A Deep Pyramid Deformable Part Model for Face Detection

Rajeev Ranjan, Vishal M. Patel, Rama Chellappa
Center for Automation Research
University of Maryland, College Park, MD 20742
{rranjan1, pvishalm, rama}@umiacs.umd.edu

Abstract

We present a face detection algorithm based on Deformable Part Models and deep pyramidal features. The proposed method called DP2MFD is able to detect faces of various sizes and poses in unconstrained conditions. It reduces the gap in training and testing of DPM on deep features by adding a normalization layer to the deep convolutional neural network (CNN). Extensive experiments on four publicly available unconstrained face detection datasets show that our method is able to capture the meaningful structure of faces and performs significantly better than many competitive face detection algorithms.

1. Introduction

Face detection is a challenging problem that has been actively researched for over two decades [37], [36]. Current methods work well on images that are captured under user controlled conditions. However, their performance degrades significantly on images that have cluttered backgrounds and have large variations in face viewpoint, expression, skin color, occlusions and cosmetics.

The seminal work of Viola and Jones [32] has made face detection feasible in real world applications. They use cascaded classifiers on Haar-like features to detect faces. The cascade structure has been a subject of extensive research since then. Cascade detectors work well on frontal faces, however, sometimes they fail to detect profile or partially occluded faces. A recently developed joint cascade-based method [1] yields improved detection performance by incorporating a face alignment step in the cascade structure. Headhunter [25] uses rigid templates along similar lines. The method based on Aggregate Channel Features (ACF) [34] deploys a cascade of channel features while Pixel Intensity Comparisons Organized (Pico) [24] uses a cascade of rejectors for improved face detection.

Most of the recent face detectors are based on the Deformable Parts Model (DPM) structure [6] where a face is defined as a collection of parts. These parts are trained side-by-side with the face using a spring-like constraint. They are fine-tuned to work efficiently with the HOG [3] features. A unified approach for face detection, pose estimation and landmark localization using the DPM framework was recently proposed in [38]. This approach defined a "part" at each facial landmark and used mixture of tree-structured models resilient to viewpoint changes. A properly trained simple DPM is shown to yield significant improvement for face detection in [25].

The key challenge in unconstrained face detection is that features like Haar wavelets and HOG do not capture the salient facial information at different poses and illumination conditions. The limitation is more due to the features used than the classifiers. However, with recent advances in deep learning techniques and the availability of GPUs, it is becoming possible to use deep Convolutional Neural Networks (CNN) for feature extraction. In has been shown in [17] that a deep CNN pretrained with a large generic dataset such as Imagenet [4], can be used as a meaningful feature extractor. The deep features thus obtained have been used extensively for object detection. For instance, Regions with CNN (R-CNN) [7] computes regions-based deep features and attains state-of-art on the Imagenet challenge. Methods like Overfeat [28] and Densenet [10] adopt a sliding window approach to detect objects from the $pool_5$ features. Deep Pyramid [8] and Spatial Pyramid [9] remove the fixed-scale input dependency from deep CNNs which makes them attractive to be integrated with DPMs. Although, a lot of research on deep learning has focused on object detection and classification, very few have used deep features for face detection which is equally challenging because of high variations in pose, ethnicity, occlusions, etc. It was shown in [5] that deep CNN features fine-tuned on faces are informative enough for face detection, and hence do not require an SVM classifier. They detect faces based on the heat map score obtained directly from the fifth convolutional layer. Although they report competitive results, detection performance for faces of various sizes and occlusions needs improvement.

In this paper, we propose a face detector which detects faces at multiple scales, poses and occlusion by efficiently

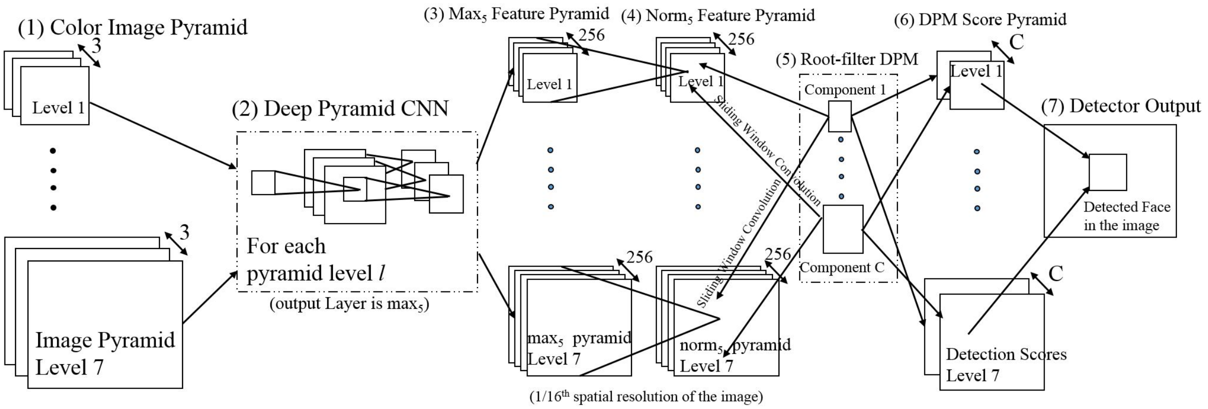


Figure 1. Overview of our approach. (1) An image pyramid is built from a color input image with level 1 being the lowest size. (2) Each pyramid level is forward propagated through a deep pyramid CNN [8] that ends at max variant of convolutional layer 5 (max_5). (3) The result is a pyramid of max_5 feature maps, each at 1/16th the spatial resolution of its corresponding image pyramid level. (4) Each max_5 level features is normalized using z-score to form $norm_5$ feature pyramid. (5) Each $norm_5$ feature level gets convoluted with every root-filter of a C-component DPM to generate a pyramid of DPM score (6). The detector outputs a bounding box for face location (7) in the image after non-maximum suppression and bounding box regression.

integrating deep pyramid features [8] with DPMs. This paper makes the following contributions:

1. We propose a novel method for training DPM for faces using deep pyramidal features.
2. We propose adding a normalization layer to the deep CNN to reduce the bias in face sizes.
3. We achieve new state-of-the-art detection performances on four challenging face detection datasets.

This paper is organized as follows. Section 2 describes our proposed face detector in detail. Section 3 provides the detection results on four challenging datasets. Finally, Section 4 concludes the paper with a brief summary and discussion.

2. Face Detection with Deep Pyramid DPM

Our proposed face detector, called Deep Pyramid Deformable Parts Model for Face Detection (DP2MFD), consists of two modules. The first one generates a seven level normalized deep feature pyramid for any input image of arbitrary size. Fixed-length features from each location in the pyramid are extracted using the sliding window approach. The second module is a linear SVM which takes these features as input to classify each location as face or non-face, based on their scores. In this section, we provide the design details of our face detector and describe its training and testing processes.

2.1. DPM Compatible Deep Feature Pyramid

We build our model using the feature pyramid network implementation provided in [8]. It takes an input image of variable size and constructs an image pyramid with seven levels. Each level is embedded in the upper left corner of a large (1713 × 1713 pixels) image and maintains a scale factor of $\sqrt{2}$ with its next lower level in the hierarchy. Using this image pyramid, the network generates a pyramid of 256 feature maps at the fifth convolution layer ($conv_5$). A 3×3 max filter is applied to the feature pyramid at a stride of one to obtain the max_5 layer which essentially incorporates the $conv_5$ "parts" information. Hence, it suffices to train a root-only DPM on the max_5 feature maps without explicitly training on DPM parts. A cell at location (j, k) in the max_5 layer corresponds to the pixel $(16j, 16k)$ in the input image, with a highly overlapping receptive field of size 163×163 pixels. Despite having a large receptive field , the features are well localized to be effective for sliding window detectors.

It has been suggested in [8] that deep feature pyramids can be used as a replacement for HOG Pyramid in DPM implementation. However, this is not entirely obvious as deep features are different than HOG features in many aspects. Firstly, the deep features from max_5 layer have a receptive field of size 163×163 pixels, unlike HOG where the receptive region is localized to a bin of 8 × 8 pixels. As a result, max_5 features at face locations in the test images would be substantially different from that of a cropped face. This prohibits us from using the deep features of cropped faces as positive training samples, which is usually the first step in training HOG-based DPM. Hence, we take a different approach of collecting positive and negative training samples from the deep feature pyramid itself. This procedure is described in detail in subsection 2.3.

Secondly, the deep pyramid features lack the normaliza-

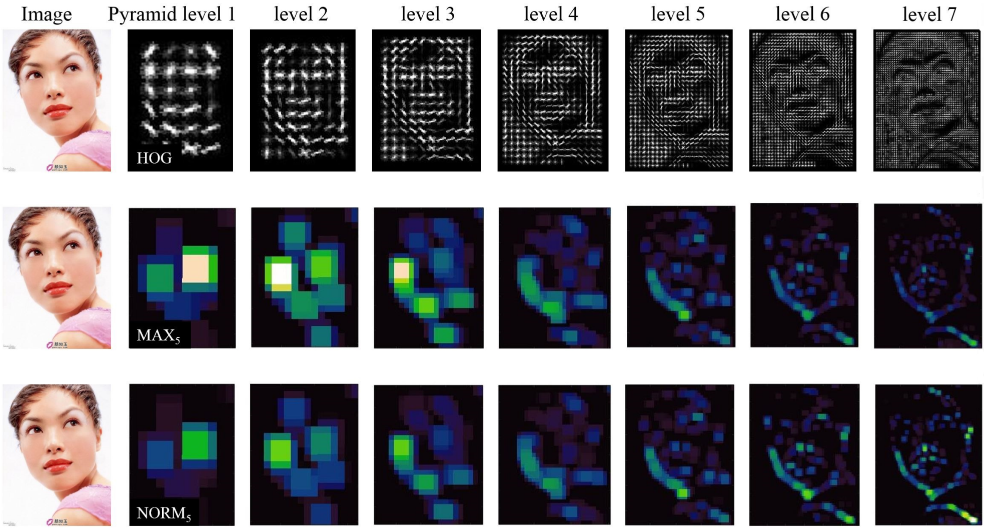


Figure 2. Comparison between HOG, max_5 and $norm_5$ feature pyramids. In contrast to max_5 features which are scale selective, $norm_5$ features have almost uniform activation intensities across all the levels.

tion attribute associated with HOG. The feature activations vary widely in magnitude across the seven pyramid levels as shown in Figure 2. Typically, the activation magnitude for a face region decreases with the size of pyramid level. As a result, a large face detected by a fixed-size sliding window at a lower pyramid level will have a high detection score compared to a small face getting detected at a higher pyramid level. In order to reduce this bias to face size, we apply a z-score normalization step on the max_5 features at each level. For a 256-dimensional feature vector $x_{i,j,k}$ at the pyramid level i and location (j, k), the normalized feature $\hat{x}_{i,j,k}$ is computed as:

$$\hat{x}_{i,j,k} = \frac{x_{i,j,k} - \mu_i}{\sigma_i}, \tag{1}$$

where μ_i is the mean feature vector, and σ_i is the standard deviation for the pyramid level i. We refer to the normalized max_5 features as "$norm_5$". A root-only DPM is trained on the $norm_5$ feature pyramid using a linear SVM. Figure 1 shows the complete overview of our model.

2.2. Testing

At test time, each image is fed to the model described above to obtain the $norm_5$ feature pyramid. They are convolved with the fixed size root-filters for each component of DPM in a sliding window fashion, to generate a detection score at every location of the pyramid. Locations having scores above a certain threshold are mapped to their corresponding regions in the image. These regions undergo a greedy non-maximum suppression to prune low scoring detection regions with Intersection-Over-Union (IOU) overlap above 0.3. In order to localize the face as accurately as possible, the selected boxes undergo bounding box regression. Owing to the subsampling factor of 16 between the input image and $norm_5$ layer, the total number of sliding windows account to approximately 25k compared to approximately 250k for the HOG pyramid, which reduces the effective test-time.

2.3. Training

For training, both positive and negative faces are sampled directly from the $norm_5$ feature pyramid. The dimensions of root filters for DPM are decided by the aspect ratio distribution for faces in the dataset. The root-filter sizes are scaled down by a factor of 8 to match the face size in the feature pyramid. Since, a given training face maps its bounding box at each pyramid level, we choose the optimal level l for the corresponding positive sample by minimizing the sum of absolute difference between the dimensions of bounding box and the root filter at each level. For a root-filter of dimension (h, w) and bounding box dimension of (b_i^y, b_i^x) for the pyramid level i, l is given by

$$l = \arg\min_i |b_i^y - h| + |b_i^x - w|. \tag{2}$$

The ground truth bounding box at level l is then resized to fit the DPM root-filter dimensions. We finally extract the "$norm_5$" feature of dimension $h \times w \times 256$ from the shifted ground truth position in the level l as a positive sample for training.

The negative samples are collected by randomly choosing root-filter sized boxes from the normalized feature pyramid. Only those boxes having IOU less than 0.3 with the ground truth face at the particular level are considered as negative samples for training.

Once the training features are extracted, we optimize a linear SVM for each component of the root-only DPM.

Since the training data is large to fit in the memory, we adopt the standard hard negative mining method [31, 6] to train the SVM. We also train a bounding box regressor to localize the detected face accurately. The procedure is similar to the bounding box regression used in R-CNN [7] , the only difference being our bounding box regressor is trained on the $norm_5$ features.

3. Experimental Results

We evaluated the proposed deep pyramid DPM face detection method on four challenging face detection datasets - Annotated Face in-the-Wild (AFW) [38], Face Detection Dataset and Benchmark (FDDB) [11], Multi-Attribute Labelled Faces (MALF) [35] and the IARPA Janus Benchmark A (IJB-A) [16], [2] dataset. We train our detector on the FDDB images using Caffe [13] for both 1-component (DP2MFD-1c) and 2-components (DP2MFD-2c) DPM. The FDDB dataset was evaluated using the 10-fold cross-validation approach. For evaluating the AFW and the MALF datasets, images from all the 10 splits of the FDDB dataset were used as training samples.

3.1. AFW Dataset Results

The AFW dataset [38] contains 205 images with 468 faces collected from Flickr. Images in this dataset contain cluttered backgrounds with large variations in both face viewpoint and appearance.

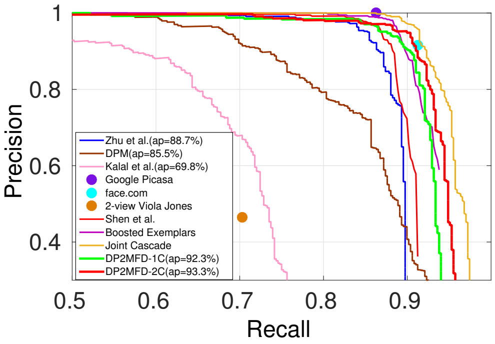


Figure 3. Performance evaluation on the AFW dataset.

The precision-recall curves [1] of different academic as well as commercial methods on the AFW dataset are shown in Figure 3. Some of the academic face detection methods compared in Figure 3 include OpenCV implementations of the 2-view Viola-Jones algorithm, DPM [6], mixture of trees (Zhu et al.) [38], boosted multi-view face detector (Kalal et al.) [14], boosted exemplar [20] and the joint cascade methods [1]. As can be seen from this figure, our method outperforms most of the academic detectors and performs comparably to a recently introduced joint cascade-based method [1] and the best commercial face detector Google Picassa. Note that the joint cascade-based method [1] uses face alignment to make the detection better and trains the model on 20,000 images. In contrast, we do not use any alignment procedure in our detection algorithm and train on only 2,500 images.

3.2. FDDB Dataset Results

The FDDB dataset [11] is the most widely used benchmark for unconstrained face detection. It consists of 2,845 images containing a total of 5,171 faces collected from news articles on the Yahoo website. All images were manually localized for generating the ground truth. The FDDB dataset has two evaluation protocols - discrete and continuous which essentially correspond to coarse match and precise match between the detection and the ground truth, respectively.

Figure 4 compares the performance of different academic and commercial detectors using the Receiver Operating Characteristic (ROC) curves on this dataset. The academic algorithms compared in Figure 4(a)-(b) include Yan et al. [33], boosted exemplar [20], SURF frontal and multiview [22], PEP adapt [19], XZJY [29], Zhu et al. [38], Segui et al. [27], Koestinger et al. [18], Li et al. [21], Jain et al. [12], Subburaman et al. [30], Viola-Jones [32], Mikolajczyk et al. [26], Kienzle et al. [15] and the commercial algorithms compared in Figure 4(c)-(d) include Face++, the Olaworks face detector, the IlluxTech frontal face detector and the Shenzhen University face detector [2].

As can be seen from this figure, our method significantly outperforms all previous academic and commercial detectors under the discrete protocol and performs comparably to the previous state-of-the-art detectors under the continuous protocol. A decrease in performance for the continuous case is mainly because of low IOU score obtained in matching our detectors' rectangular bounding box with elliptical ground truth mask for the FDDB dataset.

We also implemented an R-CNN method for face detection and evaluated it on the FDDB dataset. The R-CNN method basically selects face independent candidate regions from the input image and computes a 4096 dimensional fc_7 feature vector for each of them. An SVM trained on fc_7 features classifies each region as face or non-face based on the detection score. The method represented by "RCNN-face" performs better than most of the academic face detectors [38, 22, 19]. This shows the dominance of deep CNN features over HOG, SURF. However, RCNN-Face's performance is inferior to the DP2MFD method as the region selection process might miss a face from the image.

[1] The results of the methods other than our DP2MFD methods compared in Figure 3 were provided by the authors of [38], [1] and [20].

[2] http://vis-www.cs.umass.edu/fddb/results.html

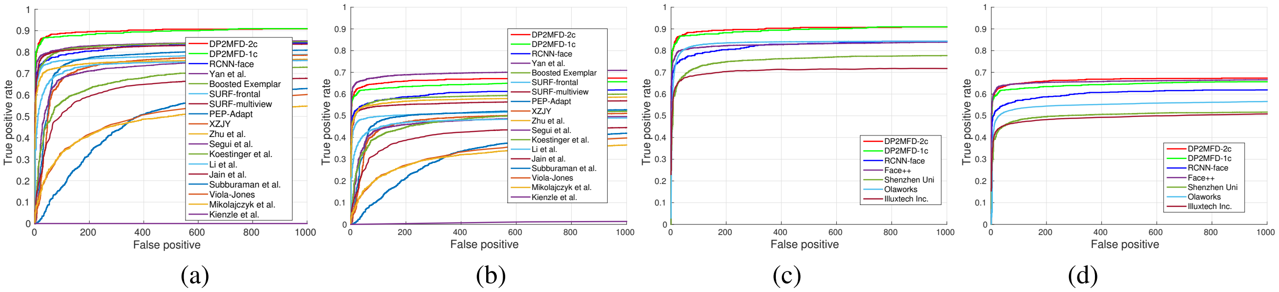


Figure 4. Performance evaluation on the FDDB dataset. (a) and (b) compare our method with previously published methods under the discrete and continuous protocols, respectively. Similarly, (c) and (d) compare our method with commercial systems under the discrete and continuous protocols, respectively.

3.3. MALF Dataset Results

The MALF dataset [35] consists of 5,250 high-resolution images containing a total of 11,931 faces. The images were collected from Flickr and image search service provided by Baidu Inc. The average image size in this dataset is 573×638. On average, each image contains 2.27 faces with 46.97% of the images contain one face, 43.41% contain 2 to 4 faces, 8.30% contain 5 to 9 faces and 1.31% images contain more than 10 faces. Since this dataset comes with multiple annotated facial attributes, evaluations on attribute-specific subsets are proposed. Different subsets are defined corresponding to different combinations of attribute labels. In particular, 'easy' subset contains faces without any large pose, occluded or exaggerated expression variations and are larger than 60×60 in size and 'hard' subset contains faces that are larger than 60×60 in size with one of extreme pose or expression or occlusion variations. Furthermore, scale-specific evaluations are also proposed in which algorithms are evaluated on two subsets - 'small' and 'large'. The 'small' subset contains images that have size smaller than 60×60 and the ''large' subset contains images that have size larger than 90×90.

The performance of different algorithms, both from academia and industry, are compared in Figure 5 by plotting the True Positive Rate vs. False Positive Per Images curves [3]. Some of the academic methods compared in Figure 5 include ACF [34], DPM [25], Exemplar method [20], Headhunter [25], TSM [38], Pico [24], NPD [23] and W. S. Boost [14]. From Figure 5(a), we see that overall the performance of our DP2MFD method is the best among the academic algorithms and is comparable to the best commercial algorithms FacePP-v2 and Picasa.

In the 'small' subset, denoted by $<$ 30 height in Figure 5(b), the performance of all algorithms drop a little but our DP2MFD method still performs the best among the other academic methods. On the 'large', 'easy, and 'hard' subsets, the DPM method [25] performs the best and our DP2MFD method performs the second best as shown in Figure 5(c), (d) and (e), respectively. The DPM and Headhunter [25] are better as they train multiple models to fully capture faces in all orientations, apart from training on more than 20,000 samples.

We provide the results of our method for the IOU of 0.35 as well as 0.5 in Figure 5. Since the non-maximum suppression ensures that no two detections can have IOU> 0.3, the decrease in performance for IOU of 0.5 is mainly due to improper bounding box localization. One of the contributing factors might be the localization limitation of CNNs due to high amount of sub-sampling. In future, we plan to analyze this issue in detail.

3.4. IJB-A Dataset Results

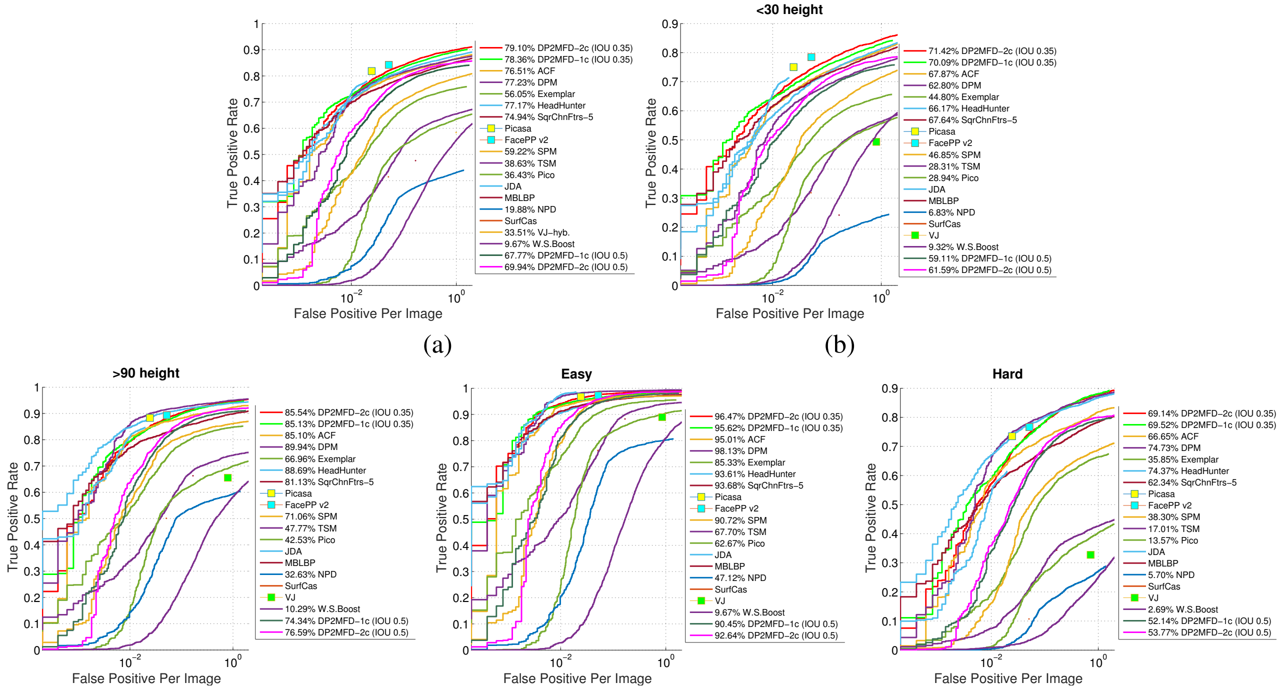
The IJB-A dataset contains images and videos from 500 subjects collected from online media [16], [2]. In total, there are 67,183 faces of which 13,741 are from images and the remaining are from videos. The locations of all faces in the IJB-A dataset were manually ground truthed by human annotators. The subjects were captured so that the dataset contains wide geographic distribution. All face bounding boxes are about 36 pixels or larger.

Nine different face detection algorithms were evaluated on this dataset in [2]. Some of the algorithms compared in [2] include one commercial off the shelf (COTS) algorithm, three government off the shelf (GOTS) algorithms, two open source face detection algorithms (OpenCV's Viola Jones and the detector provided in the Dlib library), and PittPat ver 4 and 5. In Figure 6 (a) and (b) we show the prevision vs. recall curves and the ROC curves, respectively corresponding to our method and one of the best reported methods in [2]. As can be seen from this figure, our method outperforms the best performing method reported in [2] by a large margin.

3.5. Discussion

Its clear from these results that our DP2MFD-2c method performs slightly better than the DP2MFD-1c method. This can be attributed to the fact that the aspect ratio of face

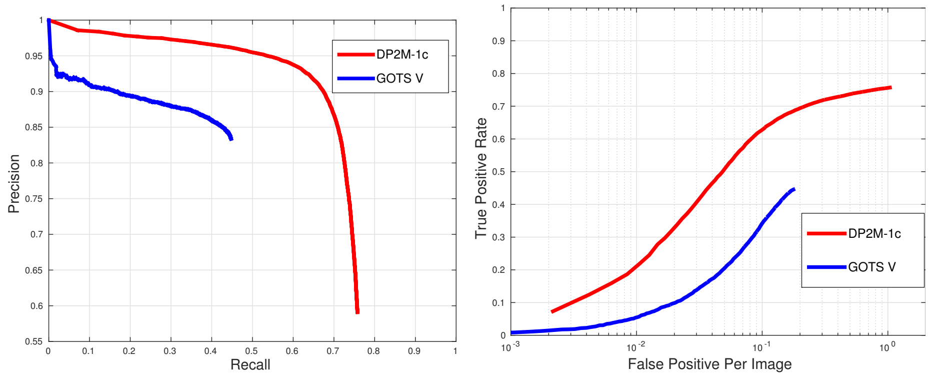
[3] The results of the methods other than our DP2MFD methods compared in Figure 5 were provided by the authors of [35].



(a) (b)

(c) (d) (e)

Figure 5. Fine-grained performance evaluation on the MALF dataset. (a) on the whole test set, (b) on the small faces sub-set, (c) on the large faces sub-set, (d) on the 'easy' faces sub-set and (e) on the 'hard' faces sub-set.



(c) (d)

Figure 6. Performance evaluation on the IJB-A dataset. (a) Precision vs. recall curves. (b) ROC curves.

doesn't change much with pose. Figure 7 shows several detection results on the four datasets. It can be seen from this figure, that our method is able to detect profile faces as well as different size faces in images with cluttered background.

3.6. Runtime

Our face detector was tested on a machine with 4 cores, 12GB RAM, and 1.6GHz processing speed. No GPU was used for processing. The model DP2MFD-1c took about 24.5s on average to evaluate a face, whereas DP2MFD-2c took about 26s. The deep pyramid feature evaluation took around 23s. It can certainly be reduced to 0.5s [8] by using Tesla K20 GPU for feature extraction.

4. Conclusions

In this paper, we presented a method for unconstrained face detection which essentially trains DPM for faces on deep feature pyramid. One of the interesting features of our algorithm is that we add a normalization layer to the deep CNN which reduces the bias in face sizes. Extensive experiments on four publicly available unconstrained face detection datasets demonstrate the effectiveness of our proposed approach.

Our future work will include a GPU implementation of our method for reducing the computing time. We will also evaluate the performance of our method on other object detection datasets.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *European Conference on Computer Vision*, volume 8694, pages 109–122. 2014.

[2] J. Cheney, B. Klein, A. K. Jain, and B. F. Klare. Unconstrained face detection: State of the art baseline and challenges. In *International Conference on Biometrics*, 2015.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. 1:886–893 vol. 1, June 2005.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[5] S. S. Farfade, M. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks, 2015. CoRR.

[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sept 2010.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[8] R. B. Girshick, F. N. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks, 2014. CoRR.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729, 2014.

[10] F. N. Iandola, M. W. Moskewicz, S. Karayev, R. B. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids, 2014. CoRR.

[11] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[12] V. Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–584, June 2011.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[14] Z. Kalal, J. Matas, and K. Mikolajczyk. Weighted sampling for large-scale boosting. In *Proceedings of the British Machine Vision Conference*, pages 42.1–42.10. BMVA Press, 2008. doi:10.5244/C.22.42.

[15] W. Kienzle, G. BakIr, M. Franz, and B. Schölkopf. Face detection: Efficient and rank deficient. In *Advances in Neural Information Processing Systems*, pages 673–680, 2005.

[16] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[18] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Robust face detection by simple means, 2012.

[19] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *IEEE International Conference on Computer Vision*, pages 793–800, Dec 2013.

[20] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1843–1850, June 2014.

[21] J. Li, T. Wang, and Y. Zhang. Face detection using surf cascade. In *IEEE International Conference on Computer Vision Workshop on Benchmarking Facial Image Analysis Technologies*, pages 2183–2190, Nov 2011.

[22] J. Li and Y. Zhang. Learning surf cascade for fast and accurate object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3468–3475, June 2013.

[23] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector, 2015. CoRR.

[24] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. A method for object detection based on pixel intensity comparisons organized in decision trees, 2014. CoRR.

[25] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, volume 8692, pages 720–735. 2014.

[26] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages 69–82, 2004.

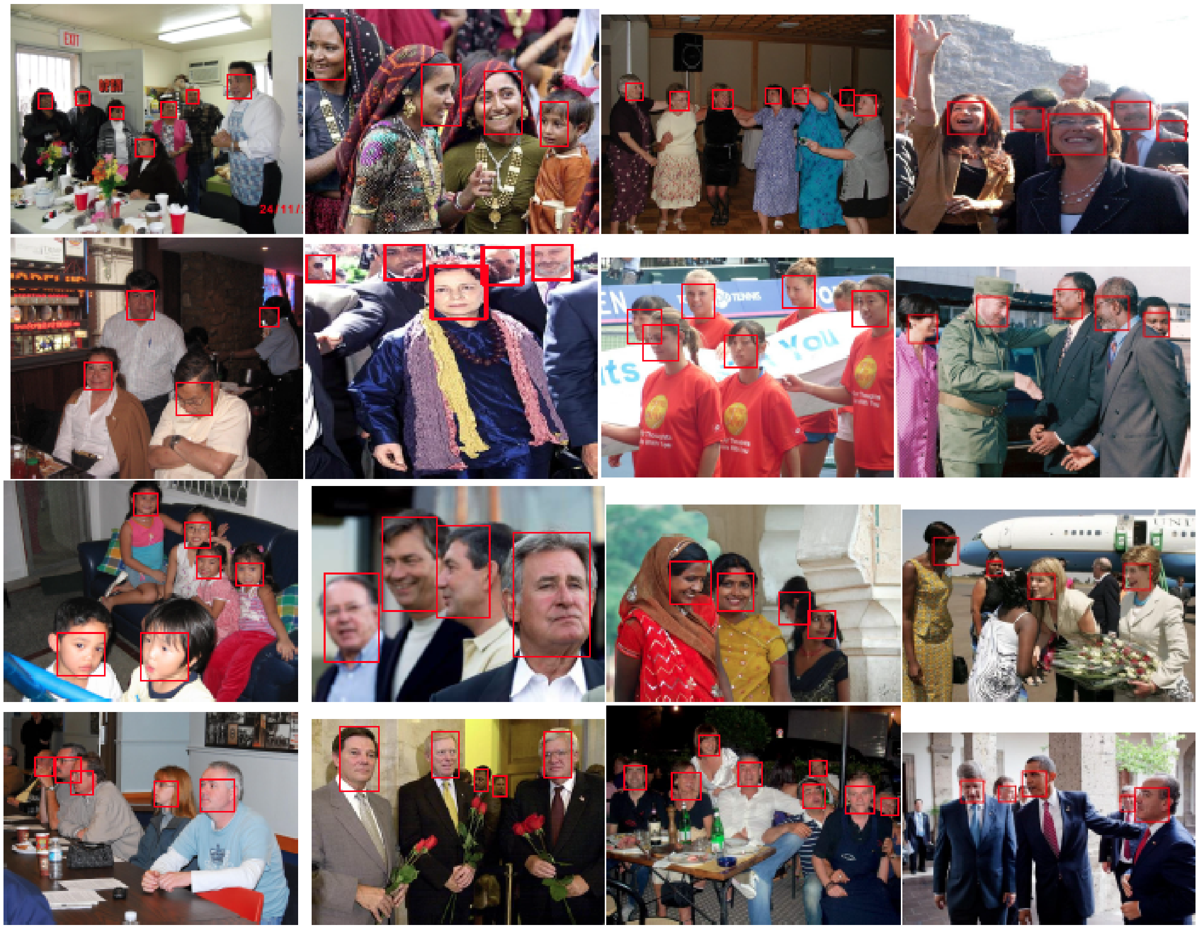


Figure 7. Qualitative results of our detector. First column - AFW dataset, Second column - FDDB dataset, Third column - MALF dataset, Fourth column - IJB-A dataset.

[27] S. Segu, M. Drozdzal, P. Radeva, and J. Vitri. An integrated approach to contextual face detection. In *International Conference on Pattern Recognition Applications and Methods*, pages 90–97, 2012.

[28] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013. CoRR.

[29] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3467, June 2013.

[30] V. Subburaman and S. Marcel. Fast bounding box estimation based face detection. In *Proc. Workshop on Face Detection of the European Conference on Computer Vision*, Heraklion, Crete, Greece, 2010.

[31] K. K. Sung and T. Poggio. Example based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:39–51, 1995.

[32] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[33] J. Yan, Z. Lei, L. Wen, and S. Li. The fastest deformable part model for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2504, June 2014.

[34] B. Yang, J. Yan, Z. Lei, and S. Li. Aggregate channel features for multi-view face detection. In *IEEE International Joint Conference on Biometrics*, pages 1–8, Sept 2014.

[35] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Fine-grained evaluation on face detection in the wild. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[36] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, Microsoft Research, 2010.

[37] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, Dec. 2003.

[38] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, June 2012.