Adaptive WildNet Face Network for Detecting Face in the Wild

Dinh-Luan Nguyen[1], Vinh-Tiep Nguyen[1], Minh-Triet Tran[1], Atsuo Yoshitaka[2]
[1]Faculty of Information Technology, University of Science, VNU-HCM
[2]School of Information Science, Japan Advanced Institute of Science and Technology

ABSTRACT

Combining Convolutional Neural Network and Deformable Part Models is a new trend in object detection area. Following this trend, we propose Adaptive WildNet Face network using Deformable Part Models structure to exploit advantages of two methods in face detection area. We evaluate the merit of our method on Face Detection Data Set and Benchmark. Experimental results show that our method achieves up to 86.22% true positive images in 1000 false positive images in FDDB. Our method becomes one of state-of-the-art methods in FDDB dataset and it opens a new way to detect faces of images in the wild.

Keywords: Convolutional Neural Network, face detection, image in the wild, deformable part models.

1. INTRODUCTION

Detection face is not a new area in object detection. However, with the variation of face's pose, lighting conditions and occlusion, especially in face image in the wild, this area is still challenging and it attracts many works to completely solve face detection problem. Some techniques are proposed to deal with this problem such as contour fragment [1], probabilistic face structure [2] and Viola-Jone's work [3]. Nevertheless, they just use hand crafted, low level feature presentation with concrete classification to get detection result. This way of calculation is bias and it does not cover all situations of face with variance conditions. Thus, there is a huge need to exploit a new way to get high level feature with agile classification for this problem.

One of promising methods to achieve high level feature automatically is using deep convolutional neural network (CNN). The more layers it has, the higher level features it returns. Although it is useful to get meaningful features, there is one thing missing in this method when applying in face in the wild, which is structure of face. Since people face is concretely structured and it has relationship between face's components, it is deficient if we just using CNN. We still need an adaptive way to make use of structure of face.

Deformable Part Models (DPM) is famous for its strong relationship between its parts. DPM with default 8 parts is widely used in pedestrian detection [4]. It exploits low level HOG features to make an input for latent SVM classifier afterwards. In spite of being well known and commonly used to deal with structured objects, DPM still just uses low level HOG features for classification process.

CNN and DPM have advantages and disadvantages which can complement each other. Based on these observations, it is propitious to combine CNN and DPM together to exploit advantages of each method. In this paper, we create a deep convolutional neural network representing face's part model to make use of high level features and structure of face.

Main contribution. There are two key ideas in our framework. First, we propose a CoarseNet network to obtain high level feature from input images. Second, combination of 4-5 part DPM-CNN is constructed to represent face's structure and relation of face's components. Our whole framework called WildNet Face is created by linking two small mentioned networks to deal with face in the wild images. We conduct experiments on standard Face Detection Data Set and Benchmark (FDDB) [5] and achieves up to 86.22% and 77.28% true positive rate at 1000 false positive images in discontinuous and continuous evaluation respectively. Our method becomes new state-of-the-art method in FDDB dataset.

The rest of our paper is organized as follows. Section 2 depicts some related works using CNN and DPM in face detection area. Our main contribution is meticulously described in section 3. Experiments and comparison with state-of-the-art techniques is given in section 4. Finally, conclusion is mentioned in section 5.

2. RELATED WORKS

Some works realizes the advantages of DPM and CNN in object detection such as DeepPyramid DPM [6], End-to-End network [7]. However, they just apply in general object detection area. Work of Girshick [6] proposes network called DeepPyramid DPM, which includes DPM-CNN network with default 8 parts. This work can exploit relation between DPM and CNN so that it is used for multiclass object detection. Meanwhile End-to-End network, which is an extension of [6], constructs a model network with 9 parts instead of 8 parts configuration. Although these methods get competitive results in object detection, they are not suitable and specific for detecting face in the wild.

Nguyen et al. [8] proposes a new model to adapt with face in the wild. It constructs model by dividing face in to small parts and using them adaptively with different face directions. This model classifies face direction into two situations, which are frontal and side-view. In frontal view, DPM with 5 parts model is used to represent face with forehead, nose, two eyes and mouth. Similarly, DPM with 4 parts model is constructed to deal with side-view face. These models compute whole face score by using hand-crafted formula, which is so bias that this equation is not optimized and does not cover all ranges of part relations. Besides, this work just improves the original DPM [4] and optimizes their model to use with HOG-DPM version so that it is concreted with HOG feature. We inherit this spirit of dividing face into two situations to create our 4,5 part DPM-CNN layer network in our framework.

Several works [9, 10] make use of advantages of CNN to exploit high level features and get promising results in face detection area. They totally give specific neural network with fine-tuned parameters to get meaningful features. Thus, their networks still miss parts for describing face's structure and relation between face's components.

All in all, although given methods have advantages in their object detection areas, object structure specifically face structure is still absent from their works. So there is a need to combine CNN and DPM together. As far as we know, our framework which is meticulously described in section 3 is one of the first works adaptively integrate DPM structure into CNN network to solve face in the wild detection problem.

3. WILDNET FACE – A DEEP ADAPTIVE NETWORK FOR FACE IN THE WILD

3.1 Network Architecture

Detail of our network is described in Figure 1. From the input image, we construct image pyramid by downscaling input image with scale factor 0.8. After that, CoarseNet comprising 4 layers of convolution and max pooling is applied to create feature map at each scale in the pyramid. Result of these consecutive layers is the pyramid feature which is an input for DPM-CNN with 4 parts later on. As we described in Section 2, Nguyen's work [8] gets the relationship between parts by using handcrafted feature while we are using neural network to exploit high level features and the hidden structure between parts. We apply max pooling on the output of 4-parts features pyramid to reduce feature's resolution. Pyramid after going through max pooling layer is used as input for 5-part DPM-CNN. Finally, we apply another max pooling layer to get the final DPM pyramid score.

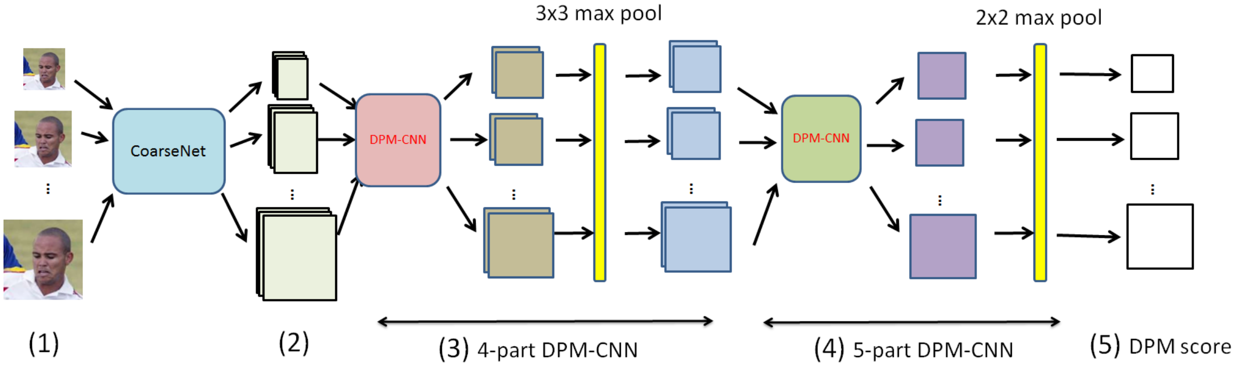

Figure 1. **WidNet Face network overview.** (1) Color images pyramid with 3-channel is an input for CoarseNet (Figure 2). (2) Feature pyramid is constructed by output of CoarseNet. This pyramid is also an input for 4-part DPM-CNN. (3) 4-part DPM-CNN without stack map process and 3x3 max pool layer are used to exploit features. (4) 5-part DPM-CNN and 2x2 max pool layer are used to get features for frontal face situation. (5) Pyramid ouput score for given pyramid image.

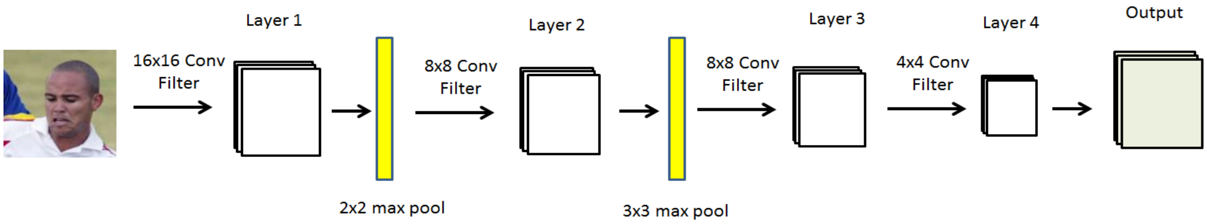

Figure 2. **Proposed CoarseNet architecture.** Input of this network is 3-channel RGB image. First and second layer have convolution filter and max pool while third and fourth layer just have convolution filter. The output of layer fourth is upscaled to get the final output of CoarseNet.

3.2 Parameters and Layers

3.2.1 Parameters

There is one problem when using deep neural network is that it lessens image feature's size after going through each layer. However, if the stride of network is too small, it leads to heavy computation in network. Thus, we specify strategy of using stride length and upscale each layer as follows. Firstly, in CoarseNet, the first, second and third layers use stride length of 1 to meticulously exploit details of input image. The fourth layer has stride length of 3 with upscaling by 2 to keep feature's size from going down quickly. Secondly, in 4-5 part DPM-CNN, we use stride length of 2 and 1 respectively in order to get more high level features and do not eliminate useful features created by gaps of stride length.

3.2.2 Layer details

We meticulously described layer details in our CoarseNet and DPM-CNN 4-5parts respectively.

Our CoarseNet includes 4 layers shown in Figure 2. In the first layer, we apply convolution filter with size 16x16. Since our WildNet has many layers, we apply 2x2 max pooling filter to decrease rate of reducing feature resolution. Similarly, 8x8 convolution and 3x3 max pooling filter are used in the second layer. Then, in the third and fourth layer, we apply 8x8 convolution and 4x4 convolution respectively and eliminate max pooling to get more meaningful features while preserving feature's size. From each image input in our CoarseNet, we have one image corresponding image feature. These images feature are stacked together to go through 4-5 parts DPM-CNN afterward.

DPM-CNN network is used with 4 and 5 parts configuration to adapt to face detection. In our framework, we eliminate stack map process in the original version of DPM-CNN because it causes single-channel in the output, which is not our goal. Thus, we still have pyramid feature with multi-channel after processing 4-part network. Again, max pooling layer is applied for whole pyramid feature with size 3x3. Finally, these image features are used in 5-part network together with preserving stack map process and fully connected layer to get single-channel score. This pyramid score goes through last 2x2 max pooling layer to get output score for our WildNet Face network.

4. EXPERIMENTS AND RESULTS

4.1 Dataset

We evaluate our method on Face Detection Data Set and Benchmark (FDDB). FDDB is a standard and famous dataset for face detection. It contains 5171 faces in 2845 images collected from newspaper, magazine with variation of background, illumination, face's pose and appearance. This dataset provides ground truth with ellipse annotation. Results of state-of-the-art methods are published in FDDB's website.

Evaluation. We use standard evaluation protocol provided in dataset for fair comparison with other methods. This evaluation has two types: continuous and discontinuous. They are different in metric used to evaluate overlap between given box and ground truth one.

4.2 Results

We run our method with configuration described in Section 3. Table 1 shows the comparison between different configurations of our network. As we can see from this result, DPM-CNN with default network is not suitable for face detection in the wild. Our proposed CoarseNet without 4-5part network is slightly better than vanilla DPM-CNN.

Table 1. Different configurations of WildNet Face network in FDDB dataset

Configuration	***True positive rate at 1000 false positive images***	
	Continuous	Discontinuous
Default DPM-CNN	53.95%	68.34%
Proposed CoarseNet	63.19%	73.23%
CoarseNet with default DPM-CNN	66.08%	77.97%
CoarseNet with 4-part DPM-CNN	68.17%	82.56%
CoarseNet with 5-part DPM-CNN	71.34%	83.02%
CoarseNet with 4,5-part DPM-CNN	75.23%	85.47%
4,5-part DPM-CNN without CoarseNet	70.02%	79.12%
Our best method	**77.28%**	**86.22%**

However, by combining with default DPM-CNN, it increases 4.74% true positive rate in discontinuous ROC (from 73.23% to 77.97%). Besides, with 4-5 part configuration in our network, our method achieve up to 85.47% in discontinuous and 75.23% in continuous true positive image.

Comparison with state-of-the-art. We compare our result with state-of-the-art results published in FDDB's website. They comprise DDFD [11], HeadHunter [12], Yan et al. [13], Joint Cascade [14] and Boosted Exemplar [15]. Our framework gets up to 86.81% in continuous ROC and 78.23% in discrete ROC. Figure 3 visualizes the ROC in discontinuous and continuous situations.

From the Figure 3, our method outperforms in continuous ROC and become the second best result in discontinuous one. We significantly increase true positive rate in continuous situation and boosts up to 2.45% in comparison with the second best method Joint Cascade (from 74.83% to 77.28%). Figure 4 shows some images, ellipse annotation in FDDB dataset and detection result of our method. Small face and low resolution together with variance in illumination leads to miss-detect in our framework.

5. CONCLUSION

In this paper, WildNet Face convolutional neural network is proposed to exploit the advantages of DPM and CNN. Our network shows that DPM and CNN can be ideally integrated together to boost up accuracy in face detection area. By using CNN to get high level features for DPM, our method becomes new state-of-the-art in FDDB dataset. Evaluation reveals that our method is superior and more robust than other state-of-the-art techniques. This work can be extended for generic object detection with variation in environmental conditions. Proposed method opens a new way to solve face detection problem in wild images.

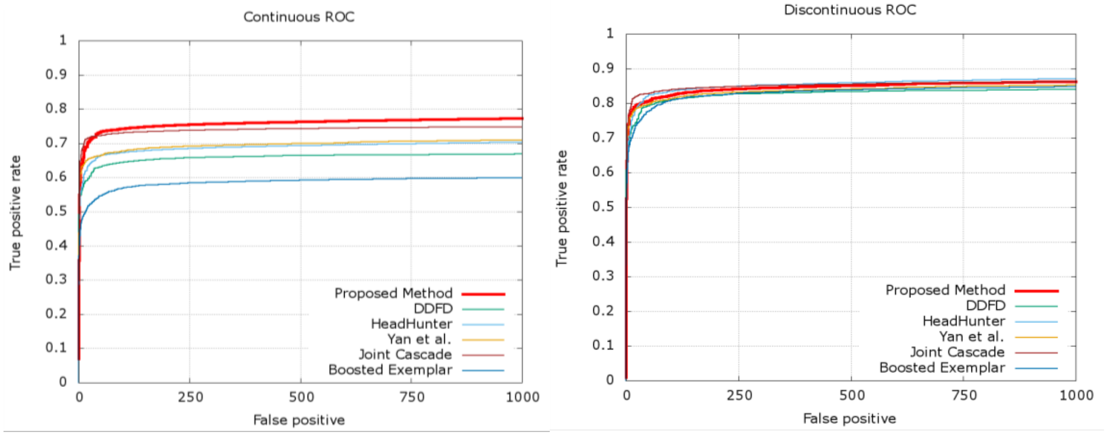

Figure 3.**Comparision with state-of-the-art on FDDB dataset.** We compare our result with state-of-the-art methods comprising DDFD [11], HeadHunter [12], Yan et al. [13], Joint Cascade [14] and Boosted Exemplar [15]. Our method becomes the state-of-the-art technique in continuous ROC and gets the second highest result in discontinuous ROC.

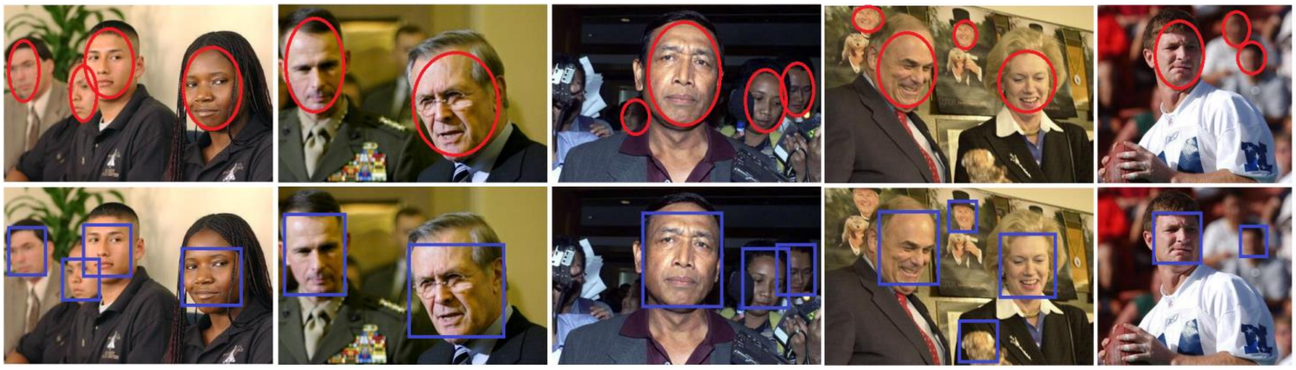

Figure 4.**Difficult images and annotation in FDDB dataset.** First row: Ellipse annotations ground truth in FDDB dataset. Second row: detection results of our method.

REFERENCES

[1] J. Shotton, A. Blake, R. Cipolla. “Multiscale Categorical Object Recognition Using Contour Fragments”,IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008,pp. 1270-1218.

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model” in Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[3] P. Viola, M. J. Jones, and D. Snow. “Detecting pedestrians usingpatterns of motion and appearance”. International Journal of Computer Vision, 2005, pp. 153–161.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models” ,Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, pp. 1627–1645.

[5] V. Jain and E. Learned-Miller. “FDDB: A benchmark for face detection in unconstrained settings”, Technical report, University of Mas-sachusetts, Amherst, 2010.

[6] R. B. Girshick, F. N. Iandola, T. Darrell, and J. Malik. “Deformable part models are convolutional neural networks”, in Computer Vision and Pattern Recognition. IEEE, 2015.

[7] L. Wan, D. Eigen, and R. Fergus. “End-to-End Integration of a Convolutional Network, Deformable Parts Model and Non-Maximum Suppression”, in Computer Vision and Pattern Recognition. IEEE, 2015.

[8] D.L. Nguyen, V.T. Nguyen, M.T. Tran, and A. Yoshitaka, “Boosting Speed and Accuracy in Deformable Part Models for Face Image in the Wild”, In International Conference on Advanced Computing and Applications, 2015.

[9] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection". In International Conference on Computer Vision, 2013, pp. 2056–2063.

[10] D.L. Nguyen, V.T. Nguyen, M.T. Tran, and A. Yoshitaka, “Deep Convolutional Neural Network in Deformable Part Models for Face Detection”, In Pacific Rim Symposium on Image and Video Technology, 2015.

[11] S. S. Farfade, Md. Saberian and Li-Jia Li. Multi-view Face Detection Using Deep Convolutional Neural Networks. In International Conference on Multimedia Retrieval, 2015.

[12] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. “Face detection without bells and whistles”. In Computer Vision and Pattern Recognition, 2014.

[13] J. Yan, Z. Lei, L. Wen and S. Z. Li. “The Fastest Deformable Part Model for Object Detection”. In Computer Vision and Pattern Recognition, 2014.

[14] D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun. “Joint Cascade Face Detection and Alignment”. In European Conference on Computer Vision, 2014.

[15] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. "Efficient boosted exemplar-based face detection". In Computer Vision and Pattern Recognition, 2014.