○ Applied Informatics

CrossMark

# Multiscale recurrent regression networks for face alignment

Caixun Wang[1,2,3], Haomiao Sun[1,2,3], Jiwen Lu[1,2,3*], Jianjiang Feng[1,2,3] and Jie Zhou[1,2,3]

*Correspondence:
lujiwen@tsinghua.edu.cn
[3] State Key Lab of Intelligent
Technologies and Systems,
Beijing 100084, People's
Republic of China
Full list of author information
is available at the end of the
article

## Abstract

In this paper, we propose an end-to-end multiscale recurrent regression networks (MSRRN) approach for face alignment. Unlike conventional face alignment methods by utilizing handcrafted features which require strong prior knowledge by hand, our MSRRN aims to seek a series of hierarchical feature transformations directly from image pixels, which exploits the nonlinear relationship between the face images to the positions of facial landmarks. To achieve this, we carefully design a recurrent regression network architecture, where the parameters across different stages are shared to memorize the shape residual descents between the initial shape and the ground-truth. To further improve the performance, our MSRNN learns to exploit the context-aware information from multiscale face inputs in a coarse-to-fine manner. Experimental results on the benchmarking face alignment datasets show the effectiveness of our approach.

**Keywords:** Face alignment, Facial landmark detection, Recurrent neural networks, Deep learning, Biometrics

## Background

Face alignment (Liu et al. 2017b, 2017c; Zhang et al. 2016; Zhu et al. 2015) has gained much attention in facial recognition (Duan et al. 2017; Lu et al. 2017) and computer vision areas (Hu et al. 2017; Liu et al. 2017a), which aims to densely localize a set of positions of semantic facial landmarks such as eyes, nose, chin, etc. While extensive efforts have been devoted to face alignment, the performance is still not satisfactory especially when face samples were captured under wild conditions, duo to large variations of diverse facial expressions, aspect ratios, and partial occlusions. Motivated by this, a robust face alignment method should be proposed to address the previous limitations.

Conventional face alignment methods can be roughly divided into two categories: the holistic models and the local models. The representative holistic models include active shape model (ASM) (Cootes et al. 1995) and active appearance model (AAM) (Cootes et al. 2001). Both methods aim at maximizing the joint posterior probability over landmarks for the given facial images. However, these methods cannot explicitly exploit the local details during facial shape refinement, which exhibits the important cues for face alignment. The representative local models, such as the local constrained model (LCM) (Cootes et al. 2012), focus on modeling the constraint shape model locally. While local features are utilized to transform and vote for facial landmark detection, the

Wang *et al. Appl Inform* (2017) 4:13

Page 2 of 11

performance is still far from being practically satisfactory in such cases, because face samples undergo large variances of aspect ratios and occlusions in the real-world application. To circumvent this problem, cascaded regression-based methods have been proposed to learn a series of nonlinear feature-to-shape mappings and further estimate the positions of facial landmarks in a coarse-to-fine manner. For example, Cao et al. (2012) proposed an explicit shape regression (ESR) model to address the cascaded regression problem by means of boosting tree. Xiong and la Torre (2013) proposed a supervised decent model (SDM) to relax the nonlinear regression optimization by leveraging a series of linear regression functions cascaded. However, the employed features are handcrafted, which may lose crucial shape-informative details, and even the performance degrades in uncontrolled environments. To address this limitation, Sun et al. (2013) proposed a deep convolutional network-cascaded (DCNC) method to predict facial landmarks by integrating both tasks of shape initialization and shape update. Zhang et al. (2014) developed a coarse-to-fine auto-encoder network (CFAN) architecture for face alignment, which exploits the image-to-shape mappings by leveraging multilayer neural networks. Nevertheless, these deep learning methods separately learned network parameters for each stage, which may lead to the suboptimal solution during back-propagation procedure.

In this study, we propose an end-to-end multiscale recurrent regression networks (MSRRN) approach for face alignment. Unlike conventional face alignment methods utilizing handcrafted features which requires strong prior knowledge by hand, our MSRRN model aims at jointly optimizing both tasks of learning shape-informative local features and localizing facial landmarks in a unified deep convolutional neural networks. As illustrated in Fig. 1, we carefully design a recurrent regression network to transform shape-index local raw patches to the spacial coordinates of facial shape, where the network parameters between stages are shared across different stages. As a result, the refinement descents for each stage are memorized, and the capacity of deep architecture is well controlled. To further improve the face alignment performance, our model is equipped with a multiscale schema to exploit the complementary information
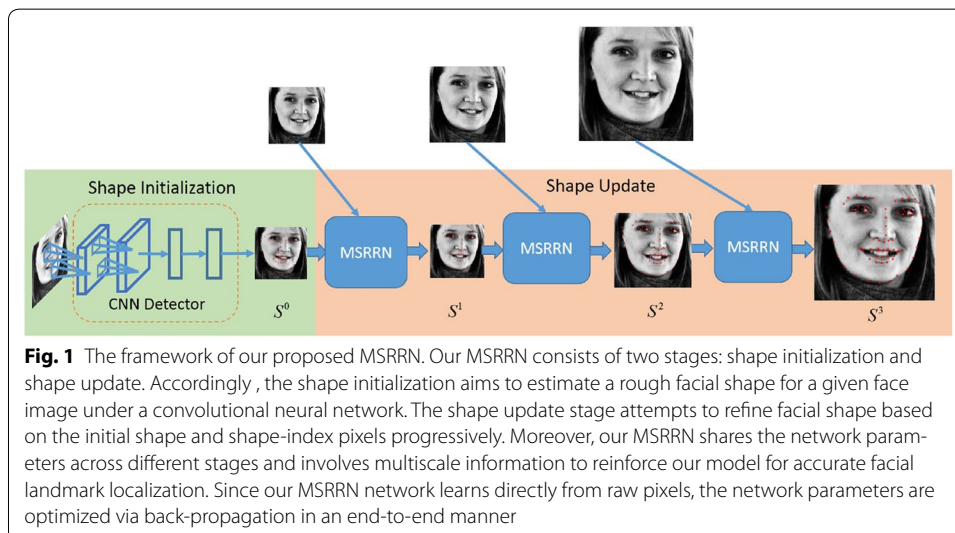


**Fig. 1** The framework of our proposed MSRRN. Our MSRRN consists of two stages: shape initialization and shape update. Accordingly , the shape initialization aims to estimate a rough facial shape for a given face image under a convolutional neural network. The shape update stage attempts to refine facial shape based on the initial shape and shape-index pixels progressively. Moreover, our MSRRN shares the network parameters across different stages and involves multiscale information to reinforce our model for accurate facial landmark localization. Since our MSRRN network learns directly from raw pixels, the network parameters are optimized via back-propagation in an end-to-end manner

Wang *et al. Appl Inform* (2017) 4:13

Page 3 of 11

from multiscale inputs during facial landmark localization. To show the effectiveness of our proposed MSRRN, we conduct experiments on the standard benchmarking dataset 300-W including the LFPW, HELEN, and IBUG datasets, where 68 landmarks were employed for evaluation. "Experimental results" show that our proposed MSRRN performs face alignment in a robust manner compared with most of the state-of-the-art methods.

## Proposed method

Unlike conventional cascaded regression-based methods which sequentially learn the stagewise shape regressors and may cause suboptimal performance during training process, we propose an end-to-end multiscale recurrent regression networks approach for face alignment. Specifically, we carefully design regression schema to model image-to-shape relationship via deep convolutional neural networks, and at the same time, learn to automatically characterize shape-sensitive local features directly from raw pixels. To achieve this goal, our model leverages a recurrent framework by sharing the network parameters in order to preserve the consistency of information across each stage. Moreover, we extend our architecture to a multiscale method, which involves obtaining complementary information from multiscale face inputs in a coarse-to-fine manner. In the next section, we will describe the formulation and optimization procedure of the proposed method.
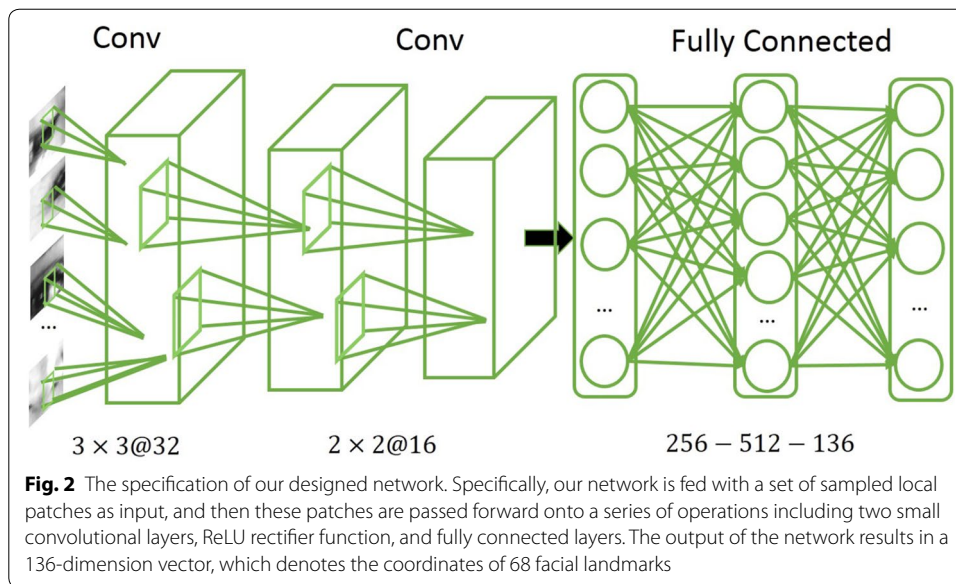
Let $S = [x_1, y_1, \ldots, x_l, y_l, \ldots, x_L, y_L]^T$ denote a facial shape with a set of facial landmarks; typically, $L$ was specified to be 68 in this paper, where $(x_l, y_l)$ represents the spatial coordinates of the $N_l$th facial landmark. Given the training set $\{(I_n, S_n)\}_{n=1}^N$ which consists of $N$ data points, where $I_n$ denotes the nth face image and $S_n$ is its ground-truth facial shape, the goal of face alignment is to estimate a shape $S$ that is as close in resemblance as possible to the real facial shape on the input face image $I$. The final shape estimation is refined based on the initial shape and facial image features progressively:

$$S^T = S^0 + \sum_{t=1}^{T} f^t(I, S^{t-1}), \tag{1}$$

where $S^0$ denotes the initial shape, $S^t$ denotes the final shape estimation, $T$ is the total stage number, and $f^t(\cdot, \cdot)$ denotes the image-to-shape mapping (basically, a regression function) for the $t$th stage.

The crucial part is to learn the regression function. The basic idea of our objective is to minimize the residual of initial shape and ground-truth shape for each stage. This enables the facial shape to be refined as close as possible to the real shape in a course-to-fine manner. In terms of the regression method, we leverage a series of nonlinear functions to transform the raw facial images to the targeted facial shape. Figure 2 shows the specification of the employed network architecture. Our network is equipped with a set of convolutional layer, pooling layer, and fully connected layers. Hence, we compute the predicted shape residual based on the deep network structure as follows:

$$f(I_i, S^{t-1}) = \text{pool}\left(\text{ReLU}(\mathbf{W} \otimes \phi_i + \mathbf{b})\right), \tag{2}$$

Wang *et al. Appl Inform* (2017) 4:13

Page 4 of 11



**Fig. 2** The specification of our designed network. Specifically, our network is fed with a set of sampled local patches as input, and then these patches are passed forward onto a series of operations including two small convolutional layers, ReLU rectifier function, and fully connected layers. The output of the network results in a 136-dimension vector, which denotes the coordinates of 68 facial landmarks

where $\otimes$ denotes the convolution operation, $\text{pool}(\cdot)$ denotes the max pooling operation, and $\text{ReLU}(\cdot)$ denotes the rectifier nonlinear function. For $\phi_i$, we utilize the shape-index local patch, which is computed as follows:

$$\phi_i = I_i \circ S^{t-1}, \tag{3}$$

where $\circ$ denotes the sampling operation resulting shape-index local patches based on the given initial shape. Note that the shape-index features implicitly exploit the shape constraints for the holistic regression.

To achieve this, we formulate our objective which aims at minimizing the following optimization:

$$\arg\min_{f} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\| f^t\left(I_i, S^{t-1}\right) - \left(S^* - S^{t-1}\right) \right\|_2^2, \tag{4}$$

where $\| \cdot \|_2^2$ denotes the $L_2$ norm to measure the distance of any two shapes in the Euclidean space, and $S^*$ represents the manually labeled landmarks.

Since conventional face alignment methods seek different parameters of $f(\cdot, \cdot)$ for different stages $t$, the learning strategy encounters parameter scalability problem. To address this issue, we propose a recurrent network architecture, where the parameters between different stages are shared. As a result, the descents of facial landmarks during each stage are memorized and involved for further refinement. Moreover, we extend our proposed model to a multiscale framework, which aims at exploiting the complementary information received from multiscale face inputs and localizing facial landmarks in a coarse-to-fine manner. Hence, we revise our formulation based on (4) as follows:

$$J = \min_{f} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\| f_{RRN}\left(I_i^{(t)}, S^{t-1}\right) - \left(S^* - S^{t-1}\right) \right\|_2^2 + \lambda \|\mathbf{W}\|_2^2. \tag{5}$$

Wang *et al. Appl Inform* (2017) 4:13

Page 5 of 11

where $\|\mathbf{W}\|_2^2$ is employed to reduce the model complexity which prevents the learning procedure from overfitting, $I^{(t)}$ denotes the scaled face input (we downscale three times during learning procedure, thus we specified $T$ to 3 in our experiments.), and $\lambda$ is the hyper-parameter to balance the objective term and the model regularization term. Note that the parameters of these regression functions $f_{\text{RRN}}(\cdot, \cdot)$ are shared across different stages.

To solve the optimization problem in (5), we leverage the stochastic gradient-decent method to compute the gradients. Specifically, for each iteration, we first pass the batched data forward onto the unrolling network, and compute the immediate and top layer results. Then we propagate these results back to the network and perform the gradients. Having obtained the gradients, the network parameters $\mathbf{W}$ and $\mathbf{b}$ are updated by averaging the gradients of different stages. The update procedure is performed using the gradient-decent algorithm as follows until convergence:

$$\mathbf{W} = \mathbf{W} - \eta \frac{\partial J}{\partial \mathbf{W}}, \tag{6}$$

$$\mathbf{b} = \mathbf{b} - \eta \frac{\partial J}{\partial \mathbf{b}}, \tag{7}$$

where $\eta$ is the learning rate, which controls the convergence speed of the objective function (5). Algorithm 1 shows the optimization procedure of MSRRN.

---

**Algorithm 1:** MSRRN

**Input**: Training set $X$, stage number $T$ ($T = 3$), iterative number $\Gamma$, and convergence error $\varepsilon$.
**Output**: Weights: $\{\mathbf{W}\}$.
**Step 1**: **Parameters Initialization**.
**Step 2**: **Optimization by back-prorogation:**
**for** $i = 1, 2, \cdots, \Gamma$ **do**
    Randomly select a batch of $X$.
    // *Forward propagation*
    **for** $1, 2, \cdots, T$ **do**
      | Perform forward propagation.
    **end**
    // *Computing gradients*
    **for** $T, T-1, \cdots, 1$ **do**
      | Perform gradients by back-propagation on $\mathbf{W}$.
    **end**
    // *Back propagation*
    Perform summation for stage-wise gradients.
    **for** $m = 1, 2, \cdots, T$ **do**
      | Update $\{\mathbf{W}\}$ by (6) and (7) by averaging gradients
    **end**
    Calculate $J_i$ using (5).
    If $i > 1$ and $|J_i - J_{i-1}| < \varepsilon$, go to **Return**.
**end**
**Return:** $\{\mathbf{W}\}$.

---

During inference process, we feed a batch of face data to the trained network, and predict the positions of facial landmarks. It is notified that we further propose a CNN, which aims to predict a initial shape for a given face image. Then the face images with

Wang *et al. Appl Inform* (2017) 4:13

Page 6 of 11

the initial shape are taken as inputs to the proposed recurrent regression network to localize the facial landmarks cascaded. The detailed procedure is shown in Fig. 1.

## Experimental results and discussions

### Datasets

To show the effectiveness of our approach, we conducted experiments on the standard benchmarking dataset 300-W (http://ibug.doc.ic.ac.uk/resources/300-W/) including LFPW, HELEN, and IBUG, where 68 landmarks were employed for evaluation. All the face images were collected from the Internet and captured under wild conditions. Specifically, the LFPW dataset consists of 811 training images and 224 testing images. The HELEN dataset contains 2000 training images and 330 testing images. There are 135 images in the IBUG dataset, which are exposed to larger variances of diverse facial expressions, aspect ratios, and partial occlusions. Note that we make the union of the LFPW testing set and HELEN testing set as the common set, and the union of the common set and IBUG samples as the full set. We also dubbed IBUG set as the challenging set.

### Implementational details and evaluation protocols

For each face image to be evaluated, we detected the face bounding box by the DLIB image processing library. Having obtained the cropped face images, we rescaled them in the sizes of $200 \times 200$, $100 \times 100$, and $50 \times 50$, respectively. Moreover, we normalized the ground-truth coordinates of facial landmarks in the range of [0, 1]. For all experiments, our network was trained on the 3148 training images of the LFPW and HELEN datasets. To further improve the performance, we augmented the training samples by adding per-pixel Gaussian noise of $\sigma = 0.5$, flipping and finally with random in-plane rotations $\pm 15°$ from a uniform distribution.

We employed twofolds of evaluation protocols: the averaged error comparisons and cumulative error distribution (CED) curves. The averaged error is computed by averaging the normalized root mean-squared error (NRMSE), which is used to measure the point-to-point distance normalized by the Euclidean distance of pupils of eyes. The CED curve is utilized to qualitatively evaluate the NRMSE errors, which demonstrates the detected image fractions with respect to the specific NRMSE values.

### Experimental results

#### Comparisons with state-of-the-art methods

In our experiments, we compared our methods with 12 state-of-the-art face alignment methods including FPLL (Zhu and Ramanan 2012), DRMF (Asthana et al. 2013), RCPR (Burgos-Artizzu et al. 2013), SDM (Xiong and la Torre 2013), GN-DPM (Tzimiropoulos and Pantic 2014), ESR (Cao et al. 2012), LBF (Ren et al. 2014), ERT (Kazemi and Sullivan 2014), CFSS (Zhu et al. 2015), CFAN (Zhang et al. 2014), BPCPR (Sun et al. 2015) and TCDCN (Zhang et al. 2016). Table 1 tabulates comparisons of the averaged errors of our method to those state-of-the-art methods, where the results were directly cropped from the original papers. From these results, we see that our model achieves significantly superior performance compared to those state-of-the-art methods. Moreover, we carefully implemented DRMF, RCPR, CFAN, TCDCN, and CFSS by following

Wang *et al. Appl Inform* (2017) 4:13

Page 7 of 11

**Table 1 Comparisons of averaged errors of our MSRRN with different face alignment methods on the 300-W dataset, where 68 landmarks were employed for evaluation**

| Method | LFPW | HELEN | CommonSet | ChallengingSet | FullSet |
|---|---|---|---|---|---|
| FPLL (Zhu and Ramanan 2012) | 8.29 | 8.16 | 8.22 | 18.33 | 10.20 |
| DRMF (Asthana et al. 2013) | 6.57 | 6.70 | 6.65 | 19.79 | 9.22 |
| RCPR (Burgos-Artizzu et al. 2013) | 6.56 | 5.93 | 6.18 | 17.26 | 8.35 |
| GN-DPM (Tzimiropoulos and Pantic 2014) | 5.92 | 5.69 | 5.78 | – | – |
| SDM (Xiong and la Torre 2013) | 5.67 | 5.50 | 5.57 | 15.40 | 7.50 |
| CFAN (Zhang et al. 2014) | 5.44 | 5.53 | 5.50 | – | – |
| ERT (Kazemi and Sullivan 2014) | – | – | – | – | 6.40 |
| BPCPR (Sun et al. 2015) | – | – | 5.24 | 16.56 | 7.46 |
| ESR (Cao et al. 2012) | – | – | 5.28 | 17.00 | 7.58 |
| LBF (Ren et al. 2014) | – | – | 4.95 | 11.98 | 6.32 |
| LBF fast (Ren et al. 2014) | – | – | 5.38 | 15.50 | 7.37 |
| Deep reg (Shi et al. 2014) | – | – | 4.51 | 13.80 | 6.31 |
| CFSS (Zhu et al. 2015) | 4.87 | 4.63 | 4.73 | 9.98 | 5.76 |
| CFSS prac (Zhu et al. 2015) | 4.90 | 4.72 | 4.73 | 10.92 | 5.99 |
| TCDCN (Zhang et al. 2016) | 4.57 | 4.60 | 4.80 | 8.60 | 5.54 |
| *MSRRN* | *3.98* | *3.71* | *3.83* | *7.25* | *4.84* |

The italic values denote the highest performance

the implementation details provided in their respective original studies. Figure 3 shows the CED curves of our method compared with the state-of-the-art methods. According to these results, we see that our method obtains very competitive performance on face alignment, which shows the effectiveness of the proposed method. In particular, our model achieves better performance compared with the handcrafted features, which shows the discriminativeness of the learned features from image pixels (Fig. 4). Moreover, our model outperforms the deep learning methods, which shows the advantages of the proposed recurrent architectures and multiscale network. Besides, we show results of some examples for face alignment in Fig. 5 and demonstrate that our proposed method performs in a robust manner with regard to varying facial expressions, aspect ratios, and diverse occlusions under the unconstrained environments.

### Performance effects with different stages

We have also conducted experiments of our MSRRN with different stages. Specifically, we first leveraged the meanshape as the baseline method. Then we compared our model with different stages, where the number of stages were specified as {1, 2, 3}. Table 2 tabulates the averaged error comparisons of our method with different stages and Figure 4 demonstrates the results. According to these results, we see that more stages improves the face alignment performance. The reason is twofold: (1) this cascaded method improves the face alignment performance in a coarse-to-fine manner, which is consistent with the results in cascaded regression methods (Cao et al. 2012; Xiong and la Torre 2013; Zhang et al. 2014; Zhu et al. 2015); and (2) our model takes advantages of multiscale information, which complements multiscale cues for facial landmark localization. Note that three stages setting is enough for the practical applications.
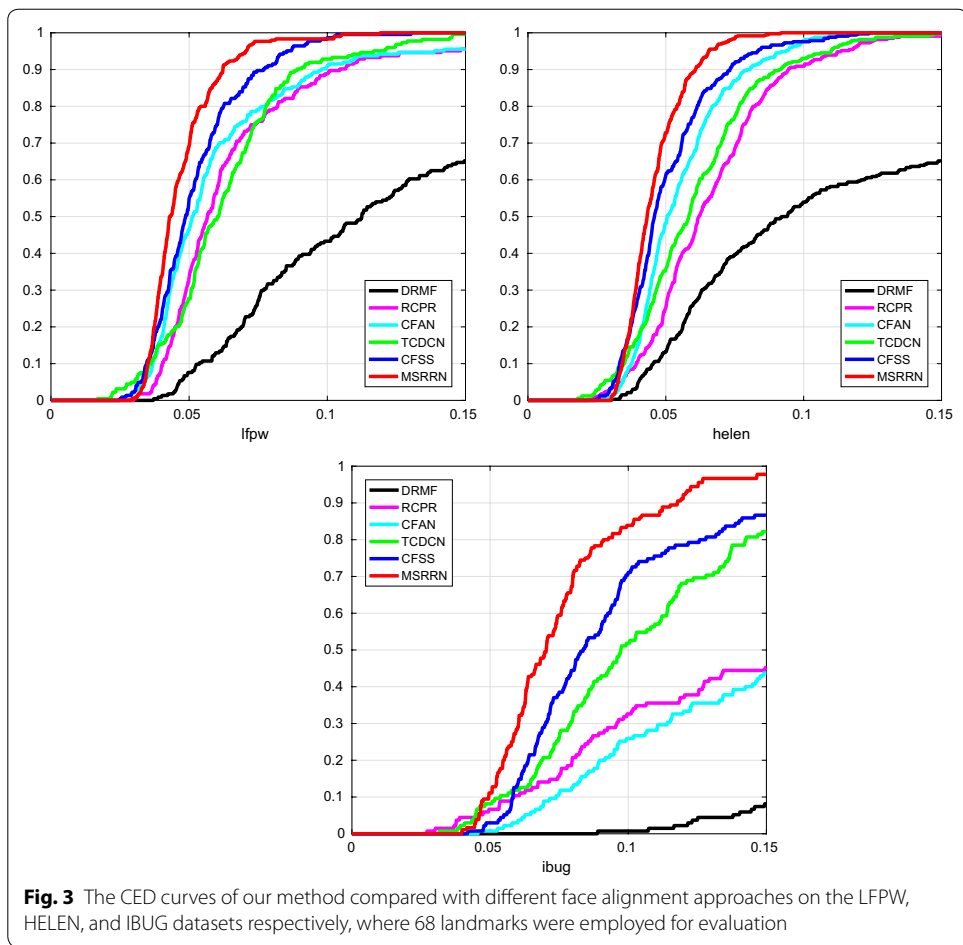
Wang *et al. Appl Inform* (2017) 4:13

Page 8 of 11



**Fig. 3** The CED curves of our method compared with different face alignment approaches on the LFPW, HELEN, and IBUG datasets respectively, where 68 landmarks were employed for evaluation
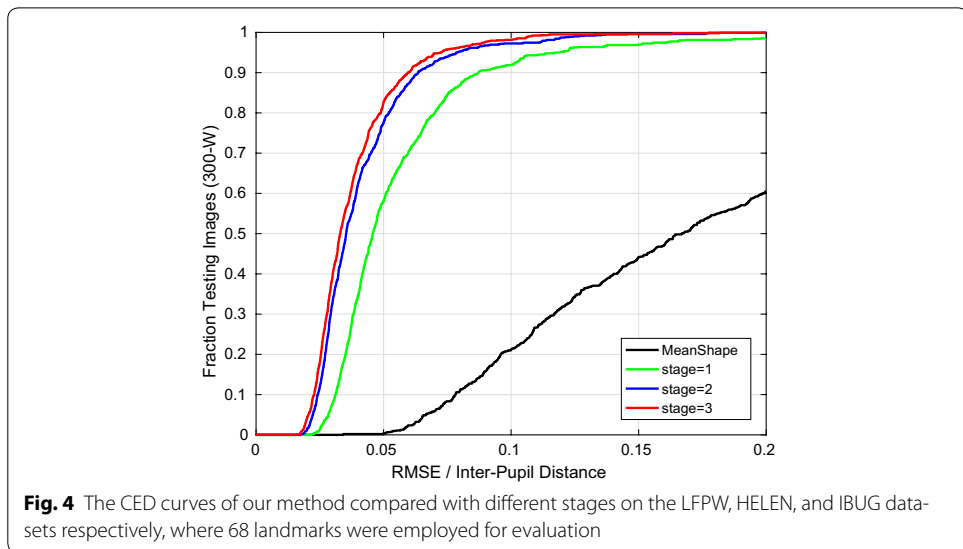


**Fig. 4** The CED curves of our method compared with different stages on the LFPW, HELEN, and IBUG datasets respectively, where 68 landmarks were employed for evaluation

Wang *et al. Appl Inform* (2017) 4:13

Page 9 of 11



**Fig. 5** Results of some examples of our method on the LFPW, HELEN, and IBUG datasets, where 68 landmarks were employed. While the face samples are exposed to diverse facial expressions, aspect ratios, and varying partial occlusions, our method performs in a robust manner with regard to these challenging situations due to the designed features in our multiscale and recurrent network model

**Table 2 Averaged error comparisons and percentages of images (CED) of our model with different stages, where the RMSEs are less than 0.05 and 0.1 of our model on the LFPW, HELEN, and IBUG datasets**

| Datasets | Meanshape | Stage = 1 | Stage = 2 | Stage = 3 |
|---|---|---|---|---|
| 300-W fullSet | 19.52 | 6.64 | 5.16 | *4.84* |

The italic values denote the highest performance

### *Computational time*

Our model was built based on the GPU-accelerated TensorFlow deep learning toolbox, which is convenient for designing and implementation of the directed acycle graph-based network architectures. The computational time required for training procedure is 10 h with a GPU of NVIDIA GTX 1080 graphic computation card for 10,000 iterations. Moreover, we tested our method on GPU during inference process, and it takes 100 frames per second, which satisfies the real-time requirements. We also tested our method with the core-i7 CPU@3.6GHZ platform. Our model runs at 10 images per second on a CPU@3.6GHZ platform.

### Conclusion

In this paper, we have proposed a multiscale recurrent regression network (MSRRN) method for face alignment. Specifically, we have leveraged feedback deep architecture to memorize the descent to pass across consecutive stages via the recurrent neural networks. Moreover, the proposed MSRRN exploits the complementary information from multiscale faces in a coarse-to-fine manner. The network parameters are optimized by the standard back-propagation algorithm. Extensive results on public benchmarking datasets have validated the effectiveness of the network decisions of making full access

Wang *et al. Appl Inform* (2017) 4:13

Page 10 of 11

to multiple scales. It is promising to apply the recurrent architectures to video-based face alignment by incorporating with the dependency information across frames.

### Authors' contributions
The basic idea and the main draft have been accomplished by the first author CW. In the meanwhile, the optimization has been implemented by both CW and HS. The final submission has been revised by JL, JF, and JZ. JL is the corresponding author. All authors read and approved the final manuscript.

### Authors' information
Caixun Wang is a Ph.D candidate and Haomiao Sun is the Bachelor candidate from the Department of Automation, Tsinghua University. Jiwen Lu and Jianjiang Feng are the associate professors, and Jie Zhou is the professor in the Department of Automation, Tsinghua University.

### Author details
[1] Department of Automation, Tsinghua University, Beijing 100084, People's Republic of China. [2] State Key Lab of Intelligent Technologies and Systems, and Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, People's Republic of China. [3] State Key Lab of Intelligent Technologies and Systems, Beijing 100084, People's Republic of China.

### Competing interests
The authors declare that they have no competing interests.

### Availability of data and materials
For the experimental evaluation, we leveraged the standard benchmarking dataset, which can be found on the following website: https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/.

## Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### References
300 faces in-the-wild challenge. http://ibug.doc.ic.ac.uk/resources/300-W/
Asthana A, Zafeiriou S, Cheng S, Pantic M (2013) Robust discriminative response map fitting with constrained local models. In: CVPR, pp 3444–3451
Burgos-Artizzu XP, Perona P, Dollár P (2013) Robust face landmark estimation under occlusion. In: ICCV, pp 1513–1520
Cao X, Wei Y, Wen F, Sun J (2012) Face alignment by explicit shape regression. In: CVPR, pp 2887–2894
Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. CVIU 61(1):38–59
Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. TPAMI 23(6):681–685
Cootes TF, Ionita MC, Lindner C, Sauer P (2012) Robust and accurate shape model fitting using random forest regression voting. In: ECCV, pp 278–291
Duan Y, Lu J, Feng J, Zhou J (2017) Context-aware local binary feature learning for face recognition. PAMI **(in Press)**
Hu J, Lu J, Tan Y (2017) Sharable and individual multi-view metric learning. PAMI **(in Press)**
Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: CVPR, pp 1867–1874
Liu H, Lu J, Feng J, Zhou J (2017a) Learning deep sharable and structural detectors for face alignment. TIP 26(4):1666–1678
Liu H, Lu J, Feng J, Zhou J (2017b) Label-sensitive deep metric learning for facial age estimation. TIFS **(in Press)**
Liu H, Lu J, Feng J, Zhou J (2017c) Two-stream transformer networks for video-based face alignment. PAMI **(in Press)**
Lu J, Liong VE, Zhou J (2017) Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition. PAMI **(in Press)**
Ren S, Cao X, Wei Y, Sun J (2014) Face alignment at 3000 FPS via regressing local binary features. In: CVPR, pp 1685–1692
Shi B, Bai X, Liu W, Wang J (2014) Deep regression for face alignment. arXiv:1409.5230
Sun Y, Wang X, Tang X (2013) Deep convolutional network cascade for facial point detection. In: CVPR, pp 3476–3483
Sun P, Min JK, Xiong G (2015) Globally tuned cascade pose regression via back propagation with application in 2D face pose estimation and heart segmentation in 3D CT images. arXiv:1507.07508

Wang *et al. Appl Inform* (2017) 4:13

Page 11 of 11

Tzimiropoulos G, Pantic M (2014) Gauss-newton deformable part models for face alignment in-the-wild. In: CVPR, pp 1851–1858

Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: CVPR, pp 532–539

Zhang J, Shan S, Kan M, Chen X (2014) Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: ECCV, pp 1–16

Zhang Z, Luo P, Loy CC, Tang X (2016) Learning deep representation for face alignment with auxiliary attributes. PAMI 38(5):918–930

Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: CVPR, pp 2879–2886

Zhu S, Li C, Loy CC, Tang X (2015) Face alignment by coarse-to-fine shape searching. In: CVPR, pp 4998–5006