Two-stage Cascade Model for Unconstrained Face Detection

Darijan Marčetić, Tomislav Hrkać, Slobodan Ribarić
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{darijan.marcetic, tomislav.hrkac, slobodan.ribaric}@fer.hr

Abstract - In this paper, we propose a two-stage model for unconstrained face detection. The first stage is based on the normalized pixel difference (NPD) method, and the second stage uses the deformable part model (DPM) method. The NPD method applied to in the wild image datasets outputs the unbalanced ratio of false positive to false negative face detection when the main goal is to achieve minimal false negative face detection. In this case, false positive face detection is typically an order of magnitude higher. The result of the NPD-based detector is forwarded to the DPM-based detector in order to reduce the number of false positive detections. In this paper, we compare the results obtained by the NPD and DPM methods on the one hand, and the proposed two-stage model on the other. The preliminary experimental results on the Annotated Faces in the Wild (AFW) and the Face Detection Dataset and Benchmark (FDDB) show that the two-stage model significantly reduces false positive detections while simultaneously the number of false negative detections is increased by only a few.

Keywords: Unconstrained Face Detection, Normalized Pixel Difference Model, Deformable Part-based Model

I. INTRODUCTION

Detection of faces in unconstrained real-world images is a challenging computer vision task. Some of the main constraining factors are the variability and diversity of the face pose, the presence of occlusions, expression variability, different illumination conditions, scale variations, and colour and texture richness. Various face detection methods have been developed, such as scanning window-based classifiers [1], part-based models [2-5], normalized pixel difference [6] and deep neural networks [7]. A generalization of the work of Viola and Jones for the efficient detection of faces with occlusions was described in [8]. The authors in [9] divided a face region into multiple patches, and mapped those weak classifiers, inspired by the Viola and Jones approach, to the patches. All of the valid face patches, with different weights, were used to make a final decision. In [10], an approach for automatic and robust face detection was presented. The method used Haar-like features and an AdaBoost trained classifier cascade, and the topology verification was based on graph matching techniques. What all the above scanning window-based methods have in common is that they use a cascade of simple weak classifiers trained by a boosting procedure.

Initially, face detection, pose estimation, and landmark localization were handled separately. These three tasks can be simultaneously handled with a unified part-based approach that uses mixtures of trees with a shared pool of part templates for face landmarks, and utilizing an encoding scheme for modes of elastic deformation, and a view-based topology for face detection and pose estimation [4].

Recently, combinations of the mentioned methods have been used to improve the performance and/or speed of the face detector. For example, in [11], the face detection algorithm, called DP2MFD, which is able to detect faces of various sizes and poses in unconstrained conditions, is proposed. It is based on deformable part models and a deep convolutional neural network. In [12], so-called joint cascade face detection and alignment combines the Viola-Jones detector with a low threshold to ensure high recall, and pose indexed features with boosted regression [13]. In [14], a two-stage cascade model for robust head-shoulder detection is introduced. It combines several methods as follows: a HOG and LBP feature-based Viola Jones type classifier in the first stage, and a Region Covariance Matrix (RCM) in the second stage. These three recent papers have shown that a combination of several different methods significantly improves face detection results compared to "classical" one-stage approaches.

In this paper, we propose a two-stage cascade model for unconstrained face detection that combines a deformable part model [4] and a normalized pixel difference model [6]. The experimental results, obtained on Annotated Faces in the Wild (AFW) [15] and Face Detection Dataset and Benchmark (FDDB) [16] testsets, have shown improved performance with regard to face detection results and satisfactory detection speed at the same time when compared to the results of each of these methods when used alone.

The rest of the paper is organized as follows. In Section II, we present a brief theoretical background of two methods which are combined in the two-stage cascade model. A description of the two-stage cascade face detector is given in Section III. We show the experimental results in Section IV and conclude our work in Section V.

II. BRIEF THEORETICAL BACKGROUND

In the proposed two-stage cascade model for unconstrained face detection, a deformable part model [4] and a normalized pixel difference model [6] are used. A short description of both methods follows.

A. *Normalized pixel difference method*

In [6] the authors proposed a method for unconstrained face detection based on the features called Normalized Pixel Difference (NPD). NPD is defined as: $f(x_i, x_j) = (x_i - x_j) / (x_i + x_j)$, where $x_i, x_j \geq 0$, are intensity values of two pixels, and f(0,0) by definition is 0. This simple NPD feature is inspired by the Weber Fraction in experimental psychology. The NPD feature has the following properties: i) the NPD is antisymmetric; ii) the sign of NPD is an indicator of the ordinal relationship; iii) the NPD is scale invariant; iv) the NPD is bounded in $[-1, 1]$. These properties are useful because they reduce feature space, encode the intrinsic structure of an object image, increase robustness to illumination changes, and the bounded property makes the NPD feature amenable for threshold learning in tree-based classifiers [6]. Due to limitations with the stump, as a basic tree classifier with only one threshold that splits a node in two leaves, the authors in [6] used for learning a quadratic splitting strategy for a deep tree, where the depth was eight.

For practical reasons, the values of the NPD features are quantized into L = 256 bins. In order to reduce redundant information contained in the NPD features, the AdaBoost algorithm is used to select the most discriminative features. Based on these features, a strong classifier is constructed. The final face detector contains 1226 trees and 46401 NPD features. The pre-computed multiscale detector templates, based on the basic 20 × 20 learned face detector, are applied to detect faces in various scales. The score *ScoreNPD*(I, S_i) is obtained based on the result of 1226 deep quadratic trees and 46401 NPD features, but the average number of feature evaluations per detection window is only 114.5.

The output of the NPD detector is represented by square regions S_i , i = 1, 2, ... n, where n is the number of detected regions of interest in an image I. For each region S_i in an image I, the score *ScoreNPD*(I, S_i) is calculated [6]. The regions S_i , i = 1, 2, ... , $j \leq n$, with scores *ScoreNPD*(I, S_i) greater than some predefined threshold θ_{NPD} are classified as faces. Originally, to achieve minimum FN, a threshold $\theta_{NPD} = 0$ is used [6].

B. *Deformable Part-based Models*

A mixture of multiscale deformable part-based models (DPMs), represented by HOG features, is used for joint face detection, pose estimation and landmark localization methods [4]. The method proposed in [4] uses a model based on mixtures of trees with a shared pool of parts represented by HOG descriptors. Facial landmarks are modelled as parts. A global mixture of parts is used to capture topological changes due to face viewpoints. A set of 13 trees was used to represent different face viewpoints. Each tree has only one root node. An energy function represented as a score *ScoreDPM*(I, L_i) = *App*(I, L_i) + *Shape*(L_i) is maximised over all potential part locations, i = 1, 2, ... m, where m is the number of detected regions of interest in an image I, and L_i is the spatial configuration of face parts. *App*(I, L_i) is the appearance evidence for placing a HOG-based template of part at a location defined by L_i. *Shape*(L_i) scores a specific spatial arrangement of parts for different face viewpoints [4]. The output of the DPM detector is represented by the candidate face part locations L_i , i = 1, 2, ... , m, which determine the square regions F_i , i = 1, 2, ... m, in an image I, and the score value *ScoreDPM(I, L_i)* [4]. The regions F_i, i = 1, 2, ... , $k \leq m$, with scores *ScoreDPM*(I, L_i) greater than some predefined threshold θ_{DPM}, are classified as faces. Originally, the threshold $\theta_{DPM} = -0.65$ was used [4].

III. TWO-STAGE CASCADE FACE DETECTOR

We propose a two-stage cascade model (2SCM) for unconstrained face detection. The first stage of the cascade is based on the NPD method, and the second stage uses the DPM method. The NPD method achieves low FN face detections for in the wild image datasets, but the number of FPs is relatively high (typically an order of magnitude higher than FN face detections) when the operating point is set to achieve minimal FN detections - the NDP detection threshold θ_{NPD} is zero [6]. Fig. 1 illustrates a typical result of the NPD detector for an image with rich texture in which the number of FP face detections is relatively high. The number of FP face detections for the NPD detector can be reduced by increasing the θ_{NPD}, but this has negative effects on FN face detections.

Figure 1. Example of the results of the NPD detector for the threshold $\theta_{NPD} = 0$. The green squares denote ground truth, the red true positive, and the blue squares denote false positive. The *ScoreNPD(I, S_i)* is given at the bottom of each square.

The output of the NPD detector is represented by square regions $S_i = (x_i, y_i, s_i)$, i = 1, 2, ..., j, where x_i and y_i are the coordinates of a square centre, s_i is the size of a region ($s_i \times s_i$), and j is the number of detected faces in an image I. Note that for S_i, i = 1, 2, ..., j, the score *ScoreNPD*(I, S_i) is greater than zero [6]. In order to reduce FP face detections, but to keep FNs as low as possible, the output of the NPD detector is conditionally forwarded to the second stage of the cascade which is realized as the DPM detector.

Depending on the value of *ScoreNPD*(I, S_i), the output of the NPD detector is forwarded to the DPM detector to classify the regions S_i as face or non-face as follows:

i) if *ScoreNPD*(I, S_i) $\geq \theta_1$, then a face is detected in image I at region S_i (true positive - TP);

ii) if *ScoreNPD*(I, S_i) $\leq \theta_2$, then a face is not detected in image I at region S_i (true negative - TN);

iii) if $ScoreNPD(I, S_i)$ falls in the interval $<\theta_2, \theta_1>$, then by applying the DPM detector to region S_i, the score $ScoreDPM(I(S_i), L^*)$ is evaluated and used for classification as follows:

a) if $ScoreDPM(I(S_i),L^*) \geq \theta_3$, a face is detected in image I at region S_i , where $I(S_i)$ represents the sub-image of an image I defined by the parameters of S_i;

b) if $ScoreDPM(I(S_i),L^*) < \theta_3$, a face is not detected in image I at region S_i.

Note that the execution time of the DPM detector is typically more than two orders of magnitude shorter than the NPD detector, which justifies the use of the DPM detector in the second stage only in a relatively small number of regions S_i, selected based on the criterion in iii) (see above) and these regions are a small fraction of the whole area of an image I.

The thresholds θ_1, θ_2 and θ_3 define the operating point of the two-stage face detector, and they are selected to maximize the sum of precision and recall, where Precision = TP / (TP + FP) and Recall = TP / (TP + FN). The thresholds are determined by the experiments performed on the Annotated Faces in the Wild (AFW) testset ($\theta_1 = 44$, $\theta_2 = 5$, $\theta_3 = -0.65$).

IV. Experimental results

Two databases were used to perform the experiments. From the Annotated Faces in the Wild (AFW) [15] testset, 205 images that contain in total 468 unconstrained faces were used. From the Face Detection Dataset and Benchmark (FDDB) [16], a subset of 2814 annotated images, which contains in total 5171 unconstrained faces, was used. For these two testsets, we performed six experiments as follows.

i) We ran the NPD detector on the AFW testset. The following results were obtained with the originally proposed threshold $\theta_{NPD} = 0$ [6]: 28 faces were not detected (FN) among 468 faces (Fig 2a), and 783 image regions were detected as a face FP among 205 images (Table I).

ii) The NPD detector was applied to the subset of the FDDB testset. The following results were obtained with the predefined threshold θ_{NPD} equal to the zero: 857 faces were not detected among 5171 faces (Fig 2e), and 1902 image regions were detected as FP among 2845 images.

Based on the above results (Table I) obtained on the AFW and the subset of the FDDB testset, we can conclude that the NPD detector in general achieves a low ratio of FN / FP face detections due to the high value of FP.

iii) The DPM detector was evaluated on the AFW testset to obtain the corresponding $ScoreDPM(I, S_i)$, and the originally proposed DPM classification threshold $\theta_{DPM} = -0.65$ proposed in [4] was used for classification. The following results were obtained: 161 faces were not detected (FN) among 468 faces (Fig 2b, Table I), and 72 image regions were detected as a face FP among 205 images.

iv) The DPM detector was then evaluated on the FDDB testset. The following results were obtained: 1250 faces were not detected (false negatives) among 5171 faces (Fig 2f, Table I), and 63 image regions were detected as a face FP among 2845 images.

From the results of experiments i-iv), we can conclude that the DPM method has much higher FN face detections than the NPD detector for in the wild image datasets, but the number of FPs is much lower than the NDP detector. This indicates that the good performance of the NPD detector regarding FN face detections can be combined with the good performance of the DPM detector with regard to FP detections.

v) The two-stage cascade model (2SCM) described in Section III is used for the AFW testset. The following results were obtained (Fig 2c, Table I): 48 = 28 + 20 faces were not detected (FN), where 28 corresponds to the FN from the first stage and the remaining 20 correspond to the FN from the second stage, among 468 faces, and only 23 image regions were detected as FP among 205 images. Note that FP was reduced more than 34 times, while FN was increased by only 1.71 times.

vi) The same cascade described in v) was used for the experiment for the FDDB testset. The following results were obtained (Fig 2g, Table I): 1081 = 857 + 224 faces were not detected (FN), where 857 corresponds to the FN from the first stage and the remaining 224 corresponds to the FN from the second stage, among 5171 faces, and only 122 image regions were detected as FP among 2845 images. In this experiment, FP was reduced by more than 15 times, while FN was increased by only 1.26 times.

The recall and precision curves of all three experiments performed on the AFW and FDDB dataset are depicted in Fig 2d and 2h respectively.

Regarding the processing time for a high resolution image, the typical ratio between the original DPM and NDP is about 60, and between the proposed two-stage cascade (2SCM) detector and the NPD it is about 2. Note that the latter ratio is the function of the number and the size of the detected candidate face regions by the NPD detector in the first stage that have a $ScoreNPD(I, S_i)$ between θ_2 and θ_1.

TABLE I. The results of the experiments performed on AFW and FDDB testsets

Testset		**AFW [15]**				**FDDB [16]**			
Number of images		205				2845			
Number of faces		468				5171			
Method	Thresholds	FN	FP	Precision	Recall	FN	FP	Precision	Recall
NPD	$\theta_{NPD} = 0$	28	783	0.360	0.940	857	1902	0.694	0.834
DPM	$\theta_{DPM} = -0.65$	161	72	0.810	0.656	1250	63	0.984	0.758
2CSM	**$\theta_1 = 44$, $\theta_2 = 5$, $\theta_3 = -0.65$**	**48 (28+20)**	**23**	**0.948**	**0.897**	**1081 (857+224)**	**122**	**0.971**	**0.791**

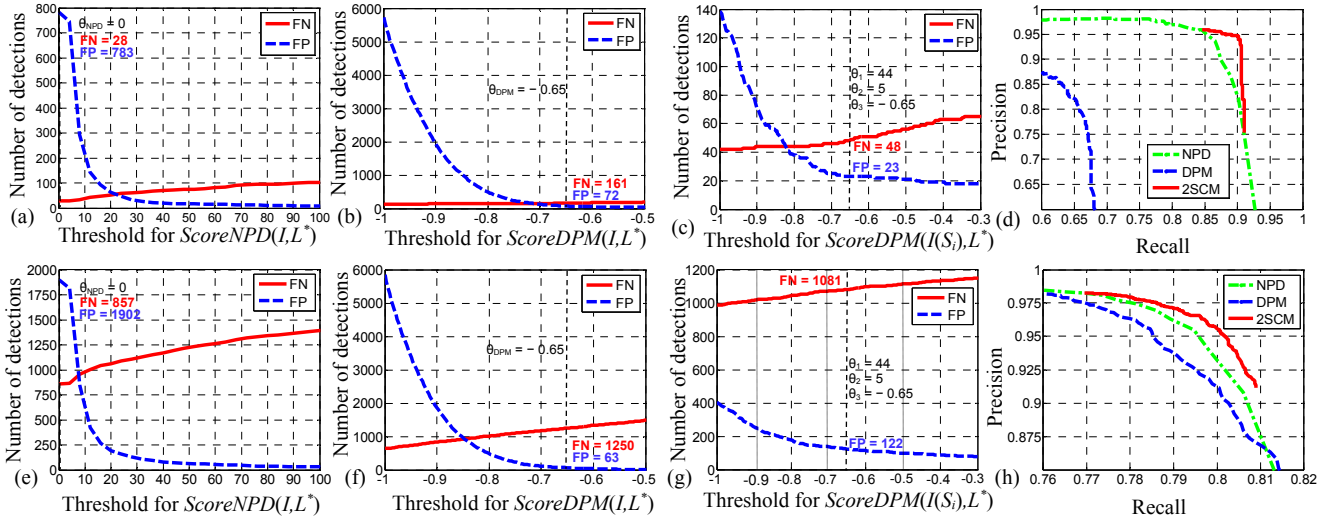

Figure 2. The first row (a)-(d) represents the results of experiments for the AFW testset. The second row (e)-(h) represents the results of the FDDB testset. The results of experiments: (a) and (e) NPD detector; (b) and (f) DPM detector; (c) and (g) the proposed two-stage detector (2SCM); (d) precision and recall of experiments for (a)-(c); (h) precision and recall of experiments for (e)-(g).

V. Conclusion

In this paper, we have proposed a two-stage cascade model (2SCM) for unconstrained face detection. The first stage of the model is based on the NPD detector, and the second is based on the DPM detector. The model is introduced in order to reduce FP face detections, but to keep FN as low as possible. This is achieved by conditionally forwarding the output of the NPD detector to the second stage of the cascade. The condition is based on a value of the NPD detector score. There are four arguments for using the proposed model: i) NPD and DPM detectors use features which are uncorrelated (NPD vs. HOG-based features); ii) the NPD detector is used in the first stage due to its high recall, and it is much faster than DPM; iii) the DPM detector is used conditionally as a post classier as it operates on the output of the DPM detector; iv) the DPM detector has higher precision than the NPD detector (Table I). Experiments performed on the AFW testset show that FP was reduced more than 34 times, while FN was increased only by 1.71 times compared with the NPD detector for the originally proposed threshold [6]. For the FDDB testset, FP was reduced by more than 15 times, while FN was increased only by 1.26 times (for the same NPD threshold). Our future work will focus on improving the cascade model in order to reduce the introduction of additional false negative face detections in the second stage.

Acknowledgments

This work has been supported by the Croatian Science Foundation under project 6733 De-identification for Privacy Protection in Surveillance Systems (DePPSS). It is also the result of activities in COST Action IC1206 "De-identification for Privacy Protection in Multimedia Content".

Literature

[1] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” inProc. IEEE Comput. Soc. Conf. Comput.Vis. Pattern Recog., 2001, pp. I-511–I-518.

[2] M. Fischler and R. Elschlager, “The representation and matching of pictorial structures,” IEEE. Trans. Computers, C-22(1), 1973, pp.67–92.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” IEEE TPAMI, 32(9), 2010, pp. 1627–1645.

[4] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 2879–2886.

[5] J. Yan, Z. Lei, L. Wen, and S. Z. Li, “The fastest deformable part model for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 2497–2504.

[6] S. Liao, A. K. Jain, S. Z. Li, “A Fast and Accurate Unconstrained Face Detector,” IEEE TPAMI, vol. 38, Issue: 2, 2016, pp. 211–223.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. “Deepface: Closing the gap to human-level performance in face verification,” In CVPR, 2014, pp. 1701–1708.

[8] Y. Lin, T. Liu, and C. Fuh, “Fast object detection with occlusions,” in Proc. Eur. Conf. Comput. Vis., 2004, pp. 402–413.

[9] J. Chen, S. Shan, S. Yang, X. Chen, and W. Gao, “Modification of the adaboost-based detector for partially occluded faces,” in Proc. 18th Int. Conf. Pattern Recog., 2006, pp. 516–519.

[10] L. Goldmann, U. Monich, and T. Sikora, “Components and their topology for robust face detection in the presence of partial occlusions,” IEEE Trans. Inf. Forensics Security, vol. 2, no. 3, 2007, pp. 559–569.

[11] R. Ranjan, V. M. Patel, and R. Chellappa. “A deep pyramid deformable part model for face detection, ” IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2015, pp. 1-8.

[12] D. Chen, S. Ren, Y. Wei, X. Cao , and J. Sun, “Joint cascade face detection and alignment, ” In Computer Vision–ECCV Springer International Publishing, 2014, pp. 109-122.

[13] P, Dollár, P. Welinder, P. Perona, “Cascaded pose regression,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1078-1085

[14] H. Ronghang, W. Ruiping, S. Shiguang, C. Xilin, “Robust Head-Shoulder Detection Using a Two-Stage Cascade Framework,” 22nd International Conference on Pattern Recognition (ICPR), 2014, pp. 2796 - 2801.

[15] The Annotated Faces in the Wild (AFW) testset, https://www.ics.uci.edu/~xzhu/face/

[16] V. Jain and E. Learned-Miller, FDDB: A Benchmark for Face Detection in Unconstrained Settings, Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst. 2010.