Single Shot Attention-Based Face Detector

Chubin Zhuang, Shifeng Zhang, Xiangyu Zhu, Zhen Lei*, and Stan Z. Li

CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China
{chubin.zhuang,shifeng.zhang,xiangyu.zhu,zlei,szli}@nlpr.ia.ac.cn

Abstract. Although face detection has taken a big step forward with the development of anchor based face detector, the issue of effective detection of faces with different scales still remains. To solve this problem, we present an one-stage face detector, named Single Shot Attention-Based Face Detector (AFD), which enables accurate detection of multi-scale faces with high efficiency, especially for small faces. Specifically, AFD consists of two inter-connected modules, namely attention proposal module (APM) and face detection module (FDM). The former aims to generate the attention region and coarsely refine the anchors. The latter takes the output from APM as input and further improve the detection results. We obtain state-of-the-art results on common face detection benchmarks, *i.e.* FDDB and WIDER FACE, and can run at 20 FPS on a Nvidia Titan X (Pascal) for VGA-resolution images.

Keywords: Face detection, Attention mechanism, Single shot

1 Introduction

Face detection is a fundamental and essential step for many face related applications, *e.g.* face recognition [1,2] and face alignment [3,4]. Since the pioneering work of Viola-Jones [5], face detection has achieved significant progress in the past few decades, especially the CNN [6] based detector. However, there are still some challenging problems: 1) The large variation of faces in cluttered backgrounds requires detectors to be more robust and accurate; 2) The large search space of possible faces further imposes a serious challenge of trade-off between accuracy and efficiency, especially for small faces.

To address these problems, recent CNN based face detectors can be divided into two categories. One is cascade based methods, such as CascadeCNN [7] and MTCNN [8]. The other is anchor based methods [9,10]. However, these two kinds of methods focus on different aspects. The former can well handle faces with diverse scales, but tends to be much time-consuming when the number of faces is large. While the latter's speed is invariant to the object number, but the performance will drop dramatically as the objects getting smaller as indicated in [11]. Therefore, efficient detection of multi-scale faces is still one of the critical issues that remains to be settled.

* Corresponding author

To solve these two conflicting issues, our core idea is introducing attention mechanism to detection and leveraging feature fusion of different layers as RefineDet [12] so as to highlight the possible facial regions and enrich the feature information to promote the performance of detection. Specifically, we present an effective face detector called Single Shot Attention-Based Face Detector (AFD), which consists of two inter-connected modules, respectively attention proposal module (APM) and face detection module (FDM). The former aims to generate the attention region and coarsely refine the anchors. The latter takes attention maps as input and dot-multiply them with the feature maps to highlight the features from the facial region. Then the feature maps from the high-level layers are integrated into the low-level layers to increase the richness of feature information and output the final detection results.

Due to the attention mechanism in APM and the fusion of features in different layers in FDM, our face detector can well address the dramatic deterioration problem of anchor based detectors as faces getting smaller, especially for small faces. Consequently, for VGA-resolution images, our face detector can run at 20 FPS on a NVIDIA Titan X (Pascal) GPU in inference. Besides, we comprehensively evaluate this detector and demonstrate state-of-the-art detection performance on several common face detection benchmark datasets, including the FDDB [13] and WIDER FACE [14]. For clarity, the main contribution of this work can be summarized as three-fold:

- We propose an attention mechanism to enhance the robustness and performance of detectors.
- We leverage feature fusion in different detection layers to enrich the feature information.
- We achieve state-of-the-art performance on common face detection benchmarks and keep the efficiency.

2 Related Work

Face detection has attracted large research attention in past few decades, which can be roughly divided into two categories. One is hand-craft based detectors, and the other is built on CNN. This section briefly reviews these two methods.

Hand-craft based methods. Following the milestone work of Viola-Jones face detector [5], most early methods pay attention to designing robust features [15] and training effective classifiers [16]. Besides, the deformable part model [17] is introduced into face detection task by [18] and achieves remarkable performance. However, these detectors highly depend on non-robust hand-craft features, thus they are efficient but not accurate enough for the large visual variation of faces.

CNN based methods. Recent years have witnessed the advance of CNN based detectors. CascadeCNN [7] employs a cascade structure to detect face in a coarse to fine way. MTCNN [8] proposes an architecture to address both detection and landmark alignment jointly. Besides, the anchor based methods originated from Faster-RCNN [19] structure have achieved great progress in past

few years. Jiang et al. [20] apply Faster R-CNN framework in face detection and achieves promising results. SSD [9] introduces multi-scale mechanism to anchor designing. S^3FD [10] proposes anchor matching strategy to improve the recall rate of small faces. FPN [21] proposes a top-down structure to use high-level semantic feature maps at different scales. FAN [22] introduces anchor-level attention to improve the detection performance. RefineDet [12] develops a single-shot inter-connected architecture to improve the performance of detector while maintain the high efficiency.

Generally, the hand-craft based methods tend to be efficient but less accurate. While the CNN based methods dominate the performance but present less efficiency. Notably, our proposed AFD is able to achieve state-of-the-art performance and keep the high efficiency.

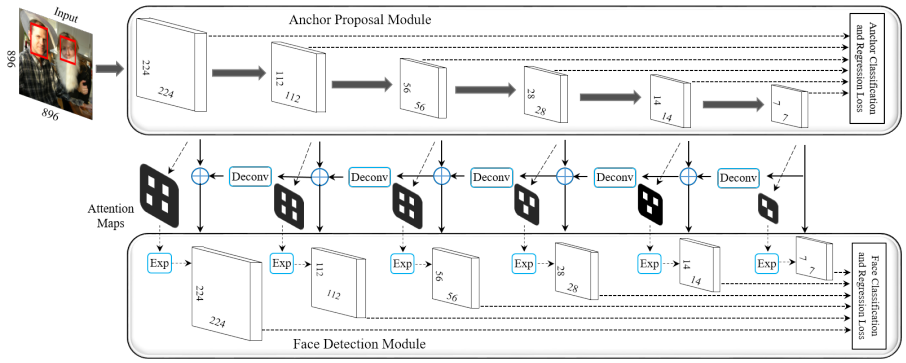


Fig. 1. The framework of AFD. We only display the layers used for detection. The parallelograms denote the attention maps associated with different feature layers and the white rectangles represent the possible facial regions.

3 Single Shot Attention-Based Face Detector

This section introduces the details of AFD. It includes three components: the overall network architecture, loss function and some implementation details.

3.1 Overall network architecture

Anchor-based object detection frameworks with reasonable design of anchors in different layers have proven to be effective to handle faces with different scales [9]. As illustrated in Fig. 1, the architecture of AFD uses the same extended VGG16 backbone and anchor design strategy as [10], which can generate feature maps at different layers and anchors with equal-proportion interval. Besides, the attention proposal module and face detection module is added on this backbone to get the final results.

Attention proposal Module (APM). We use the APM to roughly predict the locations and scores of anchors from different layers as the attention supervision information, and construct the attention maps which indicate the possible facial regions based on these supervision information. Then these hierarchical attention maps are sent to FDM to highlight the potential face areas. Besides, we also leverage the predicted information of different anchors in APM to coarsely refine the anchors as [12], which provides better initialization for the detection in the FDM.

Face Detection Module (FDM). After obtaining these attention maps, we feed the attention maps to an exponential operation to rescale the value of score from 1 to e and only take the maximum score for the overlapping parts, which will not only highlight the detection information, but also maintain more context messages. Then these attention maps are dot-multiplied with the feature maps to highlight the facial regions. Besides, we follow the design in RefineDet [12] to integrate high-level semantic features into low-level layers with higher resolution by adding the high-level features to the transferred features, which will greatly enrich the feature information of different layers. To match the dimensions between them, the deconvolution operation is used to enlarge the high-level feature maps and sum them in the element-wise way. Then we pass the refined anchor boxes to the corresponding feature maps in the FDM to further generate accurate face locations and sizes.

3.2 Loss function

The loss function for AFD consists of two parts, *i.e.*, the loss in the APM and the loss in FDM. For the APM, we assign a binary class label (of being a face or not) to each anchor and regress its location and size. After that, we send the anchors with positive confidence higher than threshold to the FDM to further predict the locations and sizes of faces. The loss function is defined as:

$$L(p_i, x_i, c_i, t_i) = \frac{1}{N_{apm}}(\sum_i L_b(p_i, l_i^*) + \sum_i l_i^* L_r(x_i, g_i^*)) + \frac{1}{N_{fdm}}(\sum_i L_b(c_i, l_i^*) + \sum_i l_i^* L_r(t_i, g_i^*)). \quad (1)$$

where i is the index of anchor in a batch, l_i^* is the ground truth class label of anchor i, g_i^* is the ground truth location and size of anchor i. p_i and x_i are the predicted confidence and coordinates of anchor i in the APM. c_i and t_i are the predicted class and coordinates of the bounding box in the FDM. N_{apm} and N_{fdm} are the numbers of positive anchors in the APM and FDM. The binary classification loss L_b is the softmax loss over two classes (face vs. background) confidences. We use the smooth L_1 loss as the regression loss L_r.

3.3 Training and implementation details

This subsection introduces the training dataset, anchor setting strategy, hard negative mining and other implementation details.

Training dataset. Our AFD is trained end-to-end on 12,880 images from the WIDER FACE [14] training set. To increase the robustness of training data, each training image is sequentially processed by color distortion, random cropping, horizontal flipping and scale transformation, and finally gets a 896 $\times$ 896 square sub-image from original image.

Anchor setting strategy. We tile the anchors scaled from 16 to 512 pixels at different detection layers as [10], and set the corresponding stride size to be a quarter of the anchor size, which gradually doubled from 4 to 128pixels. During training, we firstly match each face to the anchor with the best jaccard overlap, and then match anchors to any faces with jaccard overlap higher than 0.45.

Hard negative mining. After anchor matching step, the positive and negative training samples are extremely imbalance because most of the anchors are negative, which will make training process slow and unstable. Thus we sort these samples by the loss values and choose the top ones to make sure that the ratio between negatives and positives is almost 3:1.

Other implementation details. We initialize the parameters of the base layers from the pre-trained VGG16, and the other additional layers are randomly initialized with the "xavier" method. We fine-tune the model using SGD with 0.9 momentum, 0.0005 weight decay and batch size 12. The maximum number of iteration is 120k and we use 10^{-3} learning rate for the first 80k iterations, and continue training for 20k iterations with 10^{-4} and 10^{-5}. Our method is implemented in Caffe [23].

4 Experiments

In this section, we first analyze our model in an ablative way, then evaluate our model on the common face detection benchmarks and introduce its runtime efficiency.

4.1 Model analysis

We evaluate our model on the FDDB dataset by extensive experiments, the experiments are carried out on the same settings, except for specified changes to the components. Firstly, we redesign the network by directly using the regularly paved anchors instead of the refined ones from the APM. Secondly, we cut off the feature fusion part in FDM, and only use the independent feature maps from different layers to detect faces. Finally, we ablate the attention proposal part.

According to Tab.1, some promising conclusions can be summed up. Firstly, we find that mAP is reduced from 98.5% to 97.9%, which indicates that the refined anchor in APM can help promote the performance of detector. Secondly, cutting off the feature fusion part in FDM will deteriorate the performance (*i.e.*, 0.9%). Finally, the attention map can help the FDM highlight the features from facial region and improve the performance by 0.7%.

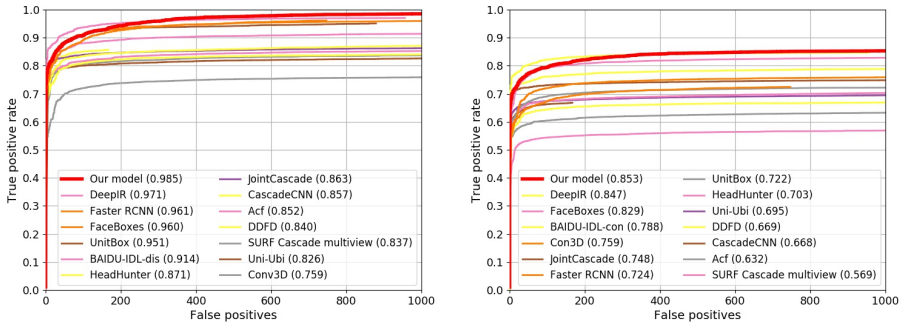
Table 1. Ablative results on FDDB (True positive rate at 1,000 false positives).

Component	AFD			
Attention?	√	√	√	
Feature fusion?	√	√		
Anchor Refined?	√			
Accuracy (mAP)	98.5	97.9	97.0	96.3

4.2 Evaluation on benchmark

We evaluate our AFD model on the common face detection benchmarks, including Face Detection Data Set and Benchmark (FDDB) [13] and WIDER FACE [14].

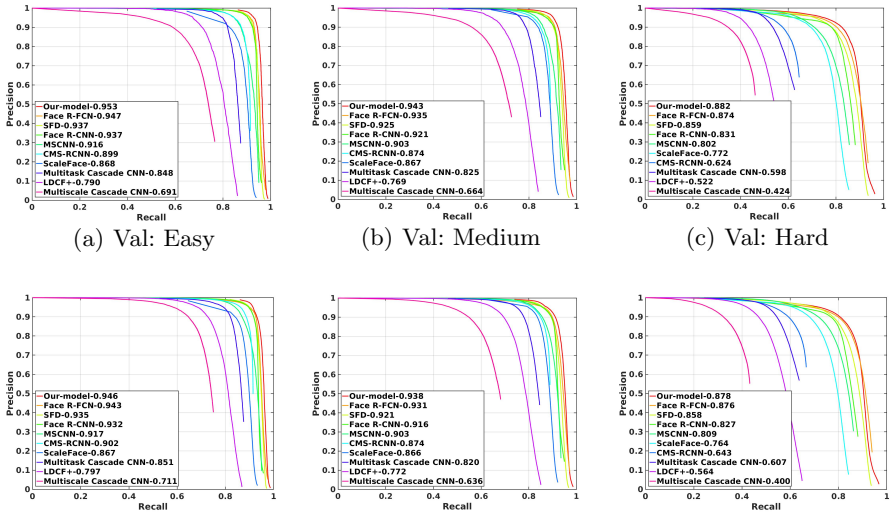
FDDB dataset. It consists of 5,171 faces in 2,845 images. Considering that FDDB uses ellipse face annotation while our model outputs rectangle bounding box. For a more fair comparison, we train an elliptical regressor to transform our predicted bounding boxes to bounding ellipses. As illustrated in Fig.2, our model achieves state-of-the-art performance.



(a) Discontinuous score curves (b) Continuous score curves

Fig. 2. Evaluation on the FDDB dataset

WIDER FACE dataset. It has 32,203 images and labels 393,703 faces with a high degree of variability in scale, pose and occlusion. These images are divided into three levels (Easy, Medium and Hard) according to the difficulties of the detection. Our model is trained only on the training set and tested on the validation and test set against recent face detection methods. As shown in Fig.3 , our model achieves state-of-the-art performance across the three subsets, *i.e.*, on validation set, 0.953 (Easy), 0.943 (Medium) and 0.882 (Hard) and 0.946(Easy), 0.938 (Medium) and 0.878 (Hard) on test set.



(a) Val: Easy (b) Val: Medium (c) Val: Hard

(d) Test: Easy (e) Test: Medium (f) Test: Hard

Fig. 3. Precision-recall curves on WIDER FACE validation and test sets

4.3 Runtime efficiency

Despite great performance, the speed of our algorithm is not compromised. The computational cost is tested on a Titan X (Pascal) and cuDNN v6.0. For a VGA-resolution image using a single GPU, our face detector can run at 20 FPS, which keeps the efficiency and owns higher accuracy.

5 Conclusion

In this paper, we present a single shot attention-based face detector, which consists of two inter-connected modules, *i.e.*, the APM and the FDM. The APM generates the attention region and coarsely refines the anchors. The FDM takes attention maps as input and dot-multiply them with the feature maps to highlight the features from the facial region. Then the feature maps from the high-level layers are integrated into the low-level layers to enrich the feature information and output the final detection results. The whole net is trained end-to-end fashion and tested on common face detection benchmarks, which achieves state-of-the-art performance and keeps high efficiency.

Acknowledgments

This work was supported by the Chinese National Natural Science Foundation Projects #61473291, #61572536, #61572501, #61573356, the National Key Research and Development Plan (Grant No.2016YFC0801002), and AuthenMetric R&D Funds.

References

1. Luan, T., Yin, X., Liu, X.: Disentangled Representation Learning GAN for Pose-Invariant Face Recognition. CVPR (2017)
2. Masi, I., Chang, F. J., Choi, J., Harel, S., Kim, J., Kim, K.G.: Learning pose-aware models for pose-invariant face recognition in the wild. PAMI (2018)
3. Xing, J., Niu, Z., Huang, J., Hu, W., Xi, Z., Yan, S.: Towards robust and accurate multi-view and partially-occluded face alignment. PAMI (2018)
4. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face Alignment Across Large Poses: A 3D Solution. IEEE Conference on Computer Vision and Pattern Recognition. CVPR (2016)
5. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV (2004)
6. Lecun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks (1995)
7. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. CVPR (2015)
8. Zhang, K., Zhang, Z., Li, Z., Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. Signal Processing Letters (2016)
9. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y.: SSD: Single Shot MultiBox Detector. ECCV (2016)
10. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S^3FD: Single Shot Scale-invariant Face Detector. ICCV (2017)
11. Huang, J., Guadarrama, S., Murphy, K., Rathod, V., Sun, C., Zhu, M., et al.: Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. CVPR (2017)
12. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. CVPR (2018)
13. Jain, V., Learned-Miller, E.: FDDB: A Benchmark for Face Detection in Unconstrained Settings. UMass Amherst Technical Report (2010)
14. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: a face detection benchmark. CVPR (2016)
15. Huang, C., Ai, H., Li, Y., Lao, S.: High-performance rotation invariant multiview face detection. PAMI (2007)
16. Li, S. Z., Zhu, L., Zhang, Z.Q., Blake, A., Zhang, H.J., Shum, H.: Statistical Learning of Multi-view Face Detection. ECCV (2002)
17. Felzenszwalb, P., Mcallester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. CVPR (2008)
18. Yan, J., Zhang, X., Lei, Z., Li, S.Z.: Face detection by structural models. Image and Vision Computing (2014)
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. NIPS (2015)
20. Jiang, H., Learned-Miller, E.: Face Detection with the Faster R-CNN. Automatic Face and Gesture Recognition (2017)
21. Lin, T. Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. CVPR (2017)
22. Wang, J., Yuan, Y., Yu, G.: Face attention network: an effective face detector for the occluded faces. arXiv: 1711.07246 (2017)
23. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. ACMMM (2014)