WIDER FACE DETECTION (TRACK 1)

Vishweswar Mohapatra
Lebbar Mohammed Abdul Rahman
Aajad Chauhan
Baizel Kurian Varghese
Dr. Tapas Badal

LEADINGINDIA.AI
NATIONWIDE AI SKILLING & RESEARCH INITIATIVE

BENNETT UNIVERSITY
TIMES OF INDIA GROUP

Abstract

- **Face detection and alignment in unconstrained environment are challenging due to various poses, illuminations and occlusions.**
- **We propose a deep cascaded multi-task framework which boost up the detection performance.**
- **In particular, our framework leverages a cascaded architecture with three stages of carefully designed deep convolutional networks to predict face and landmark location in a coarse-to-fine manner.**
- **Our method achieves superior accuracy over the state-of-the-art techniques on the challenging WIDER FACE benchmarks for face detection while keeps real time performance.**

Introduction

FACE detection is essential to many face applications, such as face recognition and facial expression analysis. However, the large visual variations of faces, such as occlusions, large pose variations and extreme lightings, impose
great challenges for these tasks in real world applications. The cascade face detector proposed by Viola and Jones utilizes Haar-Like features and AdaBoost to train cascaded classifiers, which achieves good performance with real-time efficiency. However, quite a few works indicate that this kind of detector may degrade significantly in real-world applications with larger visual variations of human faces even with more advanced features and classifiers. Recently, convolutional neural networks (CNNs) achieve remarkable progresses in a variety of computer vision tasks, such as image classification and face recognition. Inspired by the significant successes of deep learning methods in computer vision tasks, several studies utilize deep CNNs for face detection.

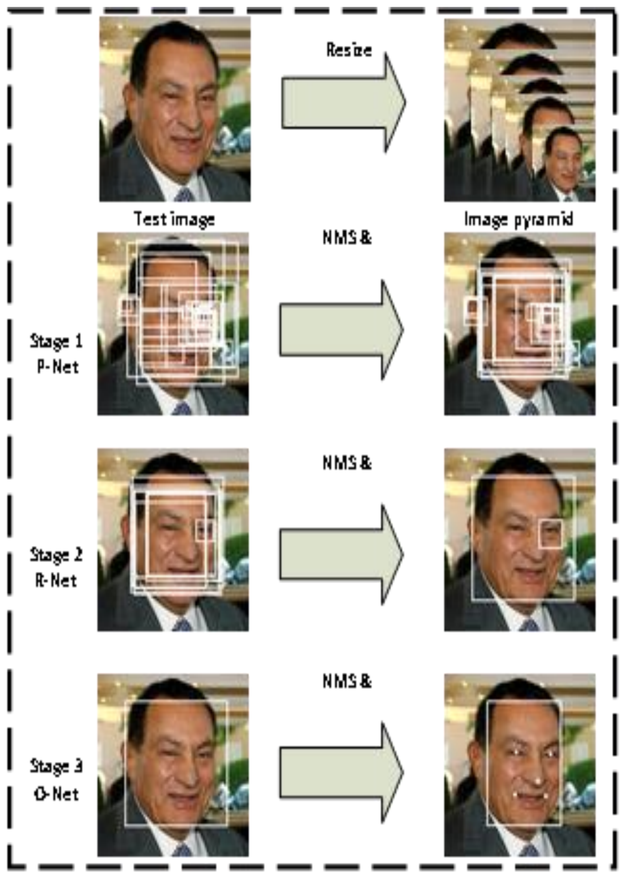

Fig. 1. Pipeline of our cascaded framework that includes three-stage multi-task deep convolutional networks. Firstly, candidate windows are produced through a fast Proposal Network (P-Net). After that, we refine these candidates in the next stage through a Refinement Network (R-Net). In the third stage, The Output Network (O-Net) produces final bounding box and facial land- marks position.

Proposed Method

In this project, we propose a new framework to integrate face detection tasks using unified cascaded CNNs by multi-task learning. The proposed CNNs consist of three stages. In the first stage, it produces candidate windows quickly through a shallow CNN. Then, it refines the windows by rejecting a large number of non-faces windows through a more complex CNN. Finally, it uses a more powerful CNN to refine the result again and output five facial landmarks positions.

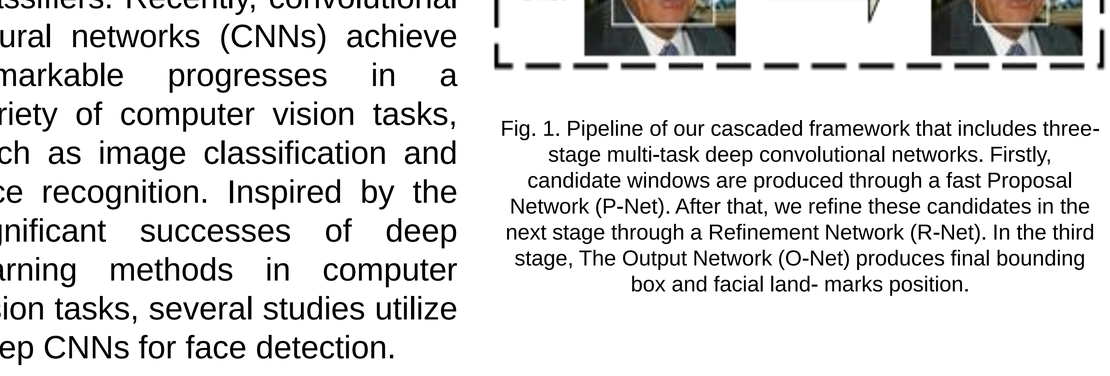

Fig. 2. The architectures of P-Net, R-Net, and O-Net, where “MP” means max pooling and “Conv” means convolution. The step size in convolution and pooling is 1 and 2, respectively.

Experimental Results and Discussion

In this work, we make three contributions. We introduce a large-scale face detection dataset called WIDER FACE. It consists of 32, 203 images with 393, 703 labelled faces, which is 10 times larger than the current largest face detection dataset.
Total training data are composed of 3:1:1:2 (negatives/ positives/ part face/ landmark face) data. The training data collection for each network is described as follows:

1) **P-Net**: We randomly crop several patches from WIDER FACE to collect positives, negatives and part face.
2) **R-Net**: We use the first stage of our framework to detect faces from WIDER FACE to collect positives, negatives and part face.
3) **O-Net**: Similar to R-Net to collect data but we use the first two stages of our framework to detect faces and collect data.

Fig. 3. Test Images of the model with best accuracy

1) **Face classification**: The learning objective is formulated as a two-class classification problem. For each sample x_i, we use the cross-entropy loss:

$$L_i^{det} = -(y_i^{det} \log(p_i) + (1 - y_i^{det})(1 - \log(p_i)))$$

where p_i is the probability produced by the network that indicates sample x_i being a face. The notation $y_i^{det} \in \{0,1\}$ denotes the ground-truth label.

2) **Bounding box regression**: For each candidate window, we predict the offset between it and the nearest ground truth (i.e., the bounding boxes’ left, top, height, and width). The learning objective is formulated as a regression problem, and we employ the Euclidean loss for each sample x_i :

$$L_i^{box} = \left\| \hat{y}_i^{box} - y_i^{box} \right\|_2^2$$

where $\hat{y}_i^{box}$ is the regression target obtained from the network and y_i^{box} is the ground-truth coordinate. There are four coordinates, including left top, height and width, and thus $y_i^{box} \in \mathbb{R}^4$.

3) **Facial landmark localization**: Similar to bounding box regression task, facial landmark detection is formulated as a regression problem and we minimize the Euclidean loss:

$$L_i^{landmark} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2$$

where $\hat{y}_i^{landmark}$ is the facial landmark’s coordinates obtained from the network and $y_i^{landmark}$ is the ground-truth coordinate for the i-th sample. There are five facial landmarks, including left eye, right eye, nose, left mouth corner, and right mouth corner, and thus $y_i^{landmark} \in \mathbb{R}^4$.

Conclusions

Our algorithm can detect between 90.5% and 99.8% of faces in a set of 32, 203 total images, with an acceptable number of false detections. Depending on the application, the system can be made more or less conservative by varying the arbitration heuristics or thresholds used. The system has been tested on a wide variety of images, with many faces and unconstrained backgrounds.

References

- https://github.com/ipazc/mtcnn/tree/master/mtcnn
- https://github.com/wangbm/MTCNN-Tensorflow
- Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks

Team No. - 12