Image and Vision Computing 87 (2019) 37–46

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A coupled encoder–decoder network for joint face detection and landmark localization☆,☆☆

Lezi Wang[a,*], Xiang Yu[b], Thirimachos Bourlai[c], Dimitris N. Metaxas[a]

[a]Rutgers, The State University of New Jersey, USA
[b]NEC Laboratories America, Media Analytics, USA
[c]MILab, LCSEE, West Virginia University, Morgantown, WV, USA

ARTICLE INFO

Article history:
Received 1 October 2017
Received in revised form 16 May 2018
Accepted 12 September 2018
Available online 21 September 2018

Keywords:
Face detection
Landmark localization
Convolutional neural network
Deep learning

ABSTRACT

Face detection and landmark localization have been extensively investigated and are the prerequisite for many face related applications, such as face recognition and 3D face reconstruction. Most existing methods address only one of the two problems. In this paper, we propose a coupled encoder–decoder network to jointly detect faces and localize facial key points. The encoder and decoder generate response maps for facial landmark localization. Moreover, we observe that the intermediate feature maps from the encoder and decoder represent facial regions, which motivates us to build a unified framework for multi-scale cascaded face detection by coupling the feature maps. Experiments on face detection using two public benchmarks show improved results compared to the existing methods. They also demonstrate that face detection as a pre-processing step leads to increased robustness in face recognition. Finally, our experiments show that the landmark localization accuracy is consistently better than the state-of-the-art on three face-in-the-wild databases.

1. Introduction

Face detection has been one of most important and still open problems in Computer Vision and Human–Computer Interaction. Face landmark localization is a prerequisite for many facial analysis applications, such as face recognition, face modeling, and expression transfer. Many effective face detection and landmark localization algorithms have been proposed to close the gap from realistic situations [1,2,11,20,29]. However, face-in-the-wild conditions, such as large pose variation and occlusions, largely degrade the performance of the methods.

Rooting back to the seminal works, Viola–Jones face detector [33] and Active Shape Model [16] have achieved wide applications in their own fields. Many representative works have been proposed to expand and better interpret the above models, such as Deformable Part Models (DPM) [6] for face detection and Active Appearance Models (AAM) [13,17] for landmark localization. However, seldom efforts are put on jointly dealing with the two problems. A well established pipeline is that face landmark localization accepts the bounding boxes input from face detection. However, such rule is not necessarily to hold. Zhu and Ramanan [6] presented a pioneering work to jointly detect faces and facial key points using DPM. The internal correlation between facial key points and the overall face location is well captured by the deformable part model. The method is limited by its hand-crafted feature (HOG) and a predefined tree structure, which is a hard constraint and lacks flexibility in capturing the shape variations.

With the strong power of feature representation, convolutional neural networks (CNN) have shown significant advantages over traditional methods in many fields, i.e., object detection [4] and semantic segmentation [23,24]. Among those methods, the Faster-RCNN [4] demonstrates superior performance across almost all the detection tasks, i.e., ImageNet, PASCAL and KITTI, where the ROI pooling for region proposals is a key factor to achieve fast sampling and high accuracy. In face detection, the cascaded CNN [19] employs multi-scale shallow network cascade to fast localize faces. The early rejection of false alarms on smaller scale speeds up the run-time significantly. The Faster-RCNN relies firmly on the output of region proposal networks (RPN), which needs to carefully design the anchors with different aspect ratios, and to prepare unbiased training data. The cascaded CNN achieves good efficiency but the shallow structure prevents the further improvement of effectiveness.

☆ This work is supported partially by the following grants: AFOSR FA9550-15-1-0054, NSF-11733843, NSF-1703883, and ARO-MURI-68985NSMUR to D. Metaxas, and a project to T. Bourlai and D. Metaxas from the U.S. Department of Homeland Security (DHS) under Grant Award #2015-ST-061-BSH001.

☆☆ This paper has been recommended for acceptance by Hatice Gunes.

* Corresponding author.
E-mail address: lw462@cs.rutgers.edu (L. Wang).

https://doi.org/10.1016/j.imavis.2018.09.008
0262-8856/© 2018 Published by Elsevier B.V.

Sliding window on feature maps is a simple and natural improvement to the RPN, whereas the region classification net (RCN) remains the same. The ROI pooling fully supports the sliding window operation on feature maps. To seek discriminative feature maps, we investigate the well established encoder–decoder framework originated from semantic segmentation [23]. The encoder–decoder generates a set of feature maps in different scales. As we observed, the feature maps not only capture the responses for face landmark localization but also for facial regions, as shown in Fig. 1. The strong face region responses are always coupled with the strong responses of the landmarks. If we consider landmark localization as a sparse segmentation problem (classifying the landmark regions as foreground), the localization task becomes bounding box independent and the feature maps could be re-utilized for face detection.

As a consequence, we propose a novel coupled encoder–decoder network to simultaneously localize landmarks and detect faces. First, an encoder–decoder network followed by a shallow landmark regression network is set up end-to-end, where the feature maps from the intermediate convolutional layers are gathered. Second, ROI pooling is applied on the gathered feature maps in three scales to further extract features for the face region classification. Different from cascaded CNN [19], we apply sliding windows on feature maps instead of raw images. During training, the encoder–decoder and landmark regression are updated for one iteration, while the facial region classification is updated for another iteration. The alternative training is similar to the Faster-RCNN training of RPN and RCN, which achieves stable convergence.

Our contributions are summarized as the following.

- A coupled encoder–decoder structure for joint face detection and landmark localization.
- A carefully designed alternative network training for landmark localization and multi-scale cascaded facial region classification.
- A demonstration of the proposed method's advantages in both the face detection and landmark localization under face-in-the-wild conditions.

2. Related work

2.1. Face detection

Early works on face detection focus on the hand-crafted features and the face classifiers, i.e., the Viola–Jones detector utilizes the Haar feature combined with the Adaboost classifier [33], a vanilla DPM [6] was proposed to defend the model-based methods with top performance and a template based classifier was proposed as well [29]. Different from model-based methods, Shen et al. [50] propose to detect faces by image retrieval. Li et al. [48] further improve it to a boosted exemplar-based face detector. As the development of deep convolutional neural networks (CNN), there are many successful CNN-based methods with much better performance, i.e., the Cascaded CNN [19] applies the cascade of multi-resolution shallow networks to detect faces. Rather than training each cascade stage independently, in [20], the authors propose a joint training framework to learn the cascade model. The Convolutional Channel Feature [38] fully utilizes the rich features from the convolutional layers of different channels. Farfade et al. [30] proposed the multi-view deep neural network based framework to detect faces. Several state-of-the-art methods demonstrate the advantages of deep neural networks. [60] presents a method of end-to-end integration of a ConvNet and a 3D model for face detection in the wild. [31] used a CNN to detect facial parts and combine parts for holistic face detection, Ranjan et al. fuses the DPM with a deep pyramid structure [34]. [57] takes advantage of CNN futures and proposes an effective framework for finding small faces, demonstrating that both large context and scale-variant representations are crucial. [58] introduces the Single Stage Headless (SSH) face detector that, unlike two-stage proposal/classification approaches, detects faces in a single stage.

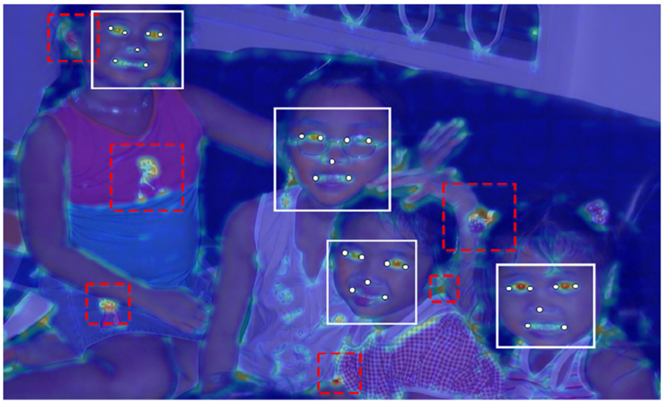

Fig. 1. An example of our coupled encoder–decoder framework result, simultaneously predicting the face regions (white bounding boxes) and the face fiducial points (white dots). The false positive responses for landmark localization are effectively suppressed by the coupled face detection task, in which regions marked by the red dash bounding boxes are classified as non-face.

2.2. Face landmark localization

The model-based methods are back-traced to the Active Shape Models [16] and Active Appearance Models [13,17]. Tons of improvements have been proposed, such as Constrained Local Model [14,59], probabilistic matching [15], and DPM [6]. The regression-based landmark localization [3,5,7,8,10,12,26,54] significantly improve the performance and run-time. In [54], multiple cascaded regressors are proposed with the capability to handle global shape variation and irregular appearance-shape relation. Those regression-based methods directly regress landmark locations from the features, reducing the complexity of model update. However, regression-based method is sensitive to the initial bounding boxes and is feature-specific where regression embedding firmly relies on the feature representations. Recently, the CNN-based approaches show more compelling performance than regression. [1] proposes to use three stages of neural networks to cooperatively localize facial landmarks. [9] applies coarse-to-fine auto-encoders for the regression of landmark positions. In [53], a lightweight and compact CNN architecture is designed for landmark localization. [55] introduces a Recurrent Attentive Refinement network for facial landmark regression where landmark locations are refined progressively. In [2], the multi-task training strongly suggests joint dealing with multiple jobs while boosting each task's performance. [11] proposes the shape basis network to fast approach the global optimal and point transformer network to refine the local shape variations. Compared to the efforts which explicitly import cascade structures, we propose a unified encoder–decoder model to incorporate both face detection and landmark localization, which boosts the learning convergence of the feature maps and the localization accuracy.

2.3. Joint face detection and landmark localization

As the first work that jointly handles face detection, landmark localization, Zhu and Ramanan [6] proposed a DPM based framework and achieved promising results in face-in-the-wild conditions. The similar structure is applied by [36] to detect and localize faces under occlusion. [36] claims to simultaneously achieve the two tasks as well. However, it is more of a joint framework of using DPM to achieve face detection and landmark localization, which lacks consideration of the interactive boosting between the two tasks. The cascaded face detection and alignment [32] jointly deals with the two tasks, where it actually regresses the landmark positions after the face detection. Under the multi-task learning frameworks, several CNN-based methods are recently proposed, i.e. [37] applied a cascaded CNN for multi-task learning, [35] integrated many tasks, face detection, landmark localization, pose estimation and gender recognition.

2.4. Encoder–decoder networks

The encoder and decoder networks are well studied in machine translation [21], where the encoder learns intermediate representation and the decoder transforms that representation. It is intensively investigated in speech recognition [22] and computer vision [23,24,52,56]. In [56], the encoder–decoder architecture is applied to estimate human pose. In [23], authors applied an encoder–decoder structure on the semantic segmentation. The proposed algorithm in [23] mitigates the limitations of the previous methods based on fully convolutional networks by integrating deconvolutional network and pixel-wise prediction, which identifies detailed structures and handles objects in multiple scales. In this work, we employ the encoder–decoder network to learn the discriminative features for describing faces which can be shared by face detection and landmark localization. The architecture differs from the parallel multi-task framework [2] in which convolutional layers are shared and the last fully connected layers are split according to different tasks.

3. Coupled encoder–decoder network

We propose the coupled encoder–decoder network in a unified framework which consists of three modules: 1) an encoder–decoder to predict facial response maps; 2) a coupled cascade face detection network sharing the feature maps with the encoder–decoder; 3) a regression network that outputs the 2D coordinates of facial landmarks.

3.1. The encoder–decoder for facial response map prediction

Semantically, landmark localization is a sparse segmentation problem. Segmenting the landmark regions are feasible without the constraint of the bounding boxes. As the encoder–decoder framework has shown strong evidence in the performance of segmentation [23], we employ it as the facial response map provider.

The network in Fig. 2 (a) takes an image $\mathbf{I} \in R^{w\times h\times 3}$ as input and a corresponding label map $\mathbf{Z} \in R^{w\times h\times 1}$ as ground truth. Each pixel in $\mathbf{Z}$ is a discrete label $\{0,1,255\}$ that marks the presence of facial landmarks, where 0 denotes a non-landmark region, 1 for landmark and 255 set as ignore label for uncertain areas.

The encoder incorporates a set of convolutional layers, pooling layers and batch normalization layers [25], which is to encode the input $\mathbf{I}$ into a feature space $\mathbf{C}$:

$$\mathbf{C} = f_{ENC}(\mathbf{I};\theta_{ENC}), \mathbf{C} \in \in^{w_c\times h_c\times d_c} \tag{1}$$

where $\mathbf{C}$ denotes the encoded $w_c \times h_c \times d_c$ feature maps and θ_{ENC} denotes encoder parameters. Symmetrically, the decoder module involves a set of unpooling, convolution and batch normalization to transform the feature maps $\mathbf{C}$ to the 2-channel response maps $\mathbf{M}$ in the same size of the image:

$$\mathbf{M} = f_{DENC}(\mathbf{C};\theta_{DENC}), \mathbf{M} \in \in^{w\times h\times 2} \tag{2}$$

where θ_{DENC} denotes the decoder parameters. The objective is formulated as a pixel-wise two-class classification problem with cross-entropy loss:

$$\mathcal{L}^{map} = \frac{1}{N_{px}}\sum_{i=1}^{N_{px}} y_i^m \log\left(p_i^m\right) + \left(1-y_i^m\right)\log\left(1-p_i^m\right) \tag{3}$$

where N_{px} denotes the number of pixels (ignored pixels are excluded); $p_i^m = g(f_{DENC})$ is the probability of i-th pixel belonging to the landmark region and y_i^m is the ground truth.

The response map $\mathbf{M}$ plays a significant role in the whole framework for two reasons. First, it provides the accurate confidence maps of the foreground (fiducial point regions) and background. A shallow regression model is able to regress the coordinates in favor of the spatial information preserved by the encoder–decoder. Second, $\mathbf{M}$ provides the facial region information for the fiducial point segmentation task, which could be re-utilized for the face detection.

3.2. The coupled feature map cascade for face detection

In [19], authors propose a cascade framework consisting of 12, 24 and 48 nets, which early rejects non-face regions in the lower scale net (12 nets) and passes the detected proposals to networks in the larger scale (24 and 48 nets) for aggregation. However, sliding window on the original image is time-consuming, which may restrict [19] to adopt deeper networks and higher image resolution. As in Fig. 2(b), the same scale intermediate feature maps from f_{ENC} and f_{DENC} are concatenated as the feature maps for face detection. The sliding window is applied to the feature maps instead of original images, avoiding redundant convolutional computation. ROI pooling [4] is applied to map each sub-region of feature maps into a feature vector in the fixed dimension.

The learning objective is formulated as a binary classification as well as the bounding box coordinates localization.

$$\begin{aligned}
\mathcal{L}^{det} &= \mathcal{L}^{cls} + \gamma\mathcal{L}^{loc}\\
\mathcal{L}^{cls} &= y_i^{cls}\log\left(p_i^f\right) + \left(1-y_i^{cls}\right)\left(1-\log\left(p_i^f\right)\right)\\
\mathcal{L}^{loc} &= \left\|(x_1,x_2)-(x_1^*,x_2^*)\right\|_2^2
\end{aligned} \tag{4}$$

where the classification loss $\mathcal{L}^{cls}$ is defined as the cross-entropy loss over the probability p_i^f of the i-th window being a face and $y_i^{cls} \in \{0,1\}$ denotes the ground truth. The loss of bounding box localization $\mathcal{L}^{loc}$ is defined as the euclidean distance between the ground truth bounding box denoted as $\left(x_1^*,x_2^*\right)$ upper-left and bottom-right corner points and the predicted two points (x_1,x_2). The regularization factor γ is set up to balance the penalty from the two branches. We apply 0.01 as the typical value in our framework.

3.3. Landmark localization from response maps

As shown in Fig. 3, the model f_{REG} combines the response maps $\mathbf{M}$ and feature maps in the last layer of f_{ENC} to predict landmark coordinates. Only the foreground channel of response map $\mathbf{M}$ is used for landmark localization. According to the detected window, ROI pooling is applied on the foreground channel and on the last encoder's feature layer, yielding feature vectors $\mathbf{f}_l$ and $\mathbf{f}_g$ respectively. We concatenate f_l and f_g for landmark localization, taking advantage of both local and global information. The concatenated feature is fed to a shallow regression network in which the last layer is a fully connected layer with $2N \times 1$ neurons, which outputs 2D coordinates of N facial

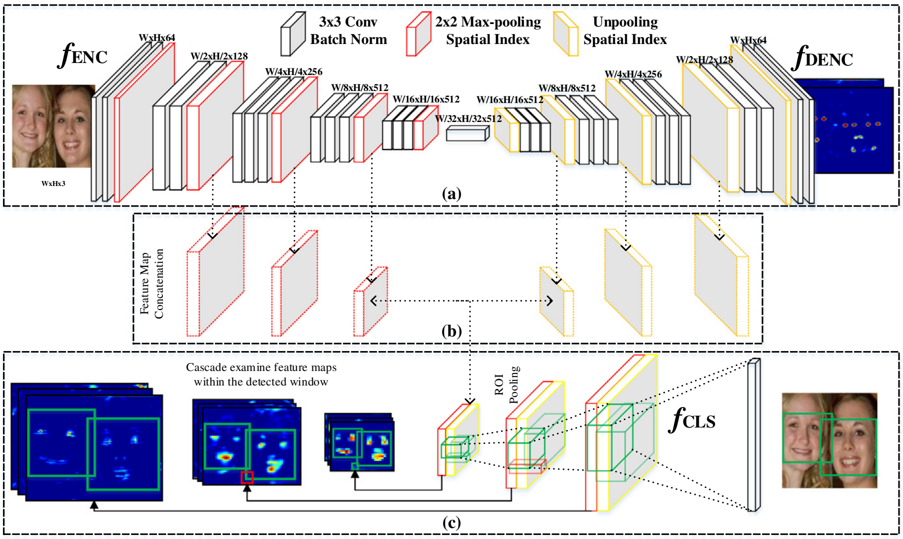

Fig. 2. The illustration of coupled encoder and decoder network. (a) illustrates the encoder and decoder layers (f_{ENC} and f_{DENC}), which consists of convolution (Conv), max pooling, up-sample pooling and fully convolutional layers. (b) shows the coupling structure in which we collect the feature maps from f_{ENC} and f_{DENC} for procedure (c). (c) shows the coupled cascade face detection net f_{CLS}, where the sliding window and ROI pooling are applied on the feature maps to generate the feature representations. The proposals, classified as positive (green bounding boxes) in all three stages, are collected and non-maximum suppressed as face regions.

key points. The landmark localization is formulated as a regression problem with Euclidean loss:

$$\mathcal{L}^{reg} = \sum_{i=1}^{N} \left(\left(sx_i - sx_i^{gt}\right)/w\right)^2 + \left(\left(sy_i - sy_i^{gt}\right)/h\right)^2 \tag{5}$$

where (sx, sy) is the coordinate of detected facial points and (sx^{gt}, sy^{gt}) is the ground truth. The distance in x and y-axis is normalized by window width w and height h, respectively.

4. Implementation details

In this section, we describe the architectures and training procedure for the three proposed modules.

4.1. Network architectures

4.1.1. Encoder f_{ENC} and decoder f_{DENC}

The encoder is designed based on a modification of the VGG-16 network [18]. There are 13 convolutional layers with 3 × 3 filters corresponding to the first 13 convolutional layers in VGG-16. The fully connected layers are replaced by fully convolutional layers, which preserve spatial information. The f_{ENC} contains 5 max-pooling layers in 2 × 2 size and a constant stride of 2. A 2-bit code strategy introduced by [45] is applied to record the spatial information of the maximum activation. At the corresponding unpooling layer, such spatial information is utilized to recover each activation back to its original location. The f_{DENC} is in a mirrored configuration of the f_{ENC} except replacing max pooling with unpooling layers. The decoder outputs a 2-channel response map which is fed to a softmax classifier to predict pixel-wise confidence. Batch normalization [25] and rectified linear unit (ReLU) [44] are applied after each convolutional layer to reduce internal shift within a mini batch.

4.1.2. Coupled face detection f_{CLS}

Fig. 2 (b) demonstrates that the feature maps from both encoder and decoder in the same scale are concatenated, which occurs in three scales: 1) *Conv*2_2 and *Deconv*2_2; 2) *Conv*3_3 and *Deconv*3_3; 3) *Conv*4_3 and *Deconv*4_3. In 1), face detection begins with dense scanning over the feature maps. In this sale, scanning by a 5 × 5 window with 1-pixel stride is equivalent to a 40 × 40 window with stride of 8 on the original image, obtaining $\lfloor(W - 40)/8\rfloor + 1) \times (\lfloor(H - 40)/8\rfloor + 1)$ candidates. ROI pooling is applied to map each window to a 256-d feature vector (128-d for $Conv_{4\_3}$ and $Deconv_{4\_3}$ respectively). The feature vector is fed into

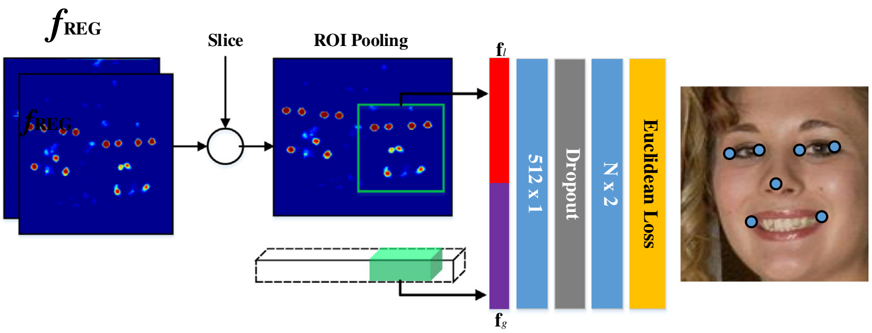

Fig. 3. The architecture of f_{REG}. ROI pooling is applied on the foreground channel and the last encoder's feature layer, yielding feature vectors $\mathbf{f}_l$ and $\mathbf{f}_g$. f_{REG} utilizes the concatenation of $\mathbf{f}_l$ and $\mathbf{f}_g$ to predict 2D coordinates for N landmarks.

fully connected layers of 128 neurons followed by a softmax classifier, generating confidence score for the specific window. A threshold T_1 is set to reject non-face areas. NMS is applied on highly overlapped proposals to reduce the output windows. In the second scale, the ROI pooling transforms regions preserved in previous stage into a 512-d feature vector. A threshold T_2 is set to filter out non-face regions further. In the last scale, we continue to examine the preserved windows. The dimension of feature vectors generated by ROI pooling is 512 and fully connected layers have 256 neurons. A threshold T_3 is set to reject false alarms. The three stage box calibration networks introduced by [19] are removed as the spatial alignment is naturally incorporated by the feature maps.

4.1.3. Landmark regression f_{REG}

As shown in Fig. 3, the f_{REG} applies fully connected layers of 512 neurons to directly regress the input to the 2D coordinates. The network input is a combination of the response map M and the feature maps given by f_{ENC}. According to detected windows, ROI pooling transforms each region of interest into a feature vector with dimension of 512. The dropout layer with 0.5 probability is also applied. In our task, N number of landmarks is set as 7.

4.2. Training

In our framework, we apply an alternative training procedure for the coupled structure. First, the model of f_{ENC} and f_{DENC} followed by the f_{REG} are trained end-to-end, in which the gradients could be back-propagated without any gradient interception. Then, the coupled feature maps are concatenated. ROI pooling is applied on the feature maps to generate features for the facial region classification and bounding box localization. In the last, f_{REG} is fined tuned with cropped sample images according to the windows given by f_{CLS}. The first step is considered the mainstream, while the second step is based on the feature maps generated in the first one. By alternatively optimize each part, the two objectives are optimized simultaneously.

In the first step, the convolutional parameters are initialized by weights of VGG-16 trained on large datasets for object classification. The rest parameters are set with Gaussian Distribution. In this step, the data augmentation is performed, including horizontal flip, central rotation ($\pm 10^\circ$) and scaling (0.8–1.2), yielding 24 variations for one image.

The second step involves a cascade of the three stages of f_{CLS}. We crop patches by sliding window to collect positive and negative samples to train the first stage. Patches of Intersection-of-Union (IoU) larger than 0.6 to ground truth are labeled as positive. The negative samples are regions of IoU less than 0.2. Additionally, we add more negative samples by collecting around 2000 background images, from which we randomly sample 100,000 non-face patches. The detector of the first stage is applied to mine positive and negative samples for the second stage. The non-face regions with confidence score given by the detector higher than threshold T_1 become negative samples. Similarly, the detectors of the first and second stages are both used to mine the training samples for the third stage.

The face detection follows the cascaded structure in [19]. T_1 in the first stage is set to keep 98% recall on the validation set, which rejects 85% false positive windows. Threshold T_2 in the second stage is set to keep 95% recall on the validation set.

Sliding window is applied to feature maps with various sizes for multi-scale detection, e.g.,a 5×5 or 10×10 window with 1-pixel stride on $Conv_{3\_3}$ feature map is equivalent to 40×40 or 80×80 window on original image with stride of 8.

The last is a fine-tuned step for f_{REG} where training samples are the cropped images according to the windows given by f_{CLS}. In this step, f_{REG} shares the feature maps with f_{ENC} and f_{DENC}. In the training stage, given image size of 224×224 and batch size of 64, one iteration of f_{ENC} and f_{DENC} takes 0.3 s, and f_{CLS} or f_{REG} takes 1 s.

5. Experiments

5.1. Experimental setup

For landmark localization, the training data consists of images from training set of LFPW [43] (LFPW-train) and Helen [41] (Helen-train). The evaluation set contains AFW [6], testing set of LFPW (LFPW-test) and Helen (Helen-test). We follow the annotation rule in [40] for 68 facial points to generate 7 landmarks to locate eye corners, mouth corners and nose tip. The facial images used in our experiments cover large head pose variations, expressions, variations of background and occlusions.

For face detection, we apply a commonly used wild face dataset, WIDER FACE [39], for training set. The testing is conducted on two mostly deployed pubic benchmarks, FDDB [28] and AFW [6]. WIDER FACE consists of 393,703 labeled face bounding boxes in 32,203 images. FDDB dataset contains the annotations for 5171 faces in a set of 2845 images and AFW [6] is a 205-image dataset with 468 faces annotated. Images of the three datasets contain cluttered backgrounds and large variations in viewpoints and appearance.

5.2. Evaluation of face landmark localization

We first evaluate the coupled encoder–decoder network for face landmark localization. The localization accuracy is measured by the pixel distance between detected points and the ground truth. We follow the evaluation metric in [2] where the pixel distance is normalized by the inter-ocular distance. As illustrated in Fig. 4, the performance of our model is compared with four methods including 1) SDM [5]; 2)DLIB [26,27]; 3) TCDCN [2]; 4) CoR [3]; 5) HPM [36].

The bottom row of Fig. 4 illustrates statistical curves of mean localization errors on three datasets, Helen-test (left column), AFW(middle column) and LFPW-test(right column). According to the bottom row of Fig 4, the accuracy of our approach is better than the other four methods on datasets of Helen-test and LFPW-test. Regarding of AFW, our coupled encoder and decoder model is comparable to DLIB, but still better than the other three.

The top row of Fig. 4 shows the localization errors for the seven facial components. According to the two histograms of Helen-test and AFW, the accuracy of our approach is comparable to DLIB with respect to left eye's left corner and right eye's right corner. In the LFPW-test set, our approach is comparable to SDM for right eye's right corner and DLIB for left eye's right corner. Regarding the rest facial components, the localization accuracy of our approach is higher than the four methods. The localization accuracy of our coupled encoder–decoder network for mouth corners and nose tip is higher than other methods by a significant margin.

The evaluation conducted on the three benchmarks demonstrates the superior performance of our coupled encoder–decoder network than the state-of-the-art methods. It can be interpreted as the fact that our model captures the more discriminative features and the segmentation scheme is more effective than the regression-based methods. With the powerful feature representation, a shallow regression network in our work is applied to localize face landmarks precisely.

5.3. Evaluation of face detection

We compare our coupled face detector with the state-of-the-art approaches on FDDB and AFW benchmarks. For FDDB, we compare our performance directly with the published methods listed in FDDB platform [28]. Two evaluation protocols are provided by [28], discontinuous score and continues score. Continuous score heavily relies on annotations of training set. We do not follow the eclipse labeling style for the faces, so we only report discontinuous score, where detected regions of IoU larger than 0.5 to the ground truth are

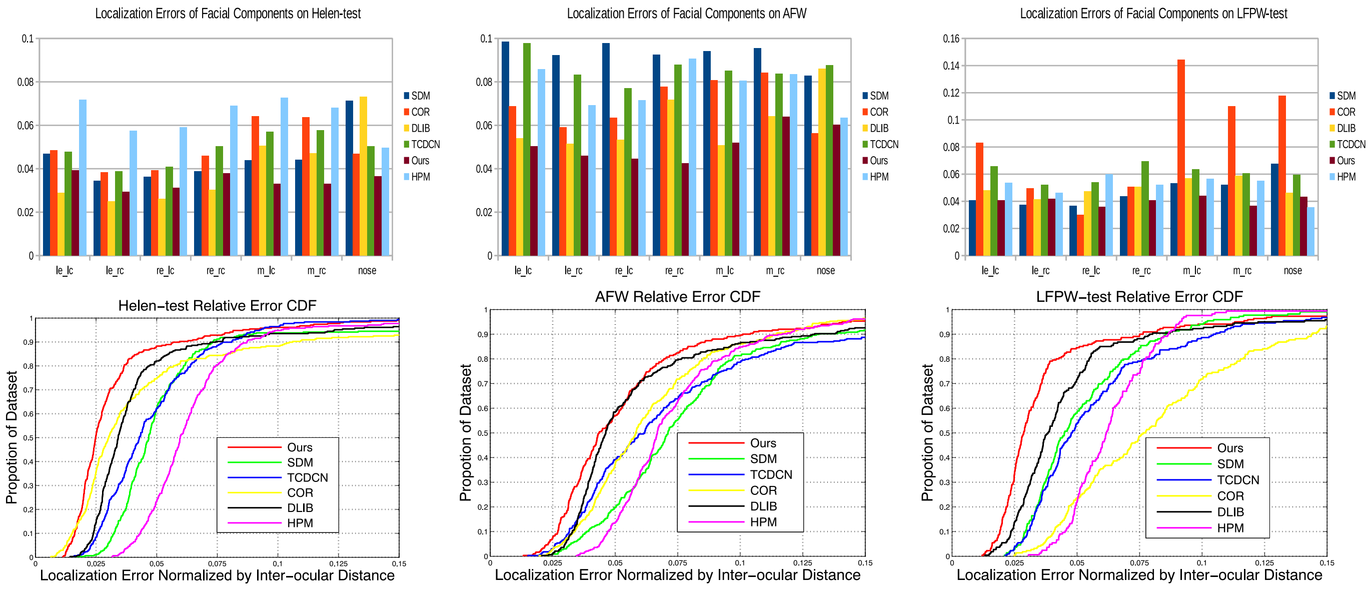

Fig. 4. The face landmark localization results on three benchmarks, Helen (left column), AFW (middle column) and LFPW (right column). The top row illustrates the localization errors for seven facial components, left corner of left eye (le_lc), right corner of left eye (le_rc), left corner of right eye (re_lc), right corner of right eye (re_rc), mouth left corner (m_lc), mouth right corner (m_rc) and nose tip. The bottom row shows the cumulative curves of relative mean errors, where the horizontal axis is the normalized distance with respect to the inter-ocular distance and the vertical axis is the proportion of images in the dataset.

regarded as true positives. For AFW, we use the toolbox provided by [29] to evaluate the detection performance.

The evaluation on FDDB and AFW is illustrated in Fig. 5. In both datasets, our performances are favorably comparable to the state-of-art methods. More worth to highlight, solving the same problem as ours, both Joint Cascade [32], TSM [6] and HPM [36] are proposed to jointly detect faces and localize landmarks. Our approach achieves higher accuracy than the two methods on FDDB. While on AFW, our detection accuracy is consistently higher than TSM by a significant margin. Even without box calibration networks, our structure demonstrates the better performance than the previous cascade CNN [19] on FDDB. Two main reasons may lead to the performance boosting: 1) the features captured by the coupled encoder–decoder are more discriminative for describing faces; 2) our approach takes advantage of multi-task training which brings mutual benefits among different tasks.

Fig. 6 shows that there are several faces miss detected in both benchmarks. For face detection as a single task, in order to achieve top performance, the detectors are designed to be able to find face regions even in small size, low quality and heavy occlusion, as depicted in Fig. 6. Our approach tries to jointly detect faces and localize the landmarks, where the encoder–decoder are trained with faces to be reasonable clear and in prosper size, without covering such extreme cases. The miss rate indicates our future effort can be put on exploration of dealing with the facial images in low quality.

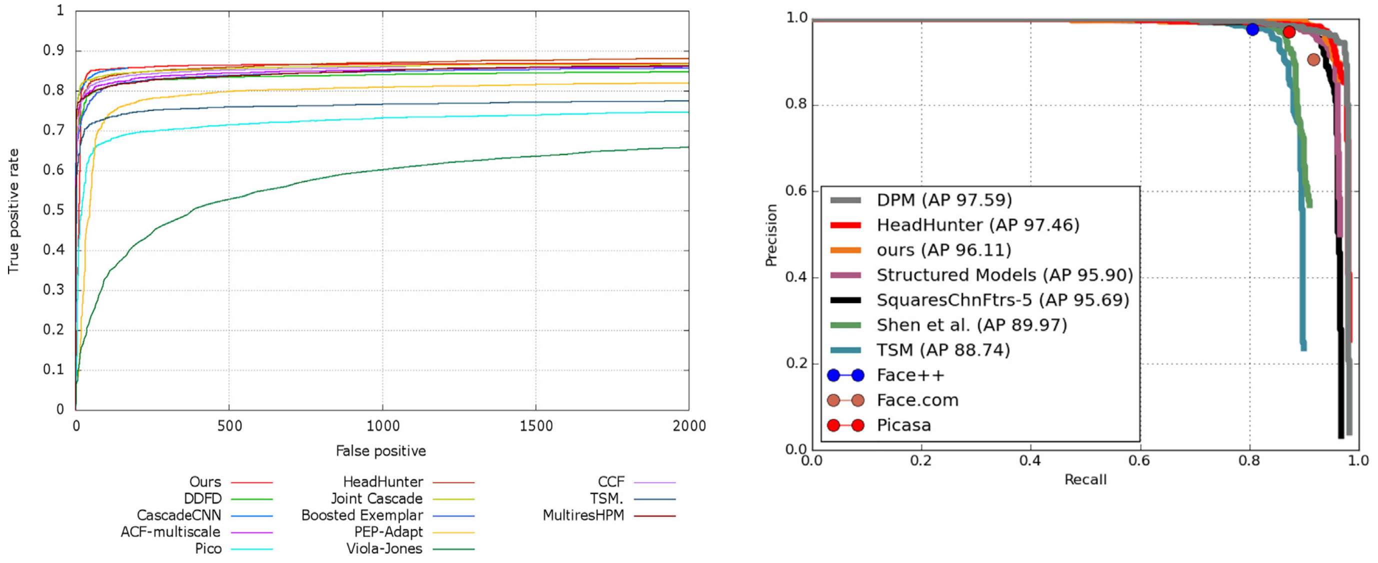

Fig. 5. The evaluation of face detection on datasets of FDDB (left) and AFW (right). On the FDDB dataset we compare our performance with the state-of-the-art methods including CasecadeCNN [19], Joint Cascade [32], DDFD [30], ACF-multi-scale [46], PEP-Adapt [47], Boosted Exemplar [48], HeadHunter [29],Pico [49], Viola–Jones [33], TSM [6], and HPM [36]; on the AFW dataset, comparison methods includes HeadHunter [29], DPM [29], SquaresChnFtrs-5 [29], Shen et al. [50], TSM [6], and three commercial applications, Face++, face.com and Piscasa.

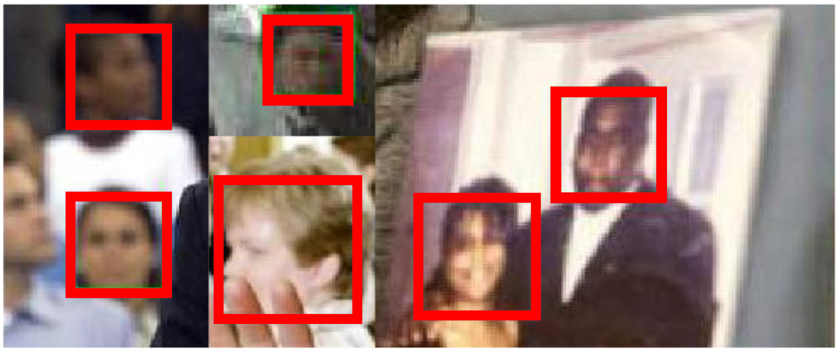

Fig. 6. The examples of faces in the dataset with low resolution, where face landmark localization is challenging.

5.3.1. Robust pre-processing for face recognition

In the study [51], a Face Recognition(FR) dataset is proposed where faces are collected with multi-sensor mobile devices in challenge conditions. The challenges for mobile based FR are variation in illumination conditions, poor face image quality (due to various factors including noise and blurriness due to movement of hand-held device during collection), variations in face pose and camera sensor quality. Those in-the-wild conditions make face area localization inaccurate, which causes recognition performance degrade. In this section, we apply the f_{CLS} module as the FR pre-processing. The detection accuracy up to 98.4% of detecting faces within 1 meters distance to the sensor demonstrates the robustness of our approach to handle cases in the mobile condition.

In [51], the multi-sensor (MS) face image database is collected using a set of cell phone devices including Samsung S4 Zoom, Nokia 1020, Samsung S5 and iPhone 5S. The visible band face database is collected indoors, outdoors, at standoff distances of 1m, 5 m and 10 m respectively, and with different pose angles as shown in Fig. 7.

We conducted the experiments for faces photographed in all the three distance settings, 3459 frames are sampled under each setting. In total, we uni-sampled 10377 frames from the videos and processed the images using f_{CLS} module. The number of successful and failure cases are 8860 and 1517, respectively. The success rate is up to 85.4% and 98.4% for 1 m setting:

(i) the numbers of failure cases are 54, 360, 1103 for distance settings of 1m, 5 m and 10 m respectively;

(ii) most of the failure cases are people photographed in 10 m and 5 meters; it is due to missing the faces which are in the

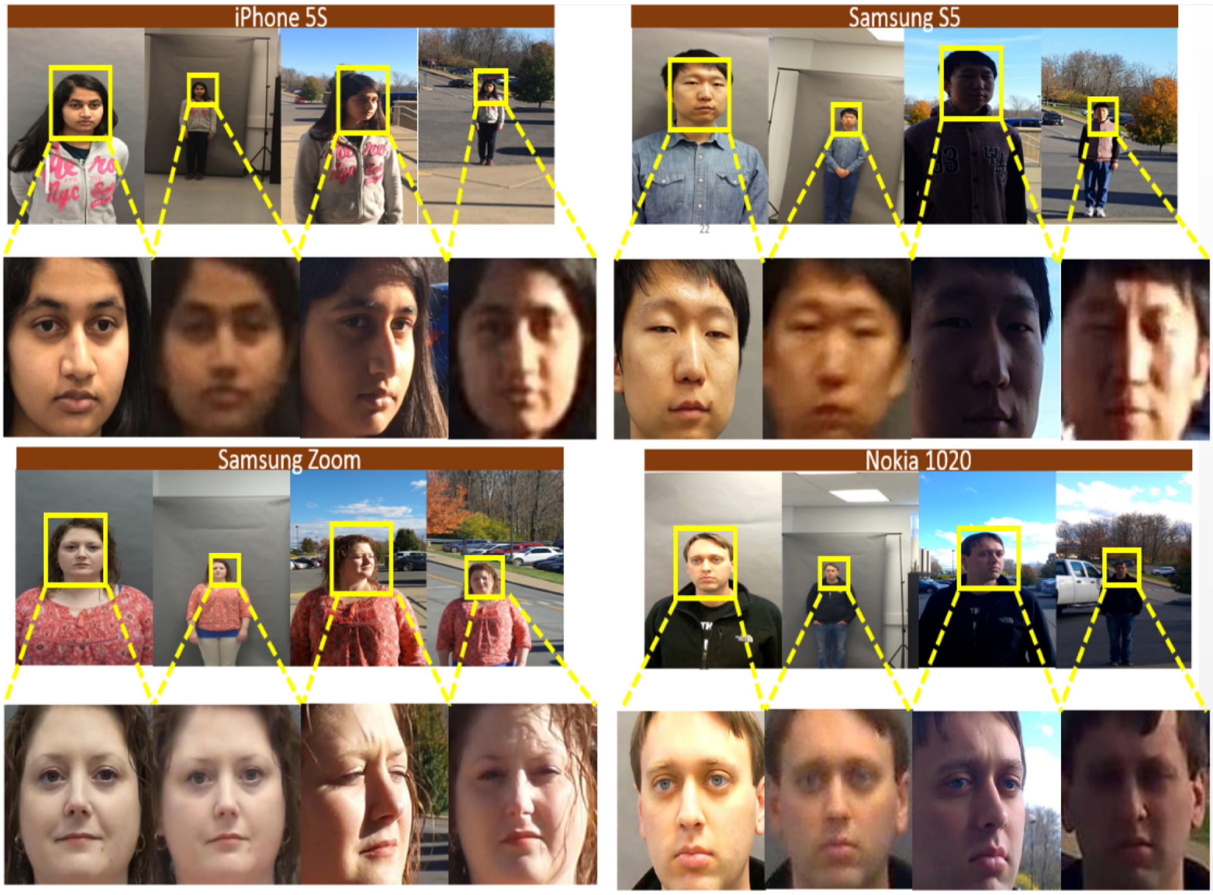

Fig. 7. Face examples of the multi-sensor database, which are collected using various cell phones and photographed in different distances.

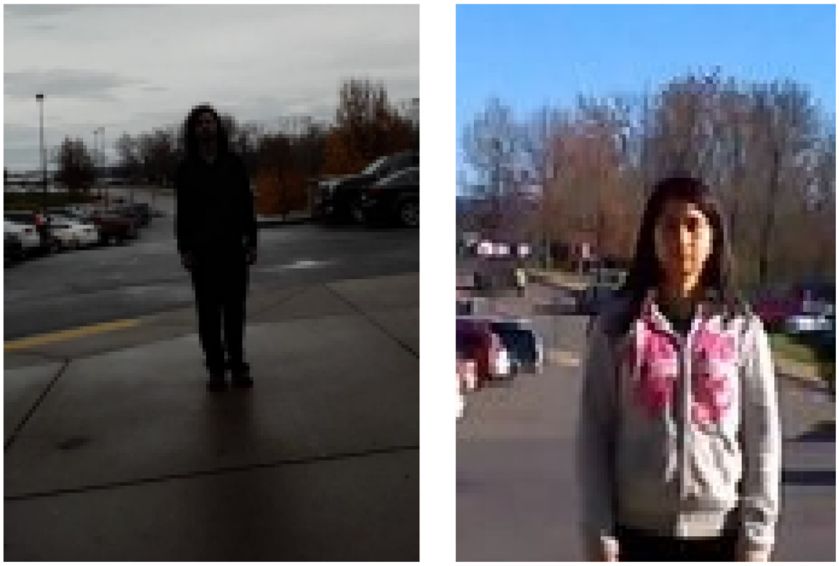

Fig. 8. Failure cases of finding the face area for FR under challenge conditions.

uneven/low illumination (Fig. 8 left) or blur (Fig. 8 right), and tiny (photographed in long distance);

(iii) the failure cases of close faces (1m) are caused by sampling the dark frames and camera pointing to wrong direction where the faces are not or partially shown in the frames.

The results of applying the f_{CLS} module to detect the faces photographed by mobile devices demonstrate the robustness of our approach as to be the pre-processing step for Face Recognition. In FR, users are most likely to present their faces closed to the sensor and f_{CLS} finds all the visible faces in the case of short distance.

5.4. Qualitative results

Fig. 9 shows qualitative results of joint face detection and key points localization, which are performed simultaneously under deep neural network frameworks. Different from other landmark localization methods, we are able to localize landmarks without a face bounding box prior and generate the 7-point landmarks. In our work, we aim to generate the semantic feature maps to boost the performance of face detection and landmark localization. By carefully defining the landmark positions on top of the feature maps, the landmarks could have semantic meanings, i.e., the 7-point setup as denoting the eye centers, nose tip and the mouth corners. If we

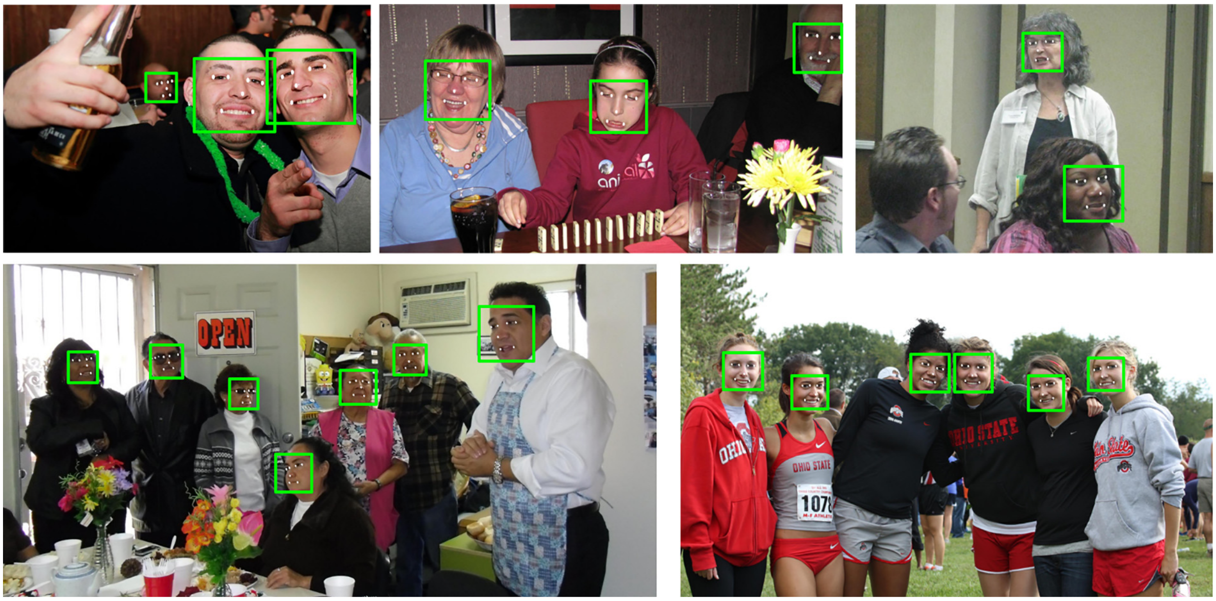

Fig. 9. Qualitative results of face detection and landmark localization.

adopt the 68-point annotation as the other landmarks settings, some landmarks such as along the profile may be less meaningful.

In our work, the two tasks would boost each other in generating better features. The encoder–decoder framework is proposed to generate the feature map, which indicates semantic facial structures as shown in Fig. 1 the heat maps. This semantic highlight is also observed in face detection. Thus, setting up a coupled structure for face detection and landmark localization, the response map constrained from face detection would be also beneficial for the landmarks.

6. Conclusion

In this paper, we proposed a coupled encoder–decoder neural network to jointly detect faces and localize landmarks. The encoder–decoder provides the discriminative feature maps for landmark localization. Further, we observe that the feature maps is also effective for the task of face detection, which enables a unified coupled structure as proposed in our method. The performance on both of the two tasks are very competitive while sometimes better than some of the state-of-the-art methods. The training of the overall framework is alternative optimization. Future work will focus on how to formulate the two tasks as a single optimization problem.

References

[1] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2013) 3476–3483.
[2] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, Inproceedings of European Conference on Computer Vision (ECCV), 2014.
[3] X. Yu, Z. Lin, J. Brandt, D.N. Metaxas, Consensus of regression for occlusion-robust facial feature localization, ECCV, 2014.
[4] S. Ren, K. He, R. Girshick, J. Sun, Faster r-CNN: towards real-time object detection with region proposal networks, Neural Information Processing Systems (NIPS), 2015.
[5] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2013) 532–539.
[6] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 2012, pp. 2879–2886.
[7] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Int. J. Comput. Vis. 107 (2) (2014) 177–190.
[8] G. Trigeorgis, P. Snape, M.A. Nicolaou, E. Antonakos, S. Zafeiriou, Mnemonic descent method: a recurrent process applied for end-to-end face alignment, CVPR, 2016.
[9] J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment, European Conference on Computer Vision, Springer International Publishing. 2014, pp. 1–16.
[10] G. Tzimiropoulos, Project-out cascaded regression with an application to face alignment, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 2015, pp. 3659–3667.
[11] X. Yu, F. Zhou, M. Chandraker, Deep deformation network for object landmark localization, 14th European Conference on Computer Vision, Netherland, Amsterdam, 2016.
[12] Y. Wu, Q. Ji, Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[13] J. Alabort-i-Medina, S. Zafeiriou, A unified framework for compositional fitting of active appearance models, Int. J. Comput. Vis. (2016) 1–39.
[14] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shifts, Int. J. Comput. Vis. 91 (2) (2011) 200–215.
[15] Y. Wu, Z. Wang, Q. Ji, Hierarchical probabilistic model for facial feature detection, IEEE Computer Vision and Pattern Recognition (CVPR), 2014.
[16] S. Milborrow, F. Nicolls, Locating facial features with an extended active shape model, ECCV, 2008. pp. 504–513.
[17] X. Gao, Y. Su, X. Li, D. Tao, A review of active appearance models, IEEE Trans. Syst. Man Cybern. 40 (2) (March 2010) 145–158.
[18] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, CoRR. 2014, abs/1409.1556.
[19] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. pp. 5325–5334.
[20] H. Qin, J. Yan, X. Li, X. Hu, Joint training of cascaded CNN for face detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[21] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, CoRR. 2014, abs/1409.1259.
[22] L. Lu, X. Zhang, K. Cho, S. Renals, A study of the recurrent neural network encoder–decoder for large vocabulary speech recognition, INTERSPEECH, 2015.
[23] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, Proceedings of the IEEE International Conference on Computer Vision, 2015.
[24] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A Deep Convolutional Encoder–Decoder Architecture for Image Segmentation, CoRR. 2015.
[25] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, CoRR. 2015, abs/1502.03167.
[26] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, CVPR, 2014.
[27] D.E. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[28] V. Jain, E.G. Learned-Miller, FDDB: a benchmark for face detection in unconstrained settings, Technical Report UMCS-2010-009, University of Massachusetts, Amherst, 2010.
[29] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, ECCV, 2014.
[30] S.S. Farfade, M.J. Saberian, L.-J. Li, Multi-view face detection using deep convolutional neural networks, Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM. 2015.
[31] S. Yang, P. Luo, C.C. Loy, X. Tang, From facial parts responses to face detection: a deep learning approach, Proceedings of the IEEE International Conference on Computer Vision, 2015.
[32] D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascade face detection and alignment, European Conference on Computer Vision, Springer International Publishing. 2014,
[33] P.A. Viola, M.J. Jones, Rapid object detection using a boosted cascade of simple features, Computer Vision and Pattern Recognition, 2001. pp. 511–518.
[34] R. Ranjan, V.M. Patel, R. Chellappa, A deep pyramid deformable part model for face detection, 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE. 2015.
[35] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: A Deep Multitask Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition, arXiv preprint. 2016, arXiv:1603.01249.
[36] G. Ghiasi, C.C. Fowlkes, Occlusion Coherence: Detecting and Localizing Occluded Faces, arXiv preprint. 2015, arXiv:1506.08347.
[37] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks, arXiv preprint. 2016, arXiv:1604.02878.
[38] B. Yang, J. Yan, Z. Lei, S.Z. Li, Convolutional channel features, IEEE International Conference on Computer Vision (ICCV), 2015.
[39] S. Yang, P. Luo, C.C. Loy, X. Tang, FACE: a face detection benchmark, IEEE Conference on Computer Vision and Pattern Recognition (CVPR, 2016.
[40] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: the first facial landmark localization challenge, ICCVW, 2013.
[41] V. Le, J. Brandt, Z. Lin, L. Boudev, T.S. Huang, Interactive facial feature localization, ECCV, 2012.
[43] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2011.
[44] V. Nair, G.E. Hinton, Rectified Linear units improve restricted Boltzmann machines, International Conference on Machine Learning (ICML), 2010.
[45] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, ECCV, 2014. pp. 818–833.
[46] B. Yang, J. Yan, Z. Lei, S.Z. Li, Aggregate Channel Features for Multi-view Face Detection, arXiv preprint. 2014, arXiv:1407.4023.
[47] H. Li, G. Hua, Z. Lin, J. Brandt, J. Yang, Probabilistic elastic part model for unsupervised face detector adaptation, Proc. IEEE International Conference on Computer Vision, 2013.
[48] H. Li, Z. Lin, J. Brandt, X. Shen, G. Hua, Efficient boosted exemplar-based face detection, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[49] N. Markuš, M. Frljak, I.S. Pandzic, J. Ahlberg, R. Forchheimer, A method for object detection based on pixel intensity comparisons, In 2nd Croatian Computer Vision Workshop, vol. 8, 2013, May.
[50] X. Shen, Z. Lin, J. Brandt, Y. Wu, Detecting and aligning faces by image retrieval, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. pp. 3460–3467.
[51] N. Narang, M. Martin, D.N. Metaxas, T. Bourlai, Learning Deep Features for Hierarchical Classification of Mobile Phone Face Datasets in Heterogeneous Environments, FG. 2017, 186–193.
[52] X. Peng, R.S. Feris, X. Wang, et al. A recurrent encoder–decoder network for sequential face alignment, European Conference on Computer Vision, Springer International Publishing. 2016, pp. 38–56.
[53] A. Bulat, G. Tzimiropoulos, Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, International Conference on Computer Vision, 2017.
[54] S. Zhu, et al. Unconstrained face alignment via cascaded compositional learning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[55] S. Xiao, et al. Robust facial landmark detection via recurrent attentive-refinement networks, European Conference on Computer Vision, Springer, Cham, 2016.

[56] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, European Conference on Computer Vision, Springer, Cham, 2016.

[57] P. Hu, D. Ramanan, Finding tiny faces, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. 2017.

[58] M. Najibi, et al. Ssh: Single stage headless face detector, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[59] T. Baltruaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. 2016.

[60] Y. Li, et al. Face detection with end-to-end integration of a ConvNet and a 3D model, European Conference on Computer Vision, Springer, Cham, 2016.