Track Facial Points in Unconstrained Videos

Xi Peng\(^1\)  
\texttt{xipeng.cs@rutgers.edu}  
Qiong Hu\(^1\)  
\texttt{qionghu.cs@rutgers.edu}  
Junzhou Huang\(^2\)  
\texttt{jzhuang@uta.edu}  
Dimitris N. Metaxas\(^1\)  
\texttt{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.

© 2016. The copyright of this document resides with its authors.  
It may be distributed unchanged freely in print or electronic forms.
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 \( \{ M_s^s, V_s^s \} \), where \( s \) denotes shape. An instance shape can be modeled as \( s(p) = M_s^f + V_s^f p \), where \( p \) is the shape representation.

The appearance representation is learned using local response maps [29]. Given a image \( I \) and the shape representation \( p \), the local response map around the \( l \)-th landmark is \( A_l(p, I) = 1/(1 + \exp(a_l \phi(s(p), I) + b_l)) \), where \( \{a_l, b_l\}_{l=1}^L \) are patch experts learned by 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 \( \{\Delta p_{ij}\} \) around the ground-truth \( p_i \) as illustrated in Figure 1. The perturbed response maps are arranged as a tensor \( T_l = \{ A_l(p_i^l + \Delta p_{ij}, I_l) \}_{i,j} \), where \( i \) and \( j \) count images and perturbations respectively. Similar to the shape representation, we apply PCA on \( T_l \) to obtain the mean and eigenvectors \( \{ M_s^a, V_s^a \} \), where \( a \) denotes appearance. The appearance representation of the \( l \)-th landmark can be calculated by fast projection \( x_l = (V_s^a)^{-1}(A_l(p, I) - M_s^a) \).

Now we can model the shape and appearance of an instance face using \( p \) and \( x(p, I) = [x_1^T, \cdots, x_L^T]^T \). 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 \( x(p, I) \) to the shape update \( \Delta p \). We refine the shape representation \( p \) from an initial guess \( p^0 \) to the ground truth step by step:

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

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

\[
\arg\min_{R^k, b^k} \sum_{i=1}^M \sum_{j=1}^N \| \Delta p_{ij} - x(p_{ij}^k, I_i)R^k - b^k \|^2. \tag{2}
\]

Let \( \hat{R}^k = \left[ R^k T b^k T \right]^T \) and \( \hat{x} = \left[ x(p^k, I_i) T \right]^T \), the regressor can be computed with a closed-form solution \( \hat{R}^k = [\hat{x}^T \hat{x} + \lambda I]^{-1} \hat{x}^T \Delta p^k \).
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 \( \{ M_s, V_s \} \) and \( \{ M_{al}, V_{al} \}_{l=1}^{L} \) 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 \( [T_A, T_B] = U' \Sigma' V_T' \).

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:

\[
[U E] \begin{bmatrix} \Sigma & U^T \hat{T}_B \\ 0 & E(\hat{T}_B - UU^T \hat{T}_B) \end{bmatrix} \begin{bmatrix} V^T & 0 \\ 0 & I \end{bmatrix},
\]

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

\[
T_C = \bar{U} \Sigma \bar{V}_T, T_C = \begin{bmatrix} \Sigma & U^T \hat{T}_B \\ 0 & E(\hat{T}_B - UU^T \hat{T}_B) \end{bmatrix}.
\]

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

\[
M_{AB} = \frac{m}{m+n} M_A + \frac{n}{m+n} M_B, \\
U' = [U \ E] \bar{U}, \ \Sigma' = \bar{\Sigma}.
\]

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 \( \hat{R}^k \) needs to be recomputed after \( \hat{R}^{k-1} \). To address this issue, we follow [1] to decouple the dependence in the cascade by directly sample \( p^k \) from a norm distribution \( p^k \sim \mathcal{N}(p^*, \Lambda^k) \), where \( \Lambda^k \) 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 \( \mathbf{x}_A \) and \( \mathbf{R}_A \) 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 \( \mathbf{x}_B \). \( \mathbf{R}_A \) can be adapted to \( \mathbf{R}_{AB} \) by:

\[
\mathbf{R}_{AB} = \mathbf{R}_A - P_{AB} \mathbf{R}_A + (P_A - P_{AB} P_A) (\mathbf{x}_B)^T \Delta \mathbf{p}_B,
\]

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

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 \( \mathbf{x}_B \) 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.

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. Finally, 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 angles [23]: left $[-90^\circ, -30^\circ]$, frontal $[-30^\circ, 30^\circ]$ and right $(30^\circ, 90^\circ)$. 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 $11 \times 11$ and $21 \times 21$ respectively.

We sampled 10 perturbations for each training image with the standard deviations of $\pm 0.1$ for scaling, $\pm 10^\circ$ for rotation, and $\pm 10$ 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.

<table>
<thead>
<tr>
<th>Update</th>
<th>&lt; 0.04</th>
<th>&lt; 0.06</th>
<th>&lt; 0.08</th>
<th>&lt; 0.04</th>
<th>&lt; 0.06</th>
<th>&lt; 0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep.</td>
<td>78.2%</td>
<td>93.1%</td>
<td>95.6%</td>
<td>62.9%</td>
<td>78.0%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Fit.</td>
<td>84.0%</td>
<td>96.4%</td>
<td>98.5%</td>
<td>58.2%</td>
<td>73.3%</td>
<td>91.2%</td>
</tr>
<tr>
<td>Rep. &amp; Fit.</td>
<td>91.8%</td>
<td>97.0%</td>
<td>99.4%</td>
<td>68.7%</td>
<td>84.6%</td>
<td>93.5%</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>94.4%</td>
<td>91.7%</td>
<td>85.3%</td>
<td>88.1%</td>
</tr>
</tbody>
</table>

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 average fitting errors of different methods on four datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RLMS [29]</td>
<td>4.11%</td>
<td>5.68%</td>
<td>7.79%</td>
<td>7.19%</td>
</tr>
<tr>
<td>DRMF [2]</td>
<td>3.75%</td>
<td>5.17%</td>
<td>6.25%</td>
<td>6.03%</td>
</tr>
<tr>
<td>IFA [1]</td>
<td>3.52%</td>
<td>4.54%</td>
<td>5.71%</td>
<td>5.48%</td>
</tr>
<tr>
<td>ESR [6]</td>
<td>3.49%</td>
<td>4.85%</td>
<td>5.85%</td>
<td>5.61%</td>
</tr>
<tr>
<td>OURS</td>
<td><strong>3.36%</strong></td>
<td><strong>4.41%</strong></td>
<td><strong>5.38%</strong></td>
<td><strong>5.23%</strong></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>116ms</td>
<td>55ms</td>
<td>89ms</td>
<td>76ms</td>
<td>218ms</td>
</tr>
</tbody>
</table>

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


