Multiple Task-driven Face Detection Based on Super-resolution Pyramid Network

Jianjun Li[1], Juxian Wang[1], Xingchen Chen[2], Zhenxing Luo[3], Zhugang Song[3]

[1]School of Computer Science and Engineering, Hangzhou Dianzi University, China
[2] Department of Computer Science, Shananxi University of Technology, China
[3] Science and Technology on Communication Information Security Control Laboratory, the 36th Institute of CETC of China
jianjun.li@hdu.edu.cn, wjx2018@gmail.com, cxc2019@gmail.com, zx.luo@jec.com.cn, z.song@jec.com.cn

Abstract

Although research in face detection and recognition has achieved tremendous progress through the various frameworks that are being put forward every year, face detection under complex circumstances is still a challenging issue. Multiple task-driven face detection has wide applications, such as crowd number estimation, face recognition attendance and so on. In this paper, we propose a multiple task-driven cascade detection networks based on super-resolution Pyramid, to effectively tackle the following challenges in face detection: low-resolution faces under the lens; faces from blur, illumination, scale, pose, expression and occlusion. Our method integrates the advantages of the super-resolution technology and an efficient image pyramid structure. The design of this structure not only recover high frequency information lost in the sampling process, but also can handle multi-scale invariants. Also, facial landmarks play non-negligible roles during detection. Our method achieves state-of-the-art results over prior arts on both the WIDER FACE dataset and the Face Detection Dataset and Benchmark (FDDB), and our results show a higher average detection precision of 90%. Notably, we demonstrate superior performance and robustness in a challenging environment.

Keywords: Face detection, Super-resolution, Cascaded conventional neural network, Facial landmarks

1 Introduction

With the rapid development of deep learning technology, face detection in the field of computer vision has been greatly improved. However, there are still some problems in practical applications, especially in public video surveillance. Due to the far distant between faces and lens, the low resolution of video and serious face occlusion etc., the image resolution is low and the identification difficulty is increased. In order to solve the above problems, it is necessary to reconstruct low-quality, low-resolution video images into high-quality, high-resolution face images, to improve image recognition accuracy.

The current image super-resolution techniques mainly include two methods: the reconstruction-based image super-resolution method and the learning-based image super-resolution method. The reconstruction-based super-resolution method, while relatively mature, requires additional degradation models that conform to the actual imaging conditions, and accurately estimate the sub-pixel motion in the image sequence, which increases the difficulty in the process and does not make good use of the prior knowledge of the image. Compared with the reconstruction-based super-resolution method, learning-based super-resolution can reduce the computational complexity and pertinently recover the object by learning prior knowledge.

Besides solving the universality issues of face detection, such as heavy blur, overlap, extreme illumination, small objects and irregular posture etc., our network architecture also effectively solve the problems of low resolution, small target face and serious face occlusion from surveillance video in public contexts. In this paper, we named the proposed model Super-Resolution Pyramid Convolution Neural Network (SRPN-CNN). This framework constructs a super-resolution image pyramid (SRPN) based on multiple task-driven detection networks, which distinguishes it from other prior image pyramids in many aspects. On the contrary, the resolution of the image was taken into consideration during the construction of the image pyramid based on super-resolution and it has the function of noise reduction as well. The multiple task-driven detection network aims to detect faces through the cascade networks: the first of which mainly implements face detection. The second network employs facial landmarks to assist with final face detection. In this stage, it is conducive to reduce the probability of false detections. The aim of increasing the accuracy was achieved accordingly.

*Corresponding Author: Jianjun Li; Email: jianjun.li@hdu.edu.cn
DOI: 10.3966/160792642019072004024

Overall, the contributions of this paper are mainly summarized in the following three aspects:

As stated above, the key issue is: Can we propose a model that addresses the comprehensive problem? In this paper, our network architecture effectively solves the complex issues of heavy blur, overlap, extreme illumination, small objects and irregular posture, etc. We named the proposed model Super-Resolution Pyramid Convolution Neural Network (SRPN-CNN). This framework integrates two cascaded networks: the first of which completes the up-sampling operation at different factors to directly build an exquisite image pyramid, which distinguishes it from other prior image pyramids in many aspects. On the contrary, the resolution of the image was taken into consideration during the construction of the image pyramid based on super-resolution and it has the function of noise reduction as well. The second network aims to detect faces through a deep convolutional neural network. Finally, we attempt to use facial landmarks to assist with final face detection. In this stage, it is conducive to reduce the probability of false detections. The aim of increasing the accuracy was achieved accordingly. Overall, the contributions of this paper are mainly summarized in the following three aspects:

(1) We put forward a new image Pyramid based on super-resolution reconstruction technology. At each scale of the image pyramid, our pyramid network recovers the lost high-frequency information in the process of image sampling and greatly restore the image. In addition, the image obtained with SRPN has the effect of sharpening and denoising. We conducted extensive experiments using this model and the experimental results demonstrate that our proposed pyramid model improves the facial recognition performance as shown in Table 2 of Section 3.2.

(2) We propose a multiple task-driven face detection algorithm based on ResNet (MRF-CNN). Our cascaded network we designed reduces the computation cost in both forward and backward propagation, accelerating convergence rate. In addition, we explore the relationship between face detection methods and facial landmarks locating approaches. Analysis results show that facial landmarks accuracy contribute more improvements of face detection accuracy. We designed a powerful CNN network to locate facial landmarks accurately in order to reduce the false positive rate as much as possible.

(3) Many experiments were conducted on challenging face datasets, and the results demonstrate that our proposed multiple task-driven cascaded network outperformed all compared methods. Each characteristic of our cascaded networks optimized our overall framework to better adapt to complex circumstances. For example, our approach not only is robust against blurring and tiny objects engendered by the lens, but also has strong capability of detection in occlusion, changing illumination or posture, etc. However, existing methods based CNN do not provide such flexibility.

The rest of the paper is organized as follows: In Section 2, we briefly introduce the related research in the area of face detection and image resolution. Section 3 presents the framework of our work and provides a more detailed description for each module. Section 4 provides the experimental results, and conclusions are given in Section 5.

2 Related Works

2.1 Multi-scale Representation

Multi-scale representation is very important in image processing. David Lowe [1] proposed a scale invariant feature transform algorithm (SIFT) that keeps the invariance of image translation, rotation, zoom, and affine transformation. Speed up robust features (SURF) proposed by H. Bay [2] is similar as the above SIFT method. In Ramanan et al.'s work [3], their algorithm mainly uses multiscale representation and deformable part model for object detection and their result outperforms the best results in the 2007 challenge. Researchers have shifted from visual geometric restoration to more object recognition problems since the emergence of Bag of Words (BoW), spatial pyramids and vector quantization. Single Shot MultiBox Detector (SSD) [4] is the recent technique that obtains predicting category scores and a series of fixed-size bounding boxes. This network is different from the basic CNN network and it adds additional auxiliary structures. However, all these multi-scale representations describe local features of images in a simple form at different scales. Our work aims to choose the appropriate scale invariant method to optimize our model so that our model can be better applied to the multi-scale face detections.

2.2 Image Super-resolution

According to whether training samples are dependent on each other, super resolution algorithms of a single image can be categorized into two types: methods based on enhanced edges (non-example-based) and methods based on learning (example-based). The example-based method has been a hot spot in recent years. It extracts the high-frequency information model from the training samples and is followed by predicting the information of test samples using the machine learning method in order to achieve an improved image resolution.

Most of the learning-based super-resolution methods are patch-based methods, which generate small image patches from the input image and calculates the higher-resolution image patches corresponding to the lower-resolution image patches. It was first proposed in Pentland's work [5]. These studies vary depending on how to construct a model and how to select a training

set. Freeman et al. [6] proposed a Markov random field (MRF) to deal with low-level vision tasks. Chang et al. [7] introduced a neighbor embedding technique, which uses the training samples effectively. However, these algorithms are less efficient. Researchers have therefore proposed a series of improved algorithms [8-9] to accelerate the operational speed. The pioneering work of Yang et al. [10-11] assumed that the input of low resolution patches could be represented by a sparse linear expression.

The super-resolution reconstruction method based on learning for face images [12-13] is also a related research hot spot that can improve the accuracy of face recognition. Hennings Yeomans et al.'s [14] work proposed to simultaneously perform both the super-resolution reconstruction and the face recognition, and obtain both the result of face recognition and the super-resolution of the face image. However, none of them applied the three-channel image super-resolution technique into face detection.

2.3 Face Detection and Facial Landmarks' Location

Face detection is the key step to face recognition and an indispensable part of face detection applications. However, it also meets with many challenges, such as blur, occlusion, extreme lighting, and large pose variation, in real applications. Early detection methods based on geometric features [15-18] have characteristics of small storage and immunity to illumination interference, but require high quality of image and high accuracy of feature points. These external conditions are often used as an aid to guarantee the accuracy of the detection. Compared with these traditional methods [19-23], the method based on machine learning has its unique advantages in face detection and recognition. When implemented with GPU, it can significantly improve detection speed. In recent years, more complex networks, such as VGGNet [24], GoogleNet [25], ResNet [26], etc. have been applied to face detection.

The location of facial landmarks is not only a crucial problem in face recognition research field, but also a basic problem in the field of graphics and computer vision, their purpose is to locate landmark information on the images that correspond to facial features such as the eyes, nose and mouth. The basic idea of traditional location algorithms, such as active shape model ASM [27] and active appearance model AAM [28], is that they combine the texture features of faces with the position constraints among the feature points. Certainly, there are template fitting approaches as well, such as the methods of [29-31]. Recently, Zhang et al. [32] use the structure of multiple deep convolutional neural network to enhance performance through multi-task learning.

However, most of the available face detection and face alignment methods have not exploited the inherent correlation between the above-described two tasks. We investigate to associating facial landmarks' location and face detection.

3 The Proposed Approach

In this section, we will introduce the architecture of multiple task-driven cascaded networks for face detection and describe the characteristic of each proposed approach in detail.

3.1 Overall Framework

The overall framework mainly includes three parts: the construction of the super-resolution pyramid, rough face detection based on MR-net and face refinement based on F-net, as shown in Figure 1.

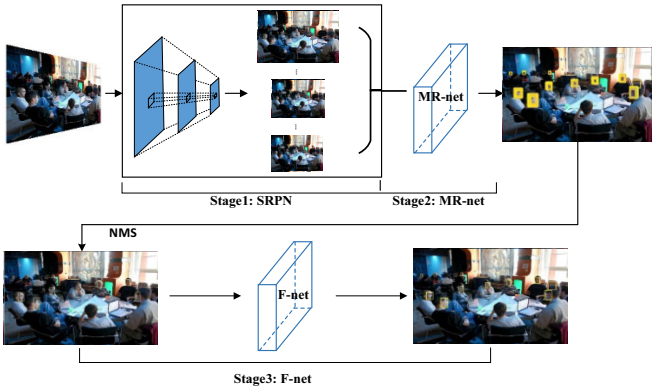


Figure 1. Overview of our detection pipeline that includes three stages

Stage 1: We exploit a deep convolutional neural network to obtain a super-resolution image pyramid, which is the input for the next network. We named it the super-resolution pyramid network (SRPN), and it consists of two sub-modules. The first sub-module utilizes bi-cubic interpolation algorithm to construct a low-resolution image pyramid. The second sub-module uses a convolutional network to recover the local detail of the low-resolution face image.

Stage 2: Multi-scale images generated by the pyramid network are fed to the second network for multi-tasking face detection. We named this network as Multitasking Residual Network (MR-net). The MR-net utilizes the advantages of low-dimensional features and high-dimensional features from different layers of the network to simultaneously implement face classification and regression. we obtain detection bounding boxes for each scale and then merge them back into the original scale. After that, the final detection bounding boxes will be extracted by employing non-maximum suppression (NMS).

Stage 3: This stage uses a facial attribute as an auxiliary task to enhance face detection performance by exploiting another different convolutional neural network. In this stage, the feature point network(F-net) outputs five facial landmarks to verify whether the face

is true or not. The advantage of this network is that it reduces the false positive rate.

With these well-designed networks, our method obviously outperforms other methods on "easy" and "hard" sets, despite the "medium" set results being slightly worse than those of the CMS-RCNN method as shown in Table 1.

Table 1. Performance of our approach on validation set of WIDER FACE [33]. Underline indicates the best performance

Method	easy	medium	hard
ACF [34]	0.659	0.541	0.273
Two-stage CNN [33]	0.681	0.618	0.323
Multi-scale Cascade CNN [35]	0.691	0.634	0.345
LDCF+ [36]	0.790	0.769	0.522
Multi-task Cascade CNN [32]	0.848	0.825	0.598
CMS-RCNN [37]	0.899	0.874	0.624
Ours	0.9	0.872	0.710

3.2 Network Architectures

3.2.1 Super-resolution Pyramid Network

In the target recognition, bilinear interpolation used in the image pyramid is generally common, this interpolation method is fast and simple in operation, but it causes the important details of the enlarged image to be lost and the image becomes blurred. We present a new image pyramid based on super-resolution to solve the above problems, mainly by combining the super-resolution technique with a coarse image pyramid. The image super-resolution technique [38] is integrated into our system by employing the convolutional network to reconstruct a high-resolution image without extra pre/post-processing operations. Therefore, it decreases the computational complexity compared with other similar methods [11, 39]. Our present structure SRPN, applied to face detection, has the advantages of image deblurring and noise reduction, resulting in good robustness and simplicity. The structure of super-resolution pyramid network as show in Figure2. Further, one could investigate different scaling factors to cope with the scale invariance. Our precision of detection improves up to 90% with SRPN, as shown in Table 2.

3.2.2 Multiple Task-driven Cascaded Networks

Multiple task-driven cascaded networks consist of MR-net and F-net. Our cascade network architecture is showed in Figure 3. The MR-net based on 101-residual network is used to achieve face classification and regression task, according to its characteristics of different layers. The last network of our cascade network is F-net, the function this network assists the results of detection with more accurately. The multiple task-driven cascaded networks we designed has the following advantages:

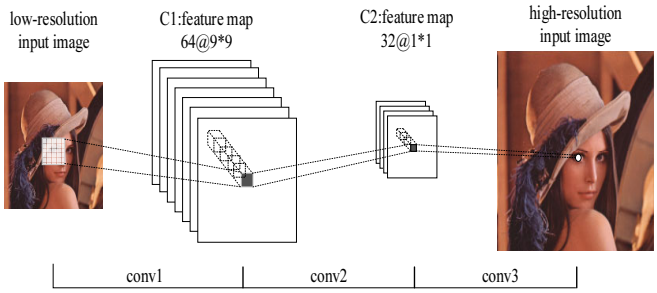
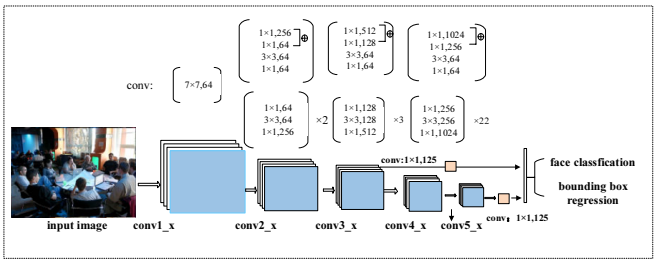


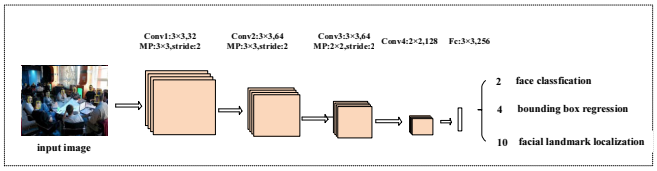
Figure 2. The statistic of average object sizes on the training dataset. (a) The distribution of WIRED FACE dataset, different colors represent different ranges of face pixels. (b) The distribution of pre-trained ImageNet dataset

Table 2. Comparison of our pyramid structure with SRPN and without SRPN. Obviously, the results with SRPN shows better performance on the datasets. Underline indicates the best performance

Method	Easy	Medium	Hard
With SRPN	0.9	0.872	0.710
Without SRPN	0.886	0.855	0.688



MR-net



F-net

Figure 3. The architectures of MR-net and F-net, where "MP" indicates max pooling and "Conv" indicates convolution

(1) The MR-net network can reduce the computation cost and speed up the convergence. First, since the front of the network has eliminated a lot of non-face, therefore, it reduces the cost of computation for regression task in forward propagation. Second, the classification task only needs to update the gradient of the corresponding part in backpropagation, thus the gradient calculation is greatly reduced and the convergence speed is accelerated.

(2) As shown in Figure 3, there is a modification using 1×1 convolution instead of full connection. This design allows the input of our approach to be an arbitrary size.

(3) The F-net network refine the output result of MR-net and eliminate false face. Therefore, face detection to be more accurate.

3.3 Implementation

For training MR-net, we define the class label (positive or negative samples) to each object: (i) Positives: Regions that the intersection-over-union (IOU) overlap higher than 70% with any ground-truth boxes; (ii) Negatives: we assign a negative label to a non-face background if its IOU overlap is lower than 30% with any ground-truth boxes (others are ignored). With these definitions, we minimize an objective function. We use log loss function for face classification and Huber loss for bounding box regression for each sample x_i. Obviously, we need an aggregate loss and it is defined as:

$$L\left(p_i, \hat{y}_i^{box}\right) = \frac{1}{N_{\det}} \sum_i L^{\det}\left(p_i, y_i^{\det}\right) + \lambda \frac{1}{N_{reg}} \sum_i p_i L^{box}\left(\hat{y}_i^{box}, y_i^{box}\right) \quad (1)$$

Here i is the index of object in mini-batch and p_i is the predicted probability of object i being a face. The notation $y_i^{gt} \in \{0, 1\}$ denotes the ground-truth label. r_i is a vector representing the four parameterized coordinates of the predicted bounding box, including left top, height and width. r_i^{gt} is the ground-truth box coordinates.

For training F-net, we use AFLW dataset [40] for extracting facial landmark, and loss function of this network for the classification and regression tasks are the same as that of the MR-net, and square sum loss function as the loss function of facial features localization. The formula is defined as follows:

$$L_i^{landmark} = \| \hat{y}_i^{landmark} - y_i^{landmark} \|_2^2 \quad (2)$$

We implemented the proposed face detector on the Caffe library [41] platform. The pre-trained ImageNet [42] model was used for fine-tuning on the WIDER FACE training set. We randomly resized the training data to a specific resolution and randomly cropped one 227×227 region from the re-scaled image and we defined a learning rate of 10^5 and a momentum of 0.9.

4 Experiments

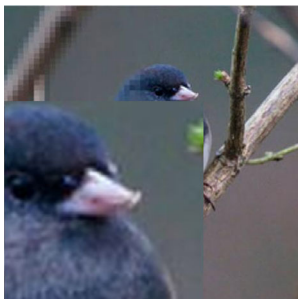
In this section, we first explore the designs of image pyramid and study the relationship between the performance of face detection and scaling factors. Then we evaluate the detection performance against other methods and qualitative results on WIDER FACE and FDDB [43]. Finally, we investigate the impact of using the localization of facial landmarks, which assists with face detection to reduce the false positive rate.

4.1 Comparison of Various Interpolation Results

In this experiment, we compared the result via SRPN method with the existing three interpolation methods, and achieve 2x magnification. The existing methods include: the nearest neighbor interpolation, bilinear interpolation and bi-cubic interpolation. As shown in Figure 4: (a) shows the result obtained by the nearest neighbor interpolation method to enlarge 2 times; (b) is the result obtained by bilinear interpolation method; (c) and (d) generated by the bi-cubic interpolation algorithm and the SRPN algorithm respectively.

(a) respectively represent the nearest neighbor, bilinear, bi-cubic and image super resolution interpolation method

(b) respectively represent the nearest neighbor, bilinear, bi-cubic and image super resolution interpolation method



(c) respectively represent the nearest neighbor, bilinear, bi-cubic and image super resolution interpolation method

(d) respectively represent the nearest neighbor, bilinear, bi-cubic and image super resolution interpolation method

Figure 4. Comparison of the results after magnification of various interpolation methods

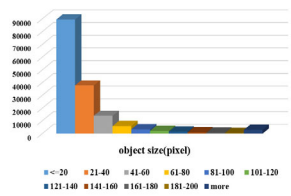
Observation of the experimental results, we found that: the beak edge from (a) exist jagged phenomenon, and (b) overcome this phenomenon, but its edge and feather texture are vaguer than (c). In comparison, the effect of (d) is best obviously. From the beak edge of four visualizations, the images obtained by the others are indistinct, and the image obtained by SRPN is

relatively clear. The effect of SRPN preferably maintain the sharp edges, retain more details and de-noising. Therefore, super-resolution technology integrated to the image Pyramid is the best choice.

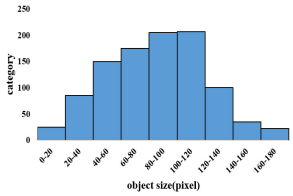
4.2 Scaling Factor and Performance Trade-offs

4.2.1 Data Analysis

During the training phase, the data analysis is essential to obtain a better training model. As shown in Figure 5(a), we found that more than 78% of faces, on the WIDER FACE dataset, had an average size between 8 and 40 pixels, approximately. Smaller objects obviously outnumber larger objects, and this is the phenomenon of imbalanced data. Therefore, we added one step for data augmentation. In the section, we use simple amplification method: cropping, scaling and rotation. Certainly, it is necessary to consider the distribution of a pre-trained dataset (ImageNet), as shown in Figure 5(b).



(a) The distribution of WIRED FACE dataset, different colors represent different ranges of face pixels



(b) The distribution of pre-trained ImageNet dataset

Figure 5. The statistic of average object sizes on the training dataset

4.2.2 The Sensitivity Analysis of the Scaling Factor

Next, we conducted an experiment to explore the following problems: First, is there a relationship between object resolution factors and template sizes? Here, we used t (h, w) to refer to a template, in which "h" and "w" refer to height and width of the template, respectively. In order to find diverse sizes of objects o $(\frac{h}{\sigma}, \frac{w}{\sigma})$ in the test picture, we must specify the range of the upsampling factor " σ ". For example, we assumed that the template size is 200×100, and that there are many faces with different sizes, such as 100×100, 200×100, 134×67. To achieve more accurate results, the corresponding factor was applied to the sample images for different resolutions. The second question was what size is more conducive to detection? We conducted relevant experiments to determine this scope. The results showed that the maximum value of the upsampling factor is 2. As shown in Figure 6, the factor 2, expressed in blue bars, works the best, and the effect of factors below 2 is worse. Our method used a series of discrete factors to adapt the template to find the final targets.

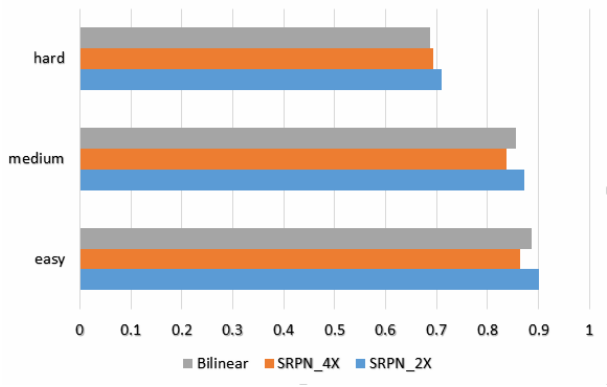


Figure 6. Constructing super-resolution image pyramid helps to improve the performance of face detection, especially for find small and blur face. The improvement with SRPN is improved by 2.2% on "hard" set. Because "hard" set has more blur, smaller faces. The gray columnar represents bilinear interpolation, while the blue columnar represents scaling factor 2, the orange represents factor 4

4.3 Evaluation on Face Detection

In this section, we visualize the qualitative results of our detectors in Figure 7. We chose challenging samples with high occlusion, exaggerated expressions, and other cases (atypical pose, blur, illumination, etc.). The results show that our method has both better robustness and higher accuracy. We include a detailed comparison for the following methods, and datasets are divided into five groups according to attributes. As the experimental results shown in Table 3, our model works well in any case of challenging datasets, which proves that characteristics of multi-task cascaded networks play a pivotal role as well.

(a) Post (b) Illumination (c) Scale

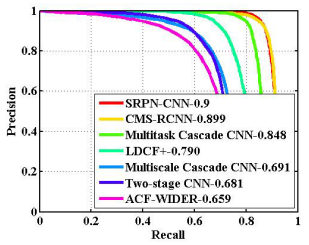
(d) Occlusion (e) Expression (f) Blur

Figure 7. Visualized results for each challenging situation with our multi-task network framework

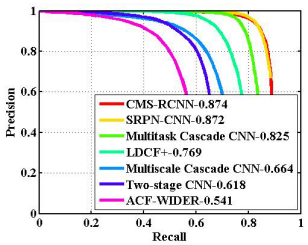
Table 3. Experimental result validate that our method has notable robustness for various situations. The Seetaface [44] approach is an open source C++ face recognition engine

method	blur	occlusion	expression	illumination	pose
Ours	0.891	0.669	0.952	0.854	0.908
MTCNN [32]	0.840	0.603	0.933	0.779	0.795
seetaface [44]	0.556	0.190	0.933	0.742	0.833

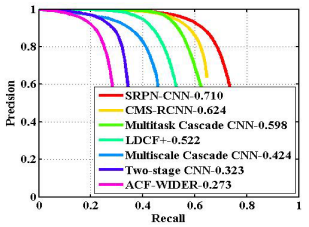
As shown in Figure 8, compared with other state-of-the-art face detectors [32-37], our face detector achieves excellent results on the WIDER FACE. Notably, when comparing ours with the Multitask Cascade CNN algorithms, our performance improves by 10.4% on the "hard" set. Our method obviously can improve precision and recall rate on the hard and easy sets, respectively. As shown in Figure 8(d), we adopt the area under the curve (AUC) as an evaluation indicator. Comparison of these detection algorithms: [32, 34, 45-51], our results show AUC of 95.6% and excelled than the MTCNN on the FDDB dataset.



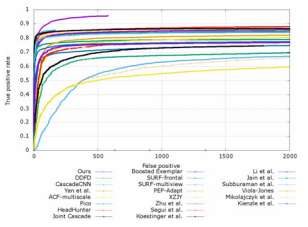
(a) Precision recall curves on three subsets of WIDER FACE validation set



(b) Precision recall curves on three subsets of WIDER FACE validation set



(c) Precision recall curves on three subsets of WIDER FACE validation set

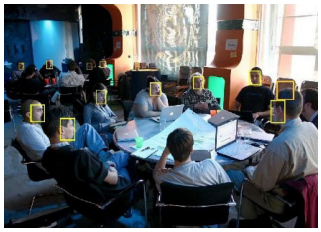


(d) Evaluation on FDDB

Figure 8.

4.4 The Joint Effectiveness of Face Detection and Facial Landmarks

Facial landmarks can be used for auxiliary detection to reduce the error rate, especially when applied to simple backgrounds, such as indoor rooms, conference rooms, classrooms, etc. Some related experiments have been conducted to evaluate the joint contribution of detection and facial landmarks. We evaluate the performance of two different approaches (with and without the localization of facial landmarks) on the WIDER FACE dataset. The results show that it is beneficial for face detection with the jointed localization of facial landmarks, and some bounding boxes that were detected by mistake were eliminated, as shown in Figure 9. The left-hand image was not done through the processing steps of facial landmarks. However, there are some wrong boxes, such as the photo frame on the table and the chair beside the window. On the contrary, the right-hand image has jointed detection and facial landmarks and shows significant improvements.



(a) The result of test image without facial landmarks

(b) The result of test image with facial landmarks

Figure 9. Observing the result of facial landmarks

5 Conclusion

In this paper, we proposed a framework based on cascaded CNNs for multiple task-driven face detection. The SRPN-CNN framework better leverages the resolution and the scale of an image. Moreover, in the final stage, we exploited the inherent correlation between the face detection and facial landmarks to further reduce the false positive rate. Extensive evaluations on the challenging benchmarks for face detection demonstrate that our methods have achieved superior performance than other state-of-the-art methods. In the future, we will explore different training strategies and consider model optimization in order to achieve faster speeds and to further improve the performance.

Acknowledgments

This work was supported by the National Natural Science Fund of China (No. 61871170), Pre-research fund of China (No. 6140137050202) and the National Equipment Development Pre-research fund (No. 6140137050202).

References

[1] D. G. Lowe, Object Recognition from Local Scale-invariant Features, *International Conference on Computer Vision (ICCV)*, Kerkyra, Corfu, Greece, 1999, pp. 1150-1157.

[2] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up

Robust Features, *Computer Vision and Image Understanding (CVIU)*, Vol. 110, No. 3, pp. 346-359, April, 2008.

[3] P. Felzenszwalb, D. McAllester, D. Ramanan, A Discriminatively Trained, Multiscale, Deformable Part Model, *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, 2008, pp. 1-8.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, SSD: Single Shot Multibox Detector, *European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 21-37.

[5] A. Pentland, B. Horowitz, A Practical Approach to Fractal-based Image Compression, *Data Compression Conference*, Snowbird, UT, 1991, pp. 176-185.

[6] W. T. Freeman, E. C. Pasztor, Learning Low-level Vision, *International Journal of Computer Vision*, Vol. 40, No. 1, pp. 25-47, October, 2000.

[7] H. Chang, D. Y. Yeung, Y. Xiong, Super-resolution through Neighbor Embedding, *Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, 2004, pp. 1-1.

[8] J. Sun, N.-N. Zheng, H. Tao, H.-Y. Shum, Image Hallucination with Primal Sketch Priors, *Computer Vision and Pattern Recognition (CVPR)*, Madison, WI, 2003, pp. 729-736.

[9] W. Fan, D.-Y. Yeung, Image Hallucination Using Neighbor Embedding over Visual Primitive Manifolds, *Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, 2007, pp. 1-7.

[10] J. Yang, J. Wright, T. Huang, Y. Ma, Image Super-resolution as Sparse Representation of Raw Image Patches, *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, 2008, pp. 1-8.

[11] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image Super-resolution via Sparse Representation, *IEEE Transactions on Image Processing*, Vol. 19, No. 11, pp. 2861-2873, November, 2010.

[12] C. Liu, H.-Y. Shum, W. T. Freeman, Face Hallucination: Theory and Practice, *International Journal of Computer Vision*, Vol. 75, No. 1, pp. 115, October, 2007.

[13] W. Zhang, W. K. Cham, Learning-based Face Hallucination in DCT Domain, *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, 2008, pp. 1-8.

[14] P. H. Hennings-Yeomans, S. Baker, B. V. Kumar, Simultaneous Super-resolution and Feature Extraction for Recognition of Low-resolution Faces, *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, 2008. pp. 1-7.

[15] N. Roeder, X. Li, Accuracy Analysis for Facial Feature Detection, *Pattern Recognition*, Vol. 29, No. 1, pp. 143-157, January, 1996.

[16] A. Yuille, Deformable Templates for Face Recognition, *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 59-70, January, 1991.

[17] K.-M. Lam, H. Yan, Locating and Extracting the Eye in Human Face Images, *Pattern Recognition*, Vol. 29, No. 5, pp. 771-779, May, 1996.

[18] J. Y. Deng, F. Lai, Region-based Template Deformation and Masking for Eye-feature Extraction and Description, *Pattern Recognition*, Vol. 30, No. 3, pp. 403-419, March, 1997.

[19] Z. P. Liu, Linear Discriminant Analysis, *Chicago*, Vol. 3, No. 6, pp. 27-33, June, 2013.

[20] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, Q. Dai, Supervised Hash Coding with Deep Neural Network for Environment Perception of Intelligent Vehicles, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 3, No. 1, pp. 284-295, January, 2018.

[21] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, Q. Dai, Effective Uyghur Language Text Detection in Complex Background Images for Traffic Prompt Identification, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 3, No. 1, pp. 220-229, January, 2018.

[22] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, Efficient Parallel Framework for HEVC Motion Estimation on Many-core Processors, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 24, No. 12, pp. 2077-2089, December, 2014.

[23] P. Viola, M. Jones, Rapid Object Detection Using a Boosted Cascade of Simple Features, *Computer Vision and Pattern Recognition (CVPR)*, Kauai, HI, 2001, pp. 511-518.

[24] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image recognition, *Computer Science*, Vol. 1, No. 1, September, 2014.

[25] C. Szegedy, W. Liu, Y. Jia, Going Deeper with Convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9.

[26] K. He, X. Zhang, S. Ren, Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778.

[27] T. F Cootes, C. J. Taylor, D. H. Cooper, Active Shape Models-their Training and Application, *Computer Vision and Image Understanding (CVIU)*, Vol. 61, No. 1, pp. 38-59, January, 1995.

[28] T. F Cootes, G. J. Edwards, C. J. Taylor, Active Appearance Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 6, pp. 681-685, June, 2001.

[29] X. Yu, J. Huang, S. Zhang, W. Yan, D. N. Metaxas, Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model, *International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013, pp. 1944-1951.

[30] X. Yu, J. Huang, S. Zhang, Pose-Free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model, *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, 2014, pp. 1944-1951.

[31] X. Zhu, D. Ramanan, Face Detection, Pose Estimation, and Landmark Localization in the Wild, *IEEE Conference on Computer Vision and Pattern Recognition, Providence (CVPR)*, RI, 2012, pp. 2879-2886.

[32] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks, *IEEE Signal Processing Letters*, Vol. 23, No. 10, pp. 1499-1503, October, 2016.

[33] S. Yang, P. Luo, C. C. Loy, Wider Face: A Face Detection Benchmark, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 5525-5533.

[34] B. Yang, J. Yan, Z. Lei, Aggregate Channel Features for Multi-view Face Detection, *2014 IEEE International Joint Conference on Biometrics (IJCB)*, Clearwater, FL, 2014, pp. 1-8.

[35] S. Yang, P. Luo, C. C. Loy, X. Tang, From Facial Parts Responses to Face Detection: A Deep Learning Approach, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3676-3684.

[36] E. Ohn-Bar, M. M. Trivedi, To Boost or not to Boost? On the Limits of Boosted Trees for Object Detection, *International Conference on Pattern Recognition*, Mexico, 2016, pp. 3350-3355.

[37] C. Zhu, Y. Zheng, K. Lu, CMS-RCNN: Contextual Multi-scale Region-based CNN for Unconstrained Face Detection, *Deep Learning for Biometrics*. Springer, Cham, 2017.

[38] C. Dong, C. C. Loy, K. He, X. Tang, Image Super-resolution Using Deep Convolutional Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 38, No. 2, pp. 295-307, February, 2016.

[39] K. I. Kim, Y. Kwon, Single-image Super-resolution Using Sparse Regression and Natural Image Prior, *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, Vol. 32, No. 6, pp. 1127-1133, June, 2010.

[40] M. Köstinger, P. Wohlhart, P. M. Roth, Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization, *IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, 2012, pp. 2144-2151.

[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding, *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL, 2014, pp. 675-678.

[42] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *International Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, 2012, pp. 1097-1105.

[43] V. Jain, E. Learned-Miller, FDDB: A Benchmark for Face Detection in Unconstrained Settings, *University of Massachusetts, Amherst*, Tech. Rep. UM-CS-2010-009, July, 2010.

[44] S. G. Shan, SeetaFace Engine, https://github.com/seetaface/SeetaFaceEngine, September, 2017.

[45] R. Ranjan, V. M. Patel, R. Chellappa, A Deep Pyramid Deformable Part Model for Face Detection, *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Arlington, VA, 2015, pp. 1-8.

[46] M. Mathias, R. Benenson, M. Pedersoli, Face Detection without Bells and Whistles, *European Conference on Computer Vision, Zurich (ECCV)*, Zurich, Switzerland, 2014, pp. 720-735.

[47] D. Chen, S. Ren, Y. Wei, Joint Cascade Face Detection and Alignment, *European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, pp. 109-122.

[48] S. S. Farfade, M. J. Saberian, and L. J. Li, Multi-view Face Detection Using Deep Convolutional Neural Networks, *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, Shanghai, China, 2015, pp. 643-650.

[49] J. Yan, Z. Lei, L. Wen, The Fastest Deformable Part Model for Object Detection, *Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, 2014, pp. 2497-2504.

[50] B. Yang, J. Yan, Z. Lei, Convolutional Channel Features, *2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 82-90.

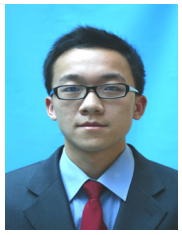
[51] H. Li, Z. Lin, X. Shen, A Convolutional Neural Network Cascade for Face Detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 5325-5334.

Biographies

Jianjun Li received the B.Sc. degree in information engineering from Xi'an University of Electronic Science and Technology, Xi'an, China, and the M.Sc. and Ph.D degrees in electrical and computer from The University of Western Ontario and University of Windsor, Canada separately. He is currently working at HangZhou Dianzi University as a chair professor. His research interests include micro-electronics, audio, video and image processing algorithms and implementation.

Juxian Wang is studying computer science and technology at HangZhou Dianzi University, HangZhou, China. She is currently working at Institute of graphics and image. His research interests include object recognition, face recognition and image processing algorithms and implementation.

Xingchen Chen is an undergraduate student at Shananxi University of Technology. His major is computer science and technology. His research interests include image modification, deep learning and artificial intelligence.



Zhenxing Luo received the B.S. degree in telecommunication engineering from Hangzhou Dianzi University, Hangzhou, Zhejiang, in 2006 and the M.S. degree in Signal and Information Processing from Hangzhou Dianzi University, Hangzhou, Zhejiang, in 2009. He is currently pursuing

the Ph.D. degree in Communication and Information System at Xian Dianzi University, Xi'an, Shaan, China. From 2009 to 2018, he was a Research Assistant with the Science and Technology on Communication Security Control Laboratory, Jiaxing, Zhejiang. His research interest includes the development and fundamental study of signal processing and machine learning techniques used in spectrum management.

Zhugang Song is a professor from the 36th Institute of CECT, Jiaxing, Zhejiang, China. He received his bachelor's degree in computational mathematics and application software from Fudan University in 1991. His current research interests include communication signal processin