Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks

Shengtao Xiao[1]([✉]), Jiashi Feng[1], Junliang Xing[3], Hanjiang Lai[1],
Shuicheng Yan[1,2], and Ashraf Kassim[1]

[1] Department of Electrical and Computer Engineering,
National University of Singapore, Singapore, Singapore
`xiao_shengtao@u.nus.edu`, {`elefjia,eleyans,ashraf`}`@nus.edu.sg`,
`laihanj3@mail.sysu.edu.cn`
[2] Artificial Intelligence Institute, 360, Beijing, China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
`jlxing@nlpr.ia.ac.cn`

Abstract. In this work, we introduce a novel **R**ecurrent **A**ttentive-**R**efinement (RAR) network for facial landmark detection under unconstrained conditions, suffering from challenges like facial occlusions and/or pose variations. RAR follows the pipeline of cascaded regressions that refines landmark locations progressively. However, instead of updating all the landmark locations together, RAR refines the landmark locations sequentially at each recurrent stage. In this way, more reliable landmark points are refined earlier and help to infer locations of other challenging landmarks that may stay with occlusions and/or extreme poses. RAR can thus effectively control detection errors from those challenging landmarks and improve overall performance even in presence of heavy occlusions and/or extreme conditions. To determine the sequence of landmarks, RAR employs an attentive-refinement mechanism. The attention LSTM (A-LSTM) and refinement LSTM (R-LSTM) models are introduced in RAR. At each recurrent stage, A-LSTM implicitly identifies a reliable landmark as the attention center. Following the sequence of attention centers, R-LSTM sequentially refines the landmarks near or correlated with the attention centers and provides ultimate detection results finally. To further enhance algorithmic robustness, instead of using mean shape for initialization, RAR adaptively determines the initialization by selecting from a pool of shape centers clustered from all training shapes. As an end-to-end trainable model, RAR demonstrates superior performance in detecting challenging landmarks in comprehensive experiments and it also establishes new state-of-the-arts on the 300-W, COFW and AFLW benchmark datasets.

Keywords: Facial landmark detection · Occlusion · Face alignment · Recurrent neural network

1   Introduction

In facial landmark detection, a set of pre-defined key points on a human face are automatically localized to solve various face analysis problems from face recognition [1] and face morphing [2,3] to 3D face modelling [4]. Among recent research efforts to develop more accurate models for localizing facial landmark points under *unconstrained* conditions [5–12], cascaded regression based approaches [8–12] have demonstrated state-of-the-art performance in both efficiency and accuracy, even in challenging scenarios.

Cascaded regression methods progressively refine landmark detections through multiple cascading stages beginning with the extraction of visual features from current predicted landmarks that are used to update estimates of the face shape[1], which gives rise to new landmarks that are fed into the next stage as inputs. In this way, landmark detection is progressively refined until convergence. As the performance of these cascaded regression methods heavily depends on the quality of the initial locations of landmarks as well as the visual features, recent efforts have focused on enhancing robustness of detection methods e.g., smart restarts [13] and coarse-to-fine searching [12,14].
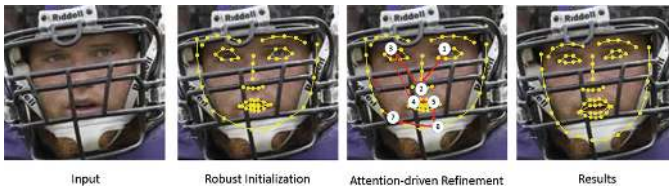

Fig. 1. Illustration of our proposed Recurrent Attention-Refinement (RAR) network. Given an input face image, our model first produces a robust initial estimate of the face shape specified by landmarks. RAR identifies a proper sequence of attention centers which steer the refinement process and make the result robust to challenging conditions.

Recently, deep learning methods [15–18] have been successfully applied to learn discriminative features for face analysis and demonstrated good performance in detecting landmarks under moderate conditions. However, their performance is still "fragile" under extreme scenarios such as severe occlusion or large pose variations.

In this work, we propose a novel recurrent neural network-based facial landmark detection model, called recurrent Attentive-Refinement network (RAR), to work under unconstrained conditions. RAR follows a pipeline similar to cascaded regression methods that refines landmark detection results progressively via multi-stage predictions. However, while existing cascading methods update

---

[1] The face shape depicts global spatial configuration of all the landmark points for a face. Throughout the paper, we use shape to denote the collection of all the landmarks.

all landmark locations concurrently and globally, RAR refines landmark locations in a sequential manner at each recurrent stage as illustrated in Fig. 1.

Given an input face image, to obtain a good initial estimate for landmark locations, RAR employs a robust initialization strategy that refines a preliminary landmark detection result by fitting it to a population prior on human face shapes. Then, at each recurrent stage, RAR adopts a sequential decision making policy to update the landmark points. Reliable information is collected from earlier landmarks in the sequence which is then used to help detect other challenging landmarks selected later. To automatically identify the sequence of landmarks and refine them progressively, RAR employs two LSTM based components – an attention LSTM (A-LSTM) and a refinement LSTM (R-LSTM) – that work collaboratively. At each recurrent stage, A-LSTM selects one landmark point with highest reliability as an *attention center*[2] and R-LSTM refines those landmarks that are close to the attention center. In this way, reliable information from the attention center is communicated to other landmarks to better refine their locations. Landmark points that are occluded or noisy will be selected by A-LSTM very late, and so their impact is effectively alleviated. Finally, context information provided by other landmarks enables the challenging landmarks to be also detected accurately. Therefore, RAR can provide accurate landmark detection results even in presence of heavy occlusion or other extreme conditions. This sequential detection procedure adopted by RAR is similar to the process how people annotate landmarks of a face image: "easy" landmarks with strong discriminative visual features are usually annotated first and "difficult" landmarks are annotated later with the reference from earlier annotated landmarks.

The main contributions of this paper can be summarized as follows:

- We propose to reform the regression-based face landmark detection in a sequential manner which is more robust to extreme face conditions;
- We present a recurrent attentive-refinement network to realize our sequential formulation which seamlessly incorporates an attention LSTM and a refinement LSTM to perform robust face landmark detection;
- We also develop a robust method to estimate the initial facial shapes which works well even under very challenging conditions;
- Our framework provides new state-of-the-art performance on 300-W, COFW and AFLW sets and significantly outperforms all existing methods.

2   Related Work

2.1   Regression Based Face Feature Points Detection

Regression based face landmark detection models [9,12,19,20] directly learn a mapping function from the feature space to the shape space. To improve accuracy, the shape indexed features are often employed [10] and the regression

---

[2] This name is inspired by the process how humans annotate facial landmarks manually: one prefers to annotate the most clear and reliable landmark points first and then infer the position of other landmark points according to overall face shape.

process is often implemented in a cascade manner that learns a series of projection functions to iteratively update the positions. The face shape output at convergence is then regarded as the landmark detection result. Denote the face shape represented by L landmarks as the $S \in \mathbb{R}^{L \times 2}$, and the regression process can generally be formulated as

$$\hat{S} = \lim_{\Delta S_t \to 0} S_t = \lim_{\Delta S_t \to 0} \{S_{t-1} + \Delta S_t\} = \lim_{\Delta S_t \to 0} \{S_{t-1} + f(\Phi(I, S_{t-1}))\}, \quad (1)$$

where $\Phi(I, S_t)$ is the shape indexed feature extractor and f is the regression function, which is usually modelled through a linear projection process, *i.e.*, $\Delta S_t = f(\Phi(I, S_t)) = W_t \Phi(I, S_t)$. Here W_t is the projection matrix which needs to be learned as the model parameters. Given a training set $\{I_n, S_n^\star\}_{n=1}^{N}$ with N samples, each of which consists of a face image I_n and an annotated true face shape $S_n^\star$, the optimal projection matrix can be obtained by minimizing the following objective function:

$$\min_{W_t} \sum_{n=1}^{N} \|S_n^\star - \hat{S}_{t,n} - W_t \Phi(I_n, \hat{S}_{t,n})\|_2^2. \quad (2)$$

To improve the effectiveness of the learned model, some regularizations can be imposed on the model parameters to avoid over-fitting [12,19] and more complex non-linear mapping functions have also been employed [21].

2.2   Recurrent Neural Network

Recurrent neural network (RNN) has drawn great interests from researchers in the field of computer vision recently. Long short term memory (LSTM) [22] is a typical recurrent neural network which has achieved great success in many sequential data analysis applications, [23,24]. The computation within an LSTM can be described as follows:

$$f_{t+1} = \sigma(W_f \cdot [C_t, h_t, \Phi_t] + b_f), \quad (3)$$
$$i_{t+1} = \sigma(W_i \cdot [C_t, h_t, \Phi_t] + b_i), \quad (4)$$
$$o_{t+1} = \sigma(W_o \cdot [C_t, h_t, \Phi_t] + b_o), \quad (5)$$
$$\tilde{C}_{t+1} = \tanh(W_C \cdot [h_t, \Phi_t] + b_C), \quad (6)$$

where C_t, h_t and Φ_t are the inputs to the LSTM. Ws and bs are model parameters. σ is the sigmoid activation function. f, i, o are the forgetting, input and output gates of a standard LSTM unit [22] which control the contribution of historical information to current decision. The outputs of an LTSM are

$$C_{t+1} = f_{t+1} C_t + i_{t+1} \tilde{C}_{t+1}, \quad (7)$$
$$h_{t+1} = o_{t+1} \tanh(C_{t+1}). \quad (8)$$

For clarity, we denote the output of LSTM by $h_{t+1} = \text{LSTM}(\Phi_t)$ with Φ_t being the only external signal that is passed into the LSTM.

3   Recurrent Attentive-Refinement Network for Landmark Detection

3.1   Overview of RAR Network

We first provide an overview on the framework of our proposed RAR network in Fig. 2, before introducing each of its components in details. As shown in the figure, our proposed model first directly predicts the locations of all landmarks via a convolutional neural network (CNN). We develop a robust initialization module to alleviate the interference of noisy detection from conv8 and ensures a good starting face shape for the following regression task.

We then extract shape-indexed features [17] from convolutional layers. After that, these features along with the initial landmark estimation are fed into the recurrent attentive-refinement network for progressively updating the landmarks. At each recurrent step, two LSTM units are employed. The first one is an Attention LSTM (A-LSTM) that determines which region to be updated first by selecting an attention center among existing feature points, according to the current global features and memory information. Then, starting with the selected attention center, landmarks around the center will be refined with high priority by an Refinement LSTM (R-LSTM). Other landmarks can also be fine tuned once an attention center close to them is selected. Repeating the attentive-refinement
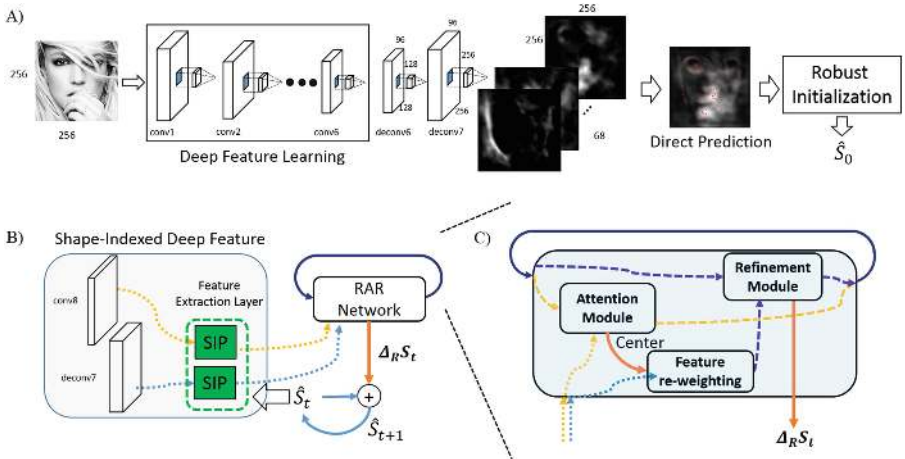

Fig. 2. The proposed framework for facial landmarks detection. (A) Deep convolutional neural network is employed to perform softmax regression to the landmark locations. A robust initialization module is introduced to select a good initial shape for further refinement. (B) Recurrent attentive-refinement network (RAR) takes shape-indexed deep features and past information as inputs and recurrently revises the landmark locations. (C) Within the RAR unit, an attention module generates an attention center at each step and re-weights regression features to encourage landmarks around the attention center to be primarily refined.

process for several times until convergence gives the final landmark detection results. We now proceed to explain each component in details.

3.2   Robust Initialization

The quality of initial landmark estimation is critical for final performance of the cascaded regression methods. Most of previous methods use an average face shape learned from the training set as the initial estimation. This may fail the regression model when processing faces with large pose and expression variations.

To get a good initial estimation of the face shape, we first design a deep CNN model inspired by [16,17] to generate detection results of all landmarks. However, detection of these landmark is often very sensitive to occlusion and it will contaminate the following shape regression steps. We therefore propose a more robust face shape initialization based on the detection results.

Intuitively, the initialized face shape should meet the following two considerations: (1) the shape should be like a human face, or in other words, the shape should satisfy a global configuration constraint on the landmarks; and (2) the initial shape should not be far away from the one detected by CNN on the raw face image, which is denoted as S_d for ease of illustration. Denote the face shape vector encoded by L landmark locations as $S = [x_1, y_1; \ldots; x_L, y_L] \in \mathbb{R}^{L \times 2}$. Based on the above two criteria, the process of looking for a good initial shape S_0 can be formulated as

$$S_0 = \operatorname*{argmin}_{S} \|S - S_d\|, \text{ s.t. } S \in \mathcal{F}, \tag{9}$$

where $\|\cdot\|$ denotes the adopted distance metric and $\mathcal{F}$ is the space of all possible face shapes.

Searching for the solution within $\mathcal{F}$ is not easy, as $\mathcal{F}$ itself is difficult to model. Fortunately, when sufficient training face images with accurate shape annotations are provided, we can take them as basis to span the space $\mathcal{F}$. Formally, given a set of shapes from m training faces, $\{S_1, \ldots, S_m\}$, any shape $S \in \mathcal{F}$ can be represented as $S = \sum_{i=1}^{m} \beta_i S_i$. The initial face shape S_0 can be estimated via

$$S_0 = \operatorname*{argmin}_{S, c_i} \|S - S_d\|, \text{ s.t. } S = \sum_{i=1}^{m} c_i S_i. \tag{10}$$

In the above formulation, both S_d and S_i could be noisy. Some landmarks in S_d may be corrupted severely due to occlusion and some sample may be wrongly labelled. We therefore further enhance the above objective by introducing the ℓ_0-induced distance metric and regularization:

$$S_0 = \operatorname*{argmin}_{S, \mathbf{c} \doteq [c_i]} \|S - S_d\|_0 + \lambda \|\mathbf{c}\|_0, \text{ s.t. } S = \sum_{i=1}^{m} c_i S_i. \tag{11}$$

The above function is our final objective for robust face shape initialization. Finding its global optimum is very time consuming due to the involved ℓ_0 norm.
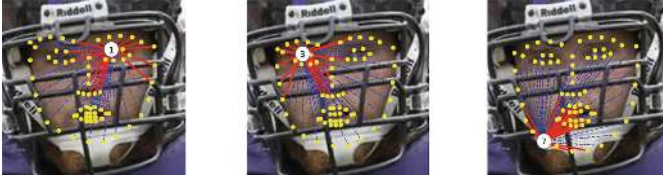
Fig. 3. This figure depicts how an attention center steers refinement of landmarks at different stages. A-LSTM selects a suitable landmark as the attention center at a recurrent step. Landmarks close (connected with red solid lines) to the attention center will be to refined significantly. Those landmarks distant (connected with green dot lines) from the attention center will be slightly refined. (Color figure online)

To ease optimization, we introduce following two simple yet effective heuristics. First, reduce the size of the problem. When m is large, the problem is extremely hard to optimize. Therefore, we first apply K-means clustering on the shapes $S_1, \ldots, S_m$ to get K representative shapes $\{\bar{S}_1, \ldots, \bar{S}_K\}$ and use these K shapes as the basis of $\mathcal{F}$. Thus the problem size is reduced from m to K. Secondly, we adopt a RANSAC flavor method to filter out significant outliers in S_d and sample some basis to evaluate the objective to find better initial shapes. The obtained face shape with the best objective value is used as the initial face shape in the following regression process.

3.3 Attention LSTM for Sequential Attention-Center Selection

Ideally, A-LSTM selects the most reliable landmark point as an attention center first. Then it proceeds to find less reliable landmarks and finally addresses the noisy landmarks (*e.g.*, occluded ones or the ones lying in the face regions with extreme illumination condition). As shown in Fig. 3, at each recurrent stage, A-LSTM selects an attention center. Locations of landmarks close to the attention center will be primarily updated at the current recurrent step and those far away from the center are slightly refined. Compared with updating all the landmark points simultaneously, treating different landmarks separately in a proper sequence can effectively alleviate the contamination from noisy landmark points and reduce the accumulative errors in the recurrent process.

A-LSTM determines which landmark points to be selected for the current step using a confidence driven strategy. By taking the features of all the landmark points and history of selections as inputs, A-LSTM estimates the confidence scores (or reliability) of all the landmark points first. The landmark having the maximal confidence score at the current step is then selected as the current attention center, $c^* \in \{1, \ldots, L\}$. This process is formally written as

$$c^* = \underset{c \in \{1, \ldots, L\}}{\arg\max} \ \text{A–LSTM} \left(\Phi(I_t, \hat{S}_t); W_a, c \right), \tag{12}$$

where the operator $\Phi(\cdot, \cdot)$ extracts shape-indexed features according to current predicted shape $\hat{S}_t$ and A-LSTM outputs L confidence scores for the landmark points, based on its input feature and parameter W_a.

Training of A-LSTM. A-LSTM aims to find a suitable selection sequence of landmarks such that the following long term attention center selection reward can be maximized:

$$\mathcal{R}_a = \sum_{t=1}^{\infty} \eta^{t-1} R(\hat{S}_{t-1}, \hat{S}_t), \tag{13}$$

where $\eta < 1$ is the discount factor and t indexes the recurrent steps. Here $R(\hat{S}_{t-1}, \hat{S}_t)$ is the intermediate reward measuring how much improvement brought by updating the shape estimate from the $\hat{S}_{t-1}$ to $\hat{S}_t$ and it is defined as

$$R(\hat{S}_{t-1}, \hat{S}_t) = \|\Gamma_t \Delta S_{t-1}\|_2^2 - \|\Gamma_t \Delta S_t\|_2^2, \tag{14}$$

with $\Delta S_t = S^\star - \hat{S}_t$ as the offset of current shape estimate from the ground truth $S^\star$. $\Gamma_t \in \mathbb{R}^L$ is the distance-based coefficient vector which re-weights each landmark point in the offset calculation in proportion to their distance from the attention center landmark $\hat{S}_t^{c^*}$ (recall c^* is attention center landmark index):

$$\Gamma_t = [\gamma_t^1, \gamma_t^2, \ldots, \gamma_t^L] \text{ with } \gamma_t^l = \kappa \exp(-\|\hat{S}_t^l - \hat{S}_t^{c^*}\|_{\ell_2}^2/(2D_t)^2), \tag{15}$$

where D_t is the inter-ocular distance based on the shape estimate $\hat{S}_t$ and $\kappa = 1/\sum_{l=1}^{L} \gamma_t^l$ is a normalization factor. Here $2D_t$ gives an estimation of the width of the face bounding box.

Training A-LSTM to maximize the long-term award $\mathcal{R}_a$ encourages the A-LSTM to make a sequence of decisions on the landmark selection such that the selected attention center would have positive impact on the overall landmark detection in the future. Here for light notations, we hide the sample index $n \in \{1 \ldots N\}$ and this notation is used throughout the entire section.

3.4 R-LSTM for Attention-Center-Driven Shape Refinement

Once A-LSTM selects one attention center landmark, the refinement component will focus on refining landmarks around the attention center. We adopt a second LSTM model to perform refinement, which is called Refinement LSTM (R-LSTM). R-LSTM will suppress refinement of landmarks far away from the attention center as their correlation to attention center is small. Thus, at each recurrent step, only a limited number of landmarks are updated significantly and the rest are slightly updated. Given the attention center from A-LSTM, we first extract attention-center aware global feature for current shape $\hat{S}_t$:

$$\Phi(I_t, \hat{S}_t) = [\gamma_t^1 \phi_t^1, \gamma_t^2 \phi_t^2, \ldots, \gamma_t^L \phi_t^L], \tag{16}$$

where γ_t^l for $l = 1, \ldots, L$ is the distance-based weighting coefficient for the l-th landmark whose computation is given in Eq. (15). The ϕ_t^l represents a shape-indexed feature extracted around the l-th landmark from the shape $\hat{S}_t$. R-LSTM takes the features and generates offset shape for update.

Training of R-LSTM. The parameters of R-LSTM are optimized through minimizing the following loss:

$$\mathcal{L}_R^t = \|\Gamma_t(\Delta_R S_t - \Delta S_t)\|_{\ell_2}^2, \text{ with } \Delta_R S_t = \alpha \Gamma_t \text{R–LSTM}\left(\Phi(\hat{S}_t)\right) \qquad (17)$$

where $\Delta S_t = S_t^\star - \hat{S}_t$ is the offset from the ground truth. R-LSTM predicts an offset $\Delta_R S_t$ specifying where the shape should be updated towards. We use a fixed scaling factor $\alpha = 128$ to rectify the outputs of R-LSTM, considering the dimension of images is 256×256 and the magnitude of R-LSTM falls in a small range of $(-1, 1)$. Without scaling, R-LSTM only provides negligible shape update at each step. We observe that the scaling factor can significantly accelerate the convergence rate for training R-LSTM. In the loss, Γ_t further ensures that RAR to focus on refining landmarks around the attention center at a certain step.

3.5   Training and Testing Strategies

Considering costs from both attention center selection and refinement, the overall cost to be optimized for training RAR is

$$\sum_{t=1}^{T} \sum_{n=1}^{N} -\gamma^{t-1} \mathcal{R}_a(\hat{S}_{t-1,n}, \hat{S}_{t,n}) + \mathcal{L}_{R,n}^t, \qquad (18)$$

where T is a pre-defined number of recurrent steps which also serves as an early-stop regularization and N is the number of training samples.

This overall objective function can be optimized in an end-to-end manner by applying the standard error back propagation method. Filters of the convolutional layers are tuned not only by the softmax regression loss from conv8 when performing direct landmark prediction but also the overall shape regression loss in Eq. (18). This ensures the learned features are much more informative for landmark detection compared with hand-crafted features, e.g. SIFT and HOG.

At the testing stage, a face image is first passed through the CNN for feature extraction. Landmark locations estimated via conv8 in the CNN, S_d, are then used to search for a good initial shape $\hat{S}_0$ as described in Sect. 3.2. After that, $\hat{S}_0$ is fed into the RAR and updated recurrently as follows:

$$\hat{S}_{t+1} = \hat{S}_t + \Gamma_t \Delta_R S_t \qquad (19)$$

where $\Delta_R S_t$ and Γ_t are the predicted offset and the distance-based weighting vector as given in Sect. 3.3.

4   Experiments

4.1   Implementation Details

Configuration. Our model is developed with the open source platform Caffe [25]. All the images including both training and testing ones are cropped

according to provided bounding boxes and scaled to 256×256 pixels. Note that in testing, before evaluation we project the detected landmark locations on the 256×256 image back to the images of the original size, in order to avoid the possible truncation error due to image scaling. We empirically set the number of recurrent regression stages as $T = 15$ as we do not observe any substantial performance enhancement by further increasing the number of recurrent steps. Our model is trained via standard stochastic gradient descent method with a momentum of 0.9, a mini-batch of 2 images and a weight decay parameter of 0.0001. The weights of LSTM are randomly initialized with a uniform distribution of $[-0.1, 0.1]$. Relevant layers in our model are initialized using the pre-trained VGG-19 model provided in [26]. All experiments are conducted using one Nvidia Titan-Z GPU. During test, it takes about 250 ms for our model to process a 256×256 face image.

Data Augmentation. Our RAR is trained on 300-W [27] training set which consists of 3,148 face images. We also generate training samples with occlusions incurred by natural objects, *e.g.*, sunglasses, medical masks, phones, hands, and cups, on the original 300-W images to introduce more occluded samples. Training samples are further augmented by rotation, scaling and mirroring. Note that in all the baselines we compare with data augmentation is also performed in different ways. In [9,19], augmentation is performed by introducing bounding box disturbances and random scaling/rotatoin to the original face images. In [28], the authors generate occluded face images with synthesized plausible coherent occlusion patterns to train an occlusion-aware model.

4.2 Benchmark Datasets

We evaluate our model on 300-W [27], Caltech Occluded Face in the Wild (COFW) [13] and Annotated Facial Landmarks in the Wild (AFLW) [29]. 300-W is a standard benchmark for facial landmark detection. The COFW consists of a large number of occluded face images. AFLW is another benchmark which contains face images with large pose variations and heavy partial occlusion.

300-W, COFW and AFLW are annotated with 68, 29 and 21 landmarks respectively. To evaluate our model on COFW, we follow the steps mentioned in [28]. We also evaluate our model for detecting five key landmark points, *i.e.* eye centers, mouth corners and nose tip, on the AFLW benchmark. This follows exactly the same settings as stated in [18]. Common evaluation metric is used, *i.e.* mean error normalized by inter-ocular distance [13,19,20].

We compare performance of our model with results from recent publications. For 300-W and AFLW, cascaded regression-based models ESR [8], SDM [9], RCPR [13], LBF [19], CFSS [12] showed great performance improvement on the benchmark over the past years. Deep learning-based methods CFAN [14] and TCDCN [18] showed slightly better performance as compared to those regression-based methods. We compare our performance on COFW with recently published algorithms RCPR, HPM [28], and RPP [30] which are designed to handle occlusion. We further compare our results with those mentioned methods on AFLW.

4.3   Results

Results on 300-W. We report the landmark detection results of our proposed model as well as results of current state-of-the-art methods on the 300-W testing set. The results are listed in Table 1. From the table, one can observe that our proposed model significantly outperforms the state-of-the-art, TCDCN [18]. Our model has improved on it for more than 10 % on the full set and 14 % on the common set. Note that TCDCN pre-trained their facial landmark detection model on the Multi-Attribute Facial Landmark database (MAFL) [18] which consists of 19,000 different face images with multiple facial attributes information and tuned their model on 300-W. On the other hand, our model is trained only on about 3,148 original face images from 300-W training set. Compared with the best ever reported regression-based method, *i.e.* CFSS [12], our model brings error reduction up to 16.3 % and 12.9 % on the challenging and common set.

Table 1. Landmark detection results on different subsets of the 300-W dataset.

Methods	Helen	LFPW	Common set	Challenging set	Full set
RCPR [13]	5.67	5.93	6.18	17.26	8.35
SDM [9]	–	–	5.57	15.40	7.50
CFAN [14]	5.44	5.53	5.50	–	–
LBF [19]	–	–	4.95	11.98	6.32
CFSS [12]	4.87	4.63	4.73	9.98	5.76
TCDCN [18]	–	–	4.8	8.6	5.54
RAR	**4.30**	**3.99**	**4.12**	**8.35**	**4.94**

Results on COFW. Table 2 shows the results of our model and baselines on the COFW dataset. It can be seen that our model outperforms all reported results on this dataset. In particular, one model gives 19.2 % performance improvement over the state-of-the-art [28]. We also report failure rates of the compared methods on this dataset in Table 2. One can observe that our model reduces the failure rate dramatically. For example, compared with the best baseline HPM, our model reduces the failure rate from 13.24 % to 4.14 %. Small failure rate also indicates the robustness of our framework to various occlusions from the dataset.

We also visualize some example detection results on COFW in top row of Fig. 4. From the examples, one can observe that our model can accurately detect the landmark points even for faces with heavy occlusion. The results clearly demonstrate the strong robustness of our model to occlusion and other extreme conditions, benefiting from the built-in attention and sequential selection model.

Results on AFLW. Table 3 shows the results of our model and baselines on the AFLW dataset. The proposed model outperformed all existing methods for at least 5 % which further verifies our model's robustness on datasets with large poses and occlusion.

Table 2. Mean error on COFW

Methods	Normalized ME	Failure rate
RCPR [13]	8.50	20.00 %
HPM [28]	7.46	13.24 %
RPP [30]	7.52	16.2 %
TCDCN [18]	8.05	-
RAR	**6.03**	**4.14 %**

Table 3. Mean error on AFLW

| Methods	Normalized ME
RCPR [13]	11.6
SDM [9]	8.5
CFAN [14]	10.94
TCDCN [18]	7.6
RAR	**7.23**

4.4   Discussion

Attention Selection and Shape Updating. It is interesting to look into how the proposed A-LSTM selects attention centers at different stages for different faces. Table 4 visualizes the frequency of different landmarks being selected as the attention center. From the results, one can observe that at the early recurrent stages, *i.e.*, S1 to S5, the A-LSTM tends to more often select landmarks from the face centers with strong discriminative features, *e.g.*, the ones on eyebrow, mouth and nose tip. Indeed, this policy — localizing central landmarks first — is essentially useful when the initial shape is not good. Global shape refinement at early stages can significantly improve the detection performance and selecting attention centers around the center of a face can help refine all the landmarks. In contrast, as shown in Table 4, the A-LSTM usually selects landmarks on the face contour at very late stages such as S11 to S15. This is reasonable as landmarks on the face contour are difficult to annotate due to their weak discriminative features and should be inferred with help from other points.

We also perform ablation studies on the effectiveness of attention LSTM and sequential selection on landmarks. In the experiments, we set the parameter γ_t^l in Eq. (16) to be 1 for all possible attention centers. By doing so, the impact of selecting attention center via A-LSTM is actually disabled as the features and training objectives are independent of the selected center now. Then we train the "attentionless" model under the same setting as above and its normalized mean error on 300-W and COFW is 5.02 and 6.11 respectively. The results are worse than the ones given by the RAR. This verifies the essential role of the attention center in the landmark prediction process. Sample images from last two columns of Fig. 4 also indicate that our model can perform better in detecting fine-grained landmarks. Since the RAR explicitly selects region of interest to refine at each step, an occluded area can be focused at certain time step and landmarks within the area will be carefully refined. However, without the attention mechanism, refinement is performed globally at every step and landmarks heavily occluded can hardly be explicitly refined.

Approaches for Estimating the Initial Shapes. Recent regression-based methods usually use mean shape [9,19] or multiple random shapes [8,13] as an initial estimate of the shapes. However, those methods hardly prevent the

Table 4. Attention center selection frequency at different stages. Y-axis represents the mean regional error of all 300-W samples calculated by taking average of weighted errors by Γ_t. The area of the red circle indicates the frequency of that landmark being selected as an attention center. Landmarks with top-10 frequencies are shown.

Stages	Center Selection Frequency VS Regional Error
S1-S5	
S6-S10	
S11-S15	
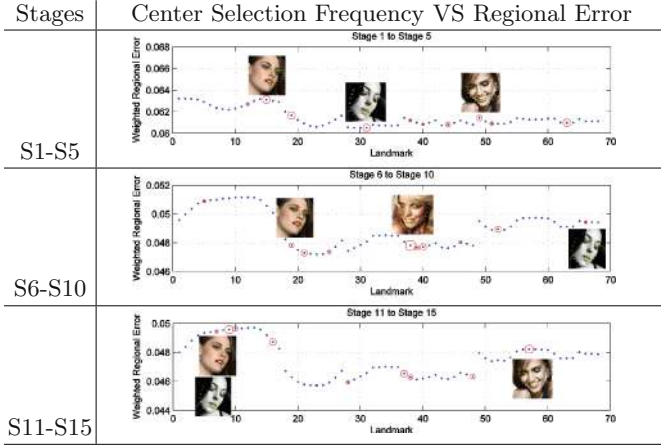
regressed shape from being trapped at a local optimum if the face pose is large. In contrast, our model directly estimates the initial shape with a softmax regression layer (*i.e.*, the Conv8 layer) and selects a good initial shape based on proposed robust initialization scheme (Sect. 3.2). This approach provides a good initial shape closer to the ground truth compared with conventional shape initialization methods, which offers a solid foundation for further shape refinement. This part investigates how the robust initialization strategy contributes to the final performance. Table 5 shows the results of four different initialization strategies including directly applying regression on the output of the conv8 layer (denoted as "direct" in the table), using mean shape and random shape as well as our proposed robust one. We also compare them with the "baseline" results that are directly output by the conv8 layer, From the results, one can observe the conv8

Fig. 4. Testing results on selected samples from the COFW testing set. Images from the top row show results of our full model. Images from the bottom row show results of other models, *i.e.* mean shape initialization(1,2), random initialization(3,4) and direct regression(5,6) and "attentionless" model(7,8).

Table 5. Mean error of RAR with different initial shape estimation approaches.

Dataset	Baseline	Direct	Mean shape	Random shape	Robust
300-W	6.24	6.66	5.26	5.22	**4.94**
COFW	30.14	11.52	6.24	6.12	**6.03**
AFLW	8.14	8.15	7.36	7.42	**7.23**

offers very bad estimation on the COFW and this indicates that direct detection is very sensitive to occlusion. Table 5 also shows directly initializing the face shape gives the worst performance. This verifies our earlier concern that noisy landmarks indeed contaminate the training process and hurt the final results.

Images from the bottom row of Fig. 4 visualize the performance differences. Direct regression can hardly guarantee a normal face shape after recurrent regression. Outlier landmarks from S_d shows direct impact over the final predicted shape. Mean shape and random shape initialization methods are more sensitive to occlusion as compared to the robust initialization method. This is possibly because too much attention is paid to correcting the initial error and occlusion is not specifically considered by the A-LSTM's under this situation (Fig. 5).

Comparison with Canonical Regression Methods. Canonical regression based methods try to optimize the shape regression objective independently at different stages [9,19]. Lacking information shared across consecutive regression stages makes those methods easy to be trapped at a bad local optimum. In contrast, the RAR employs LSTM to memorize all benefiting information from previous stages for both attention center selection and landmark refinement. This leads to superior performance of our model as shown in Tables 1 to 3.

Fig. 5. RAR shows superior results on samples from 300-W challenge set.

5   Conclusion

In this paper, we developed a facial landmark detection framework which is shown to be robust to challenging conditions via the developed recurrent

attentive-refinement network. The framework first directly detects landmarks using a CNN model. The detected landmarks are then used to initialize a good starting shape by alleviating the negative impact of noisy landmarks. Deep shape indexed features are extracted at each regression stage and passed to the A-LSTM module to select attention center at each stage. R-LSTM module then refines landmarks close to the center with high priority. This framework was extensively evaluated on the 300-W, COFW and AFLW datasets and showed significant performance improvements over the state-of-the-arts.

References

1. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. **35**(4), 399–458 (2003)
2. Liu, L., Xing, J., Liu, S., Xu, H., Zhou, X., Yan, S.: Wow! you are so beautiful today!. ACM Trans. Multimedia Comput. Commun. Appl. **11**(1s), 20 (2014)
3. Kemelmacher-Shlizerman, I., Suwajanakorn, S., Seitz, S.M.: Illumination-aware age progression. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3334–3341. IEEE (2014)
4. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. **33**(4), 43 (2014)
5. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. **91**(2), 200–215 (2011)
6. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)
7. Martins, P., Caseiro, R., Batista, J.: Generative face alignment through 2.5 d active appearance models. Comput. Vis. Image Underst. **117**(3), 250–268 (2013)
8. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. Int. J. Comput. Vis. **107**(2), 177–190 (2014)
9. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 532–539. IEEE (2013)
10. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1078–1085. IEEE (2010)
11. Lee, D., Park, H., Yoo, C.D.: Face alignment using cascade gaussian process regression trees. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 4204–4212. IEEE (2015)
12. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: CVPR, pp. 4998–5006. IEEE (2015)
13. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1513–1520. IEEE (2013)

14. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: Proceedings of European Conference on Computer Vision, pp. 1–16 (2014)
15. Luo, P., Wang, X., Tang, X.: Hierarchical face parsing via deep learning. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2480–2487. IEEE (2012)
16. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3476–3483. IEEE (2013)
17. Lai, H., Xiao, S., Cui, Z., Pan, Y., Xu, C., Yan, S.: Deep Cascaded Regression for Face Alignment. ArXiv e-prints, October 2015
18. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1 (2015)
19. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1685–1692. IEEE (2014)
20. Xing, J., Niu, Z., Huang, J., Hu, W., Yan, S.: Towards multi-view and partially-occluded face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1829–1836. IEEE (2014)
21. Sauer, P., Cootes, T.F., Taylor, C.J.: Accurate regression procedures for active appearance models. In: Proceedings of British Machine Vision Conference, pp. 1–11(2011)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
23. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649. IEEE (2013)
24. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
25. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: Proceedings of IEEE International Conference on Computer Vision Workshops. IEEE (2013)
28. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1899–1906. IEEE (2014)
29. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151. IEEE (2011)
30. Yang, H., He, X., Jia, X., Patras, I.: Robust face alignment under occlusion via regional predictive power estimation. IEEE Trans. Image Process. **24**(8), 2393–2403 (2015)