WIDER Face Challenge using Multi-Task Cascading Neural Network

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*—Nowadays, face detection is common. Its used in many areas. With the help of face detection benchmark datasets, many signs of progress have been made. Face detection methods used nowadays is not matching the real-world requirements. By using a dataset called as WIDER FACE which is very large in size than already existing datasets, we can improve the performance. Dataset has many faces which may be challenging as it includes faces under different conditions. Moreover, we can see that in face detection task, WIDER FACE dataset is best for training the model. 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 [1] is a process that is common nowadays. Almost in every area of our life, we deal with face detection. Smart phones, cameras, etc. are nowadays using this feature. Also, face detection is used in the security sector to identify criminals. So its an important feature. But the method used nowadays has some drawbacks. The existing method is not always able to detect faces. This happens if the faces are not clear if faces include masks, makeup, overexposure, part of the face.

In many tasks like analysis of expression as well as face recognition, face detection is used. The main aim of face detection is to identify faces from the image and return the location of every image. This task is difficult for computers. There are a lot of challenges to face detection. This includes changes in scale, expression, the brightness of an image, etc. 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. So we are using a cascading framework the MTCNN [2] model. This model has 3 networks in it. Each network refines the output images from the previous network and thus removes the false candidates. In the end, we get the output as bounding boxes generated with 5 facial landmarks.

II. Related Work

OpenCV is another face recognition model in which it uses the haar cascade algorithm [3] classifier to detect the faces in images. But the haar cascade mainly focuses on frontal faces. Its inefficient in detecting any other type of images which has some expression or occlusions are there. The accuracy of the OpenCV model is also low for detecting every type of face except the frontal faces.

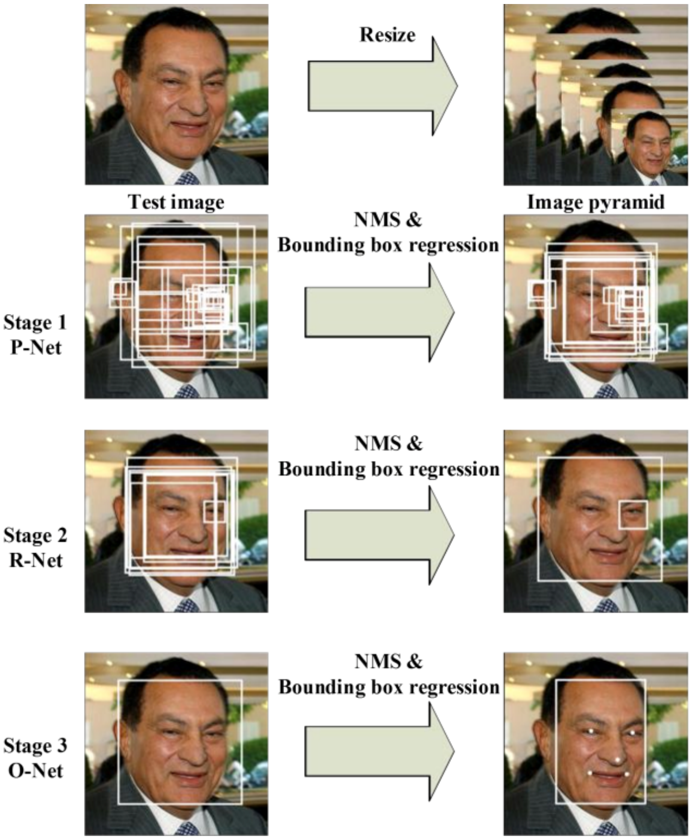

Fig. 1. Three stages of muktitask cascading neural networks

Another model VGG16 [4] which was introduced by Simonyan and Zisserman in 2014. This model performed well at that time, as the time changes challenges get tough every day and this model was not so much efficient for detecting every type of faces. The major drawback was the training time and the size of the weight parameters. This model has other issues such as vanishing gradient problem [5] as it uses Gradient-based optimization techniques.

So, now the Multi-Task cascading network uses cascading windows for computer vision tasks. This model uses the bounding box regression [6] to improve the performance of deep neural networks. Candidate windows are being made in it and being refined further which decreases the training time as well as the accuracy of the model. The face attributes are being also used to increase the performance by using the pixel value difference in it.

III. Methodology

A. Overall Structure

First, the image will be resized to different dimensions and the image pyramid is passed through the 3 neural networks of the MTCNN:

1) Phase 1: Exploiting the pyramid into a full convolutional proposal network (PNet) which will construct candidate windows for every facial attribute and do non-max suppression (NMS) [7] on it to lump together excess candidates.

2) Phase 2: The candidates are then passed through RNet for refinement of the candidates to remove the false candidate and it assesses with bounding box regression and treats it through NMS.

3) Phase 3: This phase is the same as the second phase but with more precision, it gives the output and detects the face with all 5 landmarks of the face.

TABLE I
Speed Comparison and final accuracy of MTCNN and other CNNs.

Class	CNN	300 x forward propagation	Final accuracy
Class 1	12-Net	0.038s	94.4%
	PNet	0.031s	94.6%
Class 2	24-Net	0.738s	95.1%
	RNet	0.458s	95.4%
Class 3	48-Net	3.577s	93.2%
	ONet	1.347s	95.4%

B. CNN Frameworks

In [8], several CNNs had been intended for face discovery. Regardless, the exhibition is restrained by using the accompanying certainties: (1) A few channels in convolution layers need a first rate range that can confine their discriminative potential. (2) In comparison to different multi-class object identification and classification work, face detection is a difficult parallel characterization venture, so it'd require fewer quantities of channels in keeping with layer. To this quit, it decreases the variety of channels and adjustments the 55 channel to 33 channel to decrease the figuring while incrementing the profundity to expose signs of development execution. What's more, with those headways past architecture in [9], it gets greater execution with less run time (the outcomes in preparing stage to appear in the above table. For affordable correlation it makes use of a similar making ready and approval facts in each gathering). CNN structures have appeared in fig. 2. After the convolution and absolutely affiliation layers (aside from yield layers) PReLU [9] is applied as nonlinear activation function.

C. Training

CNN detectors are trained after clouting the three phases: classification of face, bounding box regression, and localization of facial landmark.

1) Classification of face: Two-class classification problem is adopted as the learning goal. Cross-entropy loss is utilized for each part x_i:

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

where probability p_i is delivered by the network that demonstrates x_i part as a face. $y^{det} \in \{0, 1\}$ symbolizes the ground truth mark.

2) Bounding box regression: We count on the counterbalance for every candidate window amongst it and the closest ground truth (i.e., left, top, height, and width of bounding boxes). Goal of learning is defined as a regression problem, and for every part x_i we utilize Euclidean loss:

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

where $\hat{y}_i{}^{box}$ regression target got from the network and $\hat{y}_i{}^{box}$ is the ground-truth coordinate. Left, top, height and width are the four directions, and in this way $y_i{}^{box} \in \mathbb{R}^4$.

3) Localization of facial landmark : Detection of facial landmark is calculated as a regression problem like that of bounding box regression task and the Euclidean loss is limited:

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

where facial landmark's coordinate is $\hat{y}_i{}^{landmark}$, acquired from the network and ground-truth coordinate is $y_i{}^{landmark}$. Left eye, right eye, nose, left mouth corner, and right mouth corner, are five facial landmark and hence $y_i{}^{landmark} \in \mathbb{R}^4$.

IV. Experimental Results

We divided the complete dataset in this type of way that 50% is used for training, 40% for testing and the last part of the dataset is used for validation. Also, we tested the model and obtained an accuracy of 96%. Also, we were able to detect faces which other models were not able to because of undesirable conditions which we have mentioned earlier.

A. Training Data

We use four types of annotations in the training process: negative, positive, part and landmark faces [10].

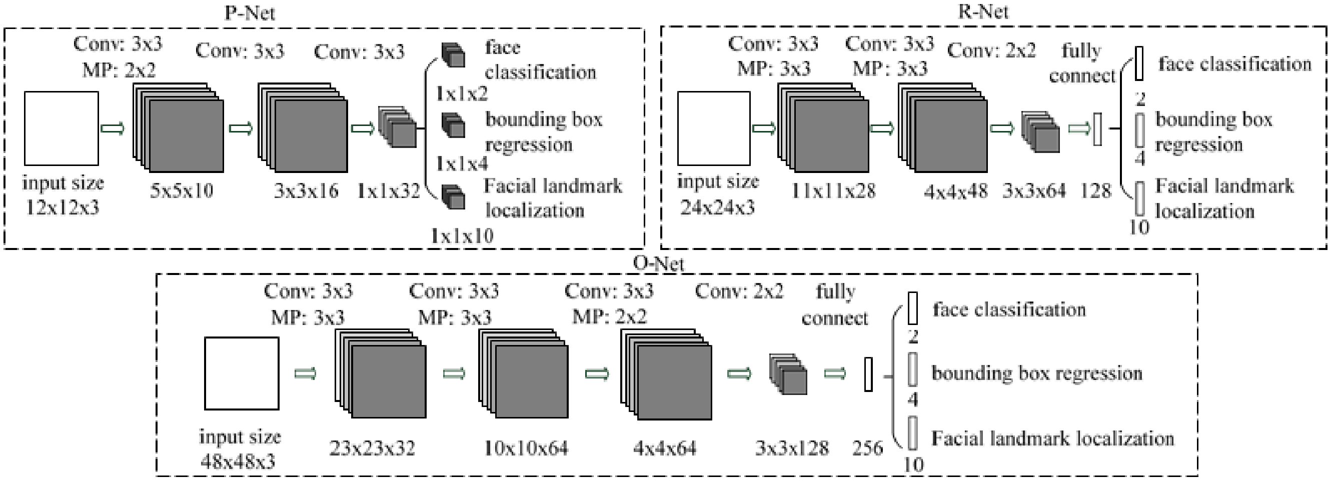

Fig. 2. PNet, RNet, and ONet Architecture where ”MP” signifies Max-Pooling and ”Conv” signifies Convolution. The step size in Convolution and Max-Pooling is 1 and 2, individually.

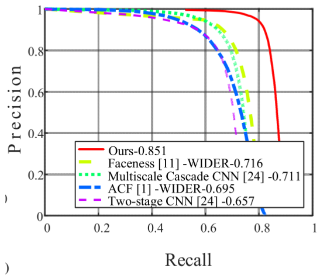

Fig. 3. MTCNN Easy Dataset Comparison

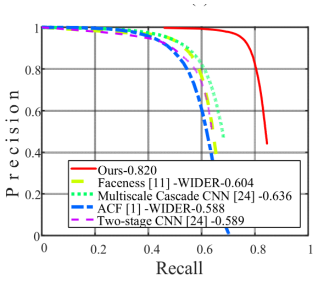

Fig. 4. MTCNN Medium Dataset Comparison

1) Negative: Regions with no faces in it. Each of the three networks created a negative folder with all the negative images in it.

2) Positive: Regions having faces in it. Each of the three networks created a positive folder with all positive images in it.

3) Part: Regions having partial images in it. Partial images are those images which have only a part of the face in it. Similar to the other two networks a folder is created with part images in it.

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

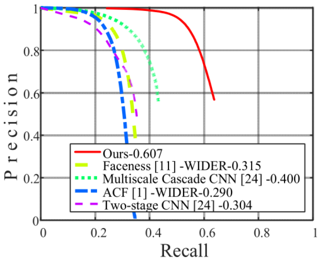

Fig. 5. MTCNN Hard Dataset Comparison

All these annotations contribute to effective face detection. Also due to these annotations, we are obtaining a pyramid structure. Also, data is collected for all of the three networks.

PNet: We crop the image randomly from the dataset [10] and generate positive, negative and part faces as well as generate the annotation files.
RNet: Stands for refinement network. It filters the bounding boxes. RNet also generates positive, negative and part faces as well as generate annotation files.
ONet: ONet is similar to RNet. The output of ONet is the facial landmarks. First 2 phases are used for detecting faces and collecting data.

After training our data using this MTCNN Model for 10 epochs in PNet and 30 epochs in RNet and 50 epochs for ONet we got the accuracy of 85.1% in easy dataset, 82% in medium dataset and 60.7% on hard dataset.

B. *Effectiveness of wider face dataset*

WIDER face dataset [8] allows us to train the model in such a way that faces with occlusion, masks, illuminations,

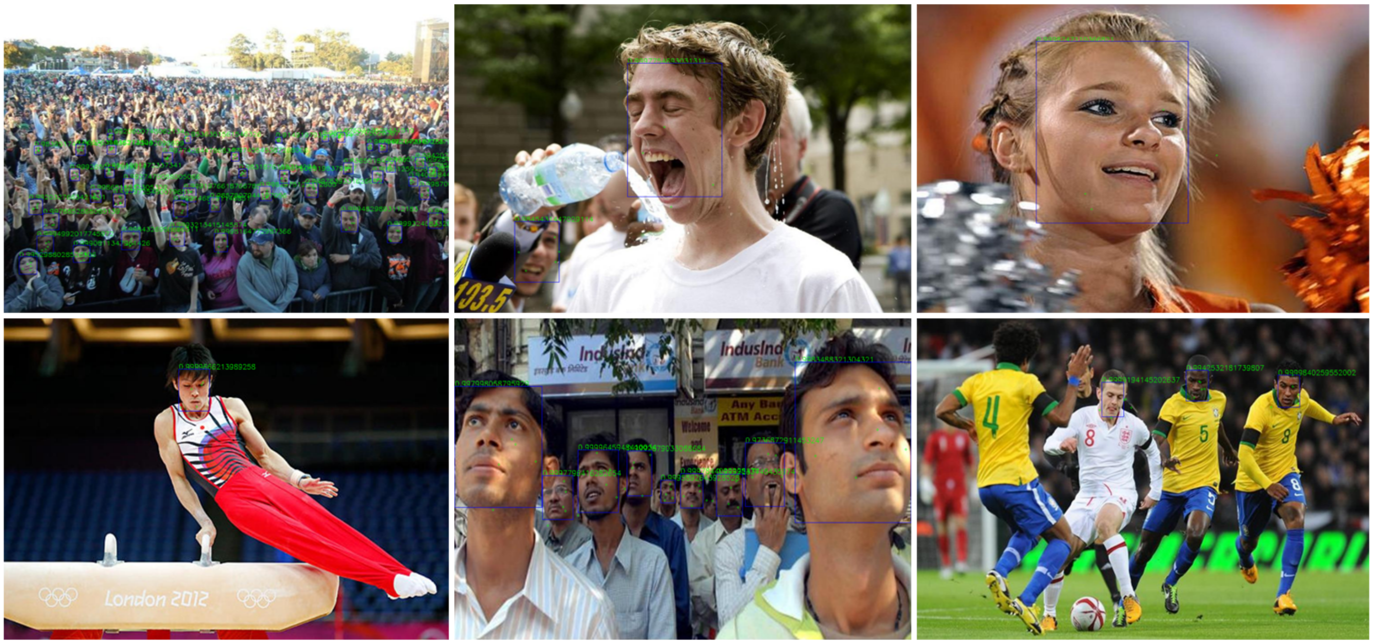

Fig. 6. Some outputs of the model with best accuracy.

varied scales, poses, etc. were detected by our trained MTCNN model.

C. *Evaluation on face detection*

The output shows that MTCNN model which was trained using WIDER FACE dataset proved efficient than other models which do face detection.

V. Conclusion

This paper utilizes multi-task cascaded convolutional neural network based structure for wider face challenge. Experimental outcomes showed that MTCNN reliably outflank different models utilized for face detection which are not fruitful in detecting faces under unsure conditions. The primary commitments for execution improvement are deliberately structured cascaded CNNs architecture and WIDER FACE Dataset [10].

References

[1] Mrs. Sunita Roy and Mr. Susanta Podder, Face detection and its applications, IJREAT International Journal of Research in Engineering Advanced Technology, Volume 1, Issue 2, April-May, 2013.

[2] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Senior Member, IEEE, and Yu Qiao, Senior Member, IEEE, Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks, IEEE Signal Processing Letters (SPL), vol. 23, no. 10, pp. 1499-1503, 2016.

[3] Varun Garg and Kritika Garg, Face Recognition Using Haar Cascade Classifier, JETIR (ISSN-2349-5162), December 2016, Volume 3, Issue 12.

[4] Karen Simonyan Andrew Zisserman, VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION, Published as a conference paper at ICLR 2015.

[5] https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

[6] Seungkwan Lee, Suha Kwak, Minsu Cho, Universal Bounding Box Regression and Its Applications, arXiv:1904.06805 [cs.CV].

[7] Rasmus Rothe, Matthieu Guillaumin, and Luc Van Gool, Non-Maximum Suppression for Object Detection by Passing Messages between Windows, Computer Vision Laboratory, ETH Zurich, Switzerland, ESAT - PSI / IBBT, K.U. Leuven, Belgium.

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

[9] 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

[10] 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