Accepted Manuscript

Face Alignment Recurrent Network

Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong

PII: S0031-3203(17)30381-3
DOI: 10.1016/j.patcog.2017.09.028
Reference: PR 6295

To appear in: *Pattern Recognition*

Received date: 28 February 2017
Revised date: 11 September 2017
Accepted date: 18 September 2017

Please cite this article as: Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong, Face Alignment Recurrent Network, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.09.028

Highlights

- This paper presents a new recurrent based facial landmark detection method for images and videos under uncontrolled conditions.
- LSTM model is employed in our network to make full use of the spatial and temporal middle stage information in a natural way.
- The results show clear improvement on public image datasets and video datasets.

Face Alignment Recurrent Network

Qiqi Hou[a,b], Jinjun Wang[a,b,*], Ruibin Bai[a,b], Sanping Zhou[a,b], Yihong Gong[a,b]

[a]*Xi'an Jiaotong University*
[b]*28 Xianning W Rd, ErHuan Lu YanXian ShangYe JingJiDai, Beilin Qu, Xian Shi, Shaanxi Sheng, China, 710048*

Abstract

This paper presents a new facial landmark detection method for images and videos under uncontrolled conditions, based on a proposed Face Alignment Recurrent Network (FARN). The network works in recurrent fashion and is end-to-end trained to help avoid over-strong early stage regressors and over-weak later stage regressors as in many existing works. Long Short Term Memory (LSTM) model is employed in our network to make full use of the spatial and temporal middle stage information in a natural way, where by spatial we mean that for each image (frame), the predicted landmark position in the current stage will be used to guide the estimation for the next stage, and by temporal we mean that the predicted landmark position in the current frame will be used to guide the estimation for the next frame, and thus providing an unified framework for facial landmark detection in both images and videos. We conduct experiments on public image datasets (COFW, Helen, 300W) as well as on video datasets (300VW), and results show clear improvement over most of the current state-of-the-art approaches. In addition, it works in 18ms per image (frame)[1].

Keywords: face alignment, recurrent network

[*] Corresponding author

Email addresses: houqiqi@xjtu.stu.edu.cn (Qiqi Hou), jinjun@xjtu.edu.cn (Jinjun Wang), bairuibin@stu.xjtu.edu.cn (Ruibin Bai), sanpingzhou@stu.xjtu.edu.cn (Sanping Zhou), ygong@mail.xjtu.edu.cn (Yihong Gong)

[1] Our source code and models will be released soon

1. Introduction

Face alignment aims at locating facial key points given a 2D image, which is a fundamental component in many computer vision tasks, such as face verification [1], face recognition [2, 3, 4, 5, 6, 7] and facial attribute inference [8]. The problem has attracted a lot of research efforts yet still remains challenging, especially when the facial images are taken under uncontrolled conditions with large variation in poses, expressions and lighting conditions. The recently shape regression approaches achieved the state-of-the-art performance [9, 10, 11, 12, 13]. To elaborate, for an image $\mathbf{I}$, beginning with an initial shape estimation $\mathbf{S}^0$, these approaches iteratively estimate the facial shape $\mathbf{S}^t$ at stage t by estimating an increment $\Delta\mathbf{S}^t$ from the previous estimation $\mathbf{S}^{t-1}$, and the process can be expressed in a generic form as follows:

$$\mathbf{S}^t = \mathbf{S}^{t-1} + \mathit{Proj}^{t-1}\big(R^t(\mathbf{I},\ \mathbf{S}^{t-1})\big)\,, \tag{1}$$

where R^t indicates a stage regressor for stage t, $R^t(\mathbf{I},\ \mathbf{S}^{t-1})$ represents a normalized shape increment, and $\mathit{Proj}^{t-1}(\cdot)$ is a back-projection function that projects the normalized shape increment $\mathit{Proj}^{t-1}\big(R^t(\mathbf{I},\ \mathbf{S}^{t-1})\big)$ back into $\mathbf{S}^{t-1}$. Note that R^t *only* employs $\mathbf{I}$ and $\mathbf{S}^{t-1}$.

With training samples $\{\mathbf{I}_i,\ \hat{\mathbf{S}}_i,\ \mathbf{S}_i^0\}_{i=1}^N$ where $\hat{\mathbf{S}}_i$ indicates the ground truth shape of image $\mathbf{I}_i$, the stage regressors $(R^1,\ \cdots,\ R^t)$ are *sequentially* trained:

$$R^t = \arg\min_R \sum_{i=1}^{N} \|\mathit{Proj}^{t-1}(\hat{\mathbf{S}}_i - \mathbf{S}_i^{t-1}) - R(\mathbf{I}_i,\ \mathbf{S}_i^{t-1})\|\,. \tag{2}$$

Although these cascade frameworks have achieved noticeable success, their limitations include:

1. The multi-stage regressors are *sequentially* learnt from the first stage regressor to the last stage regressor. Each stage regressor is learnt until the training error no longer decreases. It is observed that the cascade approach tended to learn over-strong early stage regressors and over-weak later stage regressors.
2. Each stage regressor $\{R^1,\ \cdots,\ R^t\}$ is different and is individually trained. If one of them is too weak, especially a later stage one, the final facial shape detection accuracy decreases significantly.

3. The current stage regressor R^t *only* depends on the image $\mathbf{I}$ and the estimated shape $\mathbf{S}^{t-1}$, while other useful information before stage t, such as certain middle-level features, has been omitted.
4. These cascade frameworks are developed for a single image. When they are applied to video face alignment, information among frames has been omitted.

These above limitations have motivated us to propose a new end-to-end recurrent regression approach for facial landmark detection. In this paper, we improve the existing cascade shape regression by using LSTM [14, 15] and Region Convolutional Neural Network (RCNN) [16, 17, 18] to jointly train between stages to avoid over-strong and over-weak regressors as in the cascade fashion. LSTM model makes full use of the "*spatial*" and "*temporal*" middle stage information in a natural way to provide an unified framework for facial landmark detection in both images and videos. By "spatial" we mean that for each image (frame), the location of predicted landmark position in the current stage will be used to guide the estimation for the next stage, and by "temporal" we mean that the location of predicted landmark position in the current frame will be used to guide the estimation for the next frame. We recurrent the process until we get the final face shape. Our method is based on three insights:

1. By turning the existing cascade shape regression fashion into a recurrent network fashion, the model can be optimized using different stages jointly to avoid over-strong/weak regressors, especially for the several last stage regressors;
2. By utilizing the LSTM layer, the middle level representation in a deep network brings useful information and can be modeled well for shape estimation of the next stage;
3. By naturally extending the RCNN framework from image to video, successive frames can benefit from not only the middle level information, but also from the neighboring frames's estimation.

We have conducted extensive experiments for both image and video based facial landmark detection. For image, three widely used dataset, specifically the COFW [13], the HELEN dataset [19] and the 300W dataset [20], show very competitive performance by our system compared with other existing methods. For video, we also report our results on the public datasets, specifically the 300VW dataset [21, 22, 23], and the results show that our system could achieve comparable performance even with some highly engineering-optimized systems [24, 25]. In addition, our system takes about 18ms to process one single image, and is therefore fast enough for real time video process. Besides, the system can be easily applied to facial landmark tracking problem [26].

The remaining of this paper is organized as follows: Section 2 reviews related works. Section 3 describes our system, the FARN, for both the image and video version. Section 4 presents the experimental evaluations, and Section 5 concludes our paper.

2. Related Work

Face alignment plays a fundamental role in many computer vision tasks, *e.g.* DeepFace [2], High-fidelity Pose and Espression Normalization(HPEN) [3] and Multi-Directional Multi-Level Dual-Cross Patterns(MDML-DCPs) [4]. We briefly review some related work for face alignment and LSTM in the following subsections respectively.

2.1. Face Alignment

Active Shape Models (ASM) [27] and Active Appearance Models (AAM) [28] model the face shape and appearance by optimization approaches, such as Principal Component Analysis (PCA) [29]. These methods could achieve promising results on certain datasets, while their performance severely degenerates on other more challenging ones.

Lately, cascade regression based methods showwed success on both controlled and uncontrolled face alignment. Using shape indexed features, Cascade Pose Regression (CPR) [9] and Explicit Shape Regression (ESR) [10] progressively regress the shape stage by stage over the cascade random fern regressors, which are sequentially learnt. Supervised Descent Method (SDM) [11] cascades several linear regression models and achieves the superior performance with the shape indexed SIFT features. Robust Cascade Pose Regression (RCPR) [13] improves CPR with enhanced the shape indexed features and more robust initializations. Local Binary Feature (LBF) [12] is learnt for highly accurate and fast face alignment. Furthermore, Coarse-to-Fine Shape Searching (CFSS) [30] achieve highly accurate by utilizing a coarse-to-fine shape searching method.

Deep learning methods have also beenused for face alignment [31, 32, 33, 34]. Sun et al. [31] propose a three-level Convolutional Neural Network (CNN) for landmark detection. Coarse-to-fine auto-encoder networks (CFAN) [33] cascades a few successive Stacked Auto-encoder Networks. Zhang et al. [34] also cascade several convolutional networks to get the facial landmark position and further improve the result by multi-task learning. Deep regression network coupled with sparse shape regression (DRN-SSR) [35] also cascades several regression model and they mainly focus on leveraging datasets with varying annotations for face alignment. All above deep learning methods learn their network sequentially where the middle-level features have been omitted. Tasks-Constrained Deep Convolutional Network (TCDCN) [36] detects facial landmarks in one step by utilizing auxiliary information, while our method only uses the data from the specific training set without external sources. Recurrent models also are employed in some recent works. *e.g*. Recurrent Attentive-Refinement (RAR) [37] employs an attentive-refinement mechanism to determine location of facial landmarks. X. Peng [38] uses a recurrent encoder-decoder network model for face alignment in videos.

2.2. Long Short Term Memory

Recurrent Neural Networks (RNN) have long been explored in perceptual applications. Specifically, LSTM [14] have achieved impressive results in large-scale speech recognition [39] and machine translation [40, 41] applications. The RNN models' "deep in time" property [42, 43] predated deep convolution models, such as the VGG Net [44], GoogleNet [45] and recently the deep residual network [46]. Many efforts are made by researchers to combine LSTM and computer vision tasks. To list the examples, the Long-term Recurrent Convolutional Networks (LRCN) [15] developed a recurrent convolutional network architecture for large-scale visual learning, where LRCN shows distinct advantages for video recognition, image description and video description all three tasks. LSTM is also widely employed in Visual Question Answering (VQA) [47, 48, 49, 50] tasks. Stacked Attention Networks and Spatial Memory Networks [51] uses LSTM and extract soft-attention on the image features. Multimodal Compact Bilinear pooling (MCB) [52] uses LSTM to represent sentences or phrases and CNN to represent images.

Our approach applies LSTM to make full use of the spatial and temporal middle stage information: the predicted landmark position in the current stage will be used to guide the estimation in the next stage, and the predicted landmark position in the current stage will be used to guide the estimation in the next frame. The usage of the spatial and temporal information provides more than one view to describe the data [53, 54, 55, 56, 57]. Compared to [58] which uses a large-margin Gaussian process approach to help combine multiple features together and [59] which try to accomplish multi-view learning with incomplete views by assuming that different views are generated from a shared subspace, our approach using LSTM can model the spatial and temporal information in a natural way and provide an unified framework for facial landmark detection in both images and videos.

3. Face Alignment Recurrent Network

In this section, we will introduce the formulation of recurrent regression and recurrent network.

3.1. Incorporating Multi-Stage Information using Recurrent Network

For single image, denoting a data set with N training samples as $\{\mathbf{I}_i,\ \hat{\mathbf{S}}_i,\ \mathbf{S}_i^0\}_{i=1}^N$, we can optimize the network's parameter $\boldsymbol{\theta}$ as follows:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} f(\mathbf{I}_i,\ \hat{\mathbf{S}}_i,\ \mathbf{S}_i^0,\ T,\ \boldsymbol{\theta}), \tag{3}$$

where $\hat{\mathbf{S}}_i$ indicates the ground truth shape of image $\mathbf{I}_i$, $\mathbf{S}_i^0$ indicates the initial shape, T indicates the stage number. In our experiments, mean shape $\bar{\mathbf{S}}$ is employed as the

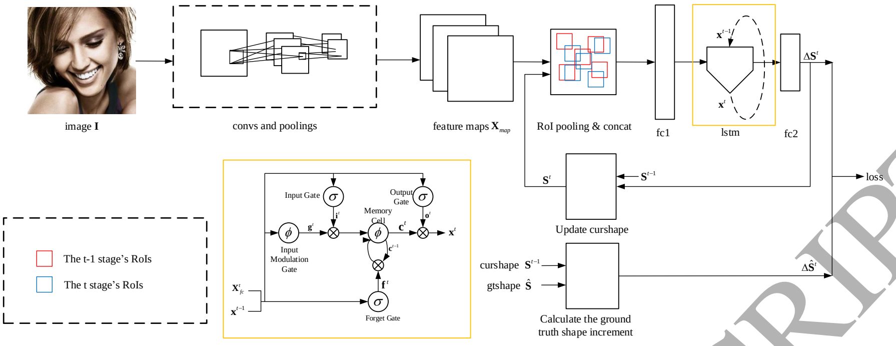


Figure 1: Face Alignment Recurrent Network(FARN) training architecture.

initial shape, which can be calculated as follow[2]

$$\bar{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{S}}_i. \tag{4}$$

We define f as

$$f \doteq \sum_{t=1}^{T}\lambda_t\sum_{i=1}^{N}\|(\hat{\mathbf{S}}_i - \mathbf{S}_i^{t-1}) - R(\mathbf{I}_i, \mathbf{S}_i^{t-1}, \mathbf{x}_i^{t-1}, \boldsymbol{\theta})\|, \tag{5}$$

where λ_t indicates the factor of each stage, R indicates the regressor with parameter $\boldsymbol{\theta}$, $\mathbf{x}_i^{t-1}$ indicates the middle level feature of stage $t-1$. Note that $\mathbf{x}_i^t$ is related to $(\mathbf{I}_i, \mathbf{S}_i^{t-1}, \mathbf{x}_i^{t-1}, \boldsymbol{\theta})$. We can get that

$$\begin{aligned}
\mathbf{x}_i^t &= g(\mathbf{I}_i, \mathbf{S}_i^{t-1}, \mathbf{x}_i^{t-1}, \boldsymbol{\theta}),\\
\mathbf{x}_i^{t-1} &= g(\mathbf{I}_i, \mathbf{S}_i^{t-2}, \mathbf{x}_i^{t-2}, \boldsymbol{\theta}),\\
&\dots\\
\mathbf{x}_i^1 &= g(\mathbf{I}_i, \mathbf{S}_i^0, \mathbf{x}_i^0, \boldsymbol{\theta}),\\
\mathbf{x}_i^0 &\doteq \mathbf{0}.
\end{aligned} \tag{6}$$

Eq.(6) means that current stage t shape $\mathbf{S}_i^t$ is not only dependent on the stage $t-1$ shape $\mathbf{S}_i^{t-1}$ and middle-level feature $\mathbf{x}_i^{t-1}$ but also all previous stage's shape and

[2] We omit the normalization term for saving space in following equations

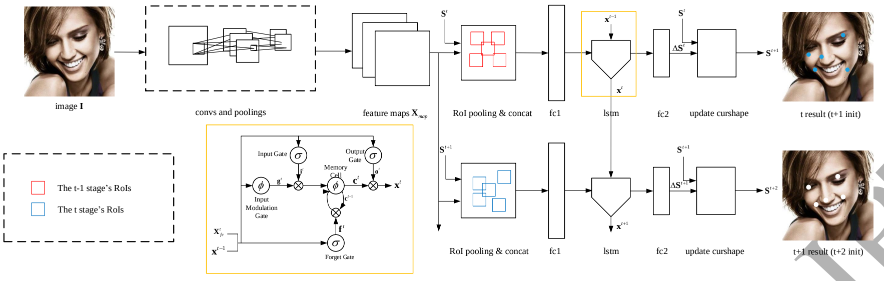


Figure 2: Face Alignment Recurrent Network(FARN) unrolled testing architecture.

middle-level information. Figure 1 and Figure 2 illustrated our workflow to apply the FARN architecture for facial landmark detection in a single image, and Algorithm 1 depicts the process in more detail.

In the training process, our approach takes the image $\mathbf{I}_i$, an initial face shape $\mathbf{S}_i^0$ and the ground truth shape $\hat{\mathbf{S}}_i$ as inputs. The network first processes the entire image with several convolutional layers and max pooling layers to produce a feature map $\mathbf{X}_{map,i}$. Then, for the initial face shape, we make Region of Interest (ROI) pooling around the region of each landmark. Then we concatenate these RoI pooling features and map them into a fully-connected layer and a LSTM layer. Finally the net will output the predicted shape increment $\Delta\mathbf{S}_i^1$ for the initial shape $\mathbf{S}_i^0$. Then we will update the initial shape $\mathbf{S}_i^0$ to $\mathbf{S}_i^1$ according to $\Delta\mathbf{S}_i^1$. We will recurrent above process but with the initial shape $\mathbf{S}_i^1$ on the convolutional feature map $\mathbf{X}_{map,i}$ to get $\mathbf{S}_i^2$. We will recurrent the process T times to get the $\mathbf{S}_i^T$, where T indicates the number of stages. Note that we compute the ground truth shape increment $\Delta\hat{\mathbf{S}}_i^t$ at each stage according to $\mathbf{S}_i^{t-1}$ and $\hat{\mathbf{S}}_i$ as the regression target of our training process. The corresponding fully connected layers and the LSTM layers of different stages share weights, and the network is end-to-end trained.

3.2. Incorporating Multi-Frame Information for Video

Similar to the above subsection, given N_V N_F-long training video samples as $\{\{\mathbf{I}_{i,f},\ \hat{\mathbf{S}}_{i,f}\}_{f=1}^{N_F},\ \mathbf{S}_i^0\}_{i=1}^{N_V}$, we define same optimized function as shown in Equation

Algorithm 1 Framework of FARN: Single image version

```
procedure TRAIN($\{\mathbf{I}_i,\ \hat{\mathbf{S}}_i,\ \mathbf{S}_i^0\}_{i=1}^N, T$)
    while in iterations do
        $\mathbf{X}_{map,i} \leftarrow \mathbf{I}_i,\ \boldsymbol{\theta}$                                ▷ convs, poolings
        $\mathbf{x}_i^0 \doteq \mathbf{0}$
        $t = 1$
        while $t <= T$ do
            $\mathbf{X}_{roi,i}^{t-1} \leftarrow \mathbf{X}_{map,i}, \mathbf{S}_i^{t-1}$                    ▷ RoI pooling, concat
            $\Delta\hat{\mathbf{S}}_i^t \leftarrow \hat{\mathbf{S}}_i,\ \mathbf{S}_i^{t-1}$                      ▷ calculate gtshape increment
            $\Delta\mathbf{S}_i^t,\ \mathbf{x}_i^t \leftarrow \mathbf{X}_{roi,i}^{t-1},\ \mathbf{x}_i^{t-1},\ \boldsymbol{\theta}$            ▷ fc, LSTM
            $\mathbf{S}_i^t \leftarrow \mathbf{S}_i^{t-1},\ \Delta\mathbf{S}_i^{t-1}$                          ▷ update curshape
        end while
        $\boldsymbol{\theta} \leftarrow \Delta\mathbf{S}_i^1, \cdots, \Delta\mathbf{S}_i^T, \Delta\hat{\mathbf{S}}_i^1, \cdots, \Delta\hat{\mathbf{S}}_i^T, \boldsymbol{\theta}$
    end while
    return $\boldsymbol{\theta}$
end procedure

procedure TEST($\mathbf{I}, T, \boldsymbol{\theta}$)
    $\mathbf{X}_{map} \leftarrow \mathbf{I},\ \boldsymbol{\theta}$                                      ▷ convs, poolings
    $\mathbf{x}_i^0 \doteq \mathbf{0}$
    $t = 1$
    while $t <= T$ do
        $\mathbf{X}_{roi}^{t-1} \leftarrow \mathbf{X}_{map}, \mathbf{S}^{t-1}$                            ▷ RoI pooling, concat
        $\Delta\mathbf{S}^t,\ \mathbf{x}^t \leftarrow \mathbf{X}_{roi}^{t-1},\ \mathbf{x}^{t-1},\ \boldsymbol{\theta}$                   ▷ fc, LSTM
        $\mathbf{S}^t \leftarrow \mathbf{S}^{t-1}, \Delta\mathbf{S}^{t-1}$                                ▷ update curshape
    end while
    return $\mathbf{S}^T$
end procedure
```

3 and Equation 5. To make fully use of the information between frames in videos, we define the f^{th} frame of the video i image $\mathbf{I}_{i,f}$'s initial shape $\mathbf{S}_{i,f}^0$ as follow

$$\mathbf{S}_{i,f}^0 = \mathbf{S}_{i,f-1}^T. \tag{7}$$

To employ the middle level information of previous frames, we define the middle

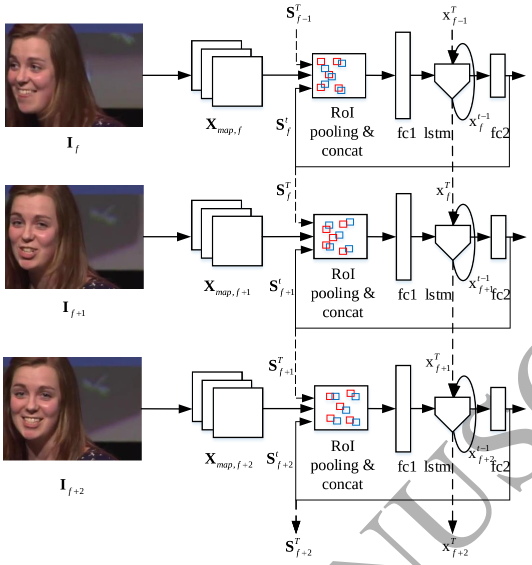


Figure 3: Extend our single image network to a video version. The shape update layer and calculate the ground truth layer are omitted in this figure.

level information as follows

$$\begin{aligned}
\mathbf{x}_{i,f}^{t} &= g(\mathbf{I}_{i,f}, \mathbf{S}_{i,f}^{t-1}, \mathbf{x}_{i,f}^{t-1}, \boldsymbol{\theta}), \\
\mathbf{x}_{i,f}^{t-1} &= g(\mathbf{I}_{i,f}, \mathbf{S}_{i,f}^{t-2}, \mathbf{x}_{i,f}^{t-2}, \boldsymbol{\theta}), \\
&\dots \\
\mathbf{x}_{i,f}^{1} &= g(\mathbf{I}_{i,f}, \mathbf{S}_{i,f}^{0}, \mathbf{x}_{i,f}^{0}, \boldsymbol{\theta}), \\
\mathbf{x}_{i,f}^{0} &= \mathbf{x}_{i,f-1}^{T}, \\
\mathbf{x}_{i,f-1}^{T} &= g(\mathbf{I}_{i,f-1}, \mathbf{S}_{i,f-1}^{T-1}, \mathbf{x}_{i,f-1}^{T-1}, \boldsymbol{\theta}), \\
&\dots \\
\mathbf{x}_{i,0}^{T} &\doteq \mathbf{0}.
\end{aligned} \tag{8}$$

As shown in Equation 8, the current stage t not only dependent on the previous stages' shape, middle level information, but also on shapes and information in the previous frames. This idea is illustrated in Figure 3, and Algorithm 2 depicts the process in detail.

Algorithm 2 Framework of FARN: Video version

```
procedure TRAIN( $\{\{\mathbf{I}_{i,f},\ \hat{\mathbf{S}}_{i,f}\}_{f=1}^{N_F},\ \mathbf{S}_i^0\}_{i=1}^{N_V}, T$)
    while in iterations do
        $\mathbf{x}_{i,0}^T \doteq \mathbf{0}$
        $f \leftarrow 1$
        while $f <= N_F$ do
            $\mathbf{X}_{map,i,f} \leftarrow \mathbf{I}_{i,f},\ \boldsymbol{\theta}$
            $\mathbf{S}_{i,f}^0 = \mathbf{S}_{i,f-1}^T$
            $\mathbf{x}_{i,f}^0 = \mathbf{x}_{i,f-1}^T$
            $t = 1$
            while $t <= T$ do
                $\mathbf{X}_{roi,i,f}^{t-1} \leftarrow \mathbf{X}_{map,i,f}, \mathbf{S}_{i,f}^{t-1}$
                $\Delta\hat{\mathbf{S}}_{i,f}^t \leftarrow \hat{\mathbf{S}}_{i,f},\ \mathbf{S}_{i,f}^{t-1}$
                $\Delta\mathbf{S}_{i,f}^t,\ \mathbf{x}_{i,f}^t \leftarrow \mathbf{X}_{roi,i,f}^{t-1},\ \mathbf{x}_{i,f}^{t-1},\ \boldsymbol{\theta}$
                $\mathbf{S}_{i,f}^t \leftarrow \mathbf{S}_{i,f}^{t-1},\ \Delta\mathbf{S}_{i,f}^{t-1}$
            end while
        end while
        $\boldsymbol{\theta} \leftarrow \Delta\mathbf{S}_{i,0}^1, \cdots, \Delta\hat{\mathbf{S}}_{i,F}^T, \boldsymbol{\theta}$
    end while
    return $\boldsymbol{\theta}$
end procedure
```

```
procedure TEST($\{\mathbf{I}_f\}_{f=1}^{N_F},\ T,\ \boldsymbol{\theta}$)
    $\mathbf{x}_0^T \doteq \mathbf{0}$
    $f \leftarrow 1$
    while $f <= N_F$ do
        $\mathbf{X}_{map,f} \leftarrow \mathbf{I}_f,\ \boldsymbol{\theta}$
        $\mathbf{S}_f^0 = \mathbf{S}_{f-1}^T$
        $\mathbf{x}_f^0 = \mathbf{x}_{f-1}^T$
        $t = 1$
        while $t <= T$ do
            $\mathbf{X}_{roi,f}^{t-1} \leftarrow \mathbf{X}_{map,f}, \mathbf{S}_f^{t-1}$
            $\Delta\hat{\mathbf{S}}_f^t \leftarrow \hat{\mathbf{S}}_f,\ \mathbf{S}_f^{t-1}$
            $\Delta\mathbf{S}_f^t,\ \mathbf{x}_f^t \leftarrow \mathbf{X}_{roi,f}^{t-1},\ \mathbf{x}_f^{t-1},\ \boldsymbol{\theta}$
            $\mathbf{S}_f^t \leftarrow \mathbf{S}_f^{t-1},\ \Delta\mathbf{S}_f^{t-1}$
        end while
    end while
    return $\{\mathbf{S}_f^T\}_{f=1}^F$
end procedure
```

4. Experiments

4.1. Datasets

Four datasets are applied for evaluating our proposed FARN model, including the COFW dataset [13], the Helen dataset [19] and the 300-W [20] for facial landmark detection in images, and the 300-VW [22, 21, 23]for videos. These datasets have the following statistics:

The first dataset, COFW, contains a large number of occluded face images. Each image has 29 landmarks. Following [13], our training set includes 845 LFPW [60] images and 500 COFW images. The remaining 507 COFW images are for the testing.

The second dataset, Helen, contains 2,300 general "n-the-wild" facial images collected from the web. Each image has 194 landmarks. Following [19], the training set contains 2,000 images and the testing set contains 300 images. The dataset is challenging in terms of computation and the large number of landmarks.

The third dataset, 300-W, contains collection of several alignment data, including AFW [61], LFPW [62], Helen [19] and XM2VTS [63] with re-annotations, as well as a new dataset called IBUG. Each image has 68 facial landmark locations. For fair comparison, we follow the protocol of [12] to construct the training partition by the training sets of LFPW, the training sets of Helen and the entire AFW with 3,148 images in total. The testing partition contains three parts: the common subset, the challenging subset and the full set, where the common subset consists of the testing set of LFPW and the testing set of Helen, with 554 images in total. The challenging test subset is the same as the IBUG set with 135 images, and finally the full set consist of both the common set and the challenging set, with 689 images in total.

The 300-VW dataset focuses on assessing the performance of face alignment system in long-term facial videos, independent of variations in pose, expression, illumination, background, occlusion and image quality. It has collected 114 facial videos in the wild. The total and average duration of the videos are 7,293 and 64 seconds respectively, and on each face there are also 68 landmark locations. All videos were captured in 30 fps, and there are 218,595 frames in total. Each video contains only one person. The training set contains 50 videos (3,063 seconds in total). The testing set (64 videos) has been divided into 3 subsets with different difficulties:

1. *Category one:* The aim of this subset is to evaluate algorithm that is suitable for facial landmark tracking system in naturalistic well-lit conditions. It contains of 31 videos. Videos in this category are recorded in well-lit conditions with various head poses and possible occlusions such as glasses and heard. Occlusions by hand or another person is not presented for the dataset.
2. *Category two:* The aim of this subset is to evaluate face alignment system that is suitable for facial motion analysis in real-world human-computer interaction applications. It contains 19 videos recorded in unconstrained conditions, such as different illuminations. People have arbitrary expressions in various head pose but without large occlusions.
3. *Category three:* The aim of this subset is to evaluate face alignment system in arbitrary conditions. It contains 14 videos recorded in completely unconstrained

conditions, including occlusions, the illuminations conditions, make-up, expressions, head pose and so on.

4.2. Evaluation metric

We evaluate our facial landmark detection system by the point-to-point Root-Mean-Square-Error (RMSE) between the face shape and the ground truth annotations. Specifically, for a face shape $\mathbf{S}_i = [x_{i,1},\ y_{i,1},\ \cdots,\ x_{i,P},\ y_{i,P}]$ and its ground truth shape $\hat{\mathbf{S}}_i = [\hat{x}_{i,1},\ \hat{y}_{i,1},\ \cdots,\ \hat{x}_{i,P},\ \hat{y}_{i,P}]$, $RMSE_i$ can be represented as follow:

$$RMSE_i = \frac{1}{Pd_i}\sum_{p=1}^{P}\sqrt{(x_{i,p}-\hat{x}_{i,p})^2+(y_{i,p}-\hat{y}_{i,p})^2}, \tag{9}$$

where P indicates the number of landmarks, d_i is normalization term and equal to the pupil distance computed as the Euclidean distance between the pupils. We use the mean RMSE as their final error.

Following the evaluation criteria of the 300-VW challenge [22, 21, 23], we use the cumulative error curve of the percentage of images as well as RMSE to evaluate the algorithms in 300-VW. Besides, we also draw the cumulative error curve for Helen and iBUG for the quantitative and qualitative evaluation.

4.3. Experiment Setting

To train our model, we augment our training data *only* with a flipping version. We use the well-known Caffe[64] to implement our network in the experiments. We have 8 convolutional layers and 2 pooling layers to generate the feature map. The recurrent resections, that are stages, are unrolled as a small network in a single layer. The model is end-to-end trained and the parameters of our network can be found in Table 1. The neural networks are trained by stochastic gradient descent with momentum set to 0.9. And we have set the learning rate for all learnable layers to 0.001, and it will decrease by timing 0.1 every 20,000 steps until 10^{-7}. The VGG16 model is used to initialize the former convolutional layers before conv 3-3. conv 3-3, fc layers and lstm layer are initialized by a zero-mean Gaussian distribution with σ set to 0.001 and biases set to 0. The size of regions around landmarks of the RoI max pooling layer has been set to 0.2 of the current shape's bounding box, and each region will be pooled into 3×3 features.

The number of stages is set to 5. The loss weight factors for each stage are set to 1, 2, 3, 3, 3. For image version, we input 1 image, 1 ground truth face shape as well as 64 initial shapes per iteration. We test our network with 16 initial shapes, and average the output as final result. For the video version, we train our network with 8 frame clips with 64 initial face shapes. We use 4-frame-clips with 16 initial face shapes and a stride of 2 frames to test our network. We obtain the final result by averaging outputs across clips. More details will be found on our codes.

Table 1: Network structure of FARN

Layer	conv1-1 to conv3-2	conv3-3	fc	lstm
Param	Same to VGG16	$8 * 3^2$	1024	256

We evaluate our system's performance on an Intel(R) Core(TM) i7-3770 CPU with 3.40 GHz and a Nvidia GeForce GTX TITAN X graphics card.

4.4. Facial Landmark Detection Accuracy

Table 2: Comparison results on COFW and Helen

Method	COFW	Method	Helen
RCPR	8.50	RCPR	6.50
HPM	7.46	ESR	5.70
RPP	7.52	SDM	5.85
TCDCN	8.05	LBF	5.41
RAR	6.03	LBF fast	5.80
FARN	**5.81**	FARN	**4.65**

In this section, we compare our approach with the state-of-the-art methods including the traditional methods (ESR [10] , SDM [11] and LBF [12]), the deep learning methods (CFAN [33], DRN-SSR [35] and TCDCN [36]) and recurrent network based method (RAR [37] and X. Peng [38]). We compare with methods that are designed to handle occlusion (RCPR [13], HPM [67] and RPP [68]).

Table.2 and 3 report the RMSE of the compared methods on the COFW, the Helen and the 300-W datasets respectively, and Fig. 4 plots the corresponding cumulative error distribution curve (test challenging subset of 300-W) datasets. It is clear that our

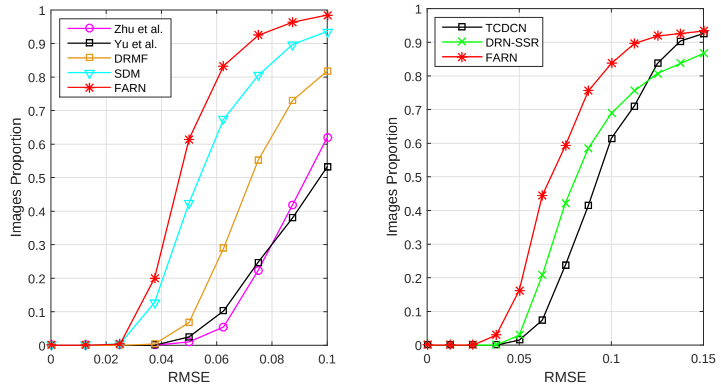
Table 3: Comparison results on 300-W

Method	Full	Common	Challenging
Zhu et al[61]	10.20	8.22	18.33
RCPR	8.35	6.18	17.26
Smith et.al[65]	-	13.30	-
GN-DPM[66]	-	5.78	-
CFAN	-	5.50	-
ESR	7.58	5.28	17.00
SDM	7.52	5.60	15.40
LBF	6.32	4.95	11.98
LBF fast	7.37	5.38	15.50
CFSS[30]	5.76	4.73	9.98
TCDCN	5.54	4.8	8.6
RAR	4.94	**4.12**	8.35
FARN	**4.88**	4.23	**7.53**

Table 4: Comparison results on 300-VW

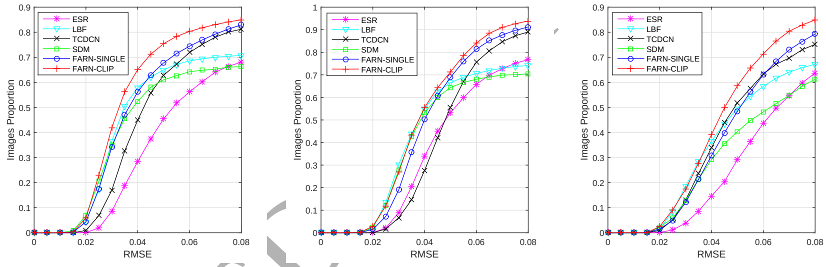
Method	Category one	Category two	Category three
SDM	14.80	11.25	13.24
ESR	18.61	12.20	15.31
LBF	9.49	7.64	8.45
TCDCN	6.85	5.29	6.57
FARN	**6.16**	**4.42**	**5.90**

model has improved a lot on all those image based methods. Compared with TCDCN which utilized pre-trained models on the Multi-Attribute Facial Landmark (*MAFL*) database, our model is trained only on the dataset's respective training set. Compared with RAR which firstly utilized a VGG19 model to generate the robust initial shape, our method uses mean shape as our initial shape and only employs the first several layers of VGG16. So our method runs much faster (18ms vs 250ms). On 300-W, RAR also generated training samples with occlusions by natural objects, e.g. sunglasses, medical masks, phones, hands and cups, as well as their rotation, scaling and mirroring, while our data is only augmented by its flipping.



(a) Helen (b) iBUG

Figure 4: Cumulative errors distribution curves of Helen and iBUG.



(a) Test Category One (b) Test Category Two (c) Test Category Three

Figure 5: Cumulative errors distribution curves of 300-VW.

Table 4[3] report the RMSE of the compared methods on 300-VW, and Figure 5 shows the cumulative error distribution curve on 300-VW. As can be seen, our method also achieves state-of-the-art performance on these datasets again.

Table 5 reports the run time of the compared deep learning methods on 300-W. FARN is accelerated by sharing computation of the conv feature map. Besides, our

[3] Following the evaluation of criteria of the 300-VW challenge [22, 21, 23], we report the 49-points RMSE. According to the criteria and the organizer, some frames in the test set were not used during the evaluation process because the face in these frames are far from frontal.

Table 5: Comparison of speed on 300W dataset.

Algorithm	Run time(ms)
Cascade CNN [31]	120
CFAN [33]	25
X.Peng [38]	30
RAR [37]	250
FARN	**18**

245 method takes the mean shape as the initial shape. Our method runs faster than those deep learning methods.

Overall, our proposed network outperform most existing works by a large margin. Specially our method has achieved significant error reduction on the challenging iBUG and 300-VW. We believe that it is due to the end-to-end training and weight sharing 250 for all stages. Some results are listed in Figure 6. Also it is observed that our system processes one 100 × 100 image in 18ms, which is fast enough for real time face alignment.

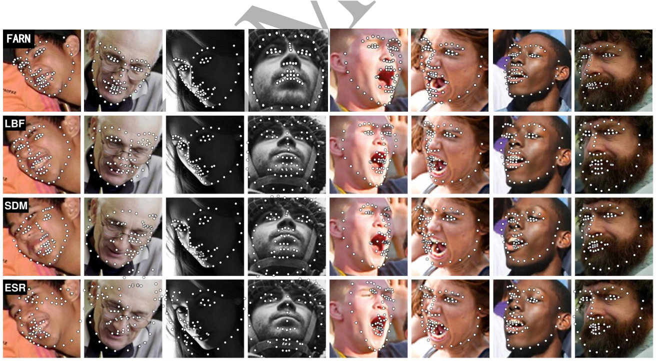


Figure 6: Comparison on the 300-W challenging dataset.

4.5. Discussion

Our proposed FARN has two key components, *i.e.* the *recurrent model* and the 255 *usage of the middle-level information*. Hence in this section we conducted additional experiments to investigate their respective performance.

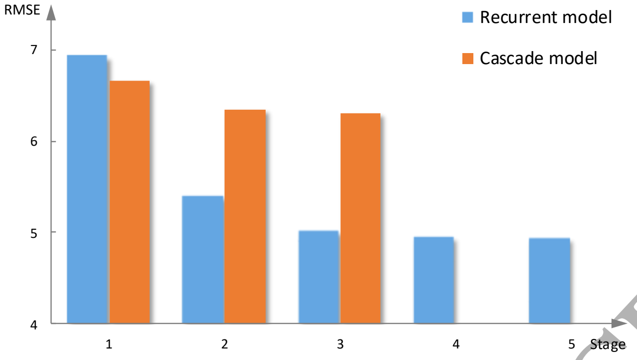


Figure 7: Comparison between recurrent model and cascade model on 300-W.

Recurrent model vs. cascade model In the baseline method, we *sequentially* train multi-stage regressors. Each stage regressor is learnt until the training error no longer decreases. With all other parameters kept the same, our recurrent model is jointly trained between stages to avoid over-strong/weak regressors. Hence it is interesting to compare whether such scheme can benefit the facial landmark detection task. Using 300-W, the results are shown in Figure 7, where we can find that our method has higher accuracy than sequentially trained methods. At the stage 1, cascade model can get greater performance than recurrent model. However, the recurrent model achieves better at stage 2 and 3. We also noticed that the sequentially trained network can be hard to converge after first 3 stages while recurrent model can achieve great performance though 5 stages. Figure 8 gives some examples. The winning performance from the recurrent model verifies that it can avoid over-strong/weak regressors as in the cascade case.

Larger receptive fields for the later stage regression The middle-level features can provide larger receptive fields for the later stage regression, which can help the network look wider and thus get better result. Take Figure 9 for example, yellow, red, blue points denote the ground truth shape, the first stage's shape, and the second stage's shape respectively. The red and blue rectangles denote the RoIs for the first and second stage respectively. If we omit the middle-level features, the second stage's receptive field is shown as Figure 9(b), which is only dependent on the second stage's shape.

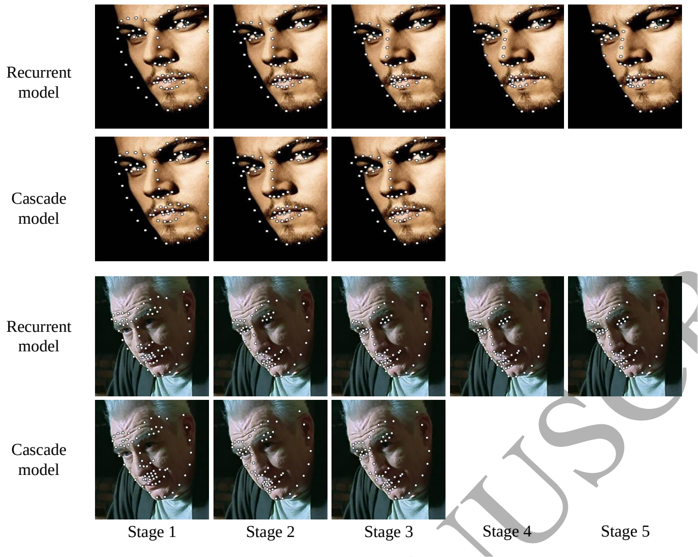


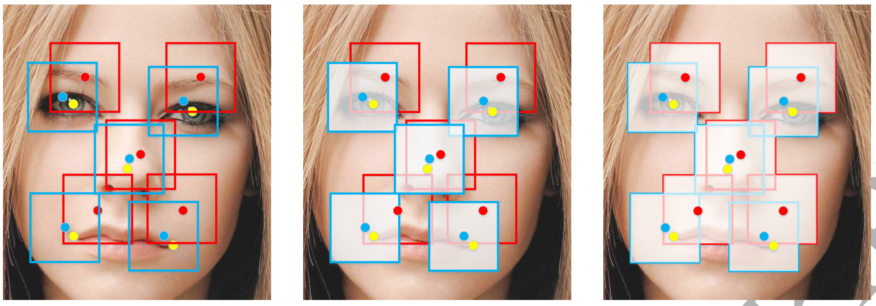
Figure 8: Examples of the comparison between recurrent model and cascade model on 300-W.

By utilizing the middle-level features, the second stage's field will not only contain the second stage's RoIs, but also contains the first stage's RoIs, like Figure 9(c). We believed that larger receptive field can improve the network's accuracy.

Table 6: Contribution of middle level information among stages

Method	COFW	Helen	300-W		
			Full	Common	Challenging
FARN_FC	6.06	5.26	5.52	4.75	8.65
FARN_LSTM	**5.81**	**4.83**	**4.88**	**4.23**	**7.53**

Usage of middle level information among stages To examine how much the middle level information among stages contribute to facial landmark detection, we use a fully-connected layer to replace the LSTM layer, and thus create the FARN_FC model. FARN_FC omit the middle-level features. From Table 6, we can find that the accuracy of FARN_LSTM is higher than FARN_FC on all datasets. We infer that the middle-level hidden features can help network predict the next stage's shape. Furthermore, in figure 10, we visualize the features of the middle level information of different stages



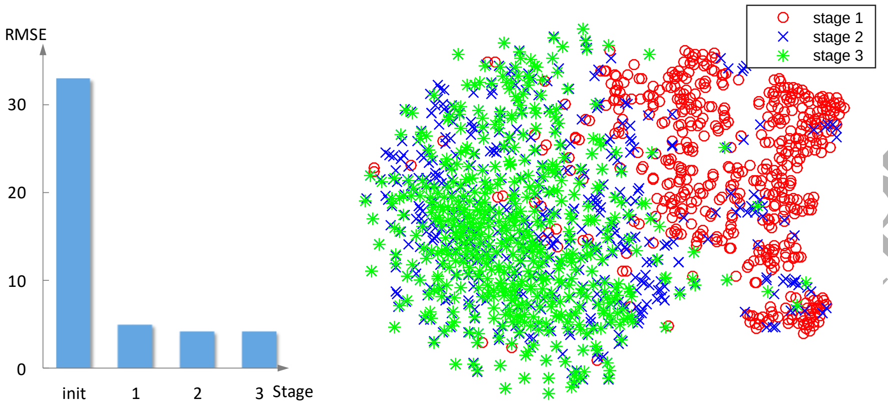
(a) Yellow, red and blue points denote the ground truth's shape, the first stage's shape and the second stage's shape respectively. Rectangles denote the RoIs.

(b) Receptive field of the second stage NOT using middle-level features

(c) Receptive field of the second stage using middle-level features

Figure 9: Comparison of the receptive field in the later stage regression.

with the help of t-Distributed Stochastic Neighbor Embedding (t-SNE) [69, 70], which is capable of giving each data point a location in a two or three-dimensional map with retaining the local structure of the data while also revealing some important global structure (such as clusters at multiple scales). The technique is a variation of Stochastic Neighbor Embedding[71] that is easy to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. Stage 1 do regression from the init shape (mean shape), while stage 2 and 3 regresses from the previous stage's shape. If the hidden features are not related the current stage's prediction, the hidden features of different stages will mix evenly with each other. However, from Figure 10(b), we can find that the distribution of stage 1's hidden features are separated to stage 2 and 3, and the distribution of stage 2 coincides with stage 3. As shown Figure 10(a), the stage 2's RMSE decrement (from stage 1 to stage 2) and stage 3's (from stage 2 to stage 3) is relatively small compared to stage 1's (from stage init to stage 1). From this, we infer that the middle-level hidden features contain information related to the current stage's prediction. Since the later stage regressor's predicted shapes tend to converge to the final result(Figure 10(a)), the later stage's hidden features will tend to coincides with each other. However, in the earlier



(a) RMSE of each stage (b) Different stages' hidden features of the LSTM layer.

Figure 10: Visualization of the hidden features of 300-W common subset. Stage 1's hidden features in (b) correspond to RMSE decrement from stage init to stage 1 in (a). Stage 2's hidden features in (b) correspond to RMSE decrement from stage 1 to stage 2 in (a). Stage 3's hidden features in (b) correspond to RMSE decrement from stage 2 to stage 3 in (a)

stage regression, the regressor often regresses to a much different shape. That will lead to the different distributions among those hidden features. Figure 11 shows some examples.

Usage of middle level information among frames To further examine how middle level information among frames help improve the facial landmark detection accuracy for videos, we create another model that does not use the information among frames. Instead. each frame in the video is processed as a single image. As can be seen from Table 7, by utilizing information between frames (the FARN_Clip), much lower estimation error than that without information from neighboring frames (the FARN_Single). Note that according to Table 4 and Figure 5, the accuracy of FARN_Single already outperform other methods. By utilizing the between frame information, we obtain another drop of estimation error.

5. Conclusion

In this paper, we introduce a FARN model for facial landmark detection for images and videos under uncontrolled conditions. FARN provides an unified framework

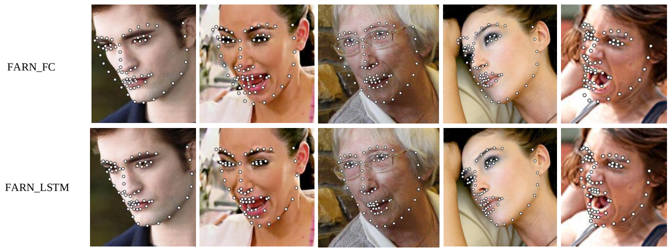


Figure 11: Examples of the comparison of FARN_FC and FARN_LSTM

Table 7: Contribution of middle level information among frames

Method	Category one	Category two	Category three
FARN_Single	6.65	4.60	6.46
FARN_Clip	**6.16**	**4.42**	**5.90**

to make full use of the spatial and temporal middle stage information to improve the accuracy in both images and videos. Experimental results from four widely adopted public datasets show clear improvement over many existing approaches. We are currently extending the system for facial landmark tracking problem.

References

[1] C. Lu, X. Tang, Surpassing human-level face verification performance on lfw with gaussianface, Computer Science.

[2] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.

[3] X. Zhu, Z. Lei, J. Yan, D. Yi, S. Z. Li, High-fidelity pose and expression normalization for face recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 787–796.

[4] C. Ding, J. Choi, D. Tao, L. S. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE transactions on pattern analysis and machine intelligence 38 (3) (2016) 518–531.

[5] K.-K. Huang, D.-Q. Dai, C.-X. Ren, Y.-F. Yu, Z.-R. Lai, Fusing landmark-based features at kernel level for face recognition, Pattern Recognition 63 (2017) 406–415.

[6] Y. Su, X. Gao, X.-C. Yin, Fast alignment for sparse representation based face recognition, Pattern Recognition 68 (2017) 211–221.

[7] S. Soltanpour, B. Boufama, Q. J. Wu, A survey of local feature methods for 3d face recognition, Pattern Recognition.

[8] N. Kumar, P. Belhumeur, S. Nayar, Facetracer: A search engine for large collections of images with faces, in: European Conference on Computer Vision, 2008, pp. 340–353.

[9] P. Dollár, P. Welinder, P. Perona, Cascaded pose regression, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 1078–1085.

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

[11] X. Xiong, F. Torre, Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 532–539.

[12] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1685–1692.

[13] X. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1513–1520.

[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[16] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[17] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[19] V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang, Interactive facial feature localization, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 679–692.

[20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: The first facial landmark localization challenge, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.

[21] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, M. Pantic, The first facial landmark tracking in-the-wild challenge: Benchmark and results, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 50–58.

[22] G. Chrysos, E. Antonakos, S. Zafeiriou, P. Snape, Offline deformable face tracking in arbitrary videos, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 1–9.

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

[24] J. Yang, J. Deng, K. Zhang, Q. Liu, Facial shape tracking via spatio-temporal cascade shape regression, in: IEEE International Conference on Computer Vision Workshop, 2015, pp. 994–1002.

[25] S. X. S. Y. A. A. Kassim, Facial landmark detection via progressive initialization, ICCV2015 workshop.

[26] X. Peng, S. Zhang, Y. Yang, D. N. Metaxas, Piefa: Personalized incremental and ensemble face alignment, in: IEEE International Conference on Computer Vision, 2015, pp. 3880–3888.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Active shape models-their training and application, Computer vision and image understanding 61 (1) (1995) 38–59.

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

[29] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[30] S. Zhu, C. Li, C. Change Loy, X. Tang, Face alignment by coarse-to-fine shape searching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4998–5006.

[31] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3476–3483.

[32] Y. Wu, Z. Wang, Q. Ji, Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3452–3459.

[33] J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 1–16.

[34] Z. Zhang, P. Luo, C. C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 94–108.

[35] S. S. X. C. Jie Zhang, Meina Kan, Leveraging datasets with varying annotations for face alignment via deep regression network, ICCV2015.

[36] Z. Zhang, P. Luo, C. Loy, X. Tang, Learning deep representation for face alignment with auxiliary attributes, IEEE Transactions on Pattern Analysis & Machine Intelligence 38 (5) (2016) 1–1.

[37] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, A. Kassim, Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks, 2016.

[38] X. Peng, R. S. Feris, X. Wang, D. N. Metaxas, A recurrent encoder-decoder network for sequential face alignment, in: European Conference on Computer Vision, 2016.

[39] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1764–1772.

[40] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in neural information processing systems, 2014, pp. 3104–3112.

[41] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint arXiv:1409.1259.

[42] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, Tech. rep., DTIC Document (1985).

[43] R. J. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, Neural computation 1 (2) (1989) 270–280.

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

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[47] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[48] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[49] I. Ilievski, S. Yan, J. Feng, A focused dynamic attention model for visual question answering, arXiv preprint arXiv:1604.01485.

[50] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (1) (2017) 32–73.

[51] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[52] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, arXiv preprint arXiv:1606.01847.

[53] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634.

[54] C. Xu, D. Tao, C. Xu, Large-margin multi-viewinformation bottleneck, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (8) (2014) 1559–1572.

[55] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE transactions on pattern analysis and machine intelligence 37 (12) (2015) 2531–2544.

[56] L. Rossi, A. Torsello, E. R. Hancock, Unfolding kernel embeddings of graphs: Enhancing class separation through manifold learning, Pattern Recognition 48 (11) (2015) 3357–3370.

[57] Z. Zhang, E. R. Hancock, Hypergraph based information-theoretic feature selection, Pattern Recognition Letters 33 (15) (2012) 1991–1999.

[58] C. Xu, D. Tao, Y. Li, C. Xu, Large-margin multi-view gaussian process for image classification, in: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, ACM, 2013, pp. 7–12.

[59] C. Xu, D. Tao, C. Xu, Multi-view learning with incomplete views, IEEE Transactions on Image Processing 24 (12) (2015) 5812–5825.

[60] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars., IEEE Transactions on Pattern Analysis & Machine Intelligence 35 (12) (2011) 545–552.

[61] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2879–2886.

[62] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (12) (2013) 2930–2940.

[63] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, Xm2vtsdb: The extended m2vts database, in: Second international conference on audio and video-based biometric person authentication, Vol. 964, Citeseer, 1999, pp. 965–966.

[64] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

[65] B. Smith, J. Brandt, Z. Lin, L. Zhang, Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1741–1748.

[66] G. Tzimiropoulos, M. Pantic, Gauss-newton deformable part models for face alignment in-the-wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1851–1858.

[67] G. Ghiasi, C. C. Fowlkes, Occlusion coherence: Detecting and localizing occluded faces, Computer Science.

[68] H. Yang, X. He, X. Jia, I. Patras, Robust face alignment under occlusion via regional predictive power estimation, IEEE Transactions on Image Processing 24 (8) (2015) 2393–403.

[69] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.

[70] L. Van Der Maaten, Accelerating t-sne using tree-based algorithms., Journal of machine learning research 15 (1) (2014) 3221–3245.

[71] G. E. Hinton, S. T. Roweis, Stochastic neighbor embedding, in: Advances in neural information processing systems, 2003, pp. 857–864.

Qiqi Hou received the B.S. degree from the Xi’an Jiaotong University of China in 2014. Now He is a Master student in the Xi’an Jiaotong University of China. His research interests include computer vision, especially detection and localization of general object and face.

Jinjun Wang got his PhD from Nanyang Technological University, Singapore in 2008. From 2006 to 2013, Prof. Wang worked in Silicon Valley, USA for leading research institutes including NEC laboratories America, Inc. and Epson Research and Development, Inc. as research scientist and senior research scientist. In 2013, Prof. Wang joined Xian Jiaotong University, School of Electronic and Information, Department of automation, Institute of Artificial Intelligence and Robotics. He was selected into the Chinese national "1000 Youth Talent" program and the Shaanxi "100 Talent" program in the same year.

Ruibin Bai received the B.S. degree from the Xi’an Jiaotong University of China in 2016. Now He is a Master student in the Xi’an Jiaotong University of China. His research interests include computer vision, deep learning.

Sanping Zhou received the Master degree from the Northwestern Polytechnical University of China in 2015. Now He joint a PhD program in the Xi’an Jiaotong University of China. His research interests include computer vision, deep learning, especially in human pose estimation.

Yihong Gong received the B.S., M.S., and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. In 1992, he joined Nanyang Technological University, Singapore, where he was an Assistant Professor with the School of Electrical and Electronic Engineering for four years. From 1996 to 1998, he was a Project Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. He was a Principal Investigator for both the Informedia Digital Video Library Project and the Experience-On-Demand Project funded by the National Science Foundation, the Defense Advanced Research Projects Agency, the National Aeronautics and Space Administration, and other government agencies. From 1999 to 2009, he was the Head of the Department of Information Analysis and Management, NEC Laboratories America, Inc., Cupertino, CA. He is currently a professor with Xi’an Jiaotong University (XJTU), China. His research interests include image and video analysis, multimedia database systems, and machine learning.