Hybrid Cascade Model for Face Detection in the Wild Based on Normalized Pixel Difference and a Deep Convolutional Neural Network

Darijan Marčetić[0000-0002-6556-665X], Martin Soldić[0000-0002-4031-0404]

and Slobodan Ribarić[0000-0002-8708-8513]

University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{darijan.marcetic, martin.soldic, slobodan.ribaric}@fer.hr

Abstract. The main precondition for applications such as face recognition and face de-identification for privacy protection is efficient face detection in real scenes. In this paper, we propose a hybrid cascade model for face detection in the wild. The cascaded two-stage model is based on the fast normalized pixel difference (NPD) detector at the first stage, and a deep convolutional neural network (CNN) at the second stage. The outputs of the NPD detector are characterized by a very small number of false negative (FN) and a much higher number of false positive face (FP) detections. The FP detections are typically an order of magnitude higher than the FN ones. This very high number of FPs has a negative impact on recognition and/or de-identification processing time and on the naturalness of the de-identified images. To reduce the large number of FP face detections, a CNN is used at the second stage. The CNN is applied only on vague face region candidates obtained by the NPD detector that have an NPD score in the interval between two experimentally determined thresholds. The experimental results on the Annotated Faces in the Wild (AFW) test set and the Face Detection Dataset and Benchmark (FDDB) show that the hybrid cascade model significantly reduces the number of FP detections while the number of FN detections are only slightly increased.

Keywords: Face Detection in the Wild, Normalized Pixel Difference Model, Deep Convolutional Neural Networks.

1 Introduction

The main precondition for successful face-based authentication (verification or identification) and face de-identification for privacy protection [1] is efficient face detection in real scenes. The detection of faces in the wild is a challenging and very hard computer vision task. Some of the main constraining factors are the variability and diversity of the face poses, occlusions, expression variability, different illumination conditions, scale variations, and the richness of colour and texture.

Recently, many methods have been proposed for face detection: the unified model for face detection, pose estimation and landmark localization called the Deformable

Parts Model (DPM) [2], a detector based on multiple registered integral image channels [3], a face detector based on Normalized Pixel Difference (NPD) [4], a deep neural network detector [5], and a very deep convolutional network detector [6]. An alternative approach to robust face detection is based on cascades consisting of multilevel homogenous or hybrid stages. In general, the first stages are very fast but less accurate in the sense of FP detections, and the following stages are used for reducing FP detections with minimal impact on FN detections. One of the earliest homogenous cascade models for face detection is described in [7]. The model was used for fast face detection and localization in images using nonlinear Support Vector Machine (SVM) cascades of a reduced set of support vectors. The cascaded model achieved a thirty-fold speed-up compared to using the single level of a SVM. In [8], a face detection algorithm, called DP2MFD, capable of detecting faces of various sizes and poses in unconstrained conditions is proposed. It consists of deep pyramid convolutional neural networks (CNNs) at the first stage, and a deformable part model (DPM) at the second stage. The inputs of the detector are a colour image resolution pyramid with seven levels. Different CNNs are used for each level. The detector output is based on a root-filter DPM and a DPM score pyramid. Extensive experiments on AFW, FDDB, MALF, IJB-A unconstrained face detection test sets have demonstrated state-of-the-art detection performance of the cascade model. A joint cascade face detection and alignment is described in [9]. It combines the Viola-Jones detector with a low threshold to ensure high recall at the first stage, and pose indexed features with boosted regression [10] are used for face detection at the second stage. A two-stage cascade model for robust head-shoulder detection is introduced in [11]. It combines several methods as follows: a histogram of gradients (HOG) and a local binary patterns (LBP) feature-based classifier at the first stage, and a Region Covariance Matrix (RCM) at the second stage. In [12], cascade architecture built on CNNs with high discriminative capability and performance is proposed. The CNN cascade operates at multiple resolutions and quickly rejects the background regions at fast low-resolution stages, and carefully evaluates a small number of challenging candidates at the last high-resolution stage. To improve localization effectiveness and reduce the number of candidates at later stages, a so-called CNN-based calibration stage is introduced. The proposed cascade model achieves state-of-the-art performance and near real time performance for VGA resolution (14 FPS on a single CPU). In [13], we proposed a two-stage cascade model for unconstrained face detection called 2SCM. The first stage is based on the NPD detector, and the second stage uses the DPM. The experimental results on the Annotated Faces in the Wild (AFW) [14] and the Face Detection Dataset and Benchmark (FDDB) [15] showed that the two-stage model significantly reduces false positive detections while simultaneously the number of false negative detections is increased by only a few. These recent papers have shown that a multi-stage organization of several detectors significantly improves face detection results compared to "classical" one-stage approaches.

In this paper, we present a modification of the two-stage cascade model 2SCM described in [13] in such a way that the second stage is implemented by a CNN instead of a DPM. This modified hybrid two-cascade model is called the HCM.

2 Theoretical background

In the proposed hybrid two-stage cascade HCM for face detection in the wild, the NPD-based detector [4] and the CNN [16] are used. A short description of both stages follows.

2.1 Normalized pixel difference

Authors [4], inspired by Weber's Fraction [17] in experimental psychology, devised an unconstrained face detector based on the normalized pixel difference. The NPD features are defined as: $f(v_{i,j}, v_{k,l}) = (v_{i,j} - v_{k,l}) / (v_{i,j} + v_{k,l})$, where $v_{i,j}$, $v_{k,l} \geq 0$ are intensity values of two pixels at the positions (i, j) and (k, l). By definition, $f(0,0)$ is 0. The NPD feature has the following properties [4]: i) the NPD is antisymmetric; ii) the sign of NPD is an indicator of the ordinal relationship; iii) the NPD is a 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 an NPD feature amenable for threshold learning in tree-based classifiers. Due to limitations with a stump, as a basic tree classifier with only one threshold that splits a node in two leaves, the authors [4] used for learning a quadratic splitting strategy for 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. Pre-computed multiscale detector templates, based on the basic 20×20 learned face detector, are applied to detect faces at various scales. The score $ScoreNPD(I, S_i)$ is obtained based on the result of the 1226 deep quadratic trees and the 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 the square regions S_i, $i = 1, 2, ... n$, where n is the number of the 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 [4]. 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 a minimum FN, a threshold $\theta_{NPD} = 0$ was used [4].

2.2 Convolution neural network

In general, a deep CNN is composed of a series of stacked stages. Each stage can be further decomposed into several stacked layers, including a convolutional layer, a rectifying unit or a non-linear activation function, a pooling layer and sometimes a normalization layer [18]. In our model, we used the deep CNN architecture proposed in [19], which is supported by the Dlib library [19]. The CNN architecture (see Table 1) is designed for face detection and localization. The localization is based on combination of the sliding window and max-margin object detection (MMOD) approaches [16]. This specific CNN consists of seven stages implemented as convolutional layers

with a different number of convolutional kernels all having the same size of the first two dimensions, i.e. (5×5). Note that the CNN does not have pooling layers. The last stage is implemented with max-margin object detection (MMOD) [16]. The output of the CNN is specific in such a way that it specifies both the size and location of faces.

An input to the CNN is an original image $M \times N$ pixels. From this image a resolution image pyramid consisting of six levels is created. The first six stages, layers 1 to 12, of the CNN have a convolutional layer followed by a rectifying unit with an activation function defined as $f(x) = max(0, x)$, where the first three and the last three stages have a 2×2 and 1×1 stride, respectively. The seventh stage consists of only the convolutional layer with one 9×9×45 kernel with a 1×1 stride (see Table 1). This last stage of the CNN is specific due to the characteristics of the face localization problem (where both the size and location of a face are required) and thus it is implemented with max-margin object detection (MMOD). Specific details of the CNN architecture and number of parameters are given in Table 1. The 180,495 parameters of the CNN were learned on a dataset of 6,975 faces [20]. This dataset is a collection of face images selected from ImageNet, AFLW, Pascal VOC, the VGG dataset, WIDER, and FaceScrub (excluding the AFW and FDDB dataset).

Table 1. Architecture of the CNN for Face Detection and Localization

Stage	Layer		Kernels			Number of Parameters
No.	No.	Type	Number	Size	Stride	
1	1	con	16	5×5×3	2×2	16×5×5×3=1,200
	2	relu	-	-	-	0
2	3	con	32	5×5×16	2×2	32×5×5×16=12,800
	4	relu	-	-	-	0
3	5	con	32	5×5×32	2×2	32×5×5×32=25,600
	6	relu	-	-	-	0
4	7	con	45	5×5×32	1×1	45×5×5×32=36,000
	8	relu	-	-	-	0
5	9	con	45	5×5×45	1×1	45×5×5×45=50,625
	10	relu	-	-	-	0
6	11	con	45	5×5×45	1×1	45×5×5×45=50,625
	12	relu	-	-	-	0
7	13	con	1	9×9×45	1×1	1×9×9×45=3,645
	14	MMOD	-	-	-	0
Total number of parameters						180,495

3 Hybrid Cascade Face Detector

The first stage of the HCM is based on the NPD, and the second stage is the CNN. The NPD is very fast and it achieves low FN face detections for unconstrained scenes. The drawback of the NPD is a high number of FPs (typically an order of magnitude higher than FN face detections) when the operating point is set to achieve minimal FN detections. The originally proposed threshold θ_{NPD} for NDP detection is zero [4]. Fig. 1 illustrates a typical result of the NPD detector for an image with a rich texture in which the number of FP face detections is relatively high.

Fig. 1. Example of the results of an NPD detector (applied on the AFW test set) for the threshold $\theta_{NPD} = 0$. The green squares denote ground truth, the red true positive, and the blue squares denote false positive. The values *NS* corresponding to *ScoreNPD*(I, S_i) are given at the bottom of each square.

The number of FP face detections for the NPD detector can be reduced by increasing the θ_{NPD}, but this has negative effects on FNs.

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 region centre, s_i is the size of the region ($s_i \times s_i$), and j is the number of detected faces in an image I. Note that for all S_i, $i = 1, 2, ..., j$, the score *ScoreNPD* (I, S_i) is greater than zero [4].

In order to reduce FP face detections, but to keep the FNs as low as possible, the outputs of the NPD detector that have a *ScoreNPD*(I, S_i) in the interval corresponding to vague face region candidates are forwarded to the CNN detector to classify the regions S_i as face or non-face.

The procedure of the HCM is described as follows:

NPD decision stage

For every output square region S_i of the NPD in an image I

i) **IF** $ScoreNPD(I,S_i) \in [0, \theta_1]$, **THEN** the square region S_i is classified as non-face region and it is labelled as a non-face. The square region S_i is not forwarded to the CNN stage;

ii) **IF** $ScoreNPD(I,S_i) \in [\theta_2, \infty]$, **THEN** the square region S_i is classified as a face region and it is labelled as a face. The square region S_i is not forwarded to the CNN stage;

iii) **IF** $ScoreNPD(I,S_i) \in <\theta_1, \theta_2>$, i.e. *ScoreNPD* falls in an interval corresponding to the vague face region candidates, **THEN** the square region S_i is forwarded to the CNN detector;

CNN decision stage

iv) expand the square region S_i and resize S_i to uniform size;

v) **IF** the output of the CNN, called the confidence value $ConValCNN(I,S_i)$, is higher than θ_3, **THEN** the vague face region candidate S_i with the original dimensions is labelled as a face;

vi) otherwise the vague face region candidate S_i is labelled as a non-face.

Note that $ConValCNN(I,S_i)$ expresses the confidence that a face is detected in an image I at a region S_i and is defined in [16], [19].

The first two thresholds θ_1 and θ_2 define three intervals for the NPD score $ScoreNPD(I,S_i)$. The third threshold θ_3 defines two intervals for the CNN confidence value $ConValCNN(I,S_i)$. These thresholds are determined experimentally on a subset of AFW as $\theta_1 = 5$, $\theta_2 = 67$, $\theta_3 = 0.4$. They define the operating point of the HCM face detector, and are selected to maximize a sum of Precision and Recall, where Precision = TP/(TP + FP) and Recall = TP/(TP + FN), where TP is the number of correctly detected faces. All S_i which are inputs to the CNN stage are expanded by 75% of the original size in each direction and then resized to 225×225. In general, the CNN detector implemented on a single CPU (for high resolution images, e.g. 10 M pixels or more) is typically about an order of magnitude slower than the NPD detector.

This shortcoming of the CNN is circumvented in the HCM in such a way that the CNN is applied *only* to vague face candidate regions S_i (all scaled to small resolution ~50 K pixels). These characteristics justify using the CNN detector at the second stage only on *a relatively small number of scaled regions* S_i, selected based on the criterion in iii) (see the HCM procedure), and these regions are a small fraction of the whole area of an image I. For the implementation of the NPD and CNN we used program implementations [4] and [19], respectively.

4 Experimental results

In all experiments, we selected the same two test sets that were used in our previous work to compare the results obtained by the HCM with 2SCM [13]. The test sets are the AFW [14], consisting of 205 images that contain in total 468 unconstrained faces, and the subset of the FDDB [15] which contains 2,845 annotated images, with 5,171 unconstrained faces in total. On these two test sets, we performed five experiments as follows:

i) The NPD detector ($ScoreNPD(I,S_i) > \theta_{NPD}$, where the threshold $\theta_{NPD} = 0$ as originally proposed in [4]) achieved the results given in Table 2 and in Figs 5(a) and (e).

ii) The DPM detector (threshold $\theta_{DPM} = -0.65$ [2]) achieved the results given in Table 2 and in Figs 5(b) and (f).

Based on the above results of the first two experiments, we can see that the NPD detector, in general, achieved relatively low FN but high FP face detections in comparison with the DPM.

iii) The CNN detector ($ConValCNN(I,S_i) > \theta_{CNN}$, where the threshold $\theta_{CNN} = 0.4$ [16]) achieved the results given in Table 2 and in Figs 5(c) and (g). The results showed that the CNN outperforms both NPD and DPM detectors in the sense of FNs and FPs. Note that all 23 FPs (for the AFW test set) are due to superficial manual annotation. Fig. 2 depicts all 23 FPs for the AFW (first row) and all 29 FPs for FDDB (second row) which are not manually annotated as faces thus true FPs is 0 for both test sets.

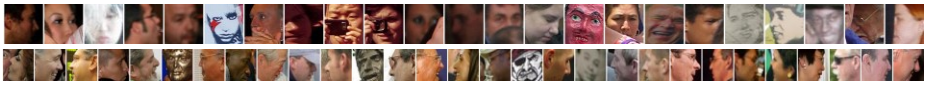

Fig. 2. Faces detected by the CNN in the AFW (first row) and FDDB test sets (second row) which are not manually annotated as faces.

Note that the face detection time of the CNN is greater than that of the NPD and the DPM when a single CPU is used. Using a GPU, the face detection time of the CNN is comparable with that of the NPD (on a single CPU). For example, the single CPU processing time for the NPD detector is 12.59 s for an image resolution 2138×2811, while for the CNN detector, the processing time for a single CPU and the GPU (384 cores) is 143.54 s and 3.55 s, respectively. The single CPU processing time for the NPD detector is 0.94 s for an image resolution 768×1024, while for the CNN detector, the processing time for a single CPU and the GPU (384 cores) is 17.05 s and 2.10 s, respectively.

iv) The two-stage cascade model 2SCM described in [13] achieved the results given in Table 2 and in Figs 5(d) and (h).

v) The proposed HCM described in Section 3, achieved the following results (Table 2, Figs 5(i) and (j)): on the AFW test set 28 faces were not detected (FNs) among 468 faces in 205 images. Note that the 783 FPs from the first stage of the HCM are reduced to only 4 true FPs at the second stage. The remaining FPs, i.e. 19 = 8 + 11, where all FPs that are the result of poor manual annotation (see Fig. 3) - 8 FPs are resulted in the first stage (which are not forwarded to the CNN stage), and 11 FPs are obtained at the second stage of the HCM (from vague face candidate regions). Similar arguments are

valid for the FDDB where from 54 FPs there are 14 (i.e. 13 +1) true FPs while the remaining 40 (i.e. 28 + 12) FPs are the result of poor manual annotation.

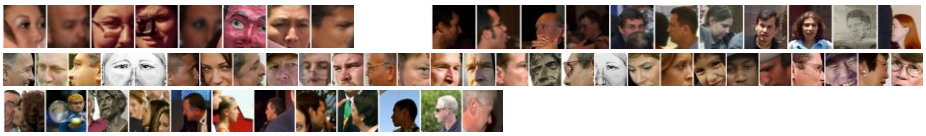

Fig. 3. Images, detected by the HCM, which are not manually annotated as faces (ground truth). Faces detected at the first stage (first 8 images) and at the second stage (11 images) in the AFW test set (first row); Faces detected by the HCM at the first stage (28 images in the second row) and 12 images in the third row for FDDB test set.

When HCM is compared with 2SCM on the AFW test set, the FNs at the second stage are reduced by more than 1.71 times (from 48 to 28 FPs), while the FPs are the same (23 FPs). Note that most FP face detections (19 from 23 FPs for the second stage of HCM) are due to the poor annotation of the AFW test set by humans as was noted in [21]. The CNN can detect "difficult faces" that are missed by most human annotators, which increases FPs. The results obtained for the FDDB test set are given in Table 3. When the HCM is compared with 2SCM on the FDDB test set, the FNs are reduced by more than 1.19 times (from 1081 to 901 FNs), while the FPs are decreased by 2.25 times (from 122 to 54 FNs). The results in terms of precision and recall are given in Table 3.

Our proposed HCM face detector achieves face detection results comparable to the results of the state-of-the-art CNN detector. The main advantage of the HCM is that it has CPU face detection time that is about 15 times faster than the CNN run on a single CPU (image resolution 768x1024). CPU face detection time for the HCM is comparable with the CNN detection time when using a GPU (384 cores). An illustration of results of the HCM applied on image from Fig. 1. is depicted in Fig. 4.

Table 2. Results of the experiments performed on AFW test set.

Test set			**AFW [14]**			
Number of images			205			
Number of faces			468			
Method		Thresholds	FN	FP	Precision	Recall
1	NPD	$\theta_{NPD} = 0$	28	783	0.360	0.940
2	DPM	$\theta_{DPM} = -0.65$	161	72	0.810	0.656
3	CNN	$\theta_{CNN} = 0.4$	7	23 ($0^{1)}$)	0.952 ($1^{1)}$)	0.985
4	2SCM	$\theta_1 = 44$, $\theta_2 = 5$, $\theta_3 = -0.65$	48 (28+20)	23	0.948	0.897
5	**HCM**	$\theta_1 = 5$, $\theta_2 = 67$, $\theta_3 = 0.4$	**28**	**23** **12($4^{2)}$) + 11($0^{2)}$)**	**0.952** **(0.991$^{2)}$)**	**0.940**

[1)] Please see iii) Section 4,

[2)] See v) Section 4.

Table 3. Results of the experiments performed on FDDB test set

Test set			FDDB [15]			
Number of images			2845			
Number of faces			5171			
Method		Thresholds	FN	FP	Precision	Recall
1	NPD	$\theta_{NPD} = 0$	857	1902	0.694	0.834
2	DPM	$\theta_{DPM} = -0.65$	1250	63	0.984	0.758
3	CNN	$\theta_{CNN} = 0.4$	778	29 ($0^{1)}$)	0.993 ($1^{1)}$)	0.850
4	2SCM	$\theta_1 = 44$, $\theta_2 = 5$, $\theta_3 = -0.65$	1081 (857+224)	122	0.971	0.791
5	**HCM**	$\theta_1 = 5$, $\theta_2 = 67$, $\theta_3 = 0.4$	**901**	**54** **41($13^{2)}$)+13($1^{2)}$)**	**0.988** **($0.997^{2)}$)**	**0.826**

[1)] Please see iii) in Section 4,

[2)] See v) Section 4.

Fig. 4. Example of the results of the HCM (applied on the AFW test set). The green squares denote ground truth, the red true positive, and the blue squares denote false positive. The values NS and CS correspond to *ScoreNPD*(I, S_i) and *ConValCNN*(I,S_i), respectively.

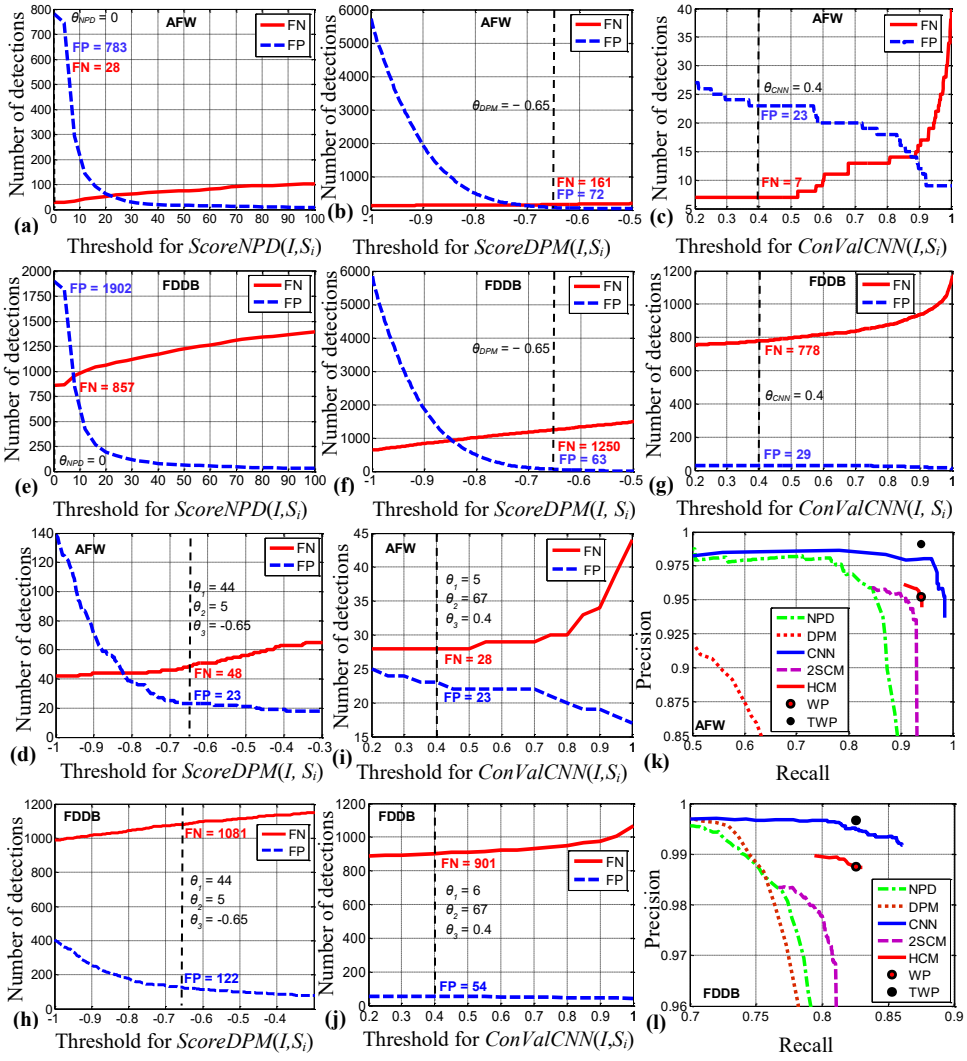

Fig. 5. The results of experiments for the AFW and FDDB test sets. The results of experiments: (a) and (e) NPD detector; (b) and (f) DPM detector; (c) and (g) CNN detector; (d) and (h) 2SCM detector; (i) and (j) the proposed HCM detector; (k) precision and recall of experiments for (a)-(d), (i) for AFW test set; (l) precision and recall of experiments for (e)-(h), (j) for FDDB test set. Working point WP and true working point TWP correspond to results obtained with HCM parameters (thresholds: $\theta_1 = 5$, $\theta_2 = 67$, $\theta_3 = 0.4$) for ground truth and manually verified annotation, respectively.

5 Conclusion

In this paper, we have proposed the hybrid cascade model HCM for unconstrained face detection. The first stage of the model is based on the NPD detector, and the second on the deep CNN-based detector. The model is introduced to reduce FP face detections, but to keep FNs as low as possible. This is achieved by conditionally forwarding the outputs of the NPD detector that represent vague face candidate regions to the CNN stage. The condition for forwarding is based on a value of the NPD detector score. The arguments for using the proposed model are: 1) the NPD detector is used at the first stage of the HCM because it is much faster (about 15 times) than the CNN for face detection and localization on a single CPU; 2) the CNN detector is used conditionally as *a post classier* and operates *only on a small number* of rescaled vague face candidate regions which are the outputs of the NPD detector. This enables effective implementation of a second stage of the HCM. The achieved time performance is comparable to the NPD which is considered one of the fastest state-of-the-art detectors [4]. The HCM is suitable for mobile and embedded platforms. 3) NPD and CNN detectors use features which are uncorrelated (normalized pixel differences vs. convolution-based features) what makes the HCM more robust then the NPD.

Experiments performed on the AFW test set showed that FPs were reduced by more than 34 times (from 783 to 23 FPs), while FNs were not increased compared with the NPD detector for the originally proposed threshold [4]. For the FDDB test set, FPs were reduced by more than 35 times (from 1902 to 54 FPs), while FNs were increased 1.05 times (from 857 to 901 FNs) for the same NPD threshold. We are aware that the CNN-based detector applied on the whole image outperforms (in terms of the Recall and Precision) the HCM, but the CNN computational load is about 15 times higher than the computational load of the HCM. Our future work will aim to improve the cascade model to decrease the number of false negative face detections and introduce the stage for face pose estimation.

Acknowledgments

This work has been supported by the Croatian Science Foundation under project 6733 De-identification for Privacy Protection in Surveillance Systems (DePPSS).

References

1. Ribarić, S., Ariyaeeinia A., and Pavešić, N.: De-identification for privacy protection in multimedia content: A survey. In: Signal Processing: Image Communication, vol. 47, pp. 131–151, (2016).
2. Zhu, X., and Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 2879–2886, (2012).
3. Dollár, P., Tu, Z., Perona, P.,and Belongie, S.: Integral Channel Features. In: Proc. British Machine Vision Conf., pp. 1–11, (2009).
4. Liao, S., Jain, A. K., Li, S. Z.: A Fast and Accurate Unconstrained Face Detector. IEEE TPAMI, vol. 38, Issue: 2, pp. 211–223, (2016).

5. Taigman, Y., Yang, M., Ranzato, M., and Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708, (2014).
6. Simonyan K., and Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. International Conference on Learning Representations, http://arxiv.org/abs/1409.1556, (2014).
7. Romdhani, S., Torr, P., Schölkopf B., and Blake, A.: Efficient face detection by a cascaded support–vector machine expansion. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 460, No. 2051, pp. 3283–3297, (2004).
8. Ranjan, R., Patel, V. M., and Chellappa. R.: A deep pyramid deformable part model for face detection. In: IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8, (2015).
9. Chen, D., Ren, S., Wei, Y., Cao, X., and Sun, J.: Joint cascade face detection and alignment. In: Computer Vision–ECCV Springer International Publishing, pp. 109–122, (2014).
10. Dollár, P., Welinder P., Perona, P.: Cascaded pose regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1078–1085, (2010).
11. Ronghang, H., Ruiping, W., Shiguang, S., Xilin, C.: Robust Head-Shoulder Detection Using a Two-Stage Cascade Framework. In: 22nd International Conference on Pattern Recognition (ICPR), pp. 2796–2801, (2014).
12. Li, H., Lin, Z., Shen, X., Brandt J., and Hua, G.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334, (2015).
13. Marčetić, D., Hrkać, T., and Ribarić, S.: Two-stage cascade model for unconstrained face detection.. In: IEEE International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), pp. 1–4, (2016).
14. The Annotated Faces in the Wild (AFW) testset, https://www.ics.uci.edu/~xzhu/face/, last accessed 2017/03/21.
15. Jain, V., and Learned-Miller, E.: FDDB: A Benchmark for Face Detection in Unconstrained Settings. In: Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst., (2010).
16. King, D. E.: Max-margin object detection. In: arXiv preprint arXiv:1502.00046, (2015).
17. Weber, E. H.: Tastsinn und Gemeingefühl. In: R. Wagner (Ed.), Hand- wörterbuch der Physiologie, Vol. III, pp. 481–588, (1846).
18. Wu, Z., Huang, Y., Wang, L. , Wang, X., and Tan, T.: A comprehensive study on cross-view gait based human identification with deep cnns. In: IEEE TPAMI, vol .39, issue 2, pp. 209–226, (2017).
19. King, D. E.: Dlib-ml: A Machine Learning Toolkit. In: Journal of Machine Learning Research 10, pp. 1755–1758, (2009).
20. http://dlib.net/files/data/dlib_face_detection_dataset-2016-09-30.tar.gz, last accessed 2017/03/21.
21. Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L.: Face detection without bells and whistles. In: European Conference on Computer Vision, pp. 720-735, (2014).