A Convolutional Neural Network Cascade for Face Detection

Haoxiang Li[†], Zhe Lin[‡], Xiaohui Shen[‡], Jonathan Brandt[‡], Gang Hua[†]
[†]Stevens Institute of Technology [‡]Adobe Research

In real-world face detection, large visual variations, such as those due to pose, expression, and lighting, demand an advanced discriminative model to accurately differentiate faces from the backgrounds. Consequently, effective models for the problem tend to be computationally prohibitive. A classic strategy in this setting to address the speed/accuracy tradeoff is the detection cascade, which applies complex models only to the most promising parts of the image. However, detection cascades are typically built on very simple features, which can ultimately limit discrimination effectiveness. In this paper, we propose a cascade architecture built on convolutional neural networks (CNNs) with very powerful discriminative capability, while maintaining high performance.

Specifically, we propose a CNN cascade which operates at multiple resolutions, quickly rejects the background regions in the fast low resolution stages, and carefully evaluates a small number of challenging candidates in the last high resolution stage. For improved robustness, the later detection stages effectively operate simultaneously at multiple resolutions. To improve localization effectiveness, and reduce the number of candidates at later stages, we introduce CNN-based calibration stages which operate on the output of each detection stage. The output of each calibration stage is used to adjust the detection window position for input to the subsequent stage. The proposed method runs at 14 FPS on a single CPU core for VGA-resolution images and 100 FPS using a GPU. State-of-the-art detection performance on two public face detection benchmarks demonstrates the effectiveness of the proposed face detection method.

Compared with previous work applying neural networks for face detection [5, 6, 7], the proposed method with the interweaving detection and calibration CNNs could improve the detection speed as well as the bounding box quality.

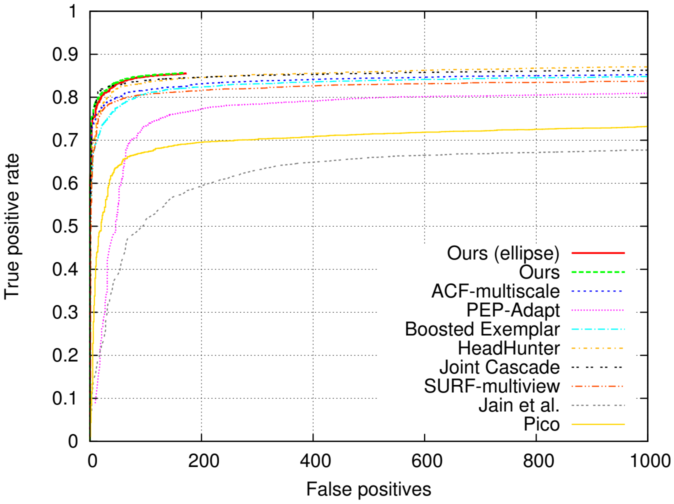

Figure 1: Performance evaluation on the FDDB dataset (discontinuous scores) comparing with ACF-multiscale [8], Boosted Exemplar [3], HeadHunter [4], Joint Cascade [1] and etc.

For a clear explanation, we describe a specific design of the proposed detector in this paper. The overall test pipeline of the detector is shown in Figure 2. The 12-*net* is a very shallow binary classification CNN to densely scan the input image for 12 × 12 detection windows. Given a test image, the 12-*net* scans the whole image densely across different scales to quickly reject more than 90% of the detection windows. The remaining detection windows are processed by the 12-*calibration-net* one by one as 12 × 12 images to adjust its size and location to approach a potential face nearby.

Non-maximum suppression (NMS) is applied to eliminate highly overlapped detection windows. The remaining detection windows are cropped out and resized into 24 × 24 as input images for the 24-*net* to further reject nearly 90% of the remaining detection windows. Similar to the previous process, the remaining detection windows are adjusted by the 24-*calibration-net* and we apply NMS to further reduce the number of detection windows.

The last 48-*net* accepts the passed detection windows as 48 × 48 images to evaluate the detection windows. NMS eliminates overlapped detection windows with an Intersection-Over-Union (IoU) ratio exceeding a pre-set threshold. The 48-*calibration-net* is then applied to calibrate the residual detection bounding boxes as the outputs.

We verify the proposed detector on two public face detection benchmarks. On the Annotated Faces in the Wild (AFW) [9] test set, our detector is comparable to the state-of-the-art. On the challenging Face Detection Data Set and Benchmark (FDDB) dataset [2], our detector outperforms the state-of-the-art methods in the discontinuous score evaluation, as shown in Figure 1.

[1] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *ECCV*. 2014.

[2] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, 2010.

[3] Haoxiang Li, Zhe Lin, J. Brandt, Xiaohui Shen, and Gang Hua. Efficient boosted exemplar-based face detection. In *CVPR*, 2014.

[4] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *ECCV*. 2014.

[5] Margarita Osadchy, Yann Le Cun, Matthew L. Miller, and Pietro Perona. Synergistic face detection and pose estimation with energy-based model. In *NIPS*, 2005.

[6] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *CVPR*, 1996.

[7] Régis Vaillant, Christophe Monrocq, and Yann Le Cun. Original approach for the localisation of objects in images. *IEE Proceedings-Vision, Image and Signal Processing*, 1994.

[8] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Aggregate channel features for multi-view face detection. *arXiv:1407.4023*, 2014.

[9] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

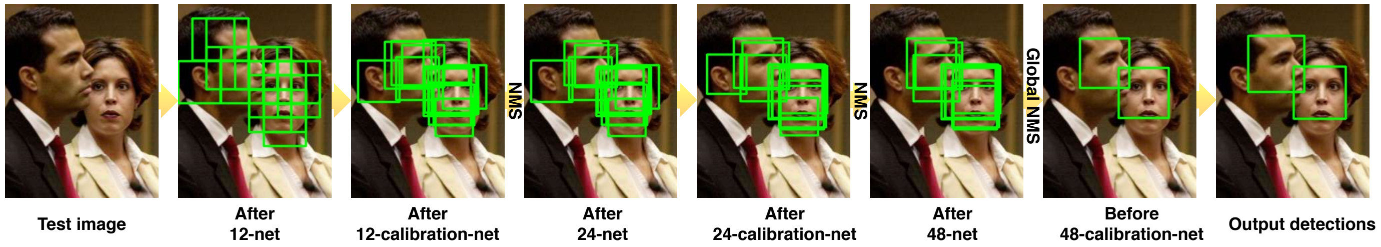

Figure 2: Test pipeline: From left to right, we show how the detection windows (green squares) are reduced and calibrated from stage to stage.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.