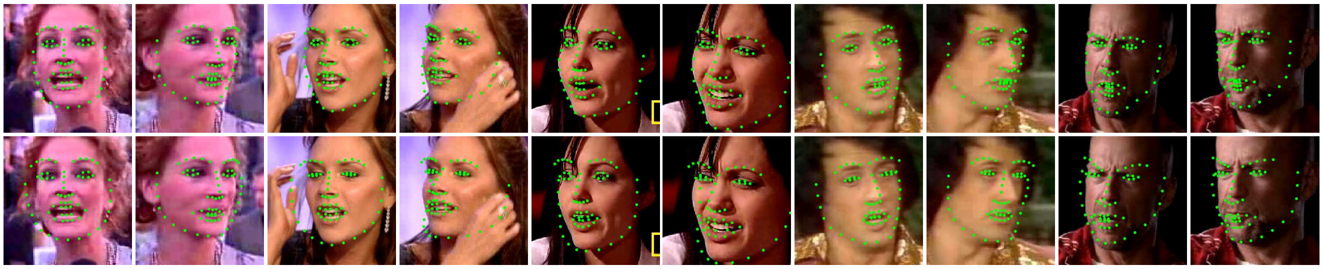


Figure 6: Examples on YouTube Celebrities dataset: first row, PIEFA-adap.; second row, RAPS (column 1-2), A-RASL (3-4), RLMS (5-6), RLB (7-8) and IFA (9-10). PIEFA-adapt. outperforms others with respect to different challenges. There are consistant improvement of fitting accuracy especially for lardmarks around eyes, mouth and face contour.

grows explosively when the number of frames increases, since they perform joint alignment simultaneously for all frames in the batch. Our approach, on the other hand, has relatively constant time cost since the ensemble alignment is performed in each, instead of all frames. **(3)** A-RASL and RAPS consume more memory to process more frames, while our approach has constant memory usage no matter how many frames in the batch. These results prove that the proposed incremental ensemble alignment outperforms traditional joint alignment methods w.r.t. fitting accuracy and efficiency. Instead of loading all frames in a batch manner, our approach can process each frame in a streaming manner with constant computational cost, which is favored by real-time and large-scale applications.

5.4. Comparison on *YouTube Celebrities* Dataset

To further investigate the performance in wild conditions, we compare our approach with state-of-the-arts on *YouTube Celebrities* dataset [18]. This database is collected from internet under low resolution settings and presents challenges in multiple aspects, *e.g.*, pose, expression, illumination and occlusion. We pick out 6[1] clips and manually label 50% frames for quantitative evaluation. Besides the aforementioned 4 methods, *i.e.*, ESR, SDM, RAPS and A-RASL, we add another 3 methods into the comparison: **(1)** RLMS [28], **(2)** RLB [23], and **(3)** IFA [2]. They all have public codes available, provided by their authors.

We report the average Norm RMSE in Table 1 and show some fitting comparisons in Figure 4. The results show that our approach achieves the best performance and has consistent fitting improvement in challenging conditions.

Besides our methods, the recently proposed RLB has the second best performance in this dataset. The combination of part-based representation (shape indexed feature) and regression-based fitting strategy guarantee its robust and efficient performance in challenging conditions. Moreover, regression-based methods, *e.g.*, SDM, have better performance than optimization-based methods, *e.g.*, RLMS, which demonstrate the validity to learn the gradient descent in a data-driven manner. We also notice that A-RASL has the lowest fitting accuracy in general. A possible reason is holistic FDMs and optimization-based methods are less flexible and more susceptible to extensive variations.

To sum up, the experiments prove that the proposed approach can effectively overcome the pool initialization of existing generic methods. The proposed incremental framework can process large-scale and real-time data with constant computational cost, which is a significant merit compared with traditional joint methods. Moreover, the proposed incremental adaptation can achieve personalized modeling, while the drifting issue is significantly mitigated, even in wild conditions.

6. Conclusion

In this paper, we propose a novel approach for sequential face alignment. It can effectively address limitations of generic and joint alignment methods. Extensive experiments on challenging datasets validated our approach in different aspects and demonstrated its superior performance compared with state-of-the-arts. We plan to incorporate learning-based temporal information and feature representation in our future work to further improve the fitting accuracy and efficiency.

[1] 1) 0292_02_002_*angelina_jolie*, 2) 0502_01_005_*bruce_willis*, 3) 1198_01_012_*julia_roberts*, 4) 1621_02_017_*ronald_reagan*, 5) 1786_02_006_*sylvester_stallone*, and 6) 1847_01_005_*victoria_beckham*

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. *Robust Discriminative Response Map Fitting with Constrained Local Models*. In *CVPR*, 2013.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. *Incremental Face Alignment in the Wild*. In *CVPR*, 2014.

[3] A. Beck and M. Teboulle. *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*. *SIAM JIS*, 2(1):183-202, 2009.

[4] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, and N. Kumar. *Localizing Parts of Faces using a Consensus of Exemplars*. In *CVPR*, 2011.

[5] M. Brand. *Fast Low-Rank Modifications of the Thin Singular Value Decomposition*. *LAIA*, 415(1):20-30, 2006.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. *Face Alignment by Explicit Shape Regression*. In *CVPR*, 2012.

[7] X. Cheng, S. Sridharan, J. Saraghi, and S. Lucey. *Anchored Deformable Face Ensemble Alignment*. In *ECCVW*, 2012.

[8] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey. *Rank Minimization across Appearance and Shape for AAM Ensemble Fitting*. In *ICCV*, 2013.

[9] T. Cootes, G. Edwards, and C. Taylor. *Active Appearance Models*. *TPAMI*, 23(6):681-685, 2001.

[10] D. DeCarlo, D. Metaxas. *Optical Flow Constraints on Deformable Models with Applications to Face Tracking*. *IJCV*, 38(2):99-127, 2000.

[11] A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[12] FGNet. *Talking Face Video*, 2004.

[13] A. Edelman, T.A. Arias, and S.T. Smith. *The Geometry of Algorithm with Orthogonality Constraints*. *SIAM JIS*, 20(2):303-353, 1998.

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. *Multi-PIE*. *IVC*, 28(5):807-813, 2010.

[15] G. Guo and X. Wang. *Kinship Measurement on Salient Facial Features*. *TIP*, 61(8):2322-2325, 2012.

[16] D.G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. *IJCV*, 2(60):91-110, 2004.

[17] J. He, L. Balzano, and A. Szlam. *Incremental Gradient on the Grassmannian for Online Foreground and Background Separation in Subsampled Video*. In *CVPR*, 2012.

[18] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. *Face Tracking and Recognition with Visual Constraints in Real-world Videos.* In *CVPR*, 2008.

[19] V. Le, J. Brandt, Z. Lin, L.D. Bourdev, and T.S. Huang. *Interactive Facial Feature Localization*. In *ECCV*, 2012.

[20] Z. Lin, M. Chen, L. Wu, and Y. Ma. *The Augmented Lgrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices*. *UIUC Report*, 2009.

[21] X. Mei and H. Ling. *Robust Visual Tracking using ℓ_1 Minimization*. In *ICCV*, 2009.

[22] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, *3D Facial Landmark Detection under Large Yaw and Expression Variations*. *TPAMI*, 35(7,): 1552-1564, 2013.

[23] S. Ren, X. Cao, Y. Wei, and J. Sun. *Face Alignment at 3000 FPS via Regressing Local Binary Features*. In *CVPR*, 2014.

[24] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. *RASL: Robust Alignment by Sparse and Low-Rank Decomposition for Linearly Correlated Images.* *TPAMI*, 34(11):2233-2246, 2012.

[25] X. Peng, J. Huang, Q. Hu, S. Zhang, and D. Metaxas. *Three-Dimensional Head Pose Estimation in-the-Wild*. In *FG*, 2015.

[26] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. *300 Faces in-the-Wild challenge: The First Facial Landmark Localization Challenge.* In *ICCVW*, 2013.

[27] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. *RAPS: Robust and Efficient Automatic Construction of Person-Specific Deformable Models*. In *CVPR*, 2014.

[28] J. Saragih, S. Lucey, and J. Cohn. *Deformable Model Fitting by Regularized Landmark Mean-shift*. *IJCV*, 91(2):200-215, 2011.

[29] J. Sung and D. Kim. *Adaptive Active Appearance Model with Incremental Learning*. *PRL*, 30(4):359-367, 2009.

[30] G. Tzimiropoulos and M. Pantic. *Gauss-Newton Deformable Part Models for Face Alignment in-the-Wild*. In *CVPR*, 2014.

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. *DeepFace: Closing the Gap to Human-Level Performance in Face Verification*. *CVPR*, 2014.

[32] M. Tang and X. Peng. *Robust Tracking With Discriminative Ranking Lists*. *TIP*, 21(7)3273-3281, 2012.

[33] X. Xiong and F. De la Torre. *Supervised Descent Method and its Applications to Face Alignment*. In *CVPR*, 2013.

[34] J. Yan, Z. Lei, D. Yi, and S. Li. *Learn to Combine Multiple Hypotheses for Accurate Face Alignment*. In *CVPR*, 2013.

[35] L. Zafeiriou, E. Antonakos, S Zafeiriou, and M. Pantic. *Joint Unsupervised Face Alignment and Behaviour Analysis*. In *ECCV*, 2014.

[36] C. Zhao, W.K. Chem, and X. Wang. *Joint Face Alignment with a Generic Deformable Face Model*. In *CVPR*, 2011.

[37] T. Zhang, S. Liu, N. Ahuja, M. Yang, and B. Ghanem. *Robust Visual Tracking via Consistent Low-Rank Sparse Learning*. *IJCV*, 111(2):171-190, 2015.

[38] X. Zhu and D. Ramanan. *Face Detection, Pose Estimation and Landmark Localization in the Wild*. In *CVPR*, 2012.

[39] S. Zhu, C. Li, C. Loy, and X. Tang. *Face Alignment by Coarse-to-Fine Shape Searching*. In *CVPR*, 2015.

[40] 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.