Track Facial Points in Unconstrained Videos

Xi Peng[1]
xipeng.cs@rutgers.edu
Qiong Hu[1]
qionghu.cs@rutgers.edu
Junzhou Huang[2]
jzhuang@uta.edu
Dimitris N. Metaxas[1]
dnm@cs.rutgers.edu

[1] Department of Computer Science
Rutgers University
New Jersey, USA

[2] Department of Computer Science
The University of Texas at Arlington
Texas, USA

Abstract

Tracking Facial Points in unconstrained videos is challenging due to the non-rigid deformation that changes over time. In this paper, we propose to exploit incremental learning for person-specific alignment in wild conditions. Our approach takes advantage of part-based representation and cascade regression for robust and efficient alignment on each frame. Unlike existing methods that usually rely on models trained offline, we incrementally update the representation subspace and the cascade of regressors in a unified framework to achieve personalized modeling on the fly. To alleviate the drifting issue, the fitting results are evaluated using a deep neural network, where well-aligned faces are picked out to incrementally update the representation and fitting models. Both image and video datasets are employed to valid the proposed method. The results demonstrate the superior performance of our approach compared with existing approaches in terms of fitting accuracy and efficiency.

1 Introduction

Fitting facial landmarks on sequential images plays a fundamental role in many computer vision tasks, such as face recognition [20, 34], expression analysis [12, 18], and facial unit detection [40, 43]. It is a challenging task since the face undergoes drastic non-rigid deformations caused by extensive pose and expression variations, as well as unconstrained imaging conditions like illuminations changes and partial occlusions.

Despite the long history of research in rigid and non-rigid face tracking [4, 21], current efforts have mostly focused on face alignment on a single image [6, 29, 32, 36, 39, 44, 46, 47, 48, 49]. They have shown great success with impressive results in standard benchmark datasets [27, 45]. However, when it comes to sequential images, many of them suffer from significant performance degradation especially in real-world scenarios under wild conditions [30]. They usually rely on models trained offline on still images and perform sequential alignment in a tracking-by-detection manner [7, 30, 35]. They lack the capability to capture neither the specifics of the tracked subject nor the imaging continuity in successive frames. To this end, personalized modeling rather than generic detection is preferred.

Pages 129.1-129.13
DOI: https://dx.doi.org/10.5244/C.30.129

One rational way to achieve personalized modeling is to perform joint face alignment [25, 28], which takes the advantage of the shape and appearance consistency in a sequence to minimize fitting errors of all frames at the same time. However, these methods are restricted to offline tasks since they usually require all images are available before image congealing [25]. They also suffer from low-efficiency issue which severely impedes their performance on real-time or large-scale tasks [24].

To avoid these limitations, other approaches attempt to incrementally construct personalized models instead of joint alignment. They either adapt the holistic face representation using incremental subspace learning [33] or update the cascade mapping using online regression [1]. However, how to jointly update the both in a unified framework still remains an open question without investigation. Besides, former approaches often employ holistic models to facilitate the adaptation [33], which is inferior to part-based models in challenging conditions [29, 49]. Moreover, many of them attempt to achieve personalized modeling without correction, which may inevitably result in model drifting.

In this paper, we further exploit person-specific modeling for sequential face alignment to address aforementioned issues. We first learn the part-based representation to model the facial shape and appearance respectively, as well as a cascade of nonlinear mappings from the facial appearance to the shape. The representation subspace and mapping parameters are then incrementally updated in a unified framework to achieve personalized modeling on the fly. In summary, our work makes the following **contributions**:

- We propose a novel approach for sequential face alignment. The person-specific modeling is investigated by incrementally learning the representation subspace and the cascade of regressors in a unified framework.
- The proposed part-based representation together with the cascade regression guarantees robust alignment in unconstrained conditions. More importantly, they are crucial to efficiently construct personalized models for real-time or large-scale applications.
- We propose to leverage deep neural networks for efficient and robust fitting evaluation. It significantly alleviates the drifting issue which would severely deteriorate learned models and inevitably lead to failure.

To fully evaluate the performance of our approach, we employed both image and video datasets in the experiments and compared our method with the state of the arts in terms of fitting accuracy and efficiency. We conducted detailed experimental analysis to validate each component of our approach. The results demonstrate that the proposed incremental learning can significantly improve the fitting accuracy with an affordable computational cost, especially in unconstrained videos with extensive variations.

2 Related Work

Face alignment in a single image has attracted intensive research interest for decades. Generally speaking, existing methods usually accomplish the task by learning a nonlinear mapping, which can be either regressors [39, 44, 48] or neural networks [32, 46, 47], from the facial representation, which is either holistic [5, 9] or part-based [29, 49], to landmark coordinates.

It has been proved that the part-based rather than the holistic representation is more robust to the extensive variations in unconstrained settings. For instance, Saragih *et al.* [29] proposed the *regularized landmark mean-shift* (RLMS) to maximize the joint probability of

the reconstructed shape based on a set of response maps extracted around each landmark using expectation maximization. Asthana et al. [2] proposed the *discriminative response map fitting* (DRMF) to learn boosted mappings from the joint response maps to shape parameters. Cao et al. [6] combined a two-level regression to achieve *explicit shape regression* (ESR) using shape-indexed features. Xiong et al. [44] proposed *supervised descent method* (SDM) to learn a sequence of descent directions using nonlinear least squares.

More recently, *deep neural networks* (DNNs) based methods have made significant progress towards systems that work in real-world scenarios [37, 38]. For example, Sun *et al.* [32] proposed to concatenate three-level convolutional neural networks to refine the fitting results from the initial estimation. Zhang *et al.* [46] employed the similar idea of the coarse-to-fine framework but using auto-encoder networks instead of CNNs. Zhang *et al.* [47] showed that learning face alignment together with other correlated tasks, such as identity recognition and pose estimation, can improve the landmark detection accuracy.

The aforementioned methods have shown impressive results in standard benchmark datasets [27]. However, they still suffer from limited performance in the sequential task as they completely rely on static models trained offline. To address this limitation, efforts of constructing person-specific models are made to improve the performance of sequential face alignment.

Some of them achieve person-specific modeling via joint face alignment. A representative example was proposed in [28], which used a clean face subspace trained offline to minimize fitting errors of all frames at the same time. However, these methods are usually limited to offline tasks due to their intensive computational costs. Others attempt to incrementally construct personalized models on the fly. For instance, Sung *et al.* [33] proposed to employ incremental principle component analysis to adapt the holistic AAMs to achieve personalized representation. Asthana *et al.* [1] further explored SDM in *incremental face alignment* (IFA) by simultaneously updating regressors in the cascade using incremental least squares. However, faithful personalized models can hardly be achieved without joint adaptation of the representation and fitting models in a unified framework. More importantly, blind model adaptation without correction would inevitably result in model drifting. How to effectively detect misalignment is still a challenging question that is seldom investigated. To address this issue, we propose a deep neural network for robust fitting evaluation to pick out well-aligned faces from misalignment, which are then used to incrementally update the representation subspace and fitting strategy for robust person-specific modeling on the fly.

3 Our Approach

In this paper, we propose a novel approach for face alignment in unconstrained videos. We first learn the *part-based representations* to model the facial shape and appearance respectively. The *discriminative fitting* is performed by learning a cascade of regressors that maps from the appearance representation to the shape parameters. Then personalized modeling is achieved by *incremental representation update* and *fitting adaptation in parallel*. Finally, we propose a deep fitting evaluation to alleviate the drifting issue.

3.1 Part-Based Representations

Our goal is to jointly learn *the shape representation* and *appearance representation* using part-based models. Both representations should be compact and efficient to facilitate incremental person-specific modeling.

The shape representation is learned by firstly performing Procrustes analysis [8] on training images to obtain normalized facial shapes. Then we apply principle component analysis (PCA) to obtain the mean shape and eigenvectors $\{\mathbf{M}^s, \mathbf{V}^s\}$, where s denotes shape. An instance shape can be modeled as $\mathbf{s}(\mathbf{p}) = \mathbf{M}^s + \mathbf{V}^s\mathbf{p}$, where $\mathbf{p}$ is the shape representation.

The appearance representation is learned using local response maps [29]. Given a image $\mathbf{I}$ and the shape representation $\mathbf{p}$, the local response map around the l-th landmark is $\mathbf{A}_l(\mathbf{p},\mathbf{I}) = 1/(1+exp(a_l\phi(\mathbf{s}(\mathbf{p}),\mathbf{I})+b_l))$, where $\{a_l, b_l\}_{l=1}^L$ are patch experts learned 6by cross-validation. $\phi(\cdot)$ is the feature vector with a possible choice from SIFT, HOG, LBP, etc.

To simulate the appearance variation and obtain more robust representation [2], we sample perturbations $\{\Delta\mathbf{p}_{ij}\}$ around the ground-truth $\mathbf{p}_i^*$ as illustrated in Figure 1. The perturbed response maps are arranged as a tensor $\mathcal{T}_l = \{\mathbf{A}_l(\mathbf{p}_i^* + \Delta\mathbf{p}_{ij}, \mathbf{I}_i)\}_{i,j}$, where i and j count images and perturbations respectively. Similar to the shape representation, we apply PCA on $\mathcal{T}_l$ to obtain the mean and eigenvectors $\{\mathbf{M}_l^a, \mathbf{V}_l^a\}$, where a denotes appearance. The appearance representation of the l-th landmark can be calculated by fast projection $\mathbf{x}_l = (\mathbf{V}_l^a)^{-1}(\mathbf{A}_l(\mathbf{p},\mathbf{I}) - \mathbf{M}_l^a)$.

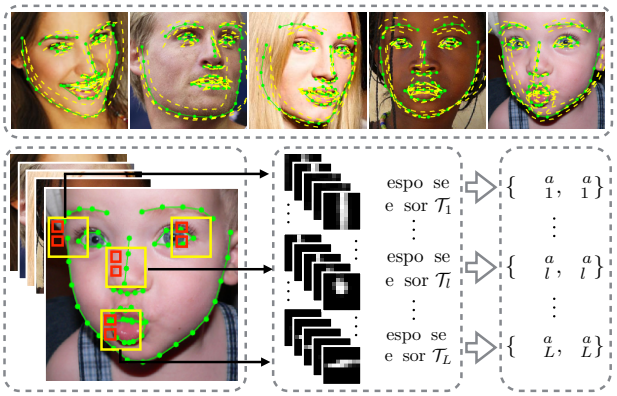

Figure 1: Top: perturbations (yellow dash) are sampled around the ground-truth shape (green dash). Bottom: response maps (yellow box) of the same landmark are arranged as tensor to learn the appearance representation.

Now we can model the shape and appearance of an instance face using $\mathbf{p}$ and $\mathbf{x}(\mathbf{p},\mathbf{I}) = \left[\mathbf{x}_1^T, \cdots; \mathbf{x}_L^T\right]^T$. The part-based representations are highly compact and efficient to compute. They are also robust to variations even for unseen images given the generative nature of parametric models [19, 22]. These merits facilitate the incremental learning for person-specific modeling which will be explained soon in Section 3.3.

3.2 Discriminative Fitting

The goal is to learn a cascade of non-linear mappings from the appearance representation $\mathbf{x}(\mathbf{p},\mathbf{I})$ to the shape update $\Delta\mathbf{p}$. We refine the shape representation $\mathbf{p}$ from an initial guess $\mathbf{p}^0$ to the ground truth step by step:

$$\mathbf{p}^{k+1} = \mathbf{p}^k + \mathbf{x}(\mathbf{p}^k,\mathbf{I})\mathbf{R}^k + \mathbf{b}^k, \tag{1}$$

where $\{\mathbf{R}^k, \mathbf{b}^k\}$ is the regressor at step k and $\mathbf{p}^*$ is the ground truth. Let $\Delta\mathbf{p}_{ij}^k = \mathbf{p}_i^\star - \mathbf{p}_{ij}^k$, the regressors can be computed by solving the least square problem [44]:

$$\arg\min_{\mathbf{R}^k,\mathbf{b}^k} \sum_{i=1}^{M}\sum_{j=1}^{N} ||\Delta\mathbf{p}_{ij}^k - \mathbf{x}(\mathbf{p}_{ij}^k,\mathbf{I}_i)\mathbf{R}^k - \mathbf{b}^k||^2. \tag{2}$$

Let $\tilde{\mathbf{R}}^k = \left[\mathbf{R}^{kT}\mathbf{b}^{kT}\right]^T$ and $\tilde{\mathbf{x}} = \left[\mathbf{x}(\mathbf{p}^k,\mathbf{I}_i)^T\mathbf{1}\right]^T$, the regressor can be computed with a closed-form solution $\tilde{\mathbf{R}}^k = \left[\tilde{\mathbf{x}}^T\tilde{\mathbf{x}} + \lambda\mathbf{I}\right]^{-1}\tilde{\mathbf{x}}^T\Delta\mathbf{p}^k$.

Former approaches [2, 6] employed boosted regressors for discriminative fitting. However, it is difficult to perform incremental learning under the boosting framework due to the heavy computational load to update a large number of week regressors. In contrast, the cascade of regressors is easy to train, fast in test, and can be effectively adapted in parallel on the fly. We leave the details in Section 3.4.

3.3 Incremental Representation Update

To achieve personalized representations of shape and appearance, our goal is to incrementally update the offline trained subspace $\{\mathbf{M}^s, \mathbf{V}^s\}$ and $\{\mathbf{M}^a_l, \mathbf{V}^a_l\}_{l=1}^L$ in a unified framework. Suppose the offline model is trained on m offline data T_A with mean M_A and eigenvectors V_A, where the SVD of T_A is $T_A = U\Sigma V^T$. Given n new online observations T_B with mean M_B, our task is equivalent to efficiently compute the SVD of the concatenation $\begin{bmatrix} T_A & T_B \end{bmatrix} = U'\Sigma' V'^T$.

It is infeasible to directly calculate the SVD as the entire offline training data need to be stored and reused online, which is extremely computationally expensive. Instead, we follow the *sequential Karhunen-Loeve* (SKL) algorithm [15, 26] to formulate the concatenation as:

$$\begin{bmatrix} U & E \end{bmatrix} \begin{bmatrix} \Sigma & U^T\hat{T}_B \\ \mathbf{0} & E(\hat{T}_B - UU^T\hat{T}_B) \end{bmatrix} \begin{bmatrix} V^T & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}, \tag{3}$$

where $\hat{T}_B = \left[T_B \ \sqrt{\frac{mn}{m+n}}(V_B - V_A) \right]$, $E = orth(\hat{T}_B - UU^T\hat{T}_B)$. Now we only need to perform SVD on the middle term instead of the entire concatenation:

$$T_C = \tilde{U}\tilde{\Sigma}\tilde{V}^T, \ T_C = \begin{bmatrix} \Sigma & U^T\hat{T}_B \\ \mathbf{0} & E(\hat{T}_B - UU^T\hat{T}_B) \end{bmatrix}. \tag{4}$$

By inserting T_C back to Equation 3, we have $\begin{bmatrix} T_A & T_B \end{bmatrix} = \left(\begin{bmatrix} U & E \end{bmatrix} \tilde{U} \right) \tilde{\Sigma} \left(\tilde{V}^T \begin{bmatrix} V^T & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \right)$. The mean and eigenvectors can be instantly updated:

$$\begin{aligned} M_{AB} &= \frac{m}{m+n}M_A + \frac{n}{m+n}M_B, \\ U' &= \begin{bmatrix} U & E \end{bmatrix} \tilde{U}, \ \Sigma' = \tilde{\Sigma}. \end{aligned} \tag{5}$$

Compared with the naive approach, the incremental subspace learning can significantly reduce the space complexity from $O(d(m+n))$ to $O(dn)$ and cut down the computational complexity from $O(d(m+n)^2)$ to $O(dn^2)$, where $m \gg n$ and d denotes the length of a single observation. It guarantees efficient modeling of the personalized representations.

3.4 Fitting Adaptation in Parallel

Once the shape and appearance representations are updated, we need to update the cascade of regressors instantly to catch up the online changes. However, adapting the cascade of regressors in a sequential order would be slow since $\tilde{\mathbf{R}}^k$ needs to be recomputed after $\tilde{\mathbf{R}}^{k-1}$. To address this issue, we follow [1] to decouple the dependence in the cascade by directly sample $\mathbf{p}^k$ from a norm distribution $\mathbf{p}^k \sim \mathcal{N}(\mathbf{p}^\star, \Lambda^k)$, where Λ^k is the shape variations learned offline. Once the cascade is flatten into independent mappings, all regressors can be simultaneously updated in parallel.

During the offline training, we compute $\tilde{\mathbf{x}}_A$ and $\tilde{\mathbf{R}}_A$ following the definition given in Section 3.2. During the online testing, we sample $\Delta\mathbf{p}_B$ based on the norm distribution and re-compute the new appearance representation $\tilde{\mathbf{x}}_B$. $\tilde{\mathbf{R}}_A$ can be adapted to $\tilde{\mathbf{R}}_{AB}$ by:

$$\tilde{\mathbf{R}}_{AB} = \tilde{\mathbf{R}}_A - P_{AB}\tilde{\mathbf{R}}_A + (P_A - P_{AB}P_A)(\tilde{\mathbf{x}}_B)^T\Delta\mathbf{p}_B, \tag{6}$$

where $P_A = \left[(\tilde{\mathbf{x}}_A)^T\tilde{\mathbf{R}}_A + \lambda I\right]^{-1}$, $P_B = \left[\tilde{\mathbf{x}}_B P_A \tilde{\mathbf{x}}_B^T + I\right]^{-1}$, and $P_{AB} = P_A\tilde{\mathbf{x}}_{AB}^T P_B\tilde{\mathbf{x}}_B$.

Given the fact that $d \gg n$, the computational cost of the matrix inversion in Equation 6 is significantly reduced from $O(d^3)$ to $O(n^3)$ by decoupling regressors in the cascade. It is also highly memory-efficient since we can pre-compute P_A offline and only a small number of online observations $\tilde{\mathbf{x}}_B$ need to be maintained for incremental adaptation.

3.5 Deep Fitting Evaluation

It is crucial to evaluate the fitting results since blind adaptation using erroneous fittings would inevitably result in model drifting. To address this issue, we leverage deep neural networks for robust fitting evaluation. Only well-fitted faces will be used to incrementally update the representation subspace and adapt the cascade of regressors for person-specific modeling

Our goal is to learn a deep neural network that takes the fitting results as input and outputs a binary label to indicate correct or erroneous 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 [41, 42]. However, we experienced very limited performance using this design in our experiments. The reason is that the pixel-wise spatial information diminishes significantly after a series of max-pooling operations [16]. The network can hardly learn the correct connection between the facial appearance and the landmark location.

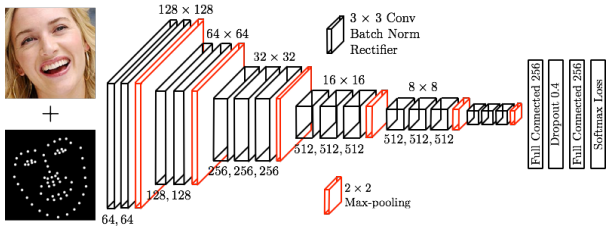

Figure 2: The architecture of the fitting evaluation network. It takes the concatenation of the face image and landmark map as input and outputs a binary label to indicate correct or erroneous alignment.

Instead, we propose to concatenate the facial image and the landmark map at the very beginning of the network as shown in Figure 2. Each pixel in the landmark map is a binary value that marks the presence of the corresponding landmark. Our network is designed based on a variant of the VGG-16 networks [31] which has a reduced number of fully connected neurons. We can, therefore, initialize the training process from weights trained on large datasets for object classification. 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 calculate cross-entropy loss for backpropagation.

The proposed deep fitting evaluation significantly outperforms former approach [1] that employs global and local handcrafted features for error detection, which will be discussed soon in Section 4.2. It is also very efficient, which takes less than 10ms to process one image using a single K40 GPU accelerator.

4 Experiments

We first introduce the datasets used in our experiments as well as detailed settings. Then we perform algorithm validation and discussion to evaluate the proposed method in different aspects. Finially, we compare our approach with state of the arts in different datasets to demonstrate its superior performance.

4.1 Datasets and Settings

Both image and video datasets were used to conduct the experiments. The image datasets were mainly used to train the representation subspace and the cascade of regressors offline, while the video datasets were used to evaluate the performance of the proposed method.

Four image datasets were used for offline training: **(1) MultiPIE** [11], **(2) LFPW** [3], **(3) Helen** [14], and **(4) AFLW** [13]. From each of them, we collected 1300, 1035, 2330, and 4050 images of total 8715 images with 68-landmark annotations [27].

Four video datasets were used for online testing: **(1) FGNET** [10], **(2) ASLV** [17], **(3) 300-VW** [30], and **(4) YtbVW** [24]. From each of them, we collected 5, 10, 20, and 6 videos of more than 30,000 frames for the evaluation. These videos present unconstrained challenges, such as pose/expression variations, illumination changes, and partial occlusions.

We trained multi-view models based on different yaw angers [23]: left $[-90°,-30°)$, frontal $[-30°,30°]$ and right $(30°,90°]$. All training images were registered to a reference 2D facial shape with an interocular distance of 50 pixels to remove any 2D rigid movement. We employed HoG features to best balance the fitting accuracy and efficiency. The size of the patch expert and the local support window were set to 11×11 and 21×21 respectively. We sampled 10 perturbations for each training image with the standard deviations of ±0.1 for scaling, $\pm10°$ for rotation, and ±10 pixels for translation. Normalized Root Mean Square Error (Norm RMSE) was used in all experiments.

4.2 Algorithm Validation and Discussion

We conducted following experiments to validate the proposed approach in different aspects: person-specific modeling, joint adaptation, and deep fitting evaluation.

Validation of person-specific modeling. To investigate the contribution of the proposed personalized modeling, we first trained the representation and fitting models using MultiPIE and then collected two clips from FGNET and ASLV for testing. Each clip contains 300 frames with intensive pose and expression variations. The testing was performed under two different settings: (1) incrementally update the representation and fitting models, and (2) without any model adaptation. The frame-wise Norm RMSE in Figure 3 shows that both settings have comparable accuracy at the beginning. The online version outperforms the offline version once the model adaptation was performed. The superior performance becomes more significant when intensive variations and partial occlusions exist (around frame 200 of FGNET and frame 150 of ASLV).

Validation of joint adaptation. To investigate the joint adaptation of representation and fitting models, we first trained the offline models using all the training images and then carried out experiments on the full sets of FGNET and ASLV under three different settings: (1) update the representation model, (2) update the fitting model, and (3) update both models. The cumulative fitting errors are recorded in Table 1. The results indicate that only adapt the fitting strategy has better performance than only update the representation subspace.

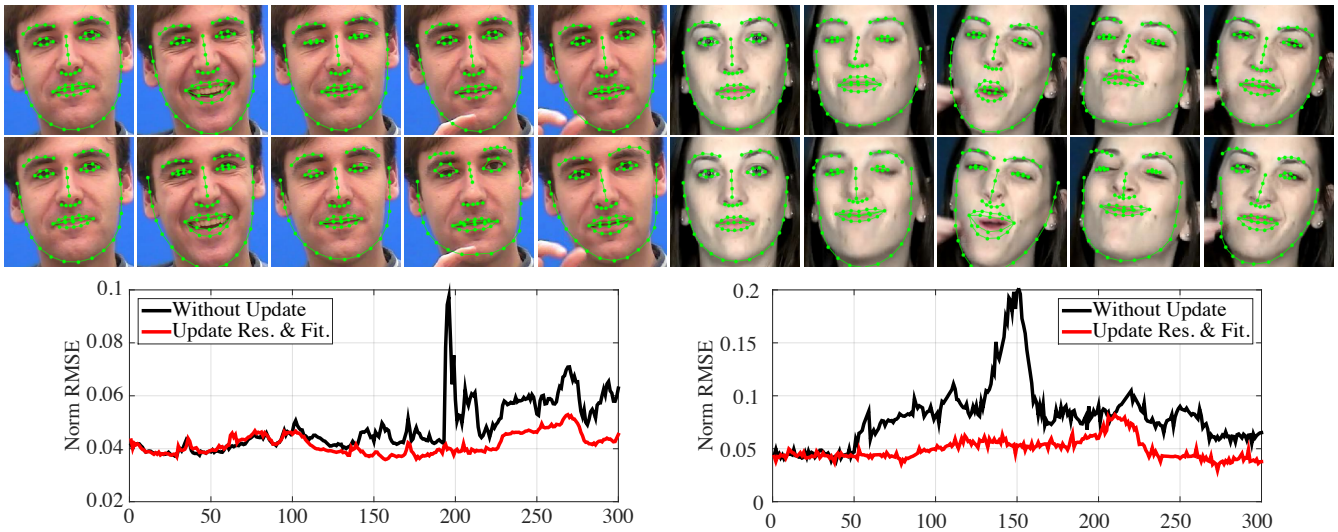

Figure 3: Average fitting errors with and without the model adaptation on FGNET and ASLV.

However, to achieve the best performance, it is necessary to jointly update both models in a unified framework. In this case, a more faithful personalized modeling can be expected by jointly update the representation subspace and adapt the fitting strategy.

Table 1: Cumulative error distributions on FGNET and ASLV with different settings.

Update	< 0.04	< 0.06	< 0.08	< 0.04	< 0.06	< 0.08
Rep.	78.2%	93.1%	95.6%	62.9%	78.0%	89.4%
Fit.	84.0%	96.4%	98.5%	58.2%	73.3%	91.2%
Rep. & Fit.	**91.8%**	**97.0%**	**99.4%**	**68.7%**	**84.6%**	**93.5%**

Validation of deep fitting evaluation. The online fitting evaluation is crucial in our incremental learning framework. Adaptation using erroneous fittings will drift the offline learned models and eventually lead to failure. To evaluate the performance the proposed deep fitting evaluation, we trained the network using image datasets and test the evaluation accuracy on videos. We sampled 5 perturbations for each image, where the ground-truth shape and perturbed shape were labeled as positive and negative respectively. Note that we used more negative rather than positive samples to train the network as misalignment detection is the major target here. Table 2 shows the average misalignment detection accuracy in different video datasets. Our approach achieves around 90% accuracy in general. It can robustly detect erroneous fittings in challenging conditions and pick out well-aligned faces for online adaptation, which can significantly alleviate the drifting issue.

Table 2: Misalignment detection accuracy on different video datasets.

	FGNET [10]	ASLV [17]	300-VW [30]	YtbVW [24]
Accuracy	94.4%	91.7%	85.3%	88.1%

4.3 Comparison with Previous Work

We compare our approach with four approaches that reported state-of-arts performance: **(1)** regularized landmark mean-shift (RLMS) [29], **(2)** discriminative response map fitting (DRMF) [2], **(3)** incremental face alignment (IFA) [1], and **(4)** explicit shape regression face alignment (ESR) [6]. For a fair comparison, we tested these methods in a tracking-by-detection manner.

Comparison of fitting accuracy. We compared our approach with the four methods on different video datasets. The average fitting errors are compared in Table 3. We have following observations. First, our approach has the lowest fitting errors and outperforms others with substantial margins, which demonstrates the superior performance of our approach in unconstrained videos. Second, compared with the performance on FGNET and ASLV, the advantage of our approach is more significant on 300-VW and YtbVW which present dynamic head movements, expression variations, illumination changes and partial occlusions. This result proves that the proposed person-specific alignment can better handle unconstrained data than other generic methods. Third, we also notice that ESR and IFA have better performance than RLMS and DRMF. A possible reason is that the explicit 2D shape used in ESR and IFA is more flexible than the constrained 3D shape used in RLMS and DRMF, which enables more accurate fittings when large pose and violent expression exist. However, they are still inferior to ours since they rely on offline models and lack the capability to capture the intensive online changes.

Table 3: Comparison of the averate fitting errors of different methods on four datasets.

	FGNET [10]	ASLV [17]	300-VW [30]	YtbVW [24]
RLMS [29]	4.11%	5.68%	7.79%	7.19%
DRMF [2]	3.75%	5.17%	6.25%	6.03%
IFA [1]	3.52%	4.54%	5.71%	5.48%
ESR [6]	3.49%	4.85%	5.85%	5.61%
OURS	**3.36%**	**4.41%**	**5.38%**	**5.23%**

Comparison of running time. We compared the average running time per frame of different methods and report the results in Table 4. For each method, the average speed was evaluated using the same 1000 frames. We tested the proposed method with either turning off or on the model adaptation. The results demonstrate that when the model adaptation is turned off, our approach is much more efficient than RLMS, and has comparable performance as DRMF and ESR. It slows down obviously when the incremental model adaptation is turned on. The reason is we apply the deep fitting evaluation at each frame, and perform the evaluation and model adaptation in a sequential order. The testing speed can be significantly improved with better implementation technique such as applying batch evaluation and model adaptation in parallel threads. We leave this as our future work.

Table 4: Comparison of the average running time per frame of different methods.

RLMS [29]	DRMF [2]	ESR[6]	OURS (off)	OURS (on)
116ms	55ms	89ms	76ms	218ms

5 Conclusion

In this paper, we propose a novel approach to track facial points in unconstrained videos. We investigate incremental learning to update the representation subspace and simultaneously adapt the cascade of regressors to achieve person-specific modeling. To address the drifting issue, we propose to leverage the deep neural network for robust fitting evaluation. Experiments on both image and video datasets have validated our approach in different aspects and demonstrated its superior performance compared with the state of the arts in terms of fitting accuracy and testing speed.

References

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

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013.

[3] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 35, pages 2930–2940, December 2013.

[4] M. Black and Y. Yacoob. 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)*, pages 374–381, 1995.

[5] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25 (9):1063–1074, Sep 2003.

[6] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[7] Grigoris G. Chrysos, Epameinondas Antonakos, Stefanos Zafeiriou, and Patrick Snape. Offline deformable face tracking in arbitrary videos. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 954–962, 2015.

[8] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995. ISSN 1077-3142.

[9] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, June 2001. ISSN 0162-8828.

[10] FGNet. Talking face video, 2004. URL http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html.

[11] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vision Computing (IVC)*, 28(5):807–813, May 2010. ISSN 0262-8856.

[12] Yimo Guo, Guoying Zhao, and Matti PietikÃd'inen. Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Transactions on Image Processing (TIP)*, 25(5):1977–1992, 2016.

[13] Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[14] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)*, pages 679–692, 2012.

[15] A Levey and Michael Lindenbaum. Sequential karhunen-loeve basis extraction and its application to images. *IEEE Transactions on Image Processing (TIP)*, 9(8):1371–1374, 2000.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[17] Carol Neidle, Jingjing Liu, Bo Liu, Xi Peng, Christian Vogler, and Dimitris Metaxas. Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in american sign language (asl). *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, 2014.

[18] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Output-associative {RVM} regression for dimensional and continuous emotion prediction. *Image and Vision Computing (IVC)*, 30(3):186 – 196, 2012. ISSN 0262-8856. Best of Automatic Face and Gesture Recognition 2011.

[19] A. Papaioannou and S. Zafeiriou. Principal component analysis with complex kernel: The widely linear model. *IEEE Transactions on Neural Networks and Learning Systems*, 2014.

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

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

[22] Xi Peng, Junzhou Huang, Qiong Hu, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. From circle to 3-sphere: Head pose estimation by instance parameterization. *Computer Vision and Image Understanding (CVIU)*, 136:92–102, 2015.

[23] Xi Peng, Junzhou Huang, Qiong Hu, Shaoting Zhang, and Dimitris N Metaxas. Three-dimensional head pose estimation in-the-wild. In *The IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6, 2015.

[24] Xi Peng, Shaoting Zhang, Yu Yang, and Dimitris N. Metaxas. Piefa: Personalized incremental and ensemble face alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[25] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2010.

[26] David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision (IJCV)*, 77(1-3):125–141, May 2008. ISSN 0920-5691.

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

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

[29] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision (IJCV)*, 91 (2):200–215, January 2011. ISSN 0920-5691.

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

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[32] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483, 2013.

[33] Jaewon Sung and Daijin Kim. Adaptive active appearance model with incremental learning. *Pattern Recognition Letters (PRL)*, 30(4):359 – 367, 2009.

[34] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[35] Ming Tang and Xi Peng. Robust tracking with discriminative ranking lists. *IEEE Transactions on Image Processing (TIP)*, 21(7):3273–3281, 2012.

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

[37] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern W. Schuller. A Deep Semi-NMF Model for Learning Hidden Representations. In *International Conference on Machine Learning (ICML)*, 2014.

[38] George Trigeorgis, Mihalis Nicolaou, Stefanos Zafeiriou, and Bjoern W. Schuller. Towards Deep Multimodal Alignment. In *NIPS Multimodal Machine Learning Workshop*, 2015.

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

[40] Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 149–, 2006. ISBN 0-7695-2646-2.

[41] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. Deeply-learned feature for age estimation. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 534–541. IEEE, 2015.

[42] Xiaolong Wang, Guodong Guo, Michele Merler, Noel CF Codella, MV Rohith, John R Smith, and Chandra Kambhamettu. Leveraging multiple cues for recognizing family photos. *Image and Vision Computing (IVC)*, 2016.

[43] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Xuehan-Xiong and Fernando De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[45] Heng Yang, Xuhui Jia, Chen Change Loy, and Peter Robinson. An empirical study of recent face alignment methods. *CoRR*, abs/1511.05049, 2015.

[46] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2014.

[47] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 94–108, 2014.

[48] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015.

[49] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation and landmark estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.