Face Classification: A Specialized Benchmark Study

Jiali Duan[1(✉)], Shengcai Liao[2], Shuai Zhou[3], and Stan Z. Li[2]

[1] School Of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China
jli.duan@gmail.com

[2] Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, University of Chinese Academy of Sciences, Beijing, China
{scliao,szli}@nlpr.ia.ac.cn

[3] Macau University of Science and Technology, Taipa, Macau
shuaizhou.palm@gmail.com

Abstract. Face detection evaluation generally involves three steps: block generation, face classification, and post-processing. However, firstly, face detection performance is largely influenced by block generation and post-processing, concealing the performance of face classification core module. Secondly, implementing and optimizing all the three steps results in a very heavy work, which is a big barrier for researchers who only cares about classification. Motivated by this, we conduct a specialized benchmark study in this paper, which focuses purely on face classification. We start with face proposals, and build a benchmark dataset with about 3.5 million patches for two-class face/non-face classification. Results with several baseline algorithms show that, without the help of post-processing, the performance of face classification itself is still not very satisfactory, even with a powerful CNN method. We'll release this benchmark to help assess performance of face classification only, and ease the participation of other related researchers.

Keywords: Face detection · Face classification · Benchmark evaluation

1 Introduction

Face detection is a key and fundamental problem in facial analysis as it is usually the first step to other high-level tasks such as face alignment, face recognition, face attribute analysis, etc. Therefore, a well-designed benchmark is essential to analyze the performance of face detection algorithms and advance the face detection research.

In the literature, face detection is evaluated by scanning a set of images containing faces in background, and counting true positives and false positives by matching the detected bounding boxes with the ground truth. This evaluation procedure generally involves three steps: block generation (multi-scale sliding subwindows or objectness proposals), face classification, and post-processing

Z. You et al. (Eds.): CCBR 2016, LNCS 9967, pp. 22–29, 2016.
DOI: 10.1007/978-3-319-46654-5_3

(non-maximum suppression, bounding box regression, etc.). Today, popular face detection benchmarks such as AFW [4], FDDB [2], and WIDER FACE [3] still continue to use such evaluation procedure.

However, on one hand, the performance of face detection methods is largely influenced by specific settings of block generation and post-processing, therefore, it is not easy to know the specific performance of the core module, namely the face classification part in existing methods. On the other hand, implementing all the three steps and achieving a good overall face detection performance results in a very heavy work, sometimes preventing researchers of related fields (e.g. feature and classification researchers) in introducing their ideas for face detection. For example, the AFW [4], FDDB [2], and WIDER FACE [3] benchmarks all require that researchers start from scratch, from generating blocks, classifying faces, to post-processing.

Motivated by this, we conduct a specialized large-scale benchmark study in this paper, which focuses purely on face classification. We start with face proposals, by which about 3.5 millions of face and non-face sample patches are collected. Then, we build a large-scale benchmark dataset for two-class face classification evaluation. Accordingly, we evaluate several feature extraction methods and classification algorithms and compare their performance. Our results show that, without the help of post-processing, the performance of face classification itself is still not very satisfactory, even with a powerful CNN method.

The data and evaluation code of this study will be released to the public[1] to help assess performance of face classification, and ease the participation of related researchers who want to try their algorithms for face detection. With this benchmark, researchers only need to do feature extraction and face classification per image patch, regardless the troubling block generation and post-processing tasks. Even more easily, we provide some baseline features, so that general classification researchers are able to evaluate their classification algorithms.

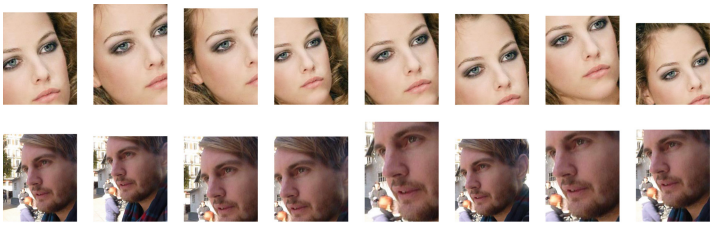
2 Face Classification Benchmark

2.1 Face Proposals

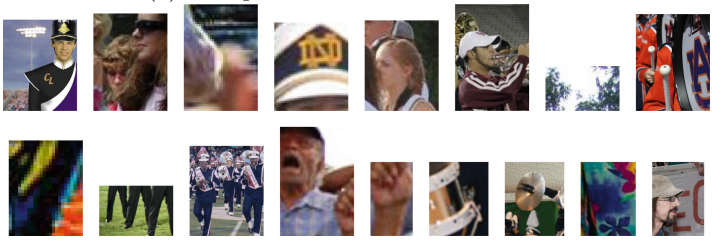
RPN network is employed to extract proposals inspired by the work of Faster R-CNN [9] and its recent application in face detection [10]. Note that generic object proposal-generating methods such as [17–19] are not very suitable for our classification benchmark because the amount of positive face patches generated is too scarce and those patches are not as discriminative as a specially trained RPN face proposal network.

We trained a 4-anchor RPN network with some slight modifications on the fc6 and fc7 InnerProduct layer of Zeiler and Fergus model [12]. The default anchor ratio was set to 1:1 and we compute anchors at 4 different scales (2,8,16,32). During training, the number of categories are modified to 2 (face and background) with 2 * 4 bounding box coordinates to be predicted. Softmax and smoothL1 loss are deployed for training classification and bounding box prediction respectively.

[1] https://davidsonic.github.io/index/ccbr_2016



(a) Face patches extracted from RPN.



(b) Non-face patches extracted from RPN.

Fig. 1. Sample proposals generated by RPN. When the IOU between a face proposal generated by RPN and a ground-truth label is greater than 0.5, we treat it as a face patch. Otherwise, when the IOU is below 0.3, we treat it as a non-face patch

All the proposals are generated from the training set of WIDER FACE [3] containing 12,880 high-resolution images through this RPN convolutional neural network. There are about 300 proposals extracted from each image. WIDER FACE contains 393,703 labelled faces collected from 32,203 images with a wide variety in scales, poses, occlusions and expressions. What's worth noticing about WIDER FACE is that the images collected are taken in crowded scenes and proves to be an effective training set and challenging evaluation set.

In our benchmark dataset, when the Intersection over Union (IOU) between a face proposal generated by RPN and a ground-truth label is greater than 0.5, we treat it as a face patch. Otherwise, when the IOU is below 0.3, we treat it as a non-face patch. As a result, we collected a face classification benchmark (FCB) database, which contains 3,558,142 proposals in total, of which 198,616 being face patches. All images are resized to 24×24 for face classification benchmark. Note that by doing so there may be some changes to the original aspect ratio. Sample face patches and non-face patches extracted from the RPN network are displayed in Fig. 1.

2.2 Benchmark Protocol

As shown in Table 1, proposals are extracted from the first 6,440 images in WIDER_FACE are used as the training set, while the remaining proposals from the next 6,440 images are used as testing set. In the testing set the size of Non-face patches are about 20 times the size of face proposals. So if an algorithm

Table 1. Details of FCB Benchmark

Dataset	#Img	#Face patches	#Non-face patches
Training	6,440	112,124	1,666,947
Testing	6,440	86,492	1,692,579
Total	12,880	198,616	3,359,526

classifies all the positive samples as negative ones, it would still get 95 % two-class classification accuracy. Under such circumstances, it would be biased in the choice of our face detection algorithms if we solely take the two-class classification accuracy as our evaluation index.

To reveal the 'true' performance of a face classifier, we put stress on performance at low False Accept Rate (FAR). Following the method in [5], the performance of the proposed algorithms are displayed via ROC curve by varying the confident score threshold, with FAR in log space being the x axis and True Positive Rate (TPR) being the y axis. Specifically, we measure the true positive rate at FAR=10^{-3} to compare the performance of different algorithms. Since there are altogether 1,779,071 proposals extracted from 6,440 images in the test set, it means we only allow for False Positive Per Image (FPPI) of 0.28 in an image, which is both challenging and persuasive in terms of real world applications.

3 Evaluation and Results

3.1 Feature Extraction and Classification Methods

Traditional Methods: Illumination changes, occlusions and pose variations are three fundamental problems for face detection under unconstrained settings. Illumination-invariance is obtained in LOMO [8] by applying the Retinex transform and the Scale Invariant Local Ternary Pattern (SILTP) for feature representation. NPD [7] gives the nice properties of scale-invariance, boundedness and its feature involves only two pixel values, hence robust to occlusion and blur or low image resolution. We use the open source code of these two feature representation methods in our experiments. Besides, LBP [14] together with its variant MB-LBP [15] are also re-implemented for evaluation. We adopt DQT+boosting [7] as our baseline classifier. We also tried SVM, but it appears to be not effective to handle this challenging problem, and it is also not efficient for our large-scale data. Therefore, we leaved SVM out finally.

CNN Methods: Convolutional Neural Network based methods have received more and more attention due to its effectiveness in computer vision tasks. In our experiments, a CIFAR-10 Net [6] based binary classification CNN and a Cascade-CNN following the paradigm of [11] have been implemented.

Several CNN structures have been explored and we picked one with the best performance based on CIFAR-10 and its detailed information is listed in Table 2.

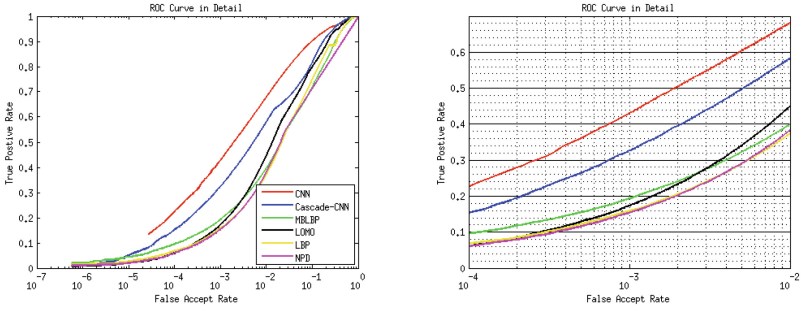
Table 2. Model structure of CIFAR10-based CNN

Layer name	Filter number	Filter size	Stride	Padding	AF
conv1	32	3×3	1	2	-
max_pool1	-	2×2	2	-	RELU
conv2	32	3×3	1	2	RELU
ave_pool2	-	2×2	2	-	-
conv3	64	3×3	1	2	RELU
ave_pool3	-	2×2	2	-	-
ip1	64	-	-	-	-
ip2	2	-	-	-	-

As for the structure of Cascade-CNN, please refer to [11]. Note that for training Cascade-CNN, hard negative samples mined by the former net are used as non-face samples to be used for the following net training and in training each net, hard positive samples mined are aggregated to face proposals for further fine-tuning.

3.2 Results and Discussion

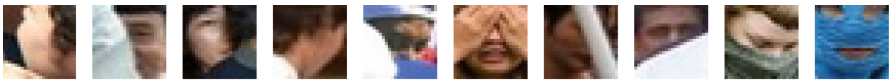
The evaluation results on the whole test set are shown in Fig. 2. From the results, we can see that the CIFAR-10 based CNN beats other methods due to its powerful representation ability. Besides, it also outperforms Cascade-CNN by a large margin. This is probably because in the process of cascade training, hard negative samples mined by the last net contain a proportion of non-face patches that actually contain faces but with a IOU less than 0.3, thus 'too hard' to identify for a shallow net structure (see Fig. 3). Another possible reason is that Cascade-CNN in [11] followed a pipeline of alternation between classification net and calibration net; the Non Maximum Suppresion (NMS) is also deployed during the Cascade-CNN training process, in contrast to the FCB evaluation, which only requires classification net.



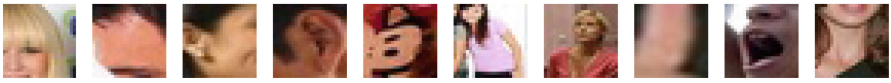
(a) ROC curve on FCB (b) enlarged

Fig. 2. Evaluation results on FCB test set, best viewed in color

However, even with the best performer CIFAR-10, the performance on FCB is still far from satisfactory as observed from Fig. 2, revealing that FCB is a large-scale and challenging benchmark on face classification. Note that we only used two CNN based methods while a more sophisticated network structure should be able to achieve better performance.



(a) Hard positive samples mined by 24-net.



(b) Hard negative samples mined by 24-net.

Fig. 3. Some hard positive samples and some hard negative samples mined by the 24-net in the Cascade-CNN method [11]

As for hand-crafted features, MB-LBP obtains about 20 % detection rate compared to LBP (about 16 %) detection rate at FAR of 10^{-3}, which is reasonable considering that MB-LBP encodes not only microstructures but also macrostructures of image patterns, thus more comprehensive and robust than LBP.

Table 3. Detection rate (%) of each algorithm at FAR $= 10^{-3}$

LOMO	LBP	MB-LBP	NPD	CNN	Cascade CNN
17.42	15.98	19.42	15.47	**43.10**	32.66

At FAR of 10^{-3}, LOMO performs slightly better than LBP, but its true positive rate increases rapidly as we move along the righthand direction of x axis. It is not until FAR of 0.3 % that LOMO outperforms MB-LBP. NPD achieves comparable performance compared to LBP. As its values are computed involving only two pixels of an image, NPD is more sensitive than other algorithms considering that face patches on FCB are not aligned. At FAR of 10^{-2}, the overall rank is almost the same except that LOMO performs better than MB-LBP. Please refer to Table 3 for more details.

Table 4 is a further illustration about the dimension of the original features and selected features by the DQT based AdaBoost, as well as the training time, testing time and platform of each algorithm employed in our evaluation. Note that LOMO requires RGB images while other methods use gray images as input. The first 4 algorithms are all combination of feature extraction and boosting with Deep Quadratic Trees (DQT) while the last two are end to end CNN

Table 4. Model details and speed of each algorithm

Algorithm	#Features	#Selected features	Training time (h)	Testing time (h)	Platform
LOMO	7,252	283,323	5.35	1.52	X5650CPU
LBP	768	20,269	1.80	0.44	X5650CPU
NPD	165,600	6,877,535	6.42	1.10	X5650CPU
MB-LBP	5,120	228,606	3.91	0.34	X5650CPU
CIFAR-10 CNN	64	64	2.20	0.50	K40-GPU
Cascade-CNN	560	560	5.85	0.60	Titan-GPU

methods, therefore the dimension of original features and selected features are the same. Besides, CNN is generally faster to train since it takes the advantage of GPU parallelization, but it's also due to this fact that CNN runs less efficiently compared to traditional methods on CPU or hand-held devices.

4 Conclusion

Face detection generally involves three steps with face classification being its core module. However, it is not easy to determine the actual performance of the face classification part due to the large influence of block generation and post-processing in traditional benchmarks. Motivated by this, we conduct a specialized benchmark study in this paper, which focuses purely on face classification. We start with face proposals by collecting about 3.5 millions of face and non-face sample patches, and build a benchmark dataset (FCB) for two-class face classification evaluation. Our results show that, without the help of post-processing, the performance of face classification itself is still not very satisfactory, even with a powerful CNN method. The data and evaluation code of this study will be released to the public to help assess performance of face classification, and ease the participation of other related researchers who want to try their algorithms for face detection.

Acknowledgements. This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects #61473291, #61572501, #61502491, #61572536, NVIDIA GPU donation program and AuthenMetric R&D Funds.

References

1. Kostinger, M., Wohlhart, P., Roth, P.M., et al.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151. IEEE (2011)
2. Jain, V., Erik Learned-Miller, F.: A benchmark for face detection in unconstrained settings. Technical Report: UM-CS-2010-009 (2010)
3. Yang, S., Luo, P., Loy, C.C., Tang, X., WIDER FACE: a face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

4. Zhu, X., Ramanan, D.: Face detection, pose estimation, landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR) (2012)
5. Dollar, P., Wojek, C., Schiele, B., et al.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
6. The CIFAR-10 dataset. https://www.cs.toronto.edu/~kriz/cifar.html
7. Liao, S., Jain, A.K., Li, S.Z.: A fast and accurate unconstrained face detector. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 211–223 (2016)
8. Liao, S., Hu, Y., Zhu, X., et al.: Person re-identification by local maximal occurrence representation, metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
9. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
10. Jiang, H., Learned-Miller, E.: Face detection with the faster R-CNN. arXiv preprint (2016). arXiv:1606.03473
11. Li, H., Lin, Z., Shen, X., et al.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334 (2015)
12. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_53
13. Mathias, M., Benenson, R., Pedersoli, M., Gool, L.: Face detection without bells and whistles. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 720–735. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_47
14. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
15. Liao, S., Zhu, X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 828–837. Springer, Heidelberg (2007). doi:10.1007/978-3-540-74549-5_87
16. Yan, J., Zhang, X., Lei, Z., et al.: Face detection by structural models. Image Vis. Comput. **32**(10), 790–799 (2014)
17. Van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., et al.: Segmentation as selective search for object recognition. In: International Conference on Computer Vision, pp. 1879–1886. IEEE (2011)
18. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_26
19. Arbelez, P., Pont-Tuset, J., Barron, J.T., et al.: Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 328–335 (2014)