Int J Comput Vis (2018) 126:184–197
https://doi.org/10.1007/s11263-017-0996-8

Toward Personalized Modeling: Incremental and Ensemble Alignment for Sequential Faces in the Wild

Xi Peng[1] · **Shaoting Zhang**[2] · **Yang Yu**[1] · **Dimitris N. Metaxas**[1]

Received: 16 February 2016 / Accepted: 2 February 2017 / Published online: 15 February 2017

Abstract Fitting facial landmarks on unconstrained videos is a challenging task with broad applications. Both *generic* and *joint* alignment methods 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, incremental subspace adaptation is performed to achieve personalized modeling in a unified framework. To alleviate the drifting issue, we leverage a very efficient fitting evaluation network to pick out well-aligned faces for robust incremental learning. Extensive experiments on both controlled and unconstrained datasets have validated our approach in different aspects and demonstrated its superior performance compared with state of the arts in terms of fitting accuracy and efficiency.

Keywords Face alignment · Personalized modeling · Incremental learning · Ensemble learning · Sparse coding

Communicated by Xiaoou Tang.

✉ Xi Peng
xpeng.cs@rutgers.edu

Shaoting Zhang
szhang16@uncc.edu

Yang Yu
yyu@cs.rutgers.edu

Dimitris N. Metaxas
dnm@cs.rutgers.edu

[1] Rutgers University, Piscataway, NJ 08854, USA

[2] The University of North Carolina at Charlotte, Charlotte, NC 28223, USA

1 Introduction

Recently, analysing image sequences in large-scale and unconstrained conditions attracts increasing interest in computer vision community. In the context of face-related topics, sequential face alignment, i.e., fitting facial landmarks on sequential images, is a crucial task with a wide range of applications, such as Face Verification (Taigman et al. 2014; Parkhi et al. 2015), Facial Action Unit (FAU) analysis (Zafeiriou et al. 2014; Cootes et al. 2001) and Human-Computer Interaction (HCI; Perakis et al. 2013; Escalera et al. 2015). It is a challenging task since the face undergoes drastic non-rigid deformations (Vogler et al. 2007) caused by extensive pose and expression variations, as well as unconstrained imaging conditions like illuminations changes and partial occlusions (Sagonas et al. 2013; Shen et al. 2015).

Despite the long history of research in rigid and non-rigid face tracking (Black and Yacoob 1995; Decarlo and Metaxas 2000; Patras and Pantic 2004; Asthana et al. 2013), it has been shown that either *generic face alignment* (Cao et al. 2014; Saragih et al. 2011; Sun et al. 2013; Trigeorgis et al. 2016; Tzimiropoulos 2015; Xiong and De la Torre 2013; Zhang et al. 2014a, b; Zhu et al. 2015; Zhu and Ramanan 2012) which aligns each frame independently in a tracking-by-detection manner, or *joint face alignment* (Zhao et al. 2011; Cheng et al. 2013; Sagonas et al. 2014), which aligns all frames simultaneously in a batch optimization manner, can be employed to fit facial landmarks on sequential images.

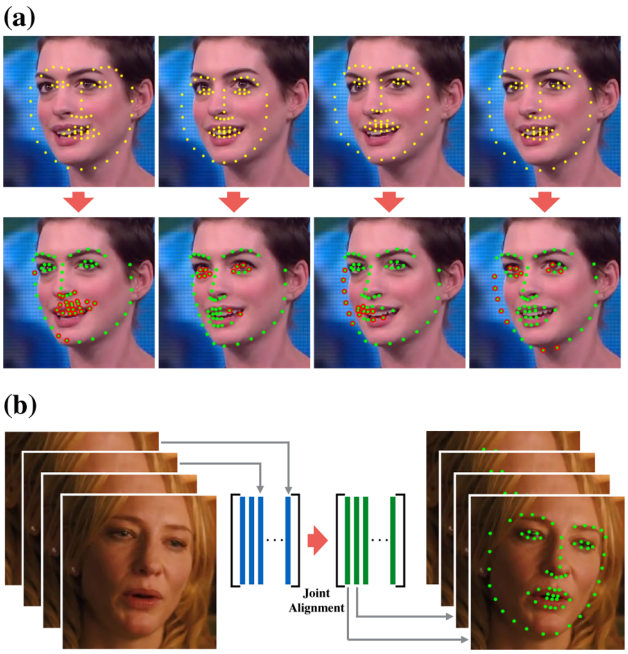

Fig. 1 Limitations of existing methods. *Yellow points*: different initial shapes. *Green points*: well-aligned landmarks. *Red points*: mis-aligned landmarks. **a** *Generic* approaches are sensitive to initializations. **b** *Joint* approaches are restricted to offline batch alignment (Color figure online)

Generic Face Alignment starts the fitting process from an initial shape, e.g., a mean face (Cao et al. 2014; Xiong and De la Torre 2013) or the result of the last frame (Asthana et al. 2013; Saragih et al. 2011), and deform the shape constrained by facial deformable models (FDMs) to minimize the reconstruction residual by either gradient descend optimization (Saragih et al. 2011; Tzimiropoulos and Pantic 2014) or cascade/boosted regression (Cao et al. 2014; Xiong and De la Torre 2013). They have shown great success on single image with respect to the efficiency, e.g., face alignment at 3000 FPS (Ren et al. 2014), and unconstrained scenarios, e.g., face alignment in the wild (Sagonas et al. 2013).

However, when it comes to sequential alignment, many of them suffer from significant limitations: (1) Many of them are heavily sensitive to initializations as illustrated in Fig. 1a. They are easily trapped into local optima when started from poor initialization. (2) They usually rely on models trained offline on still images and lack the capability to capture the personalized information and imaging continuity in consecutive frames.

Joint Face Alignment takes the advantage of the shape and appearance consistency to simultaneously minimize fitting errors for all frames (Zhao et al. 2011; Cheng et al. 2013; Sagonas et al. 2014). They are more robust to illumination changes and partial occlusions than generic methods (Peng et al. 2010). However, they still have limitations in two aspects. (1) Most of them are limited to offline tasks due to the prerequisite of all frames before batch alignment as illustrated in 1b, which severely impedes their applications on real-time or large-scale tasks. (2) Some of them attempt to achieve personalized modeling without effective correction, which may inevitably result in model drifting.

In this paper, we further improve our former work *Personalized and Incremental Ensemble Face Alignment* (PIEFA) (Peng et al. 2015) to address the aforementioned issues. Instead of a 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, incremental subspace adaptation is performed to achieve personalized modeling in a unified framework. To alleviate the drifting issue, we leverage a very efficient fitting evaluation network to pick out well-aligned faces for robust incremental learning. In summary, our work makes the following *contributions*:

- We propose a novel approach for face alignment in unconstrained videos. Our approach incorporates motion models to perform ensemble initialization, which can effectively overcome the initialization-sensitivity issue of former generic methods.
- The ensemble alignment is performed within a single frame, which is radically different from existing joint alignment methods that achieve image congealing across frames. It guarantees the online efficiency in real-time, which can be used in large-scale applications.
- A rank minimization framework with the group-sparse regularization is designed to incrementally achieve personalized modeling. In addition, we propose a novel fitting evaluation network to significantly alleviated the model drifting issue.
- We carry out extensive experiments on 4 video datasets and compare with 8 state-of-the-art methods to fully validate the proposed method. The results demonstrate that our approach can significantly improve the fitting accuracy with a constant CPU time and memory usage.

This article improves its conference version (Peng et al. 2015) in both theoretical and practical aspects. Instead of struggling with the threshold engineering, we propose to leverage a deep neural network for efficient fitting evaluation to guarantee robust person-specific modeling. We also carry out detailed complexity analysis by decomposing the proposed optimization in a step-by-step form, which proves the efficiency of our approach in terms of time and memory. Moreover, extensive experiments are performed for component-wise validations and general comparisons with state-of-the-art methods.

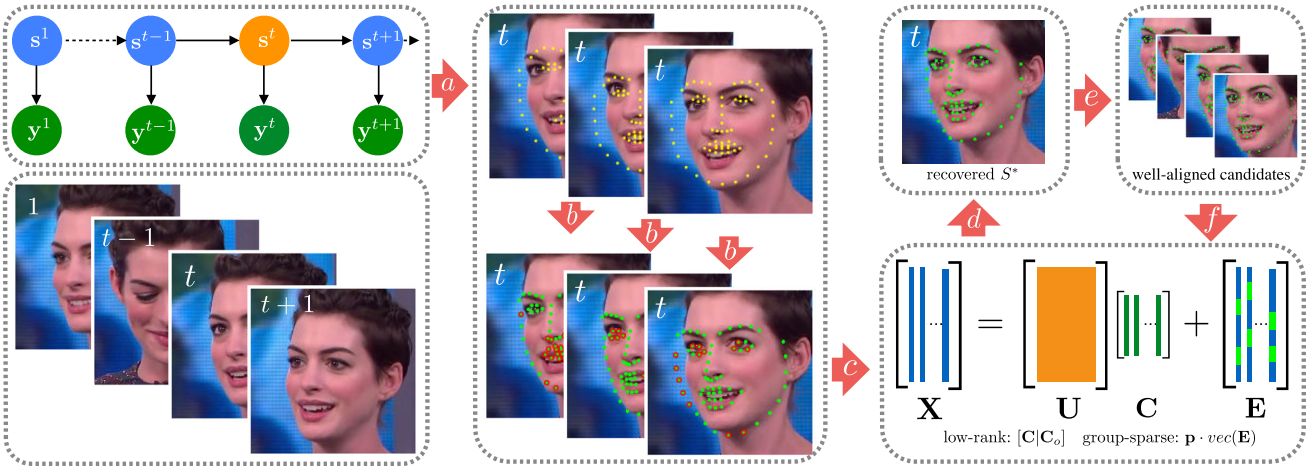

Fig. 2 Overview of our approach: **a** ensemble initialization (3.1), **b** generic face alignment in parallel, **c** constrained decompostion (3.3), **d** fitting recovery (3.4), **e** fitting evaluation and **f** personlized adaptation (3.5)

2 Related Work

Face alignment in a single image has attracted intensive research interest for decades. When it comes to the task that fitting facial landmarks on sequential images, either generic or joint alignment methods can be employed. We briefly review related work in this section.

Generic face alignment 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; Cootes et al. 2001), or *part-based models*, e.g., *constrained local models* (CLMs; Saragih et al. 2011). Among *generic face alignment* approaches, part-based FDMs combined with regression-based fitting strategies attract intensive interest. For instance, Asthana et al. (2013) proposed the *discriminative response map fitting* (DRMF) to learn boosted mappings from the joint response maps to shape parameters. Cao et al. (2014) achieved *explicit shape regression* (ESR) by combining shape-indexed feature selection and multi-layer boosted regression. Xiong and De la Torre (2013) 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 (Ren et al. 2014), however, the static FDMs and initialization-sensitivity issue severely limited their performance on streaming data.

Multiple efforts were devoted to address these limitations. For instance, Asthana et al. (2014) improved SDM by updating cascade regressors in parallel for *incremental face alignment* (IFA). Yan et al. (2013) proposed to rank and *combine multiple hypotheses* (CMH) in a structural SVM framework to address the initialization-sensitivity issue. Tzimiropoulos (2015) proposed to learn the Jacobians and descent directions in a subspace orthogonal to the facial appearance variation for *project-out cascaded regression* (POCR). Zhu et al. (2015) proposed *coarse-to-fine shape searching* (CFSS) to find a best initial shape to address the initialization-sensitivity issue. However, the temporal constraints in successive inputs, such as personalized shape, appearance and motion cues, are barely investigated.

In recent years, *deep neural networks* (DNNs) based methods attract intensive research interest as they have shown astonishing performance in both vision (Taigman et al. 2014; Schroff et al. 2015; Baró et al. 2015; Nasrollahi et al. 2015; Peng et al. 2016) and language applications (Wang et al. 2016a, b). Several DNNs based approaches have been proposed for generic face alignment. Sun et al. (2013) proposed *deep convolutional networks cascade* (DCNC) to refine the fitting results step by step from an initial guess. Zhang et al. (2014a) employed the similar idea of coarse-to-fine framework but using *auto-encoder netowrks* (CFAN) instead of CNNs. Zhang et al. (2014b) showed that learning face alignment together with other correlated tasks, such as identity recognition and pose estimation, in a uniform CNNs can improve the landmark detection accuracy (TCDCN). A major limitation of DNNs based methods is that they are extremely data-hungry and need numerous training images to avoid over-fitting. In this paper, instead of directly using deep neural networks for face alignment, we leverage CNNs for robust and efficient fitting evaluation, which is crucial to achieve faithful personalized modeling without drifting.

Joint face alignment To this end, *joint face alignment*, which takes the advantage of consistency constraints to minimize fitting errors for all frames, is mainly applied. For instance, Zhao et al. (2011) proposed to regularize the holistic texture by enforcing all frames to lie in a low-rank subspace (RASL). The drawback of this method is that it did not incorporate any face prior, which may result in arbitrary deformations. To address this problem, Cheng et al. (2012) proposed to use anchor shapes to penalize arbitrary defor-

mations (A-RASL); while Sagonas et al. (2014) proposed to employ a clean face subspace trained offline to restrict optimization directions (RAPS). 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. (2015) proposed to use dictionary learning to achieve sparse representations for rigid object tracking (Tang and Peng 2012). However, it is nontrivial to apply dictionary learning in sequential face alignment as facial appearance may undergo extensive nonrigid deformations. Moreover, it remains a challenging task to simultaneously address the initialization sensitivity issue and achieve person-specific modeling in a unified framework.

3 Proposed Approach

In this paper, we propose a novel approach for sequential face alignment. We first incorporate motion models to sample multiple initial shapes as *ensemble initialization*. Then we employ off-the-shelf generic approach to conduct batch alignment in parallel. The alignment results are then reorganized using *part-based representation*. By conducting *constrained decomposition*, we can recover the best fitting S^* via *fitting recovery*. Finally, personalized modeling is achieved by robust *personalized adaptation*. Please refer to Fig. 2 for an overview of our approach.

3.1 Ensemble Initialization

Initialization is the first and key step in landmark localization. It is relatively 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 (Doucet et al. 2001) to sample multiple initial shapes for ensemble initialization.

As illustrated in Fig. 3, 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:

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

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 (Mei and Ling 2009). The posterior state distribution is then updated at time t by:

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

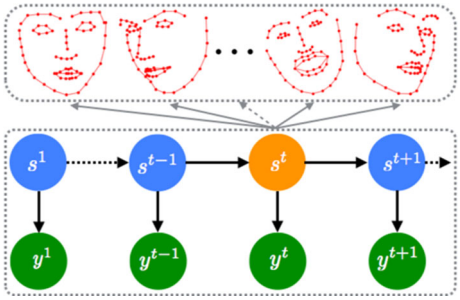

Fig. 3 Illustration of ensemble initialization. The latent state $\mathbf{s}$ controls the variations of initial shapes. The observation $\mathbf{y}$ represents the corresponding fitting results

where $p(\mathbf{y}^t|\mathbf{s}^t)$ is the observation model, which is the key component to evaluate the corresponding initial shape. We model it using group-sparse fitting errors and introduce the details in Sect. 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 (Cao et al. 2014) and SDM (Xiong and De la Torre 2013), to obtain a batch of rough fittings $\{S_1, \ldots, S_K\}$. It is worth noting that the efficiency is guaranteed since generic approaches are highly efficient (Ren et al. 2014) and we can conduct batch alignments in parallel.

To conduct batch alignment, former approaches (Cheng et al. 2012; Sagonas et al. 2014; Zhao et al. 2011) usually use holistic FDMs to parameterize the shape and appearance separately and bridge the two by image warping (Cootes et al. 2001). 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:

$$\mathbf{A} = \left[(\mathbf{x}_1 - \bar{\mathbf{x}})^T \, \mathbf{f}(\mathbf{x}_1)^T \ldots (\mathbf{x}_L - \bar{\mathbf{x}})^T \, \mathbf{f}(\mathbf{x}_L)^T\right]^T, \quad (3)$$

where $(\mathbf{x}_1 - \bar{\mathbf{x}}) \in \mathbb{R}^2$ are centralized landmark coordinates, $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d$ is the 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 $\{S^1, \ldots, S^K\}$. We propose a constrained decomposition based on the following facts and observations. **(a)** Each of $\{S^1, \ldots, S^K\}$ is aligned to the same face but with fitting errors. **(b)** With respect to the kth shape, only a small number of its landmarks are misaligned. **(c)** With respect to the lth 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, $\mathbf{X} = [\mathbf{A}^1, \ldots, \mathbf{A}^K] \in \mathbb{R}^{N\times K}$ denote the batch observation matrix. Based on the observation **(a)** we have the following low-rank constraint:

$$\arg\min_{\mathbf{C},\mathbf{E}} rank(\mathbf{C}),\ \text{s. t. } \mathbf{X} = \mathbf{UC} + \mathbf{E}, \tag{4}$$

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, $rank(\mathbf{C}) = 1$, 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 $rank([\mathbf{C}|\mathbf{C}_o])$ instead of $rank(\mathbf{C})$, 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 Eq. 4 has the group structure:

$$\mathbf{E} = \begin{bmatrix} \epsilon_1^1 & \cdots & \epsilon_1^K \\ \vdots & \ddots & \vdots \\ \epsilon_L^1 & \cdots & \epsilon_L^K \end{bmatrix}, \quad vec(\mathbf{E}) = \left[\epsilon_1^1, \ldots, \epsilon_1^K, \ldots, \epsilon_L^1, \ldots, \epsilon_L^K\right], \tag{5}$$

where $\epsilon_l^k \in \mathbb{R}^{2+d}$ is the fitting errors of the lth landmark in the kth 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:

$$\arg\min_{\mathbf{C},\mathbf{E}} \|\mathbf{p}\cdot vec(\mathbf{E})\|_{2,0},\ \text{s. t. } \mathbf{X} = \mathbf{UC} + \mathbf{E}, \tag{6}$$

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 Eqs. 4 and 6, we can achieve the object function for robust decomposition:

$$\begin{aligned} \arg\min_{\mathbf{C}_n,\mathbf{E}_v,\mathbf{C},\mathbf{E}} \quad & \|\mathbf{Z}\|_F^2 + \lambda_1 rank(\mathbf{C}_n) + \lambda_2\|\mathbf{E}_v\|_{2,0} \\ \text{subject to} \quad & \mathbf{Z} = \mathbf{X} - \mathbf{UC} - \mathbf{E}, \\ & \mathbf{C}_n = [\mathbf{C}|\mathbf{C}_o],\ \mathbf{E}_v = \mathbf{p}\cdot vec(\mathbf{E}). \end{aligned} \tag{7}$$

where λ_1 and λ_2 are non-negative parameters to balance contributions between the two constraints. We present an efficient solution for the optimization in Sect. 4.

3.4 Fitting Recovery

The error matrix $\mathbf{E}$ guarantees robust decomposition against outliers such as illumination changes and partial occlusions (Peng et al. 2010). More importantly, we can recover a well-aligned S^* from $\{S^1, \ldots, S^K\}$ using the group-sparse structure.

Equation 5 indicates that ϵ_l^k measures the fitting errors of S^k at the lth landmark. Therefore, each row of $\mathbf{E}$ models the errors distribution of all aligned shapes with respect to the same indexed landmark. Let $S^* = \{x_1^*, \ldots, x_L^*\}$, $\mathbf{x}_l^*$ can be recovered by using the intra-row ℓ_2-norm of $\mathbf{E}$ to weight the same indexed landmark of all aligned shapes:

$$\mathbf{x}_l^* = \frac{1}{\mathbf{q}_l}\Sigma_{k=1}^K e^{-\|\epsilon_l^k\|_2}\mathbf{x}_l^k,\ \text{where } \mathbf{q}_l = \Sigma_{k=1}^K e^{-\|\epsilon_l^k\|_2}. \tag{8}$$

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 Eq. 2. Considering the fact that the kth column of $\mathbf{E}$ measures the overall fitting errors of S^k, we can use inter-column ℓ_2-norm of $\mathbf{E}$ to model the observation:

$$p\left(\mathbf{y}^{t,k}|\mathbf{s}^t\right) = \frac{e^{-\mathbf{r}_k}}{\Sigma_{k=1}^K e^{-\mathbf{r}_k}},\ \text{where } \mathbf{r}_k = \Sigma_{l=1}^L\|\epsilon_l^k\|_2. \tag{9}$$

We compute $p(\mathbf{y}^{t,k}|\mathbf{s}^t)$ for each aligned shape at frame t, and apply Eq. 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 can extract the part-based representation $X^* \in \mathbb{R}^N$ to calculate $C^* \in \mathbb{R}^M$ and $E^* \in \mathbb{R}^N$ by robust decomposition:

$$\arg\min_{C^*,E^*} \|\mathbf{p} \cdot vec(E^*)\|_{2,0}, \text{ s.t. } X^* = \mathbf{U}C^* + E^*, \tag{10}$$

which can be efficiently solved by introducing the augmented Lagrangian:

$$\mathcal{L}^*(C^*, E^*, Y^*, \mu^*) = \|\mathbf{p} \cdot vec(E^*)\|_{2,0} + Y^{*T} h^* + \frac{\mu^*}{2}\|h^*\|_F^2, \tag{11}$$

where $h^* = X^* - \mathbf{U}C^* - E^*$, Y^* is Lagrange multiplier, and μ^* is penalty parameter.

To efficiently update $\mathbf{U}$, we adopt the concept of incremental subspace adaptation on Grassmannian (He et al. 2012). In our case the Grassmannian is a Riemannian manifold of all subspaces of $\mathbb{R}^N$ with fixed dimension M. According to Edelman et al. (1998), the gradient step along the geodesic of Grassmannian is:

$$\nabla\mathcal{L}^* = \left(\mathbf{I} - \mathbf{U}\mathbf{U}^T\right)\frac{d\mathcal{L}^*}{d\mathbf{U}}, \tag{12}$$

where the derivative of $\mathcal{L}^*$ with respect to $\mathbf{U}$:

$$\frac{d\mathcal{L}^*}{d\mathbf{U}} = Y^*C^{*T} - \mu^* h^* C^{*T}. \tag{13}$$

From Eqs. 12 and 13, we have $\nabla\mathcal{L}^* = \Omega C^{*T}$, where $\Omega = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)(Y^* - \mu^* h^*)$. Note that $\Omega \in \mathbb{R}^N$ and $C^* \in \mathbb{R}^M$, $\nabla\mathcal{L}^*$ has to be $rank$-1. By computing the SVD of $\nabla\mathcal{L}^*$, we can get the only non-zero singular value $\sigma = \|\Omega\|\|C^*\|$. Let $\frac{\Omega}{\|\Omega\|}$ and $\frac{C^*}{\|C^*\|}$ denote the left and right singular vectors of σ, respectively, we can add orthonormal sets $\{a_2, \ldots, a_N\}$ and $\{b_2, \ldots, b_M\}$ to Ω into the SVD:

$$\nabla\mathcal{L}^* = \left[\frac{\Omega}{\|\Omega\|}\ a_2\ \ldots\ a_N\right]\begin{bmatrix}\sigma & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\left[\frac{C^*}{\|C^*\|}\ b_2\ \ldots\ b_M\right], \tag{14}$$

Now $\mathbf{U}$ can be effectively updated with the gradient step along the geodesic of the Grassmannian:

$$\Delta\mathbf{U} = \left[(cos(\psi) - 1)\frac{\mathbf{U}C^*}{\|C^*\|} - sin(\psi)\frac{\Omega}{\|\Omega\|}\right]\frac{C^{*T}}{\|C^*\|}, \tag{15}$$

where $\psi = \eta\|\Omega\|\|C^*\|$ and η is the gradient step. The proposed incremental subspace adaption takes only $\mathcal{O}(M^2)$ operations, which is highly efficient and well suited for large-scale and real-time task.

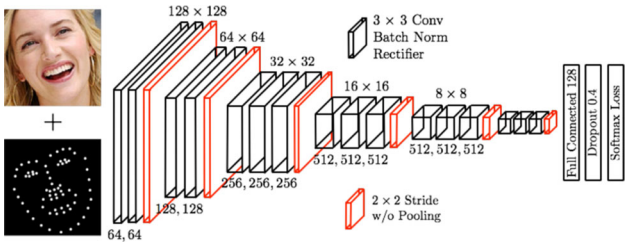

Fig. 4 The architecture of the fitting evaluation network. It takes the concatenation as input and outputs a binary label to indicate correct or erroneous alignment

3.6 Fitting Evaluation

Blind subspace adaptation without correction would inevitably result in model drifting (Sung and Kim 2009). To address this issue, we propose to leverage deep neural networks for robust fitting evaluation. The fitting network can pick out well-aligned faces to incrementally update the offline trained $\mathbf{U}$ for personalized adaptation.

Our goal is to learn a deep neural network that takes the aligned face as input and outputs a binary label to indicate well or miss alignment. To connect the facial appearance and the fitted shape, a possible solution is to directly concatenate the vector of landmark coordinates to an intermediate fully connected layer. However, we experienced limited performance using this design. The reason is the pixel-wise spatial information diminishes significantly after a series of max-pooling operations (Long et al. 2015). The network can hardly learn the correlation between the facial appearance and the landmark location.

Instead, we propose to represent the fitted shape as a landmark map and concatenate it to the facial image as shown in Fig. 4. Each pixel in the landmark map labels the presence of a corresponding landmark. Our network is designed based on a variant of the VGG-16 networks (Simonyan and Zisserman 2014). There are two changes: (1) we remove all the max pooling and use a 2-pixel stride to half the resolution of feature maps after each convolution stage; (2) we reduce the number of fully connected neurons to avoid overfitting for efficient training (Wang et al. 2015). We initialize the training process from weights trained on large datasets for object classification (Krizhevsky et al. 2012). To fine-tune the network for our task, we construct a training set $\mathcal{U} = \{(\mathbf{I}, \mathbf{S}); y\}$, where $y \in \{1, -1\}$. $\mathbf{I}$ is training images with landmark annotation. The landmark map $\mathbf{S}$ is generated using the ground-truth shape when $y = 1$, or the perturbed shape when $y = -1$. We use the cross-entropy loss for the binary classification task.

The proposed deep fitting evaluation outperforms our former approach Peng et al. (2015) that uses an offline learned threshold for error detection. It is also very efficient, which takes less than 10 ms to process one image using a single

Algorithm 1 Alternating Optimization of Eq. 16

Input: $\mathbf{X}$, $\mathbf{U}$, $\mathbf{C}_o$, $\mathbf{p}$, λ, γ
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**
  $\mathbf{C}_n \leftarrow \arg\min_{\mathbf{C}_n} \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
    $\Rightarrow \mathbf{C}_n^* = \mathcal{J}_{\frac{1}{\mu_2}}\Big[[\mathbf{C}|\mathbf{C}_o] + \frac{1}{\mu_2}\mathbf{Y}_2\Big]$,
  $\mathbf{E}_v \leftarrow \arg\min_{\mathbf{E}_v} \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
    $\Rightarrow \mathbf{E}_v^* = \mathcal{L}_{\frac{\lambda}{\mu_3}}\Big[\mathbf{p}\cdot vec(\mathbf{E}) + \frac{1}{\mu_3}\mathbf{Y_3}\Big]$,
  $\mathbf{C} \leftarrow \arg\min_{\mathbf{C}} \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
    $\Rightarrow \mathbf{C}^* = \Lambda_1\Big[\mathbf{M} + \frac{1}{\mu_1}(\mathbf{U}^T\mathbf{Y}_1 - [\mathbf{Y}_2]_{1:K})\Big]$,
    where $\Lambda_1 = (1+\frac{\mu_2}{\mu_1})^{-1}\mathbf{I}$,
      $\mathbf{M} = \mathbf{U}^T(\mathbf{X}-\mathbf{E}) + \frac{\mu_2}{\mu_1}[\mathbf{C}_n]_{1:K}$,
  $\mathbf{E} \leftarrow \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{C}_n, \mathbf{E}_v, \mathbf{C}, \mathbf{E}, \mathbf{Y}_{1-3}, \mu_{1-3})$
    $\Rightarrow \mathbf{E}^* = \Lambda_2\Big[\mathbf{W} + \frac{1}{\mu_1}\left(\mathbf{U}^T\mathbf{Y}_1 - vec^{-1}(\mathbf{Y}_3)\right)\Big]$,
    where $\Lambda_2 = (1+\frac{\mu_3}{\mu_1})^{-1}\mathbf{I}$,
      $\mathbf{W} = \mathbf{X} - \mathbf{UC} + \frac{\mu_3}{\mu_1}\left[vec^{-1}(\mathbf{p}^{-1}\mathbf{E}_v)\right]$,
  $\mathbf{Y}_1 \leftarrow \mathbf{Y}_1 + \mu_1(\mathbf{X}-\mathbf{UC}-\mathbf{E})$,
  $\mathbf{Y}_2 \leftarrow \mathbf{Y}_2 + \mu_2([\mathbf{C}|\mathbf{C}_o]-\mathbf{C}_n)$,
  $\mathbf{Y}_3 \leftarrow \mathbf{Y}_3 + \mu_3(vec(\mathbf{E})-\mathbf{E}_v)$,
  $\mu_1 \leftarrow \gamma\mu_1$, $\mu_2 \leftarrow \gamma\mu_2$, $\mu_3 \leftarrow \gamma\mu_3$.
end while

K40 GPU accelerator. More experimental validation will be discussed soon in Sect. 5.2.

4 ALM Optimization

Directly minimizing $rank(\cdot)$ and $\ell_{2,0}$-norm in Eqs. 7 and 10 is NP-hard (Peng et al. 2010). Instead, we reformulate the optimization with relaxed ℓ_*-norm and $\ell_{2,1}$-norm as:

$$\begin{aligned} \arg\min_{\mathbf{C}_n,\mathbf{E}_v,\mathbf{C},\mathbf{E}} \quad & \|\mathbf{C}_n\|_* + \lambda\|\mathbf{E}_v\|_{2,1} \\ \text{subject to} \quad & \mathbf{X} = \mathbf{UC} + \mathbf{E}, \\ & \mathbf{C}_n = [\mathbf{C}|\mathbf{C}_o]\,,\ \mathbf{E}_v = \mathbf{p}\cdot vec(\mathbf{E}). \end{aligned} \tag{16}$$

The intermediate variables $\mathbf{C}_n$ and $\mathbf{E}_v$ allows us to efficiently solve the Eqs. 7 and 10 with *Augmented Lagrange Multiplier* (ALM) method (Lin et al. 2010).

ALM method solves each variatble in an alternating manner where the convergence is guaranteed. That is, each iteration of ALM updates one variable at a time with the other variables fixed to their most recent values. Minimizing the augmented Lagrangian function $\mathcal{L}(\star)$ can be divided into multiple subproblems which are solved in alternative steps:

Step 1. Update $\mathbf{C}_n$: given Singular Value Thresholding (SVT) operator (Peng et al. 2010) $\mathcal{J}_\tau(\mathbf{X}) = \mathbf{U}\mathcal{S}_\tau(\Sigma)\mathbf{V}^T$ and $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathcal{S}_\tau(\mathbf{X}_{ij}) = \text{sign}(\mathbf{x})\max(|\mathbf{X}_{ij}| - \tau, 0)$, the subproblem is solved with closed form solution:

$$\underset{\mathbf{C}_n}{\text{argmin}}\ \frac{1}{\mu_2}\|\mathbf{C}_n\|_* + \frac{1}{2}\|\mathbf{C}_n - [\mathbf{C}|\mathbf{C}_o] - \frac{1}{\mu_2}\mathbf{Y}_2\|_F^2, \tag{17}$$

$$\Rightarrow \mathbf{C}_n^* = \mathcal{J}_{\frac{1}{\mu_2}}\Big[[\mathbf{C}|\mathbf{C}_o] + \frac{1}{\mu_2}\mathbf{Y}_2\Big] \tag{18}$$

Step 2. Update $\mathbf{E}_v$: given $\mathcal{L}_\tau(\mathbf{x}) = \max(0, 1-\frac{\tau}{\|\mathbf{x}\|_2})\mathbf{x}$, the subproblem is solved with closed form solution:

$$\underset{\mathbf{E}_v}{\text{argmin}}\ \frac{\lambda}{\mu_3}\|\mathbf{E_v}\|_{2,1} + \frac{1}{2}\|\mathbf{E_v} - \mathbf{p}\cdot vec(\mathbf{E}) - \frac{1}{\mu_3}\mathbf{Y}_3\|_F^2, \tag{19}$$

$$\Rightarrow \mathbf{E}_v^* = \mathcal{L}_{\frac{\lambda}{\mu_3}}\Big[\mathbf{p}\cdot vec(\mathbf{E}) + \frac{1}{\mu_3}\mathbf{Y_3}\Big] \tag{20}$$

Step 3. Update $\mathbf{C}$: the subproblem is solved with:

$$\begin{aligned} \underset{\mathbf{C}}{\text{argmin}}\ & tr\Big[\mathbf{Y}_1^T(\mathbf{X}-\mathbf{UC}-\mathbf{E})\Big] + \frac{\mu_1}{2}\|\mathbf{X}-\mathbf{UC}-\mathbf{E}\|_F^2 \\ & + tr\Big[\mathbf{Y}_2^T([\mathbf{C}|\mathbf{C}_o]-\mathbf{C}_n)\Big] \\ & + \frac{\mu_2}{2}\|[\mathbf{C}|\mathbf{C}_o]-\mathbf{C}_n\|_F^2 \end{aligned} \tag{21}$$

$$\begin{aligned} \Rightarrow \mathbf{C}^* = & \left(1+\frac{\mu_2}{\mu_1}\right)^{-1}\mathbf{I}\Big[\mathbf{U}^T(\mathbf{X}-\mathbf{E}) \\ & + \frac{\mu_2}{\mu_1}[\mathbf{C}_n]_{1:K} + \frac{1}{\mu_1}(\mathbf{U}^T\mathbf{Y}_1 - [\mathbf{Y}_2]_{1:K})\Big] \end{aligned} \tag{22}$$

Step 4. Update$\mathbf{E}$: the subproblem is solved with:

$$\begin{aligned} \underset{\mathbf{E}}{\text{argmin}}\ & tr\Big[\mathbf{Y}_1^T(\mathbf{X}-\mathbf{UC}-\mathbf{E})\Big] \\ & + \frac{\mu_1}{2}\|\mathbf{X}-\mathbf{UC}-\mathbf{E}\|_F^2 \\ & + tr\Big[\mathbf{Y}_3^T(\mathbf{p}\cdot vec(\mathbf{E}) - \mathbf{E}_v)\Big] \\ & + \frac{\mu_3}{2}\|\mathbf{p}\cdot vec(\mathbf{E}) - \mathbf{E}_v\|_F^2, \end{aligned} \tag{23}$$

$$\begin{aligned} \Rightarrow \mathbf{E}^* = & \left(1+\frac{\mu_3}{\mu_1}\right)^{-1} \\ & \mathbf{I}\Big[\mathbf{X}-\mathbf{UC}) + \frac{\mu_3}{\mu_1}\Big[vec^{-1}(\mathbf{p}^{-1}\cdot\mathbf{E}_v)\Big] \\ & + \frac{1}{\mu_1}\Big(\mathbf{U}^T\mathbf{Y}_1 - vec^{-1}(\mathbf{Y}_3)\Big)\Big] \end{aligned} \tag{24}$$

Step 5. Update $\mathbf{Y}_{1-3}$ **and** μ_{1-3}: the multipliers are updated with $\gamma > 1$:

$$\begin{cases} \mathbf{Y}_1 \leftarrow \mathbf{Y}_1 + \mu_1(\mathbf{X} - \mathbf{UC} - \mathbf{E}), \\ \mathbf{Y}_2 \leftarrow \mathbf{Y}_2 + \mu_2([\mathbf{C}|\mathbf{C}_o] - \mathbf{C}_n), \\ \mathbf{Y}_3 \leftarrow \mathbf{Y}_3 + \mu_3(vec(\mathbf{E}) - \mathbf{E}_v), \\ \mu_1 \leftarrow \gamma\mu_1,\ \mu_2 \leftarrow \gamma\mu_2,\ \mu_3 \leftarrow \gamma\mu_3. \end{cases} \tag{25}$$

Complexity Analysis We summarize the alternating optimization in Algorithm 1. Step 1–4 consist of simple linear algebra with an average of $\mathcal{O}(NK)$ operations. The computational bottleneck is the *singular value decomposition* (SVD) to update $\mathbf{C}_n^*$ in Step 5, which needs $\mathcal{O}(M^2K)$ operations. Following the approximation given in Zhang et al. (2015), the total complexity is $\mathcal{O}\left((M^2K + NK)\epsilon^{-0.5}\right)$, where $\mathcal{O}(\epsilon^{-0.5})$ is the iteration number.

In our case, N is the length of the concatenated part-based representation vector $\mathcal{A} \in \mathbb{R}^{L(d+2)}$ in Eq. 3. M is the number of eignvectors of the representation subspace $\mathbf{U}$. K is the number of particles for the ensemble initialization. Given the fact that $M \ll N$, the alternating iterations of ALM converge very fast with quadratic rate (Beck and Teboulle 2009). To further reduce the computational cost, we perform PCA on both of the feature space and representation space to cut down N and M. Note that we only compress the feature space but keep the (x, y) coordinates unchanged to avoid impairing the spatial information. The computation complexity can be further cut down by employing more efficient SVD Algorithm such as Wu and Stathopoulos (2015), Wu et al. (2016), given the fact that $[\mathbf{C}|\mathbf{C}_o]$ is low rank.

5 Experiments

We carried out extensive experiments to fully investigate the proposed approach. In this section, we first introduce the datasets and implementation details. Then we conduct component-wise validations using different experimental settings. Finally, we compare our approach with a bunch of state of the arts to demonstrate its superior performance in challenging conditions.

5.1 Datasets and Implementation Details

We briefly introduce the image and video datasets used in experiments and the implementation details.

Datasets The image datasets were mainly used to train the representation subspace $\mathbf{U}$ and the fitting evaluation network, while the video datasets were used to evaluate the performance and perform comparisons. Three image datasets were used for offline training. All images have 68-landmark annotations defined in Sagonas et al. (2013):

- *Multi-PIE* (Gross et al. 2010) contains images of 337 subjects under 15 view points and 7 expressions. We collected 1,300 images from this dataset.
- *LFPW* (Belhumeur et al. 2013) was recorded in wild conditions with extensive variations in both subject and imaging conditions. Totally 1,035 out of 1,400 images were collected.
- *Helen* (Le et al. 2012) was also collected in unconstrained conditions. We used all the 2,330 training and testing images. The landmark annotations are provided by Sagonas et al. (2013, 2016).

Besides the image datasets, we also employed four video datasets for online testing. All videos have the same 68-landmark annotations:

- *Talking Face* (FGNet 2004) contains 5 consecutive clips of totally 5000 frames recorded in controlled environment. We converted the annotations to the 68-landmark scheme for evaluation consistency.
- *Face Movie* (Peng et al. 2015) consists of movie clips that present unconstrained 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.
- *YtbVW* (Kim et al. 2008) was collected from internet under low resolution settings and presents challenges in multiple aspects. We collected 6 clips[1] and manually labeled 50% frames for quantitative evaluation.
- *300-VW* (Shen et al. 2015) contains 114 video clips recorded in three different wild conditions: Scenario 1, Scenario 2 and Scenario 3, corresponding to well-lit, mild unconstrained and completely unconstrained conditions, respectively. 20 videos were used in our experiments.

Although we could obtain much more training samples from videos, the variations present in video datasets are limited compared with image datasets due to the frame-wise redundancy. Therefore, we used image instead of video datasets to train the offline model.

Implementation Details To train the representation subspace $\mathbf{U}$, we first performed procrustes analysis (Saragih et al. 2011) based on a mean shape to remove any rigid 2D transformation among all images in the training set. The interocular distance is set to 50 pixels. Then, we extracted SIFT feature (Lowe 2004) around each landmark for part-based representation as it is robust to illumination and scale variations. Finally, $\mathbf{U}$ was trained by performing PCA and preserving 80% variations on the normalized training set. We used nor-

[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 (6) 1847_01_005_victoria_beckham.

Table 1 Average fitting error and time cost *w.r.t.* the number of sampled particles (initial shapes) at each frame

K	10	20	30	50	100
RMSE	5.45	5.19	4.92	4.84	4.81
Time (ms)	37	45	53	77	121

malized *root mean square error* (RMSE; Sagonas et al. 2013; Yang et al. 2015).

The latent state $\mathbf{s} \in \mathbb{R}^4$ is defined to control scale, rotation and 2D translation of the initial shape in the ensemble initialization. We set $K = 30$ as the particles sampled in each frame. The quantitative analysis of choosing K will be discussed soon in next section. To address the particles degeneration issue (Mei and Ling 2009), we employ resampling technique (Doucet et al. 2001) to initialize particles in every 50 frames.

We used Caffe (Jia et al. 2014) to train the fitting evaluation network. The network inputs were normalized to [0, 1] and mini batch size was set to 384. We optimized the network parameters by using *stochastic gradient descent* (SGD) with 0.9 momentum. The learning rate started at 0.001 and we performed the training for 10 epochs.

5.2 Algorithm Validation and Discussion

We conducted following experiments to validate the proposed approach in different aspects: particle number, error contribution, anchored representation and fitting evaluation.

Validation of Particle Number The number of sampled particles in Eq. 1 is an important parameter for ensemble initialization. To investigate the affect of different number of initial shapes, we changed K from 10 to 100 and repeated the testing on YtbVW datasets (Kim et al. 2008). We used SDM (Xiong and De la Torre 2013) to perform generic alignment in parallel.

The fitting errors and average time costs of the constrained decomposition and fitting recovery were recorded in Table 1. As K increases, the fitting error decreases while the average time cost grows approximate linearly. However, the improvement of the fitting accuracy is limited after $K \geq 50$ but the computational cost of each frame increases significantly. Therefore, we empirically set $K = 30$ in all experiments to achieve a good trade-off between the fitting accuracy and efficiency.

Validation of Error Contribution In Eq. 6, parameter ρ balances the error contributions between the shape and appearance in group-sparse error constraint: when $\rho \to 0$, the constrained decomposition completely counts on appearance errors; when $\rho \to \infty$, the constrained decomposition completely counts on shape errors. In our experiments, we tested ρ in the range of $[10, 10^4]$ on both controlled and unconstrained databases. We recorded average fitting errors in Fig. 5. We can see that the fitting errors changes little when ρ varies from 10^1 to 10^3, in other words, the fitting performance is insensitive to $\mathbf{p}$ in this range. Therefore, we simply set $\rho = 10^2$ in all experiments.

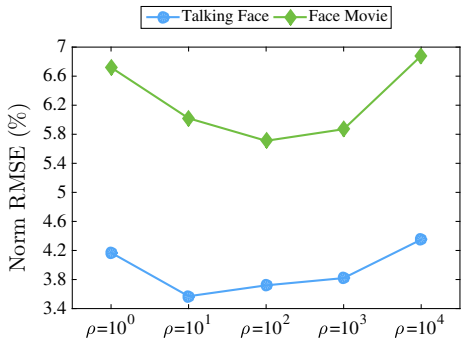

Fig. 5 Average fitting error *w.r.t.* different error contribution

Table 2 Average fitting error *w.r.t.* the number of anchored representation in the low-rank constraint

K_o	0	$K/10$	$K/5$	$K/2$
RMSE	5.21	4.92	5.04	5.37

Validation of Anchored Representation As we mentioned in Sect. 3.3, we used $\mathbf{C}_o$ as a prior to the low-rank constraint to avoid unreasonable shape deformation . More specifically, the prior information is the encoding of well-aligned candidates output by former frames. Similar to the anchored shapes used in Cheng et al. (2012) for robust alignment, we employ $\mathbf{C}_o$ as the anchored representation to incorporate additional temporal consistency to the low-rank constraint. To investigate the relation between the low-rank decomposition and the number of anchored representation, we changed K_o and reported the average fitting errors on YtbVW datasets (Kim et al. 2008) in Table 2. It is interesting to find that the fitting accuracy degrades obviously when $K_o \geq K/5$. The anchored representation contributes too much constraint on the low-rank decomposition, which drags the recovered shape away from the ground truth to the well-aligned candidates.

Validation of Deep Fitting Evaluation The online fitting evaluation is crucial to the success of personalized modeling. Adaptation using erroneous fittings will drift the offline learned subspace and eventually lead to failure. We fine-tuned the proposed fitting evaluation network using image datasets. We sampled 5 perturbations for each image, where the ground-truth shape and perturbed shape were labeled as positive and negative respectively. The perturbations were generated by scaling, rotating, and shifting the ground truth. Note that we used more negative samples than positive to guarantee the network is discriminative enough to misaligned shapes.

Table 3 Comparison of misalignment detection accuracy

	Face movie	YtbVW	300-VW
Threshold (%)	89.7	75.5	76.9
Deep (%)	94.4	88.1	85.3

To better investigate the fitting evaluation network, we compared it with the fitting evaluation method used in Peng et al. (2015), which detects misalignment using a threshold. The threshold was learned offline by exploring the structure of the decomposed group-sparse errors. The average misalignment detection accuracy are compared in Table 3. The proposed deep fitting evaluation outperforms the threshold method significantly in different video datasets. It achieves around 90% accuracy in general, which can robustly detect erroneous fittings in challenging conditions to alleviate the drifting issue.

5.3 Comparison with Previous Work

To further evaluate performance, we compared our approach with a bunch of generic methods. The methods include:

- *DRMF* (Asthana et al. 2013), discriminative response map fitting.
- *ESR* (Cao et al. 2014), explicit shape regression.
- *SDM* (Xiong and De la Torre 2013), supervised descent method.
- *IFA* (Asthana et al. 2014), incremental face alignment.
- *RLB* (Ren et al. 2014), regressing local binary features.
- *CFAN* (Zhang et al. 2014a), coarse-to-fine auto-encoder network.
- *CFSS* (Zhu et al. 2015), coarse-to-fine shape searching.
- *POCR* (Tzimiropoulos 2015), project-out cascaded regression.

We also compared the proposed method with two joint alignment methods, which includes:

- *A-RASL* (Cheng et al. 2012), anchored robust ensemble alignment.
- *RAPS* (Sagonas et al. 2014), robust person-specific deformable models.

All these methods were recently proposed and reported state-of-the-art performance in the corresponding categories. For fair comparisons, we evaluated these method in a tracking protocol: fitting result of current frame was used as the initial shape (DRMF, SDM and IFA) or the bounding box (ESR, CFSS and POCR) in the next frame.

Comparison with Generic Methods We compared our approach with generic methods on Face Movie dataset to evaluate the proposed ensemble alignment and personalized adaption methods. ESR (Cao et al. 2014) and SDM (Xiong and De la Torre 2013) were employed as the baselines for the generic alignment.

We report the average Norm RMSE of different approaches in Fig. 6. 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, such as 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.

Comparison with Joint Methods We compared our approach with joint alignment methods on Talking Face (FGNet 2004). This dataset contains 5 consecutive clips of totally 5000 frames recorded in controlled environment. We converted the original landmark annotations to the standard 68-point scheme (Sagonas et al. 2013) for evaluation consistency. We implemented two joint alignment approaches: (1) A-RASL (Cheng et al. 2012), and (2) RAPS (Sagonas et al. 2014). For fairly comparison, we trained the clean face subspace for RAPS on the training set, and used 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 Fig. 7. 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 (Sagonas et al. 2014). The ensemble initialization and person-specific modeling make our approach have the best performance in

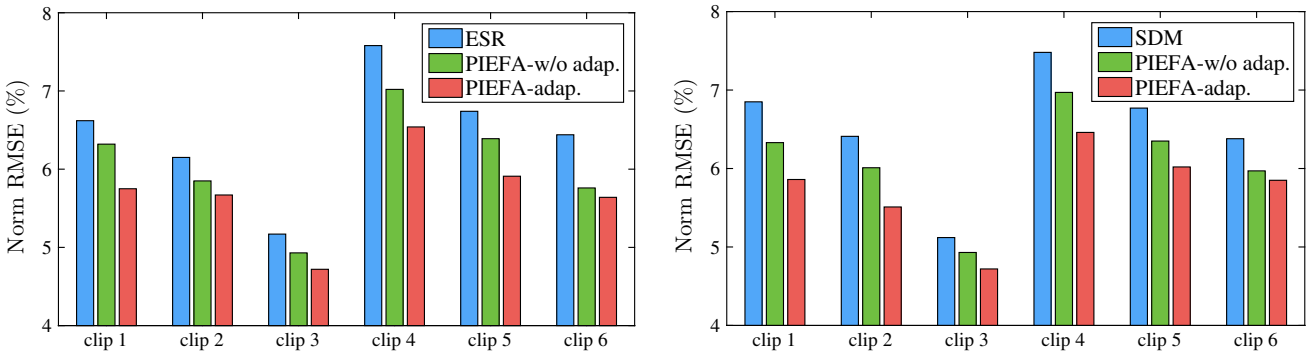

Fig. 6 Comparions with generic alignment methods on face movie (Peng et al. 2015) *w.r.t.* average fitting accuracy in each clip

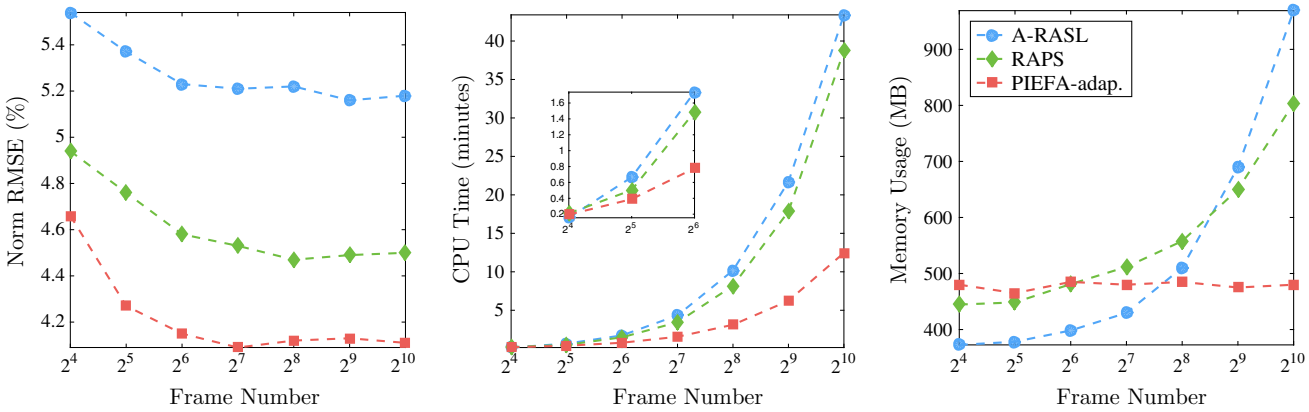

Fig. 7 Comparions with joint alignment methods on talking face (FGNet 2004) *w.r.t.* fitting accuracy, CPU time and memory usage

general *w.r.t.* both converge speed and finial accuracy. (2) The CPU time costs of both A-RASL and RAPS 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.

Comparison with the State of the Arts We compared the proposed method with the state of the arts on YtbVW (Kim et al. 2008) and 300-VW (Shen et al. 2015). The cumulative error distribution (CED) curves of different methods of different methods were compared in Fig. 8. Our approach has the most steep cumulative error distribution curve and outperforms other methods with a substantial margin. It indicates that our method has the smallest fitting errors in general, and the improvement is more significant on challenging frames. Some fitting results of different methods are shown in Fig. 9. We can see that our approach substantially improves the fittings for landmarks around mouth and face contour when extensive pose variations and expression changes exist. In these cases, the single initial shape used in generic alignment method, e.g. ESR (Cao et al. 2014) and SDM (Xiong and De la Torre 2013), is usually far away from the ground-truth, which inevitably results in local optimum and unsatisfaction. Our approach, on the other hand, takes advantage of both motion cues and person-specific information to perform ensemble initialization and robust decomposition, which effectively overcomes the imperfect initialization issue.

The recently proposed methods such as CFSS (Zhu et al. 2015) and POCR (Tzimiropoulos 2015) have better performance in general than ESR and SDM. Although similar coarse-to-fine frameworks are used in these methods for fitting optimization, CFSS and POCR can mitigate the initialization-sensitive issue by either searching the best initial shape or learning descent directions orthogonal to the appearance variation. Besides, we also observed that CNNs based methods such as CFAN (Zhang et al. 2014a) are more robust to variations than cascade regression.

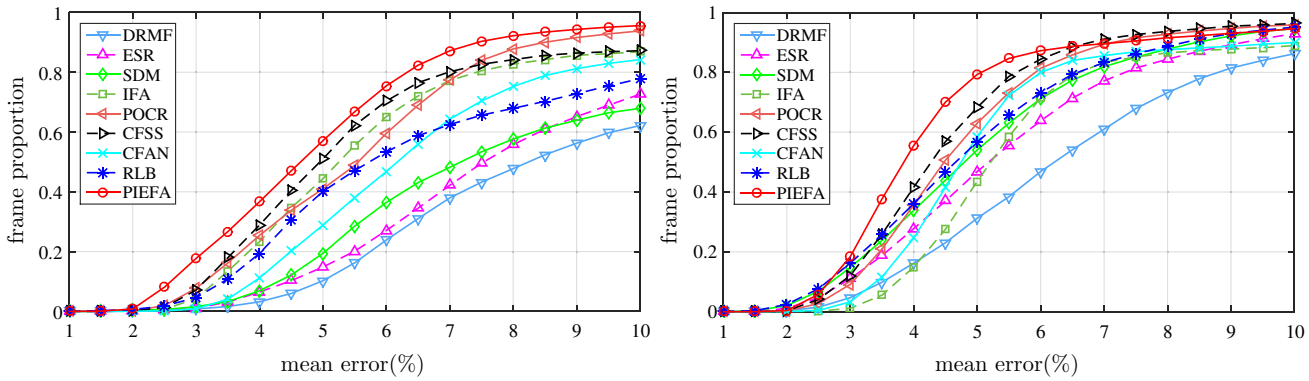

Fig. 8 Comparisons of cumulative error distribution curves with state of the arts on on YtbVW (Kim et al. 2008) and 300-VW (Shen et al. 2015)

Fig. 9 Examples of fitting results on face movie (Peng et al. 2015) and YtbVW (Kim et al. 2008): *first* and *thrid rows*, PIEFA; *second row*, ESR (*column 1–5*) and SDM (*column 6–10*); *last row*, RAPS (*column 1–2*), A-RASL (*3–4*), RLMS (*5–6*), RLB (*7–8*) and IFA (*9–10*). There is consistant fitting improvement for *lardmarks* around *eyes*, *mouth* and *face contour*

To sum up, the experiments prove that our approach can effectively overcome the pool initialization of existing generic alignment methods. The proposed incremental framework can process large-scale and real-time data with constant computational cost, which is a great merit compared with former joint alignment methods. Moreover, the proposed incremental adaptation can achieve personalized modeling in wild conditions for robust alignment, while the drifting issue can be effectively alleviated by the proposed deep fitting evaluation network.

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 deep learning based features in the future work to further improve the fitting accuracy and efficiency.

References

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3444–3451).

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Baró, X., Gonzalez, J., Fabian, J., Bautista, M. A., Oliu, M., Escalante, H. J., Guyon, I., & Escalera, S. (2015). Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1–9). IEEE.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*(1), 183–202.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*, 2930–2940.

Black, M., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 374–381).

Brand, M. (2006). Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra and Its Applications*, *415*(1), 20–30.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, *107*(2), 177–190.

Cheng, X., Fookes, C., Sridharan, S., Saragih, J., & Lucey, S. (2013). Deformable face ensemble alignment with robust grouped-l1 anchors. In: Automatic Face and Gesture Recognition (FG). In *IEEE International Conference and Workshops on* (pp. 1–7). IEEE.

Cheng, X., Sridharan, S., Saraghi, J., & Lucey, S. (2012). Anchored deformable face ensemble alignment. In *European Conference on Computer Vision* (pp. 133–142). Berlin: Springer.

Cheng, X., Sridharan, S., Saragih, J., & Lucey, S. (2013). Rank minimization across appearance and shape for aam ensemble fitting. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 577–584).

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *23*(6), 681–685.

Decarlo, D., & Metaxas, D. (2000). Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, *38*(2), 99–127.

Doucet, A., De Freitas, N., & Gordon, N. (2001). An introduction to sequential monte carlo methods. In Sequential Monte (Ed.), *Carlo methods in practice* (pp. 3–14). Berlin: Springer.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, *20*(2), 303–353.

Escalera, S., Gonzalez, J., Baró, X., Pardo, P., Fabian, J., Oliu, M., Escalante, H. J., Huerta, I., & Guyon, I. (2015). Chalearn looking at people 2015 new competitions: Age estimation and cultural event recognition. In *International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

FGNet. (2004). Talking face video.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image Vision Computing (IVC)*, *28*(5), 807–813.

He, J., Balzano, L., & Szlam, A. (2012). Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1568–1575). IEEE.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACMM* (pp. 675–678).

Kim, M., Kumar, S., Pavlovic, V., & Rowley, H. (2008). Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008*. CVPR 2008 (pp. 1–8). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)* (pp. 679–692).

Lin, Z., Chen, M., & Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Mei, X., & Ling, H. (2009). Robust visual tracking using & #x2113; 1 minimization. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 1436–1443). IEEE.

Nasrollahi, K., Escalera, S., Rasti, P., Anbarjafari, G., Baro, X., Escalante, H. J., & Moeslund, T.B. (2015). Deep learning based super-resolution for improved action recognition. In: Image Processing Theory, Tools and Applications (IPTA). In *2015 International Conference on IEEE* (pp. 67–72). IEEE.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Patras, I., & Pantic, M. (2004). Particle filtering with factorized likelihoodsfor tracking facial features. In *The IEEE International Conference on Automatic Face and Gesture Recognition (FG)* (pp. 97–102).

Peng, X., Feris, R. S., Wang, X., & Metaxas, D. N. (2016). A recurrent encoder-decoder network for sequential face alignment. In *European Conference on Computer Vision* (pp. 38–56). Berlin: Springer.

Peng, X., Zhang, S., Yang, Y., & Metaxas, D. N. (2015). Piefa: Personalized incremental and ensemble face alignment. In *The IEEE International Conference on Computer Vision (ICCV)*.

Peng, Y., Ganesh, A., Wright, J., Xu, W., & Ma, Y. (2010). RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Perakis, P., Passalis, G., Theoharis, T., & Kakadiaris, I. A. (2013). 3d facial landmark detection under large yaw and expression variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*(7), 1552–1564.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. In *Image and Vision Computing* (vol. 47, pp. 3–18). 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2014). Raps: Robust and efficient automatic construction of person-specific deformable models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1789–1796).

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision (IJCV)*, *91*(2), 200–215.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015) Facenet: A unified embedding for face recognition and clustering. In *CVPR* (pp. 815–823).

Shen, J., Zafeiriou, S., Chrysos, G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015) The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3476–3483).

Sung, J., & Kim, D. (2009). Adaptive active appearance model with incremental learning. *Pattern Recognition Letters (PRL)*, *30*(4), 359–367.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *CVPR*.

Tang, M., & Peng, X. (2012). Robust tracking with discriminative ranking lists. *IEEE Transactions on Image Processing (TIP)*, *21*(7), 3273–3281.

Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E., & Zafeiriou, S. (2016). Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *IEEE International Conference on Computer Vision Pattern Recognition (CVPR)*.

Tzimiropoulos, G. (2015). Project-out cascaded regression with an application to face alignment. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3659–3667). IEEE.

Tzimiropoulos, G., & Pantic, M. (2014). Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1851–1858).

Vogler, C., Li, Z., Kanaujia, A., Goldenstein, S., & Metaxas, D. (2007). The best of both worlds: Combining 3d deformable models with active shape models. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 1–7). IEEE.

Wang, Z., Mi, H., & Ittycheriah, A. (2016a). Semi-supervised clustering for short text via deep representation learning. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)* (pp. 31–39).

Wang, Z., Mi, H., & Ittycheriah, A. (2016b). Sentence similarity learning by lexical decomposition and composition. In *Coling 2016*.

Wang, Z., Mi, H., & Nianwen, X. (2015). Feature optimization for constituent parsing via neural networks. In *Proceedings of ACL 2015* (pp. 1138–1147).

Wu, L., Romero, E., & Stathopoulos, A. (2016). A high-performance preconditioned svd solver for accurate large-scale computations. *SIAM Journal on Scientific Computing*. arXiv:1607.01404.

Wu, L., & Stathopoulos, A. (2015). A preconditioned hybrid svd method for accurately computing singular triplets of large matrices. *SIAM Journal on Scientific Computing*, *37*(5), S365–S388.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan, J., Lei, Z., Yi, D., & Li, S. (2013). Learn to combine multiple hypotheses for accurate face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 392–396).

Yang, H., Jia, X., Loy, C. C., & Robinson, P. (2015). An empirical study of recent face alignment methods. arXiv preprint arXiv:1511.05049.

Zafeiriou, L., Antonakos, E., Zafeiriou, S., & Pantic, M. (2014). Joint unsupervised face alignment and behaviour analysis. In D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds.) *European Conference on Computer Vision (ECCV)* (pp. 167–183).

Zhang, J., Shan, S., Kan, M., & Chen, X. (2014a). Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *European Conference on Computer Vision (ECCV)* (pp. 1–16).

Zhang, T., Liu, S., Ahuja, N., Yang, M. H., & Ghanem, B. (2015). Robust visual tracking via consistent low-rank sparse learning. *International Journal of Computer Vision*, *111*(2), 171–190.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014b). Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)* (pp. 94–108).

Zhao, C., Cham, W. K., & Wang, X. (2011). Joint face alignment with a generic deformable face model. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 561–568). IEEE.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4998–5006).

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation and landmark estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.