Towards Highly Accurate and Stable Face Alignment for High-Resolution Videos

Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyou Huang, Yu Chen

†Youtu Lab, Tencent, ‡Michigan State University
§Fudan University, Nanjing University of Science and Technology
††{yingtai, easonliang, jerolinli, jasoncjwang, garyhuang} @tencent.com
‡lhx @cse.msu.edu, §15307130193 @fudan.edu.cn, ‡chenyu1523 @gmail.com
https://github.com/tyiwo/FHR_alignment

Abstract

In recent years, heatmap regression based models have shown their effectiveness in face alignment and pose estimation. However, Conventional Heatmap Regression (CHR) is not accurate nor stable when dealing with high-resolution facial videos, since it finds the maximum activated location in heatmaps which are generated from rounding coordinates, and thus leads to quantization errors when scaling back to the original high-resolution space. In this paper, we propose a Fractional Heatmap Regression (FHR) for high-resolution video-based face alignment. The proposed FHR can accurately estimate the fractional part according to the 2D Gaussian function by sampling three points in heatmaps. To further stabilize the landmarks among continuous video frames while maintaining the precise at the same time, we propose a novel stabilization loss that contains two terms to address time delay and non-smooth issues, respectively. Experiments on 300W, 300-VW and Talking Face datasets clearly demonstrate that the proposed method is more accurate and stable than the state-of-the-art models.

Introduction

Face alignment aims to estimate a set of facial landmarks given a face image or video sequence. It is a classic computer vision problem that has attributed to many advanced machine learning algorithms (Fan et al. 2018; Bulat and Tzimiropoulos 2017; Trigeorgis et al. 2016; Peng et al. 2015; 2016; Kowalski, Naruniec, and Trzcinski 2017; Chen et al. 2017; Liu et al. 2017; Hu et al. 2018). Nowadays, with the rapid development of consumer hardwares (e.g., mobile phones, digital cameras), High-Resolution (HR) video sequences can be easily collected. Estimating facial landmarks on such high-resolution facial data has tremendous applications, e.g., face makeup (Chen, Shen, and Jia 2017), editing with special effects (Korshunova et al. 2017) in live broadcast videos. However, most existing face alignment methods work on faces with medium image resolutions (Chen et al. 2017; Bulat and Tzimiropoulos 2017; Peng et al. 2016; Liu et al. 2017). Therefore, developing face alignment algorithms for high-resolution videos is at the core of this paper.

To this end, we propose an accurate and stable algorithm for high-resolution video-based face alignment, named Fractional Heatmap Regression (FHR). It is well known that heatmap regression have shown its effectiveness in landmark estimation tasks (Chen et al. 2017; Newell, Yang, and Deng 2016; Chen* et al. 2018). However, Conventional Heatmap Regression (CHR) is not accurate nor stable when dealing with high-resolution facial images, since it finds the maximum activated location in heatmaps which are generated from rounding coordinates, and thus leads to quantization errors as the heatmap resolution is much lower than the input image resolution (e.g., 128 vs. 930 shown in Fig. 1) due to the scaling operation. To address this problem, we propose a novel transformation between heatmaps and coordinates, which not only preserves the fractional part when generating heatmaps from the coordinates, but also accurately estimate the fractional part according to the 2D Gaussian function by sampling three points in heatmaps.

Using our proposed FHR, we can estimate more accurate landmarks compared to the conventional heatmap re-
gression model, and achieve state-of-the-art performance on popular video benchmarks: Talking Face (FGNET 2014) and 300-VW datasets (Shen et al. 2017) compared to recent video-based face alignment models (Liu et al. 2017; Peng et al. 2016). However, real-world applications such as face makeup in videos often demand extremely high stability, since the makeup jumps if the estimated landmarks oscillate between consecutive frames, which negatively impacts the user’s experience. To make the sequential estimations as stable as possible, we further develop a novel stabilization algorithm on the landmarks estimated by FHR, which contains two terms, a regularization term \(L_{\text{reg}}\) and a temporal coherence term \(L_{\text{tm}}\), to address two common difficulties: time delay and non-smooth problems, respectively. Specifically, \(L_{\text{reg}}\) combines traditional Euclidean loss and a novel loss account for time delay; \(L_{\text{tm}}\) generalizes the temporal coherence loss in (Cao, Hou, and Zhou 2014) to better handle nonlinear movement of facial landmark.

In summary, the main contributions of this paper are:

- A novel Fractional Heatmap Regression method for high-resolution video-based face alignment that leverages 2D Gaussian prior to preserve the fractional part of points.
- A novel stabilization algorithm to address time delay and non-smooth problems among continuous video frames.
- State-of-the-art performance, both in accuracy and stability, on the benchmarks of 300W (Sagonas et al. 2013), 300-VW (Shen et al. 2017) and Talking Face (FGNET 2014).

**Related Work**

**Heatmap Regression** Heatmap regression is one of the most widely used approaches for landmark localization tasks, which estimates a set of heatmaps rather than coordinates. Stacked Hourglass Networks (SHN) are popular architectures in heatmap regression, which have symmetric topology that capture and consolidate information across all scales of the image. Newell et al. (Newell, Yang, and Deng 2016) proposed SHN for 2D human pose estimation, which achieved remarkable results even for very challenging datasets (Andriluka et al. 2014). With the hourglass structure, Chu et al. (Chu et al. 2017) introduced multi-context attention mechanism into Convolutional Neural Networks (CNN). Apart from applications in human pose estimation, there are also several heatmap regression based models for face alignment. Chen et al. (Chen et al. 2017) proposed a structure-aware fully convolutional network to implicitly model the priors during training. Bulat et al. (Bulat and Tzimiropoulos 2017) built a powerful CNN for face alignment based on the hourglass network and a hierarchical, parallel and multi-scale block.

However, all existing models drop the fractional part of coordinates during the transformation between heatmaps and points, which brings quantization errors to high-resolution facial images. On the contrary, our proposed FHR can accurately estimate the fractional part according to the 2D Gaussian function by sampling three points in heatmaps, and thus achieves more accurate alignment.

**Video-based Face Alignment** Video-based face alignment estimates facial landmarks in video sequences (Liu 2010). Early methods (Black and Yacoob 1995; Shen et al. 2017) used incremental learning to predict landmarks on still frames in a tracking-by-detection manner. To address the issue that generic methods are sensitive to initializations, Peng et al. (Peng et al. 2015) exploited incremental learning for personalized ensemble alignment, which samples multiple initial shapes to achieve image congealing within one frame. To explicitly model the temporal dependency (Oh et al. 2015) of landmarks across frames, the authors (Peng et al. 2016) further incorporated a sequence of spatial and temporal recurrences for sequential face alignment in videos. Recently, Liu et al. (Liu et al. 2017) proposed a Two-Stream Transformer Networks (TSTN) approach, which captures the complementary information of both the spatial appearance on still frames and the temporal consistency across frames. Different from (Peng et al. 2016; Liu et al. 2017) that require temporal landmark labels across frames, our proposed method achieves state-of-the-art accuracy only by making full use of the spatial appearance on still frames, which is able to remedy the problem that labeled sequential data are very limited.

Apart from the accuracy of landmarks, stabilization is also a key metric to video-based alignment. Typically, two terms (Cao, Hou, and Zhou 2014) are adopted for stabilization, where a regularization term drives the optimized results to be more expressive, and a temporal coherence term drives the results to be more stable and smooth. However, the existing stabilization algorithm is sensitive to time delay and nonlinear movements. Our proposed algorithm takes these into account and thus are overall more robust.

**The Proposed Approach**

In this section, we introduce the details of the proposed approach based on heatmap regression. A key point of heatmap regression is the transformation between the heatmaps and coordinates. Specifically, before model training, a pre-process step is conducted to convert the coordinates to the heatmaps, which are used as the ground truth labels. After estimating the heatmaps, a post-process step is conducted to obtain the coordinates from the estimated heatmaps. In this work, we propose a novel transformation between the heatmaps and coordinates that is different from the conventional heatmap regression, which is demonstrated to be simple yet effective.

**Fractional Heatmap Regression**

As shown in Fig. 1, conventional heatmap regression mainly generates the heatmaps from integral coordinates, despite that the ground truth coordinates are usually with fractions. As a result, it causes quantization errors when scaling back to the original image resolution, since the heatmaps are usually of much lower resolution than the input image. To address this problem, our proposed fractional heatmap regression generates ground truth heatmaps based on the intact ground truth coordinates (see Fig. 1) as follows:

\[
\bar{H}_{\text{(m)}}(c) = \exp \left( -\frac{1}{2\sigma^2} ((c_x - c_{(m)\text{exp}})^2 + (c_y - c_{(m)\text{yp}})^2) \right),
\]  

(1)
we further sample another two points, e.g., \( \hat{z}_1 \) and \( \hat{z}_2 \). However, since \( \hat{z}_1 \) lies on the line \( \hat{p}^{(t-1)} \parallel \hat{p}^{(t-1)} \), it indicates a loss caused by time delay, which is more likely to happen in the stabilization process. Thus our model prefers to assign it a larger loss (equal to \( \hat{z}_1^{(t)} \)). (b) Histograms of the loss in (Cao, Hou, and Zhou 2014) (left) and our \( \mathcal{L}_{tm} \) (right). \( x^{(t-1)} \) and \( x^{(t+1)} \) are stabilization outputs of frame \( t-1 \) and \( t+1 \), respectively. \( x_1^{(t)} \) has the same loss as \( x_2^{(t)} \) in (Cao, Hou, and Zhou 2014). However, since \( x_1^{(t)} \) lies on the line \( \hat{x}^{(t-1)} \parallel \hat{x}^{(t+1)} \), we argue that movement trajectory \( \hat{x}^{(t-1)} \parallel \hat{x}^{(t+1)} \) is more smooth than trajectory \( \hat{x}^{(t-1)} \parallel \hat{x}^{(t+1)} \). Thus our model assigns it a smaller loss (equal to \( \hat{z}_3^{(t)} \)). Note that the short axis of the ellipse in (a) is \( \hat{x}^{(t-1)} \parallel \hat{x}^{(t+1)} \) while the long axis of the ellipse in (b) is \( \hat{x}^{(t-1)} \parallel \hat{x}^{(t+1)} \), thus the two terms are not in contradiction.

where \( c = (c_x, c_y) \in \Omega \) represents the coordinate, \( \Omega \) is the domain of the heatmap \( \tilde{H} \), \( \sigma \) denotes the standard deviation and \( (c_{(m)x}, c_{(m)y}) \) is the center of the 2D Gaussian in the \( m \)th heatmap.

Denoting \( \Gamma \) an input image and \( \mathcal{F} \) the deep alignment model, we can estimate the recovered heatmaps by

\[
\tilde{H} = \mathcal{F}(\Gamma),
\]

where \( \tilde{H} = [\hat{H}_1, \ldots, \hat{H}_{(m)}, \ldots, \hat{H}_{(M)}] \), and \( M \) is the number of landmarks. Given Eq. (1) and \( \hat{H}_{(m)} \), estimating the fractional coordinate \( z \) amounts to solving a binary quadratic equation, which has a closed-form solution as long as we can sample any three non-zero points from the heatmap. Specifically, we first obtain \( \hat{z} \) as the location \( c \) with the maximum likelihood, as in conventional heatmap regression:

\[
\hat{z}_m = \arg\max_c \hat{H}_{(m)}(c), \quad m = 1, \ldots, M.
\]

Conventional heatmap regression directly takes \( \hat{z}_m \) as the output, which loses the fractional part. In our method, we further sample another two points, e.g., \( \hat{z}_m = (\hat{z}_{(m)x} + 1, \hat{z}_{(m)y}), \hat{z}_m = (\hat{z}_{(m)x}, \hat{z}_{(m)y} + 1) \) near \( \hat{z}_m = (\hat{z}_{(m)x}, \hat{z}_{(m)y}) \). Let \( h_m = \tilde{H}_{(m)}(\hat{z}_m), h_{(m)} = \hat{H}_{(m)}(\hat{z}_m) \) and \( h_{(m)} = \hat{H}_{(m)}(\hat{z}_m) \), then we estimate the fractional coordinate \( z_m = (z_{(m)x}, z_{(m)y}) \) as follows:

\[
\begin{align*}
\sigma^2 &= \frac{1}{2}(\hat{z}_{(m)x}^2 - \hat{z}_{(m)x}^2)^2 + \frac{1}{2}(\hat{z}_{(m)y}^2 - \hat{z}_{(m)y}^2)^2, \\
\sigma^2 &= \frac{1}{2}(\hat{z}_{(m)x}^2 - \hat{z}_{(m)x}^2)^2 + \frac{1}{2}(\hat{z}_{(m)y}^2 - \hat{z}_{(m)y}^2)^2.
\end{align*}
\]

It should be noted that our fractional heatmap regression is applicable to any heatmap based methods. In this paper, we focus on face alignment, and adopt the stacked hourglass network (Newell, Yang, and Deng 2016) as the alignment model \( \mathcal{F} \) that minimizes the loss of \( \|\tilde{H} - \hat{H}\|^2 \) across the entire training set, where \( \tilde{H} = [\hat{H}_{(1)}, \ldots, \hat{H}_{(m)}, \ldots, \hat{H}_{(M)}] \).

**Stabilization Algorithm**

We now introduce our stabilization algorithm for video-based alignment, which takes the alignment results of \( \mathcal{F} \) in all the past frames as input, and outputs a more accurate and stable result for the current frame.

**Stabilization Model**

We denote \( z^{(t)} \) as the output of \( \mathcal{F} \) at frame \( t \) and the stabilization model as \( \mathcal{M}_\Theta \), which has parameters \( \Theta \) to be optimized. \( \mathcal{M}_\Theta \) takes \( z^{(1)}, \ldots, z^{(t)} \) as input, and outputs the stabilized landmarks of frame \( t \), which is denoted as \( x^{(t)} \). Therefore we have

\[
x^{(t)} = \mathcal{M}_\Theta \left( z^{(1)}, \ldots, z^{(t)} \right).
\]

Assume there are \( V \) videos in the training set, and the \( i \)th video has \( T_i \) frames. For frame \( t \) in the \( i \)th video, we denote its ground truth landmarks as \( p_i^{(t)} \), the output of \( \mathcal{F} \) as \( z_i^{(t)} \), and the stabilized output as \( x_i^{(t)} \) (\( p_i^{(t)}, z_i^{(t)}, x_i^{(t)} \in \mathbb{R}^{2M \times 1} \)). Here, we have \( x_i^{(t)} = \mathcal{M}_\Theta \left( z_i^{(1)}, \ldots, z_i^{(t)} \right) \).

Next, we present the specific form of \( \mathcal{M}_\Theta \) as well as its parameters \( \Theta \). Our model follows a Bayesian framework. Specifically, we model the prior distribution of \( x_i^{(t)} \) given \( z^{(1)}, \ldots, z^{(t-1)} \) as a \( K \)-component Gaussian mixture:

\[
P_r \left( x^{(t)} | z^{(1)}, \ldots, z^{(t-1)} \right) = \sum_{k=1}^{K} \alpha_k N \left( x^{(t)} ; \mu_k^{(t)} , \Sigma_k^{(t)} \right),
\]

\( ^1 \)In case \( \hat{z}_m \) is located at the edge of the heatmap, we would sample the points in the opposite directions.
where \( Pr(\cdot) \) indicates the density function, \( \mathcal{N}(\cdot; \mu, \Sigma) \) indicates a normal distribution with mean \( \mu \) and covariance \( \Sigma \). We then model the likelihood of \( x^{(t)} \) given \( x^{(t)} \) as Gaussian
\[
Pr\left(z^{(t)}|x^{(t)}\right) = \mathcal{N}\left(z^{(t)}; x^{(t)}, \Sigma_{\text{noise}}\right), \tag{7}
\]
and use Bayesian rule to get the most probable value of \( x^{(t)} \):
\[
x^{(t)} = \underset{x}{\text{argmax}} \ Pr\left(x|z^{(1)}, \ldots, z^{(t-1)}\right) Pr\left(z^{(t)}|x\right). \tag{8}
\]
Combining (6), (7) and (8), we can obtain the closed-form solution of (5). In practice, we fix \( K = 2 \) since it already achieves satisfactory results and larger \( K \) may cause both efficiency and overfitting issues. Moreover, to reflect the fact that \( x^{(t)} \) has a decreasing correlation with \( x^{(t-r)} \) when \( r \) increases, we assume that
\[
\mu_k^{(t)} = \frac{\sum_{r=1}^{t-1} \gamma^r (x^{(t-r)} - \mu_k)}{\sum_{r=1}^{t-1} \gamma^r}, \quad \Sigma_k^{(t)} = \beta_k \Sigma_k + (1 - \beta_k) \frac{\sum_{r=1}^{t-1} \gamma^r (x^{(t-r)} - \mu_k) (x^{(t-r)} - \mu_k)^T}{\sum_{r=1}^{t-1} \gamma^r} \tag{9}
\]
where \( \gamma = \{\beta_k\}_{k=1}^K \in [0, 1] \) along with \( 2M \times 2M \) positive semi-definite matrices \( \{\Sigma_k\}_{k=1}^K \) and \( \Sigma_{\text{noise}} \) are unknown model parameters. In practice Eq. (9) can be calculated recursively, whose computational complexity remains constant when \( t \) increases.

To further reduce the number of parameters, we calculate the covariance matrix \( S \) of all \( p_{i}^{(t)} - p_{i}^{(t-1)} \) in the training set, denote the matrix of eigenvectors of \( S \) as \( V \), and assume
\[
\Sigma_{\text{noise}} = V^T \Gamma_{\text{noise}} V, \quad \Sigma_k = V^T \Gamma_k V, \quad k = 1, \ldots, K, \tag{10}
\]
where \( \Gamma_{\text{noise}} \) and \( \{\Gamma_k\}_{k=1}^K \) are diagonal matrices. In summary, we have \( \Theta = \{\gamma, \{\alpha_k, \beta_k\}_{k=1}^K, \Gamma_{\text{noise}}, \{\Gamma_k\}_{k=1}^K\} \), which are optimized with the loss function in the next subsection.

**Loss Function Design** We now introduce and optimize a novel loss function so as to estimate the above stabilization model parameter \( \Theta \). Throughout this section, we denote all \( x^{(t)}_i \) as \( x^{(t)}_i(\Theta) \) to emphasize that the stabilized landmarks are functions of model parameter \( \Theta \). Our loss function is formulated as:
\[
\mathcal{L}(\Theta) = \mathcal{L}_{\text{reg}}(\Theta) + \lambda_1 \mathcal{L}_{\text{tm}}(\Theta). \tag{11}
\]
This loss has two terms. The first term, \( \mathcal{L}_{\text{reg}}(\Theta) \) is called the regularization loss, which regularizes the stabilized output to be close to the ground truth, and is defined as follows:
\[
\mathcal{L}_{\text{reg}}(\Theta) = \sum_{i=1}^{V} \sum_{t=1}^{T_i} \|x^{(t)}_i(\Theta) - p^{(t)}_i\|_2^2 + \lambda_2 \sum_{i=1}^{V} \sum_{t=1}^{T_i} \left( (p^{(t)}_i - p^{(t-1)}_i) + (x^{(t)}_i(\Theta) - p^{(t)}_i) \right)^2 / \sum_{i=1}^{V} (T_i - 1), \tag{12}
\]
where \( x^+ \) indicates the Moore-Penrose general inverse of vector/matrix \( x \). We can see that the first term of (12) is the average Euclidean distance of the ground truth and the model output. The second term aims to fit every \( x^{(t)}_i(\Theta) \) in terms of \( \alpha p^{(t-1)}_i + (1 - \alpha) p^{(t)}_i \), where \( \alpha \) is the coefficient, and estimate the expectation of \( \alpha \). If this expectation is large, it means that (1) \( x^{(t)}_i(\Theta) \) is more similar to \( p^{(t-1)}_i \) than \( p^{(t)}_i \), and 2) the model output has a significant time delay, which is undesirable. Since our stabilization model uses the alignment results of the past frames, how to avoid time delay is a critical task. Therefore we emphasize the time delay loss as an individual term in \( \mathcal{L}_{\text{reg}} \) (see Fig. 2(a)).

\[
\mathcal{L}_{\text{tm}}(\Theta) \text{ is called the smooth loss which favors the stabilized output to be smooth, and is defined as follows:}
\]
\[
\mathcal{L}_{\text{tm}}(\Theta) = \min_q \sum_{i=1}^{V} \sum_{t=1}^{T_i-1} \|x^{(t)}_i(\Theta) - q^{(t)}_i x^{(t-1)}_i(\Theta) - \alpha p^{(t)}_i + (1 - \alpha) p^{(t-1)}_i \|^2_2 + \lambda_2 \|q^{(t)}_i - q^{(t+1)}_i\|^2_2 / \sum_{i=1}^{V} (T_i - 2), \tag{13}
\]
where \( q^{(t)}_i \in \mathbb{R} \) and \( q = (q^{(2)}_i, \ldots, q^{(T-1)}_i) \). \( \mathcal{L}_{\text{tm}}(\Theta) \) can be seen as a trade-off between two stability losses, controlled by \( \lambda_2 \). When \( \lambda_2 = 0 \), it is equivalent to
\[
\min_q \sum_{i=1}^{V} \sum_{t=1}^{T_i-1} \|x^{(t)}_i(\Theta) - q^{(t)}_i x^{(t-1)}_i(\Theta) - \alpha p^{(t)}_i + (1 - \alpha) p^{(t-1)}_i \|^2_2 / \sum_{i=1}^{V} (T_i - 2) \tag{14}
\]
which is the average distance from \( x^{(t)}_i \) to \( x^{(t-1)}_i x^{(t+1)}_i \).

On the other hand, when \( \lambda_2 \to \infty \), \( \mathcal{L}_{\text{tm}}(\Theta) \) equals to
\[
\sum_{i=1}^{V} \sum_{t=1}^{T_i-1} \|x^{(t)}_i(\Theta) - \frac{1}{2} x^{(t-1)}_i(\Theta) + x^{(t+1)}_i(\Theta) \|^2_2 / \sum_{i=1}^{V} (T_i - 2) \tag{15}
\]
which is the average distance from \( x^{(t)}_i \) to the midpoint of the line \( x^{(t-1)}_i x^{(t+1)}_i \). The trade-off smooth loss \( \mathcal{L}_{\text{tm}}(\Theta) \) will cause the loss contour of \( x^{(t)}_i \) to have an ellipse with long axis \( x^{(t-1)}_i x^{(t+1)}_i \) (see Fig. 2(b)), which we argue is a more reasonable indicator of the smoothness of the movement trajectory \( x^{(t-1)}_i x^{(t+1)}_i \) (see Fig. 2(b)). With the stabilization model and the loss function, we can estimate the model parameters using the standard optimization method (Lagarias et al. 1998). The proposed stabilization algorithm is trained by various videos, which learns the statistics of different kinds of movements and thus is more robust than the traditional stabilization model (Cao, Hou, and Zhou 2014). The optimization process converges within 12 hours on a conventional laptop.

8896
Table 1: Comparisons of NRMSE, AUC and failure rate (at 8.00% NRMSE) on 300W test set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NRMSE</td>
<td>5.83</td>
<td>5.78</td>
<td>5.74</td>
<td>5.05</td>
<td>4.30</td>
<td>3.96</td>
<td>4.07</td>
<td>3.80</td>
</tr>
<tr>
<td>AUC</td>
<td>36.3</td>
<td>34.8</td>
<td>36.6</td>
<td>45.3</td>
<td>47.0</td>
<td>53.6</td>
<td>50.9</td>
<td>55.9</td>
</tr>
<tr>
<td>Failure</td>
<td>13.0</td>
<td>14.0</td>
<td>12.3</td>
<td>6.80</td>
<td>2.67</td>
<td>2.50</td>
<td>2.33</td>
<td>1.33</td>
</tr>
</tbody>
</table>

Figure 3: (a) The effect of stabilization loss for time delay; (b) The magnitude of the stability metric for FHR (green) and FHR+STA (red). (c) The orientation of the stability metric for FHR (green) and FHR+STA (red).

Table 2: RMSE comparisons on 5 face scales on 300W.

<table>
<thead>
<tr>
<th>Face scale</th>
<th>Ave. inter-ocular dis.</th>
<th>CHR/FHR (RMSE ↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very small</td>
<td>61.45</td>
<td>2.66/2.42 (0.24 ↓)</td>
</tr>
<tr>
<td>Small</td>
<td>91.50</td>
<td>3.71/3.35 (0.36 ↓)</td>
</tr>
<tr>
<td>Medium</td>
<td>127.68</td>
<td>5.06/4.55 (0.51 ↓)</td>
</tr>
<tr>
<td>Large</td>
<td>179.55</td>
<td>7.59/7.11 (0.48 ↓)</td>
</tr>
<tr>
<td>Very large</td>
<td>296.28</td>
<td>11.67/10.67 (1.00 ↓)</td>
</tr>
</tbody>
</table>

Experiments

Experimental Setup

Datasets We conduct extensive experiments on both image and video-based alignment datasets, including 300W (Sagonas et al. 2013), 300-VW (Shen et al. 2017) and Talking Face (TF) (FGNET 2014). To test on 300 images of 300W private set, we follow (Chen et al. 2017) to use 3,148 training images from LFPW, HELEN and AFW datasets. To test on 300-VW, we follow (Shen et al. 2017) to use 50 videos for training and the rest 64 videos for testing. Specifically, the 64 videos are divided to three categories: well-lit (Scenario1), mild unconstrained (Scenario2) and challenging (Scenario3) according to the difficulties. To test on 5,000 frames of TF dataset, we follow (Liu et al. 2017) to use the model trained by the training set of 300-VW dataset.

Training Setting Training faces are cropped using the detection bounding boxes, and scaled to 256 × 256 pixels. Following (Chen et al. 2017), we augment the data (e.g., scaling, rotation) for more robustness to different face boxes. We use the stacked hourglass network (Newell, Yang, and Deng 2016; Chen et al. 2017) as the alignment model. The network starts with a 7 × 7 convolutional layer with stride 2 to reduce the resolution to 128 × 128, followed by stacking 4 hourglass modules. For evaluation, we adopt the standard Normalized Root Mean Squared Error (NRMSE), Area-under-the-Curve (AUC) and the failure rate (at 8.00% NRMSE) to measure the accuracy, and use the consistency between the movement of landmarks and ground truth as the metric to measure the stability. We train the network with the Torch7 toolbox (Collobert, Kavukcuoglu, and Farabet 2011), using the RMSprop algorithm with an initial learning rate of 2.5 × 10⁻⁴, a minibatch size of 6 and σ = 3. Training our FHR on 300W takes ~7 hours on a P100 GPU.

During the stabilization training, we set λ₁ = λ₃ = 1 and λ₂ = 10 to make all terms in the stabilization loss (11) on the same order of magnitude. We estimate the average variance ρ of xₖ⁽ᵗ⁾ − pₖ⁽ᵗ⁾ across all training videos and all landmarks, and empirically set the initial value of Γnoise as ρI. Also, we initialize Γ₁ as a zero matrix O₂M ×₂M, Γ₂ as 10ρI, and γ = β₁ = β₂ = 0.5.

Ablation Study

Fractional vs. Conventional Heatmap Regression We first compare our fractional heatmap regression with the conventional version (Newell, Yang, and Deng 2016) and other state-of-the-art models (Xiong and De la Torre 2013; Zhang et al. 2014; Zhu et al. 2015; Trigeorgis et al. 2016; Kowalski, Naruniec, and Trzcinski 2017; Chen et al. 2017) on 300W test set. The training set includes 3,148 images, and the test set contains 600 images. Note that the stacked hourglass networks for our fractional method and the conventional one are the same. The only difference is the transformation between the heatmaps and the coordinates, where our method preserves the fractional part. As shown in Tab. 1, our method significantly outperforms the conventional version with an improvement of 0.27% on NRMSE, and is also better than the state-of-the-art model (Chen et al. 2017).

What’s more, the standard NRMSE cannot fully reflect the advantage of our FHR compared to CHR, since it eliminates the scaling effects by dividing the inter-ocular distance. To further demonstrate the effects of FHR, we calculate RMSE without normalization. Specifically, we collect the inter-ocular distances of all 600 images and evenly divide
Table 3: NRMSE/stability comparisons on 300-VW test set using 68 landmarks.

<table>
<thead>
<tr>
<th>Methods</th>
<th>FHR</th>
<th>FHR+STA</th>
<th>FHR+STA(λ₂ = 0)</th>
<th>FHR+STA(λ₃ = ∞)</th>
<th>FHR+STA(λ₂ = 0, λ₃ = ∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario1</td>
<td>5.07/2.79</td>
<td>4.42/1.67</td>
<td>5.55/1.64</td>
<td>4.40/1.78</td>
<td>4.49/1.68</td>
</tr>
<tr>
<td>Scenario2</td>
<td>4.34/1.85</td>
<td>4.18/1.15</td>
<td>4.74/1.16</td>
<td>4.16/1.19</td>
<td>4.33/1.17</td>
</tr>
<tr>
<td>Scenario3</td>
<td>7.36/4.48</td>
<td>5.98/2.74</td>
<td>7.58/2.57</td>
<td>5.96/2.86</td>
<td>6.74/2.76</td>
</tr>
</tbody>
</table>

Figure 4: CED curves on 300W.

Table 4: NRMSE/stability comparisons with REDnet on 300-VW test set using 7 landmarks.

<table>
<thead>
<tr>
<th>Methods</th>
<th>REDnet (2016)</th>
<th>FHR</th>
<th>FHR+STA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario1</td>
<td>8.03/10.3</td>
<td>4.44/5.49</td>
<td>3.93/3.04</td>
</tr>
<tr>
<td>Scenario2</td>
<td>10.1/9.64</td>
<td>3.96/3.55</td>
<td>3.82/3.44</td>
</tr>
<tr>
<td>Scenario3</td>
<td>16.5/15.9</td>
<td>5.45/4.72</td>
<td>4.91/4.45</td>
</tr>
</tbody>
</table>

the distances to five groups w.r.t. face scales. Tab. 2 shows that with larger scales, the gap between FHR and CHR is bigger. Especially in the largest scale, FHR achieves 1 pixel promotion for each landmark on average.

Stabilization Loss for Time Delay Here we demonstrate the effectiveness of our proposed time delay term. As in Fig. 3(a), compared to the fractional heatmap regression’s output, the stabilized output is not only more smooth, but also closer to ground truth landmarks. Besides, when the time delay term is removed (λ₂ = 0), the stabilized outputs behave lagged behind the ground truth while this phenomenon is largely suppressed by using our proposed loss (Eq. 12).

Stabilization Loss for Smooth Next, we evaluate the impact of each term in our loss function (11) on NRMSE and stability. We use the consistency between the cross-time movement of the landmarks and the ground truth as an indicator of stability. Specifically, we calculate $\Delta x_i^{(t)} = x_i^{(t)} - x_i^{(t-1)}$ and $\Delta p_i^{(t)} = p_i^{(t)} - p_i^{(t-1)}$ for every video $i$ in the test set, and calculate the average NRMSE between $\Delta x_i^{(t)}$ and $\Delta p_i^{(t)}$. Assuming that the ground truth is stable, a lower value indicates higher stability. The comparison result is shown in Tab. 3. It can be seen that dropping the time delay term (λ₂ = 0) causes a higher NRMSE, and changing the smooth loss to the one in (Cao, Hou, and Zhou 2014) (λ₃ = ∞) causes a higher stability loss. Our proposed method achieves a good balance between accuracy and stability.

Comparisons with State of the Arts

We now compare our methods FHR and FHR+STA (i.e., the stabilized version) with state-of-the-art models (Peng et al. 2016; Liu et al. 2017) on two video datasets: 300-VW and Talking Face. The comparison adopts two popular settings (i.e., 7 and 68 landmarks) used in prior works.

Comparison with 7 Landmarks First, we evaluate our method on the 300-VW (Shen et al. 2017) dataset and compare with REDnet (Peng et al. 2016), using the code released by the authors. Tab. 4 shows the results on the three test sets. Our proposed method achieves much better performance than REDnet in all cases. Especially in the hardest Scenario3, our method achieves a large improvement of ~11%, which shows the robustness of our method to large facial variations.

Then, we evaluate our method on the Talking Face (FGNET 2014) dataset compared with state-of-the-art models, such as CCFAN (Zhang et al. 2014), CFSS (Zhu et al. 2015), IFA (Asthana et al. 2014), REDnet (Peng et al. 2016) and TSTN (Liu et al. 2017). Although the annotations of the TF dataset has the same landmark number as 300-VW dataset, the definitions of landmarks are different. Therefore, following the setting in (Liu et al. 2017; Peng et al. 2016), we use 7 landmarks for fair comparisons. The results are shown in Tab. 5, in which the performance of (Zhang et al. 2014; Zhu et al. 2015; Asthana et al. 2014) are directly cited from (Peng et al. 2016; Liu et al. 2017). Since the images in TF set are collected in controlled environment and with small facial variations, all of the methods achieve relatively small errors, and our proposed method is still the best.

Comparison with 68 Landmarks Next, we evaluate our method on the 300-VW (Shen et al. 2017) dataset under the setting with 68 landmarks. The comparison methods include TSCN (Simonyan and Zisserman 2014), CFSS (Zhu et al. 2015), TCDCN (Zhang et al. 2017) and TSTN (Liu et al. 2017). We cite the results of (Simonyan and Zisserman 2014; Zhu et al. 2015; Zhang et al. 2017) from (Liu et al. 2017), and list the performance in Tab. 6. As we can see, our proposed FHR achieves the best NRMSEs in all scenarios, and our stabilized version FHR+STA further improves the performance, especially in Scenario3.

We then illustrate the Cumulative Errors Distribution (CED) curves of FHR, CHR and some SOTA methods on 300W in Fig. 4, where the gap between FHR and CHR is competitive to those gaps in prior top-tier works (e.g., CFSS&MDM). Fig. 4 also shows that our FHR contributes more to those relatively easy samples, which makes sense
since the insight of FHR is to find a more precise location near a coarse but correct coordinate, whose heatmap output may accurately model the distribution of Ground Truth (GT).

Comparison on Stabilization We further compare stabilization between our method and REDnet (Peng et al. 2016) on 300-VW (Shen et al. 2017) with 7 landmarks. As shown in Tab. 4, our FHR is much more stable than REDnet according to the metric mentioned in our ablation study. To visualize stabilization improvement, we compute \( E(||\Delta x - \Delta p||^2) \) for each landmark estimated by FHR and our proposed FHR+STA, and plot them in Fig. 3(b). Fig. 3(c) plots the difference of the stability orientation \( E(||\Delta x - \Delta p||) \) of two methods, where \( \Delta \) indicates the orientation of a vector. We provide videos in our project website, which can effectively demonstrate the stability and superiority of our method. In some continuous frames, our stabilized landmarks are even more stable than the ground truth annotations.

In addition, Fig. 5 shows that our stabilization model (i.e., STA) significantly outperforms other four baselines in both of NRMSE and stability, where all 5 methods take the same z from FHR as the input. Especially in the most challenging set Scenario3 where lots of complex movements exist, our method is much better than the closest competitor (i.e., moving average), with the improvements of NRMSE and Stability to be 1.46% and 1.22%, respectively. The reasons include: 1) moving average filter and first order smoother may cause serious time delay problem; 2) although second order and constant speed methods can handle time delay, it cannot handle multiple movement types (e.g., blinking and turning head). In contrast, our algorithm can effectively address time delay issue and multiple movement types through the Gaussian mixture setting, and hence is more precise and stable. Fig. 6 shows the stability comparisons with the second order stabilization method, which is chosen as a good baseline considering its stability performance. As we can see, our method significantly outperforms the second order method when handling complex movements, and also shows better ability for time delay issue which is very close to the GT.

Time Complexity Note that our fractional heatmap regression does not impose any additional complexity burden during training compared to the conventional heatmap regression. For inference, our method provides a closed-form solution to estimate the coordinates from the heatmaps as in Eq. (4), whose runtime is negligible. Besides, after the parameter \( \Theta \) of our stabilization model is learnt, our stabilization algorithm costs \( \sim 6s \) to process the entire 64 test videos of 300-VW, which costs \( 5 \times 10^{-5} \)s per image, and can also be ignored.

Conclusions

In this paper, a novel Fractional Heatmap Regression (FHR) is proposed for high-resolution video-based face alignment.

Table 5: NRMSE comparison with state-of-the-art methods on Talking Face dataset using 7 landmarks.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NRMSE</td>
<td>3.52</td>
<td>2.36</td>
<td>3.45</td>
<td>3.32</td>
<td>2.13</td>
<td>2.28</td>
<td>2.06</td>
<td>2.14</td>
</tr>
</tbody>
</table>

Table 6: NRMSE comparison with state-of-the-art methods on 300-VW test set using 68 landmarks.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>12.5</td>
<td>7.68</td>
<td>7.66</td>
<td>5.36</td>
<td>5.44</td>
<td>5.07</td>
<td>4.42</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>7.25</td>
<td>6.42</td>
<td>6.77</td>
<td>4.51</td>
<td>4.71</td>
<td>4.34</td>
<td>4.18</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>13.10</td>
<td>13.70</td>
<td>15.00</td>
<td>12.80</td>
<td>7.92</td>
<td>7.36</td>
<td>5.98</td>
</tr>
</tbody>
</table>
The main contribution in FHR is that we leverage 2D Gaussian generation prior to accurately estimate the fraction part of coordinates, which is ignored in conventional heatmap regression based methods. To further stabilize the landmarks among video frames, we propose a novel stabilization model that addresses the time-delay and non-smooth issues. Extensive experiments on benchmarks demonstrate our proposed method is more accurate and stable than the state of the arts.

References


