WIDER Face Detection using Multi-task Cascaded Convolutional Networks

Aajad Chauhan
Computer Science
Galgotias University

Baizel Kurian Varghese
Computer Science
Saintgits College of Engineering

L. M. Abdul Rahman
Mechanical Engineering
NIT Rourkela

Vishweswar Mohapatra
Computer Science
Babaria Institute of Technology

Dr. Tapas Badal
Assistant Professor
Bennett University

***Abstract*—In the computer vision community, face detection is one of the much-studied topics. Face detection in an uncontrolled environment is challenging because of its various poses, illuminations, and occlusions. Studies suggest that deep learning can achieve great performance. With the help of face detection benchmark datasets, many signs of progress have been made. There is a huge gap between the face detection performance that is available now that that of real-world requirements. With the help of WIDER FACE dataset which is to aid the future of the face-detection research, which is way much larger than the existing datasets. Faces in the dataset are way more challenging due to the large variations in pose, occlusion, and scale. Moreover, we can see that the WIDER FACE dataset is an effective training source for face detection. But existing face detection algorithms are not fully perfect. They have limitations. So we created a WIDER FACE detection system which will help us overcome all those issues.**

I. Introduction

FACE detection is very important to many of the face applications, such as facial expression analysis and face recognition. The large visual variations of faces, such as large pose variations, extreme lighting and occlusions impose huge challenges in real world applications. Given an image, the main aim of face detection is to get the faces in the image and if it is present, it returns the location of image and of each face.While this may seem as a simple task for human its a difficult task for computers. Challenges related with face detection can be associated with variations in scale, pose, facial expression and lighting condition. Existing face detection systems can only detect frontal face images. So they are not effective as in many cases images might not include frontal faces. So still the face has to be detected which can then be used for other tasks like recognition etc. Here, we introduce a new framework to integrate this task using unified cascaded CNNs by multitask learning. Here the CNNs consists of three stages.In the beginning stage,it gives candidate windows quickly through a hollow CNN. It then filters the windows by refusing a huge number of non faces windows by a complex CNN. In the end it uses a way more powerful CNN to clarify the result and output five facial landmarks positions. By the help of multi task learning framework, the performance of the algorithm was remarkably improved. The codes have been released in the project page. Most of the contributions of this paper can be seen here.

A. First we come up with a new cascaded based framework for wider face detection, and carefully design the lightweight CNN architecture for performance.

B. Then we propose an efficient method to manage online hard sample mining to boost up the performance.

C. Extensive experiments are operated on challenging benchmarks, to show significant improvements compared to the techniques in face detection task.

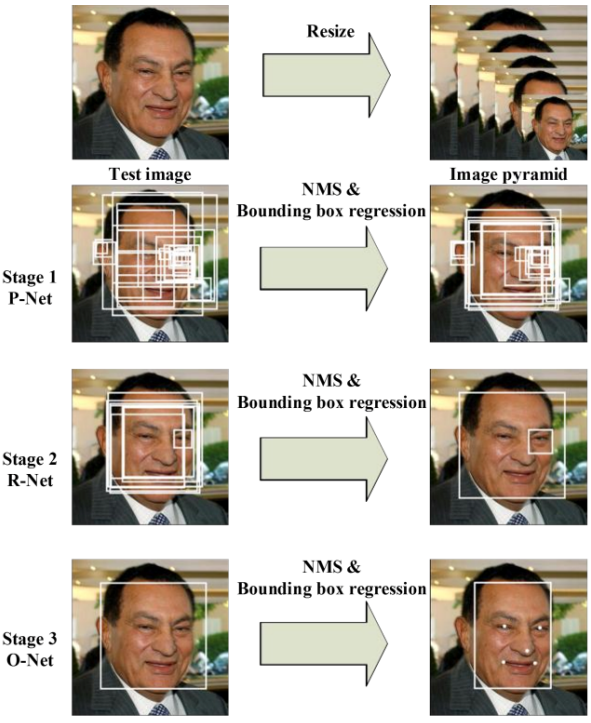

Fig. 1. Pipeline of cascaded framework that shows the three-stage multitask deep convolutional networks.

II. Related Work

Cascade face detector that was proposed by Jones and Viola [2] utilizes AdaBoost and Haar-Like features to train cascaded classifiers, which attains significant performance with real-time efficiency. But this type of detector might degrade undoubtedly in the real world applications with way larger visual variations of human faces even with more advanced features and classifiers. Adjacent to the cascade structure, introduce deformable part models (DPM) for face detection and achieve great performance. However it is way more expensive and requires expensive annotations in the training stage. Recently, convolutional neural networks (CNNs) achieved exceptional progresses in so many of the computer vision tasks, such as image classification [5] and face recognition [6]. Inspired methods of deep learning in computer viosion tasks, certain studies utilizes deep CNNs for face detection. Yang et al. [7] Deep convolutional neural networks for facial attribute recognition in order to get significant responses in face regions that further returns the candidate windows of faces. But because of its compound CNN structure, this method is time consuming in practice. [14] Li et a. used cascaded CNNs for face detection, although it requires bounding box calibration from face detection with more computational expense. Researches can be divided into two categories, regression based methods [8, 9, 12] and template fitting approaches [10, 11, 7]. Recently, Zhang et al. [16] recommended we use facial attribute recognition as an extra task to improve the performance using deep convolutional neural network. Nevertheless many of the existing works try to jointly solve them, there are still setbacks in many of these works. As, Chen et al. [13] jointly conducted detection with using features of pixel value difference. But the performance of the hand craft features are bounded. Zhang et al[15] used multi-task CNN to build up the accuracy of multi-view face detection, but the detection recall is limited by the initial detection window produced by a unsteady face detector. But, hand mining samples in training is important to improve the capability of detector. Nonetheless, traditional hard sample mining normally performs in an offline manner, that improves the manual operations.

III. Methodology

A. Overall Framework

We first initilly resize the given image to different scales to bulid an image pyramid that is the input of the following three stage cascaded:

1) Stage 1: We begin by exploting a fully convolutional network called proposal network (P- Net) in oder to obtain the candidate facial windows. We employ non maximum suppression (NMS) to combine highly overlapped candidates.

2) Stage 2: Then we feed all the candidates to R-Net that further rejects a huge number of false candidates and performs calibration with bounding box regression and then conducts NMS.

TABLE I
Comparison of speed and validation accuracy of our CNNs and previous CNNs[14].

Group	CNN	300X forward propagation	Validation accuracy
Group 1	12-Net[19]	0.038s	94.4%
	P-Net	0.031s	94.6%
Group 2	24-Net[19]	0.738s	95.1%
	R-Net	0.458s	95.4%
Group 3	48-Net[19]	3.577s	93.2%
	O-Net	1.347s	95.4%

3) Stage 3: This stage is same as the second stage, but in this stage we try to identify face regions with more supervision. In particular, the network will output five facial landmarks.

B. CNN Frameworks

In [17], multiple CNNs have been designed for face detection. Nonetheless, the performance is limited by the following facts: (1) Some filters in convolution layers lack diversity that may limit their discriminative ability. (2) Compared to other multi-class objection detection and classification tasks, face detection is a challenging binary classification task, so it may need less numbers of filters per layer. To this end, we reduce the number of filters and change the 55 filter to 33 filter to reduce the computing while increase the depth to get better performance. And with these advancements previous architecture in [14], we can get more performance with less run time (the results in training phase are shown in the above table. For fair comparison we use the same training and validation data in each group). Our CNN architectures are shown in Fig. 2. We apply PReLU [14] as nonlinearity activation function after the convolution and fully connection layers (except output layers).

C. Training

We leverage three tasks to train our CNN detectors: face/non-face classification, bounding box regression, and facial landmark localization.

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:

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

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

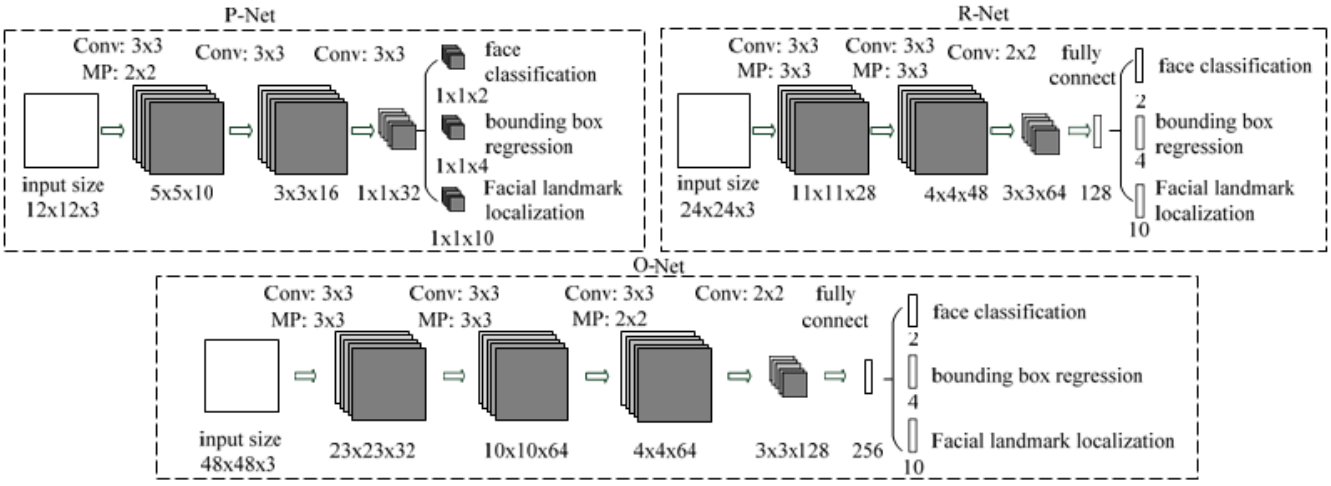

Fig. 2. The architectures of P-Net, R-Net, and O-Net, where MP means max pooling and Conv means convolution. The step size in convolution and pooling is 1 and 2, respectively.

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:

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

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

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

$$L_i^{landmark} = \left\| \hat{y}_i^{landmark} - y_i^{landmark} \right\|_2^2 \tag{3}$$

Where $\hat{y_i}^{landmark}$ is the facial landmarks coordinate obtained from the network and $y_i{}^{landmark}$ is the ground-truth coordinate. There are five facial landmarks, including left eye, right eye, nose, left mouth corner, and right mouth corner, and thus $y_i{}^{landmark} \in \mathbb{R}^4$ . .

IV. Experimental Results

In this section, we will evaluate the effectiveness of the WIDER FACE Dataset[16] and MTCNN. WIDER FACE dataset consists of 393,703 labeled face bounding boxes which are obtained from 32,203 images .Among these , 50 % of them are used for testing, 40% for training and the remaining for validation. Finally, we evaluate the computational efficiency of our face detector and we got an accuracy of 90-95%.

A. Training Data

We use four types of annotations in training process, negative,positive,part faces and landmark faces[16].

1) Negative: Regions with no faces in it.These regions have IOU(Intersection Over Union) ratio less than 0.3 with respect to ground truth faces.

2) Positive: Regions having faces in it.The IOU ratio for these regions are greater than 0.65 to a ground truth face.

3) Part faces: Part means Partial faces.It includes images with partial faces in it.For example,image with only ears,nose,etc. The IOU for these regions are between 0.4 and 0.65 with respect to a ground truth face.

4) Landmark faces: 5 landmarks are taken from every image.The landmarks are generally 2 eyes,nose and the endpoints of the mouth.The output from ONet is the 5 Landmarks.Part faces and negative are entirely different.

Negatives and positives are used for face classification tasks, positives and part faces are used for bounding box regression, and landmark faces are used for facial landmark localization. Total training data are composed of 3:1:1:2 (negatives/ positives/ part face/ landmark face) data.The training data are collected for each of the 3 Net(PNet,RNet,ONet) and are described below.

PNet: We crop the image randomly from the dataset[16] and generate positive,negative and part faces as well as generate the annotation files.

RNet: Stands for Refine network. It filters the bounding boxes. RNet also generate positive,negative and part faces as well as generate annotation files.

ONet: ONet is similar to RNet. Output of ONet is the Facial Landmarks. First 2 stages are used for detecting faces and collecting data.

B. Effectiveness of wider face dataset

Wider face dataset[17] allows us to train the model in such a way that our model will be able to detect faces with occlusion,masks,illuminations,varied scales,poses,etc.

Fig. 3. Some outputs of the model with best accuracy.

C. *Evaluation on face detection*

Output shows that our method consistently outperforms all the compared approaches by a large margin in both the benchmarks. Our model can achieve high speed in face detection.

V. Conclusion

In this paper, we have proposed a multi-task cascaded convolutional neural network based framework for wider face detection. Experimental results demonstrated that our methods consistently outperform other models used for face detection which are not successful in detecting faces under uncertain conditions. The main contributions for performance improvement are carefully designed cascaded CNNs architecture and WIDER FACE Dataset[16].

References

[1] B. Yang, J. Yan, Z. Lei, and S. Z. Li, Aggregate channel features for multi-view face detection, in IEEE International Joint Conference on Biometrics, 2014, pp. 1-8

[2] P. Viola and M. J. Jones, Robust real-time face detection. International journal of computer vision, vol. 57, no. 2, pp. 137-154, 2004

[3] M. T. Pham, Y. Gao, V. D. D. Hoang, and T. J. Cham, Fast polygonal integration and its application in extending haar-like features to improve object detection, in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 942-949.

[4] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan, Fast human detection using a cascade of histograms of oriented gradients, in IEEE Computer Conference on Computer Vision and Pattern Recognition, 2006, pp. 1491-1498.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in Advances in neural information processing systems, 2012, pp. 1097-1105

[6] Y. Sun, Y. Chen, X. Wang, and X. Tang, Deep learning face representation by joint identification-verification, in Advances in Neural Information Processing Systems, 2014, pp. 1988-1996.

[7] S. Yang, P. Luo, C. C. Loy, and X. Tang, From facial parts responses to face detection: A deep learning approach, in IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.

[8] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, Robust face landmark estimation under occlusion, in IEEE International Conference on Computer Vision, 2013, pp. 1513-1520.

[9] X. Cao, Y. Wei, F. Wen, and J. Sun, Face alignment by explicit shape regression, International Journal of Computer Vision, vol 107, no. 2, pp. 177-190, 2012.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, 2001.

[11] X. Yu, J. Huang, S. Zhang, W. Yan, and D. Metaxas, Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model, in IEEE International Conference on Computer Vision, 2013, pp. 1944-1951.

[12] J. Zhang, S. Shan, M. Kan, and X. Chen, Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment, in European Conference on Computer Vision, 2014, pp. 1-16.

[13] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, Joint cascade face detection and alignment, in European Conference on Computer Vision, 2014, pp. 109-122.

[14] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, A convolutional neural network cascade for face detection, in IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334

[15] C. Zhang, and Z. Zhang, Improving multiview face detection with multi-task deep convolutional neural networks, IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 1036-1041.

[16] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, Facial landmark detection by deep multi-task learning, in European Conference on Computer Vision, 2014, pp. 94-108

[17] S. Yang, P. Luo, C. C. Loy, and X. Tang, WIDER FACE: A Face Detection Benchmark. arXiv preprint arXiv:1511.06523.

[18] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in IEEE International Conference on Computer Vision, 2015, pp. 1026-1034.