CONTEXT-ANCHORS FOR HYBRID RESOLUTION FACE DETECTION

Tianpeng Wu[1], *Dong Liang* [1], *Jiaxing Pan*[1], *Shun'ichi Kaneko*[2]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[1]Collaborative Innovation Center of Novel Software Technology and Industrialization
[2]Graduate School of Information Science and Technology, Hokkaido University, Japan

ABSTRACT

Despite the positive trends in the development of face detection, open challenges still exist, such as the detection of degraded faces caused by small-size, defocus blur and occlusion in surveillance video. When utilizing anchor-based methods, the anchors are the basic units of training samples, and their ranges are proportional to the ranges of the original label boxes (ground truth). This paper argues that the selected range of an anchor is crucial for a detection task and proposes a face detection model CAHR (context-anchors for hybrid-resolution model) to balance the image resolution and the spatial context range for the purposes of locating small faces. In the training phase, specific size of spatial context is introduced for each anchor, and an image pyramid is employed for a dual CNNs model. In experiments, the in-depth analysis of amplification ratio of the anchor and the detection rate is revealed. The detection rate of the small faces is improved by using the proposed model. It is also validated with a massively face datasets (WIDER FACE), demonstrating its superiority to the original hybrid-resolution model (HR) and some other advanced methods.

Index Terms— anchor-based, face detection, context

1. INTRODUCTION

In recent years, anchor-based face detection methods have presented satisfactory performance on the benchmark dataset WIDER FACE [9] and FDDB. However, there are thorny issues involved the detection of degraded faces caused by small-size, defocus blur and occlusion in surveillance video. On the other hand, for a convolutional neural network itself, small-size face offers too few features within the range of facial range due to the spatial pooling process [5]. It is already proved that low-level features are helpful to detect small objects [3], while spatial contextual information is also valuable for small object detection. The role of contextual information in object detection was explored in [10, 18, 19]. In anchor-based object detection methods [2, 3, 5, 17], the context information of faces is employed via the fusion of the feature maps. The receptive field of high-level feature is usually larger than that of low-level feature, if the receptive field of low-level feature is close to that of face, the receptive field of high-level feature naturally includes the contextual information of a facial image.

For a small-face detection task, expanding the range of an anchor could capture the topological structure of human hair, shoulder and head edge, which contains more abundant information than only using facial features. However, the use of contextual information is not always beneficial. For example, in the low-density face detection task, directly enlarging the anchor would introduce background area, so that the inter-class scatter between foreground and background samples spread, which is not conducive to the convergence of model training. Another case is in the high-density face detection task, too large anchor may contain multiple faces, which will suppress the local response peak of the model, resulting in the lower accuracy of object's location.

In this paper, we propose a face detection model CAHR (context-anchors for hybrid-resolution model) to balance the image resolution and the spatial context range for the purposes of finding small faces. We explore the effect of using contextual information of different magnitude on model detection and find that a certain amount of context is beneficial, while excessive context information will reduce the proportion of face features to disturb the detection. After testing the influence of context-anchor on various scales, we finally use context-anchor on 2X and 4X enlarging image for detection. Difference from other models, CAHR employs spatial context via the expended anchor according to the specific range to each training sample. As a multi-scale face detection model, CAHR also employs an image pyramid structure and separately feeds a dual ResNet to adapt to the faces with variable sizes.

2. THE APPROACH

2.1 Anchor Box and Hybrid Resolution Model

Anchor box is firstly introduced in [1] that served as references at multiple scales and aspect ratios for object detection. This scheme can be regarded as a pyramid of regression references, which avoids enumerating images or

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

filters of multiple scales or aspect ratios, thus accelerates running speed. At each sliding-window location, it simultaneously predicts multiple region proposals, which we call anchors. An anchor is centered at the sliding window and is associated with a scale and aspect ratio. In an anchor-based model, the anchors are the basic units of training samples, and their ranges are proportional to the ranges of the original label boxes (ground truth).

Hybrid Resolution Model (HR) [3] is a multi-scale face detection model, which adopts feature map fusion to use context information. Image pyramid is adopted to find small faces. It firstly creates a coarse image pyramid, and then feeds the scaled input into a CNN to predict template responses at every resolution. In the end, it applies non-maximum suppression (NMS) at the original resolution to get the final detection results. It runs the templates tuned for 40-140px tall faces on the coarse image pyramid including 2X interpolation, while only runs the templates tuned for less than 20px tall faces on only 2X interpolated images. We choose HR model as a baseline, because it is an anchor-based multi-scale model with the state-of-the-art performance for small-face detection.

2.2 Context-anchors

2.2.1 Amplification Ratio of anchor

Previous feature fusion methods employ high-level feature layers to expand the receptive field. This is a simple but coarse way to introduce contextual information which would introduce unexpected information to affect the training. Context-anchor utilizes contextual information in the most straight-forward way - directly expands the face label box (ground truth), so that the anchor could contain the context information of a face. In the CAHR model, the training process is as follows, as illustrated in Fig.1, we define a context-anchor amplification ratio n, for each training sample, the label box of ground truth is enlarged to n times. Because anchors of the model are designed to be clustered by label box in training datasets, so that when selecting positive samples for different anchors in training, the peripheral information of a face including the head and shoulders is considered. When using context-anchor model for detection, the output bounding box should be cropped with a scaling coefficient $1/n$. The amplification ratio n is the key parameter for a context-anchor. We will discuss its optimal value in the experiment section.

2.2.2 Training Label Clustering

Clustering is used to get limited anchors' setting values, which enables the anchor to expand with the expansion of the label box. The length and width of the face labels in the training set are used as two-dimensional feature vectors input, and 25 clustering centers are obtained by K-means clustering method. These 25 clustering centers are selected as anchor's length and width, so that the model can be applied to multi-scale and multi-pose face detection.

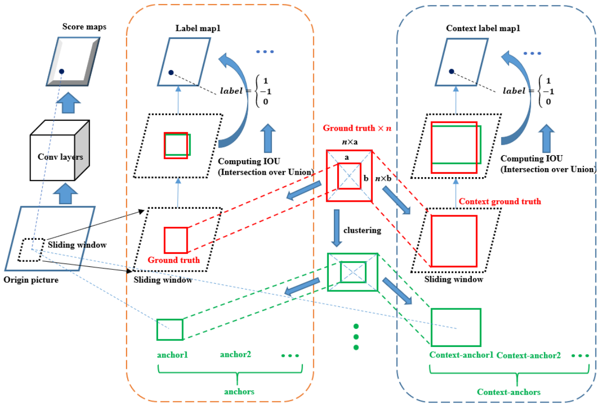


Fig.1. Context-anchors

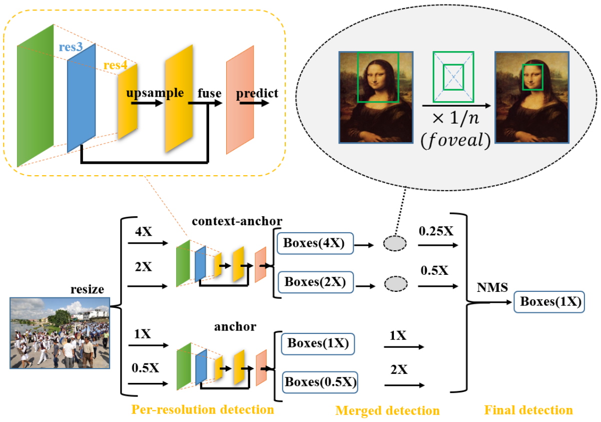


Fig.2. Overview of CAHR

2.3 Model Structure

As shown in Fig.2, CAHR firstly uses image pyramid structure to process images to 0.5X, 1X, 2X and 4X. Context-anchors are used for the 2X and 4X outputs of the image pyramid. For the 0.5X and 1X output, conventional anchors are used. This designation is to balance the image resolution and the spatial context range for the purposes of finding small faces. For the medium and large face, 1X and 0.5X image with the conventional anchors is the dominant output for the score map, while for the small or low-resolution face, 2X and 4X image with the context-anchors is the dominant output for the score map. The dual model adopts res101 network, with two-layer feature map fusion, sampled on res4 and fused with res3 feature map as score layer, and uses the single-layer score layer to predict.

3. EXPERIMENTS

3.1 Experiments of amplification ratio

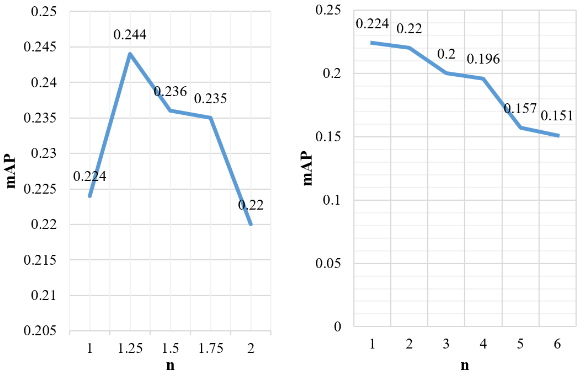


Fig.3. mAP of different amplification ratios on FIC dataset.

Training settings: Considering that context information is mainly used to deal with low-resolution face detection, first of all, we expect for a reasonable amplification ratio n through experiments. We want to explore the ability of context-anchor in the detection of faces in the crowd, those faces are hard to detect because of low-resolution. Those images are few in public datasets. We build a crowd-face-detection dataset FIC[1](Faces in the crowd), the images are selected from face dataset WIDER FACE, FDDB, AFW and internet. We choose 30 images and label 6474 faces. This dataset includes 10 grayscale images and 20 color images, the maximum number of faces in one image is 868. We train the model on this high-density human face dataset, setting batchsize as 7, learning rate as 10^{-6}, for the model obtained by 1000 iterations with different amplification ratio n.

Test Settings: When testing on FIC test set, the NMS threshold is 0.3 and the score threshold is 0.03.

As illustrated in Fig.3, the relationship of the amplification ratio of the anchor and the detection rate was revealed. When the amplification ratio n is between (1,1.75), the mAP is improved. When the amplification ratio is greater than 1.75, the mAP becomes worse with the increase of n. It indicates that the appropriate utilization of context information can improve the detection performance, while overdependence on context information will lead to even worse detection results. When $n = 1.25$, mAP got the peak. In the following experiment, 1.25 is set as a default value.

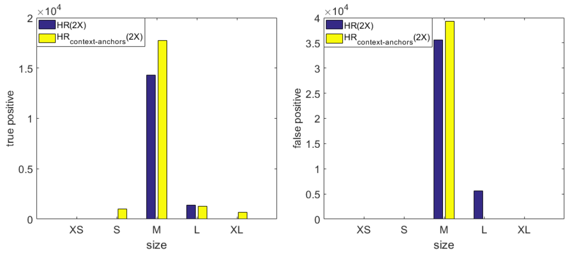
3.2 Experiments on WIDER FACE

Training Settings: In this part, the context-anchor model with the amplification ratio $n = 1.25$ is tested. The training set is WIDER FACE training set, batchsize is 7, learning rate is 10^{-4}. All the context-anchors structure model obtained by 50 iterations.

Test Settings: The test set is WIDER FACE evaluation dataset, the NMS threshold is 0.3, and the score threshold is 0.03.

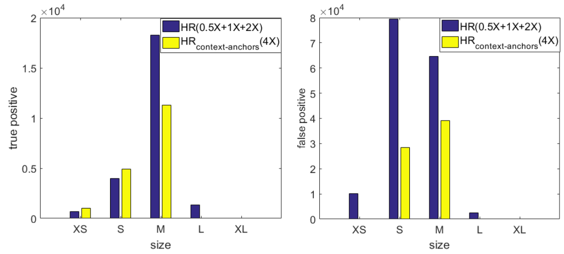
3.2.1 Experiments on context-anchors for 2X and 4X.

First, all images are enlarged to 2X. As illustrated in Fig. 4(a), XS, S, M, L and XL correspond to face scales ranging from (0, 40] , (40, 100], (100, 1500], (1500, 2000], (2000, +∞) px. The size of small face is relatively larger after using the general resize method, but the facial features are severely blurred compared with the face in the same size of the original image. Compared with the previous image 2X, using the context-anchor model, as illustrated in Fig.4(a), there is a significant improvement in mesoscale (M) face detection, which improves TP. For small scale (S) and large-scale face (XL), compared with the HR model without context-anchor, the TP of using the context-anchors model obviously increases. As illustrated in Fig.4(b), context-anchor generates slightly more FPs on Mesoscale faces (M), which is within tolerable range, considering that there are also more TP. On S, L, XL, these scales which are larger or smaller, using the context-anchor model does not produce FPs, and on these scales, all the faces found are true positive faces with such a low score threshold 0.03. This performance surpasses our expectation. As illustrated in Fig. 5(a), the context-anchors model (on image 4X) detects more small faces than original HR. As illustrated in Fig. 5(b), the context-anchors model (on



(a) (b)

Fig.4. Comparison of true positive and false positive for 2X.



(a) (b)

Fig.5. Comparison of true positive and false positive for 4X.

[1] https://github.com/AIoTP/faces_in_crowd

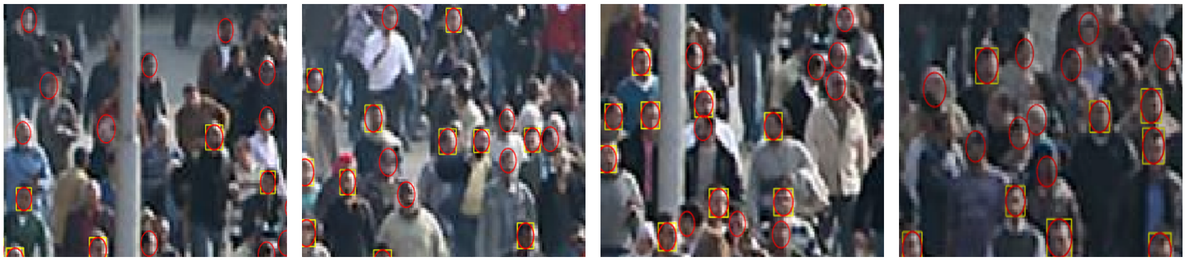
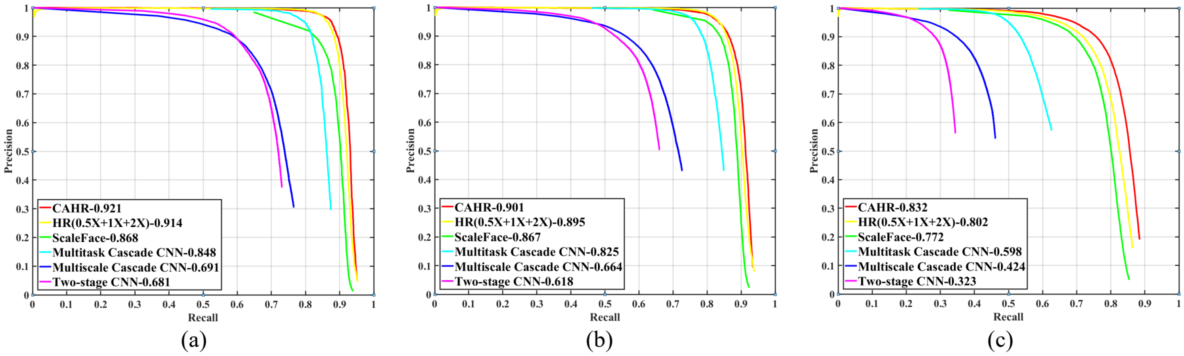


Fig.6. Visual results in crowd scenes. Yellow rectangles are results of HR(0.5X+1X+2X), and red ellipses are results of CAHR.



(a) (b) (c)

Fig.7. Precision-Recall curve on WIDER FACE evaluation dataset easy (a), medium (b), hard (c) set.

image 4X) produces less false positive, which indicates the context-anchors model (on image 4X) performs better on small face detection and provides acceptable performance on other sizes of faces. As shown in Table 1, CAHR with the combination of anchor (0.5X+1X) and context-anchor (2X+4X) achieves the best results in WIDER FACE evaluation hard set among HR-based model setting, demonstrating its superiority.

3.2.2 Overall performance of CAHR on WIDER FACE evaluation dataset

Fig.6 visualizes that CAHR detects more true faces in crowd scenes. Fig 7 presents the performance of CAHR, HR [3], ScaleFace [14], Multitask Cascade CNN [13], Multiscale Cascade CNN [15] and Two-stage CNN [9] on WIDER FACE easy, medium and hard set respectively. As shown in Fig.7, compared with HR, CAHR has the most obvious improvement in hard set with the mAP increasing 0.03.

5. CONCLUSION

We propose a context-anchor structure to introduce appropriate spatial contextual information directly to each sample of the training process. Experiments of amplification ratio indicates that the appropriate use of context information can improve the detection performance, while overdependence on contextual information will lead to even worse detection results. We propose a face detection model

TABLE 1 mAP of combinations on hard set

Anchor Setting	*Image Pyramid*	AP_{50}	AP_{75}	*AP*
--	--	**0.994**	0.980	0.802
Anchor	0.5X+1X+2X			
Context-anchor	2X	**0.994**	0.983	0.816
Anchor	0.5X+1X+2X			
Context-anchor	2X+4X	**0.994**	**0.987**	0.830
Anchor	0.5X+1X+2X			
Context-anchor	2X+4X	**0.994**	**0.987**	**0.832**
Anchor	0.5X+1X			

CAHR to evaluate performance of context-anchors on the benchmark dataset WIDER FACE. The proposed method carries strong practicality and catholicity for performance evaluation, and can be flexible to implement to other anchor-based object detection models.

6. ACKNOWLEDGE

This work is supported by the National Key R&D Program of China under Grant 2017YFB0802300, the National Natural Science Foundation of China (Grants No 61601223, 61801242), the Natural Science Foundation of Jiangsu Province (Grants No BK20150756), China Postdoctoral Science Foundation (Grants No 2015M580427), and the Fundamental Research Funds for the Central Universities (No. NS2016091).

7. REFERENCES

[1] Ren, Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence 39.6 (2017): 1137-1149.

[2] Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." european conference on computer vision (2016): 21-37.

[3] Hu, Peiyun, and Deva Ramanan. "Finding Tiny Faces." computer vision and pattern recognition (2017): 1522-1530.

[4] Hao, Zekun, et al. "Scale-Aware Face Detection." computer vision and pattern recognition (2017): 1913-1922.

[5] Zhang, Shifeng, et al. "Detecting Face with Densely Connected Face Proposal Network." chinese conference on biometric recognition (2018): 3-12.

[6] Lin, Tsungyi, et al. "Feature Pyramid Networks for Object Detection." computer vision and pattern recognition (2017): 936-944.

[7] Shrivastava, Abhinav, Abhinav Gupta, and Ross B. Girshick. "Training Region-Based Object Detectors with Online Hard Example Mining." computer vision and pattern recognition (2016): 761-769.

[8] Neubeck, Alexander, and L. Van Gool. "Efficient Non-Maximum Suppression." international conference on pattern recognition (2006): 850-855.

[9] Yang, Shuo, et al. "WIDER FACE: A Face Detection Benchmark." computer vision and pattern recognition (2016): 5525-5533.

[10] Oliva, A, and A. Torralba. "The role of context in object recognition. " Trends in Cognitive Sciences 11.12(2007):520.

[11] Tychsensmith, Lachlan, and Lars Petersson. "Improving Object Localization with Fitness NMS and Bounded IoU Loss." computer vision and pattern recognition (2018): 6877-6885.

[12] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." computer vision and pattern recognition (2005): 886-893.

[13] Zhang, Kaipeng, et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." IEEE Signal Processing Letters 23.10 (2016): 1499-1503.

[14] Yang, Shuo, et al. "Face Detection through Scale-Friendly Deep Convolutional Networks." arXiv: Computer Vision and Pattern Recognition (2017).

[15] Yang, Shuo, et al. "From Facial Parts Responses to Face Detection: A Deep Learning Approach." international conference on computer vision (2015): 3676-3684.

[16] Yang, Bin, et al. "Aggregate channel features for multi-view face detection." International Journal of Central Banking (2014): 1-8.

[17] Xiang, Wei, et al. "Context-Aware Single-Shot Detector." workshop on applications of computer vision (2018): 1784-1793.

[18] Wolf, Lior, and S. Bileschi. "A Critical View of Context." International Journal of Computer Vision 69.2(2006):251-261.

[19] Divvala, Santosh Kumar, et al. "An empirical study of context in object detection." computer vision and pattern recognition (2009): 1271-1278.

[20] Torralba, et al. "Context-based vision system for place and object recognition." international conference on computer vision (2003): 273-280.

[21] Biederman, Irving, Robert J. Mezzanotte, and Jan C. Rabinowitz. "Scene Perception" Detecting and Judging Objects Undergoing Relational Violations." Cognitive Psychology 14.2 (1982): 143-177.

[22] Zhu, Chenchen, et al. "CMS-RCNN: Contextual Multi-Scale Region-Based CNN for Unconstrained Face Detection." arXiv: Computer Vision and Pattern Recognition (2017): 57-79.

[23] Rosenfeld, Azriel, and Mark Thurston. "Edge and Curve Detection for Visual Scene Analysis." IEEE Transactions on Computers 20.5 (1971): 562-569.

[24] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." computer vision and pattern recognition (2016): 770-778.

[25] Kruppa, Hannes, and Bernt Schiele. "Using Local Context To Improve Face Detection." british machine vision conference (2003): 1-10.

[26] Tang, Xu, et al. "PyramidBox: A Context-Assisted Single Shot Face Detector." european conference on computer vision (2018): 812-828.