A HIGHLY ACCURATE FACIAL REGION NETWORK FOR UNCONSTRAINED FACE DETECTION

Han Shu, Dangdang Chen, Yali Li, Shengjin Wang

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ABSTRACT

In this paper, a new face detection method with very high accuracy is proposed. We introduce a novel facial region network to detect faces in unconstrained conditions. Firstly, a face proposal net is raised to generate possible face regions in the input image. Then, a novel weighted grid feature is applied to calculate features of face regions. Owing to that, faces with large pose variation and severe occlusion can be detected correctly. Furthermore, we use millions of general object data to pre-train the network to enhance the robustness of the extracted feature. Our method is evaluated on several public face detection datasets and achieves state-of-the-art performance on all of them. Specially, our method demonstrates a very high recall rate of 96.4% when false positives are 300 on the challenging FDDB benchmark, ranking first not only in the academic list but also in the commercial list which is much more competitive than the previous one.

Index Terms— Face detection, CNN, face proposal net, weighted grid feature

1. INTRODUCTION

Face detection has been one of the most important research fields in computer vision over the past decades. Since Viola and Jones [1] published the seminal work, Haar feature and AdaBoost classifier have been a dominating method for face detection. Many researchers have developed extended methods originating from their work. SURF [2] follows the structure and extends the feature style. Joint Cascade [3] presents a local binary feature for both face detection and alignment. Another series of classic face detection methods are based on Deformable Part Model (DPM) [4]. With the model, occluded faces can be partly handled [5]. However, those face detectors which mainly depend on hand-crafted features are confined to frontal faces or profile faces with small angles. They are incapable of detecting faces in unconstrained situations which are heavily distracted by pose, occlusion, expression and illumination.

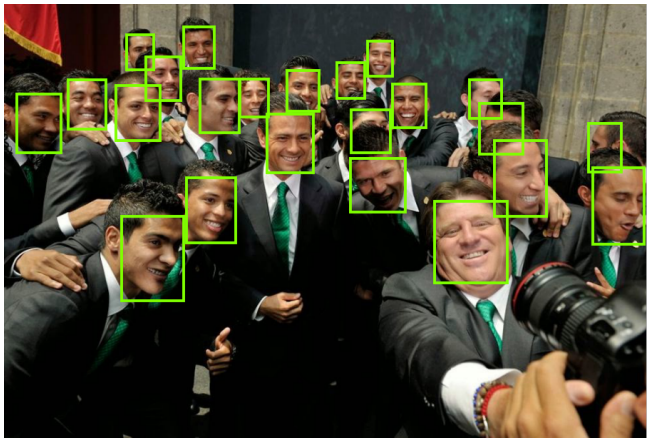


Fig. 1. All the faces in the image with various poses and severe occlusion can be accurately detected without any false alarm using our proposed method.

Recently, deep convolutional neural network (CNN) has been a great success in visual recognition [6]. Face detectors based on CNN have been proved to be more effective than the primary ones. Cascade CNN [7] exploits a cascade framework composed of CNN to get rid of false positives. DenseBox [8] introduces a fully convolutional neural network and predicts a pixel-wise bounding box of faces. Nevertheless, both methods conduct sliding window or pixel-wise non-maximum suppression (NMS) for bounding box prediction, both of which are computationally expensive for face detection task.

In more recent years, improvements in general object detection and classification have attracted more attention. R-CNN [9], Fast R-CNN [10], Faster R-CNN [11] and R-FCN [12] have made subsequent progress. The network in [11] is composed of two parts, the Region Proposal Network (RPN) and the Fast R-CNN object detector. The first part is to generate object proposals and the second one is to refine and classify the proposals. The greatest advantage of the structure

This work was supported by the Initiative Scientific Research Program of Ministry of Education under Grant No. 20141081253 and the state key development program in 13th Five-Year under Grant No. 2016YFB0801301.

is the end-to-end training and testing process of object detection and classification. It also shows superior performance when applied to face detection task [13].

Motivated by the general object detection network [12], we propose a precise and robust face detection method in this paper. The method shows a high capability of dealing with faces with large pose variation and severe occlusion as shown in Fig. 1. The main contributions are as follows: (1) A face proposal net is constructed for generating region proposals of faces, much more accurate and efficient compared with the usually conducted methods like sliding window. (2) A novel weighted grid feature which is robust to represent faces with variant poses and severe occlusion. (3) A broaden of training data. A large quantity of various general object data are utilized in the training process of the face detection network as it is derived from the general object network. Thus more valid features can be learned through the training process. Our method has shown effectiveness and achieves state-of-the-art performance on several public face detection datasets.

In Section 2, the method will be explained in details. The experimental results on public datasets will be presented and analyzed in Section 3. The work will be summarized in Section 4.

2. METHOD

2.1. Overview

Inspired by the general object detection network of R-FCN [12], we design a face region neural network for face detection. Fig. 2 depicts our proposed face detection framework. The input image is fed into a deep fully convolutional network. The ResNet [14] is selected as the baseline network due to its efficient feature extraction ability. Through the network, we can get a feature map concentrating on face regions of the input image. Then, face bounding boxes will be predicted by feeding the feature map into the face proposal net. Based on the feature map of the bounding box, the weighed grid feature is calculated to classify the region being a face or not. The entire framework is an end-to-end structure for face detection.

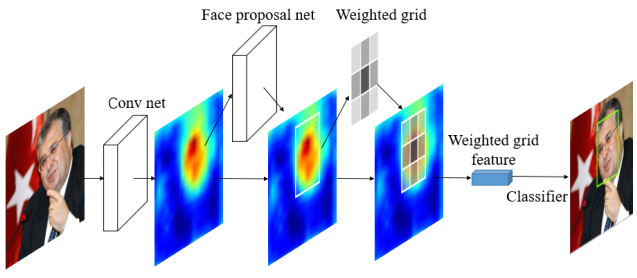


Fig. 2. Overview of the face detection network.

2.2. Face Proposal Net

Considering that face shape has a unique distribution, a region proposal network specialized for faces is constructed. We analyze the training set of WIDER FACE dataset [15], which includes 12,881 pictures and 158,988 faces with a high degree of variability in scale, pose and occlusion. For the reason of the face regions labeled by rectangle, the height/width ratios of these rectangles are counted to indicate the distribution of face shape. A normal distribution is applied to approximate the distribution. In Fig. 3, the blue bars represent the statistical ratios of height/width and the red curve represents the approximate normal distribution. The function of the approximate distribution is described as:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (1)$$

where x is the height/width ratio of face rectangles, μ is approximately 1.30, σ is approximately 0.22. It suggests that the majority of height/width ratios lie in a short interval, the center of which is approximately 1.30.

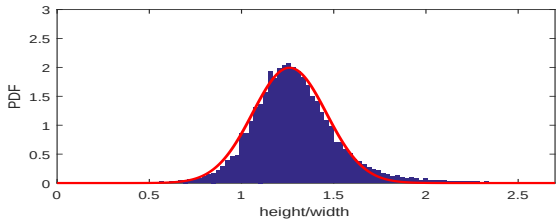


Fig. 3. The distribution of face height/width ratios.

In view of the above analysis, we put forward a face proposal net to generate face region proposals similar to [11]. Fig. 4 displays the key idea. On the feature map, a 3x3 convolutional filter is conducted which has the identical functionality as sliding window while in a much smaller range. Each location in the convolutional filter is mapped to N rectangle boxes in the original image. These boxes have the uniform height/width ratio of 4:3 which is not only convenient for calculation but also close to the statistical distribution center μ of height/width ratio. On the basis of each convolution location, a fixed-length feature is calculated to conduct a bounding box regression from a designed box to the ground truth region as well as predict the possibility of the region containing a face. To cover the majority of the facial scale range, we choose 5 sizes of 4:3 rectangle boxes with height of 16, 32, 64, 128, 256 pixels since the input image is resized to the width of 300 pixels in the implementation.

2.3. Weighted grid feature

We raise a novel weighted grid feature to calculate features of potential face regions. Different facial parts provide distinct information for humans to recognize a face. The similar idea

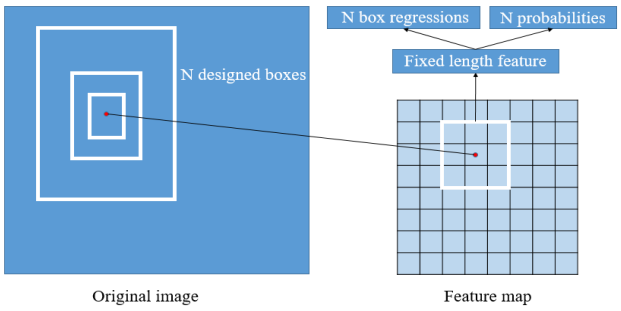
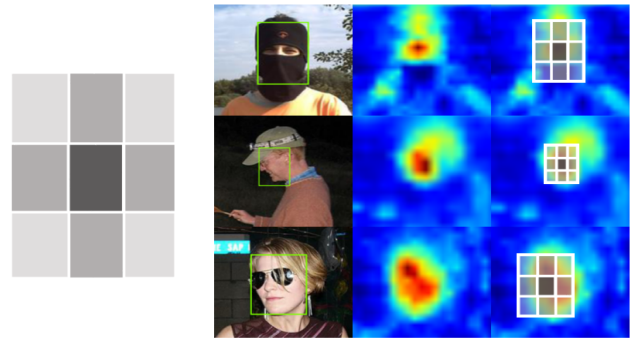


Fig. 4. The key idea of face proposal net.

goes for neural network. Each part unequally devotes to the detection of faces. Based on the analysis of the feature map, we discover that the central part of proposal region has higher response than other parts. The response declines with the distance to the central part, which is consistent with human cognition that central part attracts more attention. Therefore, the feature map of a face region can be divided into several parts and assigned with different weights.



(a) Weighted grid (b) Examples of weighted grid feature

Fig. 5. The visualization of weighted grid feature.

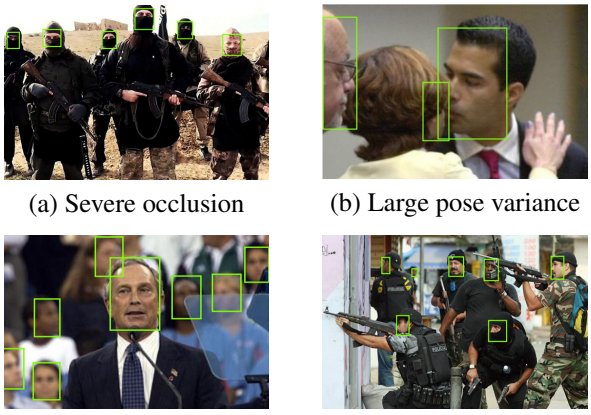
For the convenience of implementation, we divide the feature map of face regions into $n \times n$ grids and weight the grids by its distance to the center of the bounding box. Fig. 5 (a) shows the structure of the weighted grids. The darker a grid is, the greater the weight is. Then we compute the grid weighted feature defined as:

$$F = \frac{\sum_{i=1}^{n^2} w_i g(f_i)}{\sum_{i=1}^{n^2} w_i} \quad (2)$$

where F is the weighted grid feature of a predicted face region, w_i is the weight of the grid, f_i is the feature map of the grid, g can be average or maximum function. In our implementation, n is chosen to be 3. Fig. 5 (b) describes three examples of weighted grid feature on the feature map. It reveals that high response part is given greater weight when calculating features, which strengthens the significant part and suppresses the noisy ones.

3. EXPERIMENTS

As Olshausen and Field [16] have proved that complex graphics are composed of some basic structures, faces and general objects share the same basic structures which are learned in the early stage of CNN. We pre-train the face detection network with the training set of the large-scale ImageNet database [17]. There are more data to utilize. What is more, a large amount of general data can avoid a sample bias to the training set, leading to a more robust feature extraction network. Then, we fine-tune the network with the training set of WIDER FACE dataset [15]. The method is evaluated on several public face detection datasets and achieves state-of-the-art performance. Some detection results are shown in Fig. 6. It is noteworthy that our method adopts identical parameters during the test of various datasets while some methods need varied parameters to adapt to different datasets. We use the open-source deep learning framework Caffe [18] to implement our work and evaluation toolbox provided by [19] to plot curves.



(a) Severe occlusion (b) Large pose variance

(c) Low resolution and blur (d) Mixed situation

Fig. 6. Examples of detection results. Our method is capable of handling faces with severe occlusion, large pose variance, low resolution and blur.

3.1. Experiment on FDDB

Face Detection Data Set and Benchmark (FDDB) [20] includes 5171 faces in 2845 pictures. It is one of the most challenging datasets which contains severe occlusion, blur and large pose changes. Many academic institutions and commercial companies are competing on the dataset for a better result. Our method achieves the recall rate of 96.4% when false positives are 300, ranking first not only among the published methods but also among the unpublished methods. The final

Table 1. The recall rates of commercial methods on FDDB (300 false positives)

Team	Recall	Team	Recall
Ours	**0.964**	Daream	0.946
BAIDU-IDL-v4	0.960	360-NUS	0.938
Xiaomi	0.958	Uniview	0.926
MT-Face-v2	0.958	Linkface	0.908
Deep IR	0.956	Tencent Best-Image	0.860
Emotibot	0.947	Face++	0.826

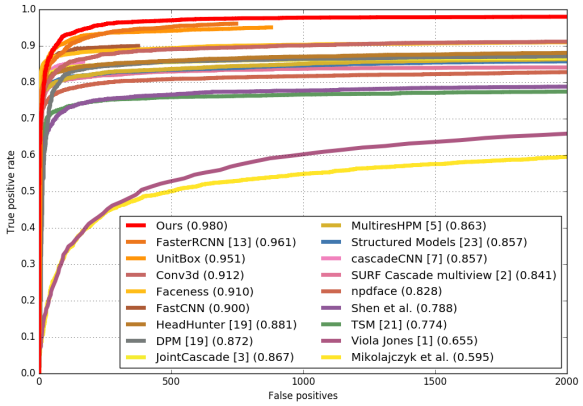


Fig. 7. Academic discrete curves on FDDB dataset.

recall rate is over 98.0%. Fig. 7 shows the true positive rate versus false positives curves of the academic methods. Table 1 shows the recall rates of competitive commercial methods when false positives are 300.

3.2. Experiment on AFW

Annotated Faces in the Wild (AFW) [21] is built using Flickr images. It has 205 images with 473 labeled faces. The same evaluation metric employed in the PASCAL VOC dataset [22] is adopted. Our method achieves the average precision (AP) of 99.03% and is the only one exceeding the AP of 99%. Fig. 8 shows the precision-recall curves on AFW dataset.

3.3. Experiment on PASCAL Faces

PASCAL Faces dataset [23] contains 851 images extracted from PASCAL VOC dataset [22] and has 1635 faces. Our method achieves the AP of 97.78% and outperforms the state-of-the-art by over 5% AP respectively. Fig. 9 shows the precision-recall curves on PASCAL Faces dataset.

3.4. Analysis on Experiments

The method is evaluated on FDDB, AFW, PASCAL Faces datasets and achieves state-of-the-art performance on all of them. Compared to other methods, our method takes advantages of prior information of face shape and concentrates on the significant facial region. Besides, millions of general object images are employed to learn a robust network. Thus, the method can handle various faces in unconstrained situations.

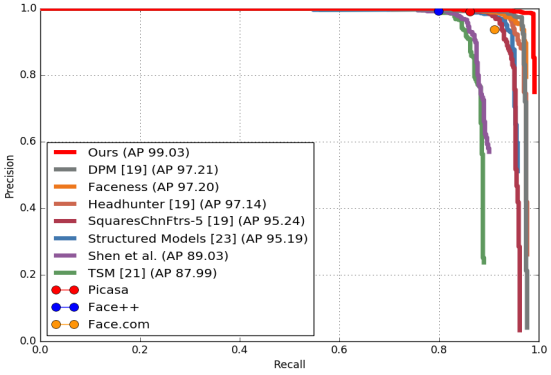


Fig. 8. Precision-recall curves on AFW dataset.

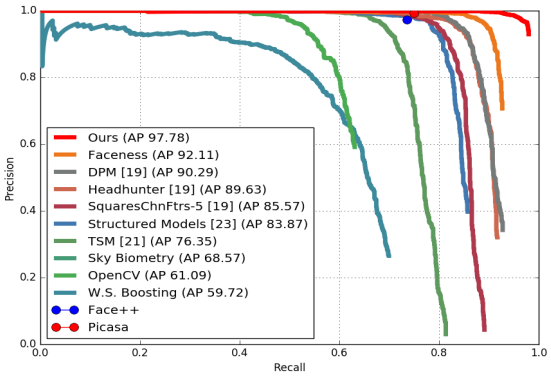


Fig. 9. Precision-recall curves on PASCAL Faces dataset.

4. CONCLUSION

In this paper, we propose a highly accurate facial region network for face detection. More specifically, we raise the face proposal net to generate potential bounding boxes on the feature map. Moreover, a novel weighted grid feature is proposed to calculate feature of the face region. Through a pre-train process with large amounts of general object data, the network can obtain more robust feature representations and detect various faces in unconstrained conditions. Benefits from the carefully designed network structure and huge amount of training data, the proposed method demonstrates high precision of face detection and achieves state-of-the-art accuracy on the challenging datasets of FDDB, AFW and PASCAL Faces.

5. REFERENCES

[1] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. IEEE, 2001, vol. 1, pp. I–511.

[2] Jianguo Li and Yimin Zhang, "Learning surf cascade for fast and accurate object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3468–3475.

[3] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun, "Joint cascade face detection and alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 109–122.

[4] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[5] Golnaz Ghiasi and Charless C Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2385–2392.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325–5334.

[8] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874*, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[12] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, "R-fcn: Object detection via region-based fully convolutional networks," *arXiv preprint arXiv:1605.06409*, 2016.

[13] Huaizu Jiang and Erik Learned-Miller, "Face detection with the faster r-cnn," *arXiv preprint arXiv:1606.03473*, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[15] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark," *arXiv preprint arXiv:1511.06523*, 2015.

[16] Bruno A Olshausen and David J Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[19] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool, "Face detection without bells and whistles," in *European Conference on Computer Vision*. Springer, 2014, pp. 720–735.

[20] Vidit Jain and Erik Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," Tech. Rep. UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[21] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[23] Junjie Yan, Xuzong Zhang, Zhen Lei, and Stan Z Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.