Face Detection in Unconstrained Environments Using Modified Multitask Cascade Convolutional Neural Network

Suchimita Bhattacharya, Manas Ghosh, and Aniruddha Dey

Abstract Due to occlusion and variation of poses, facial detection is a challenging task to accomplish. In facial detection, occlusion has always been a standing challenge. Facial occlusion like sunglasses, scarf, and mask and pose variation are crucial factors that affect the performance of face detection. Undesirably, in the real-world scenario, occlusions are a very common situation that arises in the face detection and recognition problem. To deal with this problem, we put forward the modified multitask cascade convolutional neural network (M-MTCNN) with a slight modification. MTCNN is a trainable unit which may be included in present CNN architectures. With end-to-end training supervised by only the private identity labels, Mask Net learns a correct way of adaptively generating different feature map masks for various occluded face images. This paper deals with an efficient method for the detection of numerous occluded and pose variation faces. In addition to the marking of the face with a square box, there are five landmarks drawn (the two eyes, one in the nose, and two identifying the lips). Also, using the landmark points of the eyes, we have tried to mark the eyes using the eye landmarks as the center point. Using the landmarks of the lips, we also have drawn a straight line marking the edge of the lips. The presented method has been tasted on WIDER database and obtained efficient detection of multiple occluded faces.

Keywords Convolutional network · Face detection · Partial occlusion · Pose variations · M-MTCNN

S. Bhattacharya · M. Ghosh
Department of Computer Application, RCCIIT, Kolkata, India
e-mail: manas.ghosh@rcciit.org

A. Dey (✉)
Department of Information Technology, MAKAUT, Kolkata, India

S. Bhaumik et al. (eds.), *Proceedings of International Conference on Industrial Instrumentation and Control*, Lecture Notes in Electrical Engineering 815,
https://doi.org/10.1007/978-981-16-7011-4_29

1 Introduction

Facial region detection [1–3] has gotten a lot of attention and has been widely used in various aspects of human life, such as video reconnaissance, face detection, human–machine interfaces, and picture recovery. Numerous variations in illumination, posture, impediment, and shot point present enormous challenges for face identification in real-world applications. Viola and Jones (VJ) [4] recently developed a course face location technique that was the major effort with an ongoing step and was prepared using Adaboost- and Haar-like elements. Many improved works based on VJ's suggested frameworks have been proposed in recent years [5, 6]. Li et al. [5] were the first to propose cascaded CNNs for face detection instead of Haar-like features and Adaboost of the framework.

In recent years, with the remarkable progress in convolutional neural networks (CNNs) [7], CNN has achieved path-breaking success in computer vision and multimedia, i.e., image classification objection detection [8, 9] and image retrieval [10, 11]. PyramidBox-based face detection methods based on CNN were developed [12] by Tang et al. However, the method is largely inefficient since the CNN structure is complicated. Huang et al. [13] proposed DenseBox to track face by using a single fully convolutional neural network (FCN) to predict the bounding box and the object class confidences directly. The accuracy of detection is further improved by performing localization of facial landmark. Yang et al. [14] have put forward a coarse-to-fine method, named Faceness-Net. In this method, the scores of facial parts are fused using several DCNN to obtain the face region followed by a refining network to achieve the goal face detection. Several works show that performing face detection and landmark localization concurrently improves the performance of face detection [5, 15].

Taking inspiration from cascade CNN and MTCNN, we have proposed a modified multitask cascade convolutional neural network (M-MTCNN) technique-based face detection which is presented in this paper. It is much more challenging to locate partial occluded and pose variation faces. The presented method is validated over on WIDER face database to validate severe occlusions.

In case of images of individuals in different orientations (images with half faces and other occlusions), only identifying the facial landmarks of the face is not enough, we have to clearly specify the distinct features of the face (i.e., eyes, ears, nose, and lips). So, here, in this paper, we have tried to achieve that part.

The rest of the paper is organized as follows: Section 2 defines proposed method using modified MTCNN. The experimental results on the face database are in Sect. 3. Finally, concluding remarks are summarized in Sect. 4.

2 Proposed Modified MTCNN Approach

It is suggested in traditional procedures due to the efficiency of the cascade structure. MTCNN is a face detection variation of the multitask neural network paradigm. MTCNN first uses a simple model to generate target region candidate boxes with a certain probability [12], in order to account for accuracy and performance, and then uses a more complex model for fine classification and higher precision region box regression, which is then made recursive to form a three-layer network, namely P-net, R-net, and O-net, to perform fast and efficient face detection. In the input layer, an image pyramid is utilized to change the scale of the original image. Then, using P-net, a large number of candidate target area frames is generated. The R-net is then employed for the initial selection and boundary regression of these target area frames, and the majority of the negative cases are filtered out. The remaining target area frames are then differentiated and regressed using the more composite and accurate network O-net. Here, we have proposed a modified MTCNN to find the modified landmarks of the eyes and lips areas of a face.

2.1 MTCNN Basic Network Structure

The original image is downsized after it has been processed using MTCNN to build a pyramid of images of various sizes. The different-sized images are then given to the three sub-networks for training to detect different human face sizes and detect multi-size targets. The entire process of P-net, which is the first sub-net in MTCNN, is depicted in Fig. 1. The basic structure is a completely connected convolution. The picture pyramid built in the previous stage of the method is utilized to bring out the prefatory features, scale the frame using an FCN, and produce an approximate estimate of the face candidate frame and frame regression vector. The candidate frames are then regressed by the frame, and the high-coincidence candidate frames are finally merged using the NMS algorithm. The bounding box regression adjustment window and NMS are used to filter the majority of the windows. The key difference between the FCN structure and the general convolution network is that the convolution kernel size is 11, which allows the network to take images of any size as training set samples. The deep learning training network model, in particular, requires a high number of training sets, and thus, the model training is more efficient since it eliminates the recurrent storage and convolution problems generated by the usage of pixel blocks. This feature of efficient model training can help you save time and improve your results. Its structure is more sophisticated than the above layer's P-net network structure. The constraint conditions are mostly added, and the added restriction scans the face prediction frame again.

The R-net network refines the higher layer's output window further, using the border regression process and the NMS algorithm to reject low-scoring face candidate

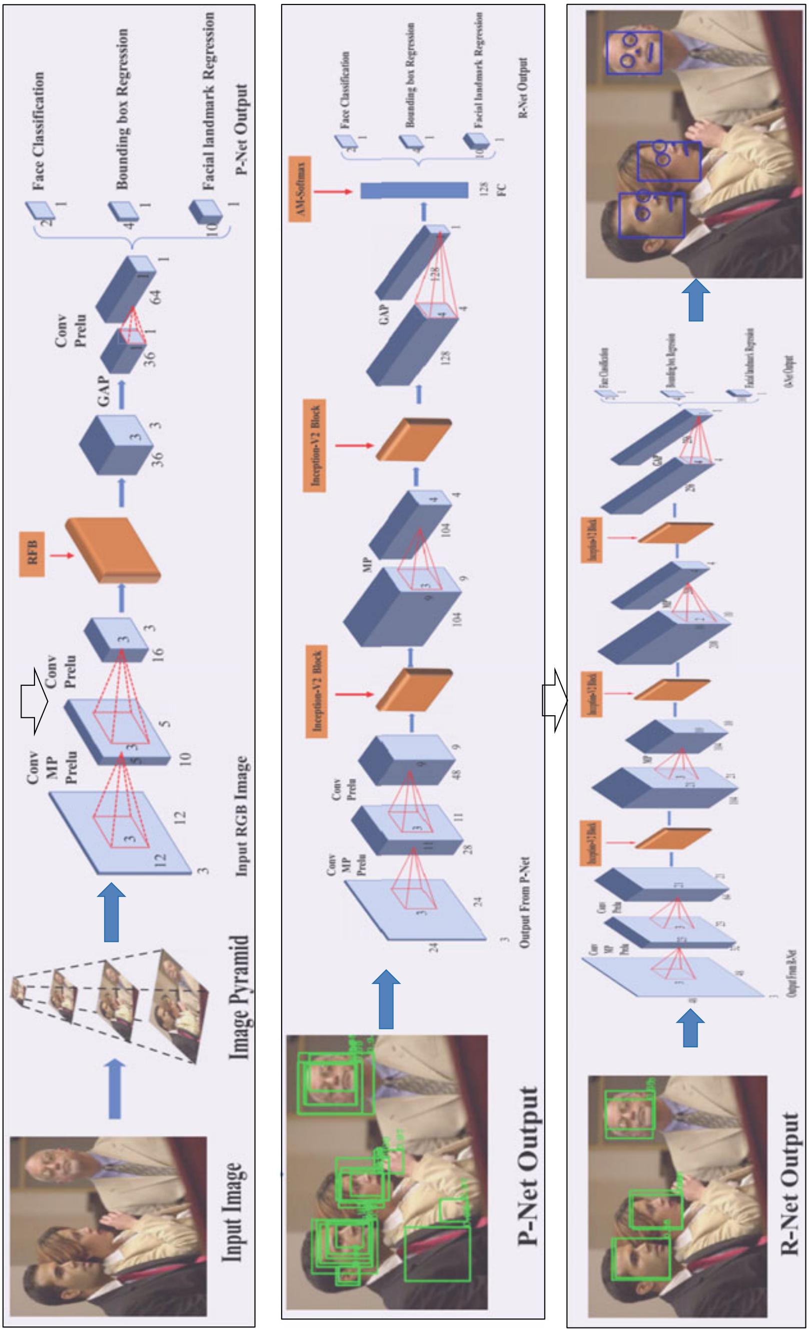

Fig. 1 Schematic diagram and the steps for the proposed method face detection method with M-MTCNN

frames and pick numerous sets of locally optimal face candidate frames. In comparison to the P-net network, R-net has an extra fully linked layer at the end, as seen in Fig. 1. The extra fully connected network's purpose is to provide a 128-dimensional vector that aids the R-net in filtering the prediction box's output.

O-net stands for output network in its entire form. The goal of the O-net network structure is to pick the best candidate frame again and produce the five feature important points that were discovered in the end. We have made changes to the landmarks, which will be addressed in the next part about the face detection procedure. O-network net's layer is deeper than the previous layer's, which explains why O-net's reaction to face identification is the best.

The size of filters is reduced from 5 × 5 to 3 × 3. The numbers of filters are also reduced. These reduction changes are done in order to lessen the computational strain. To boost the performance, the depth of the network is increased.

(a) **Proposal network (P-net)**: *classification of face*: To achieve learning, the problem is formulated as a two-class classification problem. For each sample x_i, we use the cross-entropy loss:

$$L_i^{\mathrm{Cl}} = -\left(y_i^{\mathrm{Cl}} \log(p_i) + \left(1 - y_i^{\mathrm{Cl}}\right)(1 - \log(p_i))\right) \tag{1}$$

where p_i is the probability made by the neural network that states a face sample. The symbolization representation $y_i^{\mathrm{Cl}} \in [0, 1]$ signifies the ground truth label.

(b) **Refine network (R-net)**: Using *R-net*, highly coincided detection windows in the facial image are marked; non-maximum suppression is used to scrap 90% of the coincided windows. The formula is given by:

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

(c) **Output network (O-net)**: Using O-net suggests more regulations to track the face region. Most significant of all, this stage points out five facial features. The Euclidean loss is defined below in Eq. (3); this is used to rectify the regression issue encountered during facial features detection.

$$L_i^{\mathrm{Lndmrk}} = \| \hat{y}_i^{\mathrm{Lndmrk}} - y_i^{\mathrm{Lndmrk}} \|_2^2 \tag{3}$$

In Eq. (3), the coordinates of facial features that correlate to trained network and real condition for the ith input image are denoted by $\hat{y}_i^{\mathrm{Lndmrk}}$ and y_i^{Lndmrk}. The five feature points are left eye, right eye, left mouth, right mouth, and nose. The facial features compose of these facial features.

(d) **Multi-source training**: Each CNN in the network has a different function; therefore, during the learning process, the MTCNN network is fed with different types of training images, like non-face and fragmentary aligned face. In this case, a number of the loss functions (i.e., Eqs. (1)–(3)) are not used. For instance, only the first two losses of the sample for background region are computed, and other two losses are initialized to 0. This can be carried out directly with a sample-type indicator. Then, the overall learning target can be formulated as:

$$\min \sum_{i=1}^{N} \sum_{j \in (\mathrm{Cl}, \mathrm{Rg}, \mathrm{Lndmrk})} \alpha_j \beta_i^j L_i^j \tag{4}$$

where N is the number of training samples. α_j states on the task significance. We use $\left(\alpha|\mathrm{Cl} = 1, \alpha_{\mathrm{Rg}} = 0.5, \alpha_{\mathrm{Lndmrk}} = 0.5\right)$ in *P-net* and *R-net*, while $\left(\alpha|\mathrm{Cl} = 1, \alpha_{\mathrm{Rg}} = 0.5, \alpha_{\mathrm{Lndmrk}} = 1\right)$ in *O-net* for more perfectly facial landmarks localization. $\beta_i^j \in \{0, 1\}$ denotes the sample-type indicator. In this process, it is very common to employ stochastic gradient descent to train the CNN.

2.2 Face Detection Process

In each of the three network, i.e., *P-net*, *R-net,* and *O-net*, both training and testing phases are carried out. At first, training the *P-net* in the dataset is randomly crop facial images and resized the cropped images to 12 × 12. Then, the cropped image was determined as positive or negative sample based on the *Intersection* over *Union* (IOU) ratio of the box to ground truth. Secondly, while training the *R-net*, images in the dataset are detected with a trained *P-net* model; each image generated a large number of candidate windows. For each candidate window, according to its IOU with ground truth, this candidate window was determined to be a positive and negative sample. After that, resizing these windows to 14 × 14 and train *R-net*. Finally, similar to the processing of training *R-net*, the trained *R-net* model was used to generate candidate windows, the candidate windows were determined to be positive and negative samples according to its IOU with ground truth. Finally, resized these windows to 48 × 48 and train *O-net*. The steps for training the proposed M-MTCNN are shown in Fig. 1. When conjecture was performed, first of all, image pyramid containing images of different sizes of images were generated. The candidate bounding boxes and scores were initially produced by *P-net*. And then, candidate bounding boxes with large overlap are found through NMS. Next, merge overlapped candidates of different scales. Secondly, image was detected with the *P-net* model, and the detected candidate window of the face was converted into the square boxes. Afterward, these square boxes in the original image were converted to new boxes starting at 0 coordinates and resize the new boxes to 24 × 24. Subsequently, the *R-net* model was used to detect these new boxes and get *R-net*'s candidate windows of the face and scores. After that, the overlapped candidate windows were merged with NMS. Finally, similar to *R-net*, the *O-net* model was used to detect these new boxes and output bounding boxes and scores. The steps for face detection with M-MTCNN are shown in Fig. 1. Apart from only drawing the square boxes around the faces, using the landmark points in the eyes, we have drawn circles around the eyes so that it can be used to detect the whole portion of the eyes. We have also drawn a straight line along the lips area to identify the lips area more prominently.

3 Empirical Results

The performance of the proposed modified version of the MTCNN face detection method is validated on WIDER FACE [16] dataset. The WIDER dataset is a challenging dataset and is widely used to study the problem of unrestricted face detection. It contains 393,703 faces with a high degree of variability in scale, poses, and occlusion. We choose WIDER FACE as the training datasets for training the proposed M-MTCNN. Figures 2, 3 clearly demonstrates the results of faces detection from WIDER FACE [16] dataset achieved by our algorithm. The results also contain the modified landmarks as discussed. The efficiency of the proposed method is confirmed by WIDER FACE databases which can also handle the situations like size disparities of face, occlusions, and out of plane rotation. The periodical face detection is done to ensure recovery from errors. Detection of multiple faces with pose variant can easily be handled by using M-MTCNN.

The inclusion of landmarks (circles around the eye region, and straight line specifying the lip region) in the face detection process helps in more accurate face detection process of occluded images (occlusion such as images of individuals wearing hats, mask, glasses, and sunglasses.).

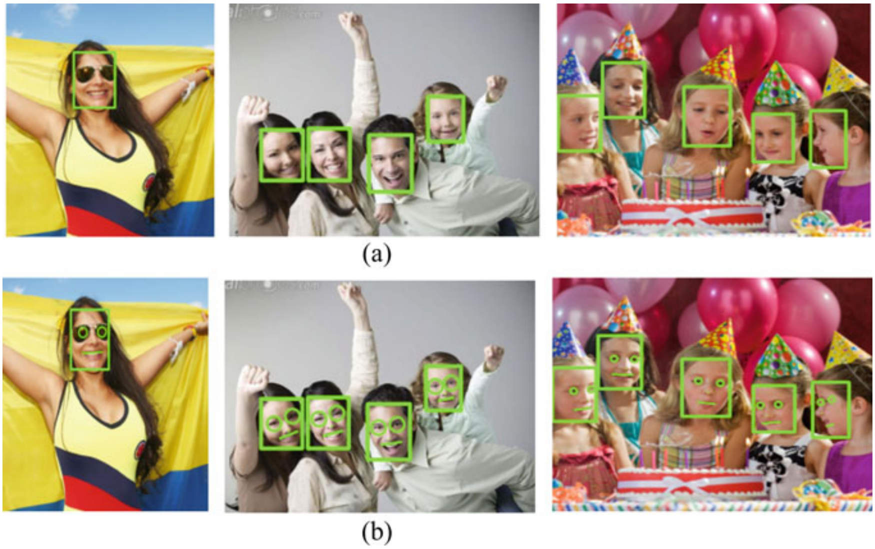

Fig. 2 Comparison with **a** MTCNN and **b** proposed MTCNN algorithm, pose, and face orientation variant locations detected by the green line rectangles using our proposed methods

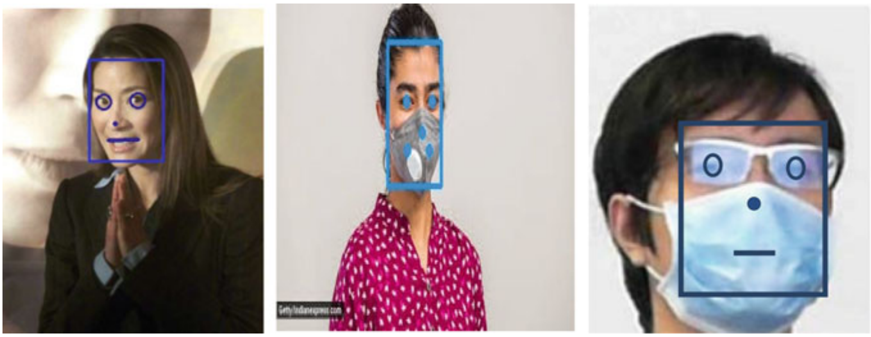

Fig. 3 Example image for face detection which contains human and picture of same human and occluded face image. The rectangles with blue line are the results using our methods

However, the detection results are strongly affected when the object environment is complex or with heavy occlusion. In order to further verify and analyze the performance of the model, we divided the test images into three categories with respect to the complexity of the environment. These three levels were easy, medium, and hard, and we carried out further experiments on these three levels. Figures 2, 3 show the results of the experiments at easy, medium, and hard levels, respectively.

4 Conclusion

We have presented an efficient modified MTCNN method to address the problems incurred during the detection of faces involving occlusion (glasses and mask scarf) and pose variance. Firstly, we have fed the modified MTCNN with input images, and then, the MTCNN can quickly generate the candidate windows. Secondly, modified MTCNN can truly detected aces that partially occluded and with different pose variations. With the modified landmarks in the detected faces. The proposed M-MTCNN model was trained with a sufficient large number of face training examples that include most partial occlusions and non-partial occlusions faces, to detect multi-view partially occluded and non-partially occluded faces efficiently. We have evaluated our proposed method on WIDER face datasets. The experimental results clearly show how accurately our model detects faces and the modified facial landmarks.

References

1. Dey, A., Chakraborty, S., Kundu, D., Ghosh, M.: Elastic window for multiple face detection and tracking from video. In: Proceeding of the CIPR 2019, pp. 487–496 (2019)
2. Dey, A.: A contour based procedure for face detection and tracking from video. In: Proceeding of the RAIT 2016, pp. 252–256 (2016)

3. Chowdhury, S., Dey, A., Sing, J.K., Basu, D.K., Nasipuri, M.: A novel elastic window for face detection and recognition from video. In: Proceeding of the ICCICN 2014, pp. 252–256 (2014)
4. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the CVPR 2001, pp. 511–518 (2001)
5. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection computer vision and pattern recognition. In: Proceedings of the CVPR 2015, pp. 5325–5333 (2015)
6. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responses to face detection: a deep learning approach. In: Proceeding of the ICCV, pp. 3676–3684 (2015)
7. Li, X., Yang, Z., Wu, H.: Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks. IEEE Access **8**, 174922–174930 (2020)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the CVPR 2014, pp. 580–587 (2014)
9. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
10. Nie, L., Wang, X., Zhang, J., He, X., Zhang, H., Hong, R., Tian, Q.: Enhancing micro-video understanding by harnessing external sounds. In: Proceedings of the ACMM 2017, pp. 1192–1200 (2017)
11. Song, X., Feng, F., Han, X., Yang, X., Liu, W., Nie, L.: Neural compatibility modeling with attentive knowledge distillation. In: Proceeding of the SIGIR'2018, pp. 5–14 (2018)
12. Tang, X., Du, D.K., He, Z., Liu, J.: PyramidBox: a context-assisted single shot face detector. In: Proceedings of the ECCV 2018, pp. 812–828 (2018)
13. Huang, L., Yang, Y., Deng, Y., Yu, Y.: DenseBox: unifying landmark localization with end to end object detection. arXiv:1509.04874
14. Yang, S., Luo, P., Loy, C.C., Tang, X.: Faceness-Net: face detection through deep facial part response. IEEE Trans. Pattern Anal. Mach. Intell. **40**(8), 1845–1859 (2018)
15. Zhan, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascade convolutional networks. IEEE Sig. Proc. Lett. **23**(10), 1499–1503 (2016)
16. Yang, S., Luo, P., Loy, C.C., Tang, X.: WIDER FACE: a face detection benchmark. In: Proceeding of the CVPR 2016, pp. 5525–5533 (2016)