SCORE-SPECIFIC NON-MAXIMUM SUPPRESSION AND COEXISTENCE PRIOR FOR MULTI-SCALE FACE DETECTION

Tianpeng Wu[1], Dong Liang [1], Jiaxing Pan[1], Han Sun[1],
Bin Kang[2], Shun'ichi Kaneko[3], Huiyu Zhou[4]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[2] Nanjing University of Posts and Telecommunications, China
[3]Graduate School of Information Science and Technology, Hokkaido University, Japan
[4]Department of Informatics, University of Leicester, United Kingdom

ABSTRACT

Face detection is an ultimate component to support various visual facial related tasks. However, detecting faces with extremely low resolution or high occlusion is still an open problem. In this paper, we propose a two-step general approach to refine the performance of modern face detectors according to human's high-level context-aware ability. First, we propose Score-specific Non-Maximum Suppression (SNMS) to preserve overlapped faces. Second, we consider the coexistence prior among faces in the scene, which could raise the sensitivity of face detection in the crowd. When integrating our approach to the existing face detectors, most of them have better results on a challenging benchmark (WIDER FACE) and a newly proposed dataset (Faces in Crowd, FIC) made by us. Codes are available on https://github.com/AIoTP/SNMSandCoexistence.

Index Terms— Face detection, Score-specific NMS, Coexistence prior, Contextual information

1. INTRODUCTION

Robust face detection in open world is an ultimate component to support various facial related problems. Modern deep learning based face detectors try to approaching the cognition of human that many detectors have surpassed human on visual detection and recognition competitions. However, because flexible attention mechanism and abundant domain knowledge [1] guide human's cognition, human have obvious advantages on the challenges of occlusion, low resolution, and extraordinary gesture [8]. In fact, the majority of samples in the training dataset are samples with normal sizes and planar [2]. The training samples in difficult situation has inadequate amount of collection, i.e. occlusion face in the crowd, faces with very large or very small size. Detection of these samples is just the trouble that most detectors have to be improved.

In multi-scale object detection, approaches based on sliding windows typically produce multiple windows with high scores close to the correct location of objects. This is a consequence of the generalization ability of object detectors, the smoothness of the response function and the visual correlation of the close-by windows. This relatively dense output is generally not satisfying for understanding the content of an image. In fact, the number of window hypotheses at this step is uncorrelated with the real number of objects in the scene. The goal of Non-Maximum Suppression (NMS) [21] is therefore to retain only one window per group, corresponding to the precise local maximum of the response function, ideally obtaining only one detection per object. Consequently, NMS has a large positive impact on performance measures that penalize false detections, which been an integral part of many detection algorithms in computer vision for almost 50 years. For an object detection task in a natural image, context is the information relevant to the detection object but not directly due to the physical appearance of the object. The ideas of using context in object detection have been studied in several recent work. [1] and [10] reviewed contextual information used in contemporary methods and analyzed the its role for challenging object detection in empirical evaluation. In their conclusions, the contextual information

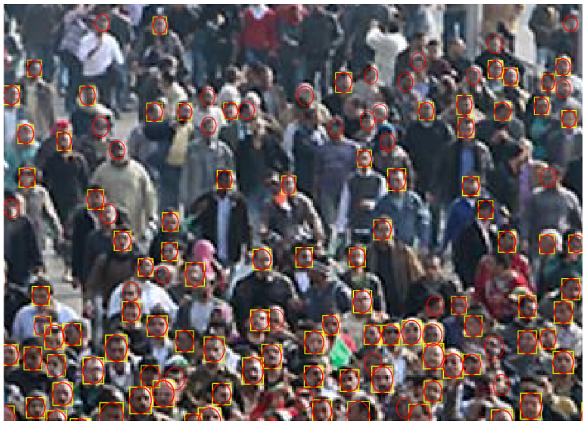

Fig. 1. A contrast of low-resolution face detection using proposed approach integrated with hybrid-resolution model (HR) [8] (red ellipses) and original HR (yellow rectangles) in crowd scene.

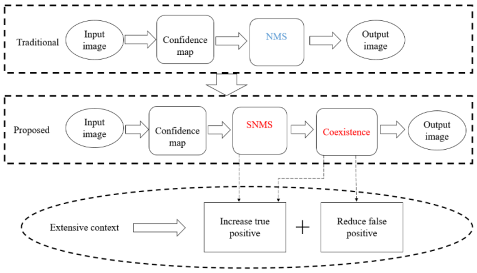

Fig. 2. Architecture of the proposed framework

not only reduces the overall detection errors, but also makes the remaining errors made by the detector more reasonable.

In this paper, we tend to explore the cues which affect human's judgement to improve recent detectors. Especially, we pay more attention on the difficult cases in multi-scale face detection task. We propose a universal strategy, which could be independent to specific training strategy and classifier, to improve the detection performance. After analyzing the results of the state-of-the-art detectors, we find that the relationship of objects could be an extensive context to enhance the confidence of human's cognition. In order to detect multi-scale occlusion objects, we propose SNMS which is a compromise formula of NMS (Non-Maximum Suppression) and Soft-NMS [3-7]. To detect the low-resolution faces in the crowd scene, we extend the contextual information to the whole scene, and propose coexistence prior to evaluate the coexistence of a face-pair and the coexistence of homogeneous faces as an extensive context. Using that high-level contextual information in the whole scene as prior obviously enhances the scene understanding capability of the existing face detectors.

2. THE APPROACH

2.1 Score-specific Non-Maximum Suppression

2.1.1 Soft-NMS

Different from traditional NMS, Soft-NMS [3] argues that the conventional NMS is too greedy, because the detection box with the maximum score is selected and all other detection boxes with a significant overlap with this box are suppressed using a pre-defined threshold. If an object lies within the predefined overlap threshold, it leads to a false negative. Soft- NMS employ an approach that suppresses bounding box through reducing its scores instead of just removing it. If the score of the bounding box is lower than threshold of score, then this bounding box will be deleted. In our early experiment, however, we found the Soft-NMS causes the increase of false positive, because some redundant boxes cannot be deleted if they have high scores. Actually, the neighboring high scores can be caused by three factors: (1) the generalization of the detector, (2) the smoothness of the response function, (3) the visual correlation of the close-by windows. Thus, we need an algorithm to effectively distinguish different situations.

2.1.2 Score-specific Non-Maximum Suppression

NMS [21] is performed by partitioning bounding-boxes into disjoint subsets using an overlap criterion. The final detections are obtained by averaging the co-ordinates of the detection boxes in the set. If b_i and b_j are two bounding boxes, IOU (intersection over union) can be expressed as follow.

$$\mathrm{o} = IOU(b_i, b_j) = (b_i \cap b_j)/(b_i \cup b_j) \qquad (1)$$

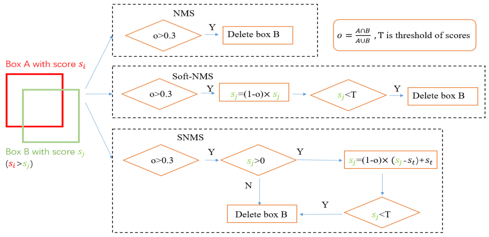

Fig.3. A comparison among NMS, Soft-NMS, and proposed Score-specific NMS. s_t is a threshold of score (in order to get boxes whose score is low).

Fig. 3 shows the comparison among NMS, Soft-NMS, and proposed Score-specific NMS. The conventional NMS preserves the detection box with the maximum score and discards all other detection boxes with a significant overlap within this box, which often leads to a false negative. More specifically, the principle of Non-Maximum Suppression (NMS) is as follows: if the ratio of the two bounding box areas, i.e. the o in the Fig. 3, exceeds the threshold (0.3 is obtained here because most algorithms using this value), then the box with the lower score is deleted directly. NMS guarantee the same face correspond to only one bounding box. This principle is important for multi-scale pyramid, as one face could be detected in different layers of the pyramid. However, in the case of overlapped two faces, this simple and crude method will cause missed detection, the face covered by a part of another face would not be detected. SNMS is proposed to refine the performance that it uses NMS when the score of bounding box to be suppressed is low and uses Soft-NMS when the score is high. For a high score box, it is more likely to be occluded face, Soft-NMS is used for this case. For low score box, NMS avoid these non-face boxes to be false positive. This gives a chance to detect faces that are covered by other faces without causing false positives as Soft-NMS does. SNMS is a compromise solution of NMS and Soft-NMS, that it provides a refine consideration of the score to avoid arbitrary discard or preserve the bounding box, which is important in a multi-scale face detection task.

2.2 Coexistence prior

2.2.1 The role of contextual information

Context acts as the key role in finding small instance in multiple scale recognition task. It is an important cue for object detection by humans, believed to reduce processing time and to help disambiguate low quality inputs by mitigating the effect of clutter, noise and ambiguous inputs. In natural images, object detection is strongly expected to fit into a certain relationship with the scene, and context gives access to that relationship. Different from adjusting the local receptive fields, our work extends the contextual information to the whole image rather than just surrounding objects. We employ semantic knowledge, the relationship of objects as higher-level extensive context, which helps to make sensitive detection when the object is ambiguously or marginally visible in crowd scene.

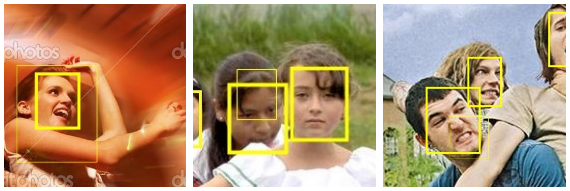

Fig. 4. Some false positive (thin yellow boxes).

2.2.2 Coexistence of a face pair

As shown in Fig. 4, some false positive appears in detection. Those false positive cannot be deleted by NMS or Soft-NMS. Because when the area of two boxes is quite different, their intersection is much smaller than union, $IOU(b_i, b_j)$ cannot reach the threshold of deleting redundant boxes by NMS. An understandable case is that a face containing another very small human face. According to the relationship between the faces, we design a principle using the coexistence of a face pair. If the two boxes are very different in size, their $IOU(b_i, b_j)$ is very small. In order to delete those false positive situations that faces should not coexist, IOB (intersection over box) is defined as follows.

$$IOB(b_i, b_j) = max\left(\frac{b_i \cap b_j}{b_i}, \frac{b_i \cap b_j}{b_j}\right) \qquad (2)$$

If IOB is more than 0.9, the box with low score will be deleted.

2.2.3 Coexistence of homogeneous faces

Because a large number of faces exist in the crowd scene, there are occlusion, low resolution and other issues. This section focuses on how to optimize the detectors for this scenario through crowd coexistence. If the scores of many faces dominate in an image, then it is reasonable to believe that size of the boxes which is similar to the sizes of these faces have high probability to be faces. To increase the scores of the boxes is effective in improving detection results. In crowd scene, this context is more reasonable. If one very small and low-resolution face is not in the crowd, human and detectors cannot detect it, however, if this face is in crowds, this kind of context could be good prior knowledge to detect an obscure face.

As is shown in Fig. 5, an easy-to-implement strategy is proposed to give detectors the ability to use coexistence of same size faces. A big number of same size faces exist in a picture, according to the ability of detector, there are some true faces with low score in this size, then the strategy is adopted to increase their scores. First, the number of same size box with high score is counted in a picture, this help us to know what size face is more possible to appear in this picture. Since this method is mainly to give chance to some true faces which have lower score than threshold of detector, so a low threshold is chosen to get more candidates, s_t is the low threshold of score. Those candidates, the number we count and previous threshold of detector are input of this method. After process of increasing score, boxes whose scores are lower than the final threshold will be deleted. The formula of this method to increase scores is as follows.

$$w_s = 0.5 + sigmoid(0.1\alpha) \qquad (3)$$

$$s_j = w_s * (s_j - s_t) + s_t \qquad (4)$$

α is the number of similar size faces in this image. s_j is the score. Two constraints are employed that w_s is belong to (1,1.5) and α is more than 5 which means this face is in crowd according to coexistence prior.

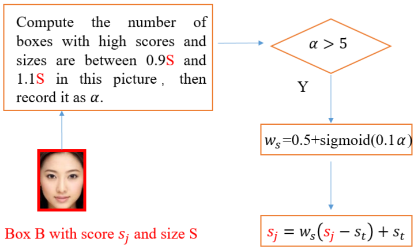
Compute the number of boxes with high scores and sizes are between 0.9S and 1.1S in this picture , then record it as α.
α > 5
Y
w_s=0.5+sigmoid(0.1α)
$s_j = w_s(s_j - s_t) + s_t$
Box B with score s_j and size S

Fig. 5. Process of the coexistence of homogeneous faces.

3. EXPERIMENTS

We first carry out Mean Average Precision analysis on WIDER FACE [2]. Individual parts of the proposed method are also evaluated on WIDER FACE. and a new dataset for crowd face detection is made by us to evaluate the performance of coexistence prior. Representative experiments on this new dataset are also presented.

3.1 Overall performance on WIDER FACE

We integrate our approach to the trained detectors: ACF [13], Two-stage CNN [2], Faceness [15], Multiscale Cascade [15], LDCF [2], Multitask Cascade [12], CMS-RCNN [10], Scale Face [14] and HR [8], and compare their performance with the original detectors. TABLE 1 shows that the proposed approach integrating into most face detectors have better performance than original methods. The best mAP (mean average precision) performance of the proposed approach presents on WIDER FACE 'hard' set, which indicates the capability of the proposed approach in challenging situations.

3.2 Separate experiments on WIDER FACE

3.2.1 Score-specific Non-Maximum Suppression

Different non-maximum suppression methods integrating into HR are tested on WIDER FACE hard set. As is shown in TABLE 2, SNMS shows better performance than Soft-NMS and traditional NMS.

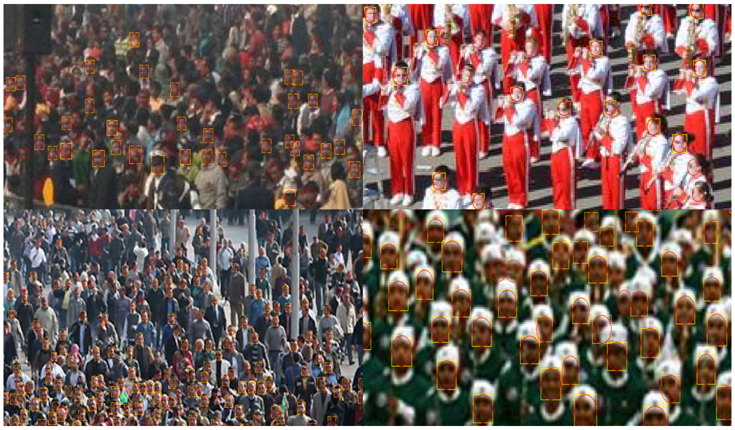

Fig. 8. Visual results on FIC. Yellow rectangles are results of original HR, and red ellipses are results of coexistence prior integrating into HR

3.2.2 Coexistence of a face pair

As is shown in Fig. 6, many false positive in all sizes of faces are removed by coexistence of a face pair (coexistence (2)). No matter what sizes of faces, this method does not delete true positive.

TABLE I. mAP ON WIDER FACE

Set	easy		medium		hard	
Method	Orignal	With proposed	Orignal	With proposed	Orignal	With proposed
ACF [13]	0.659	0.659	0.541	0.541	0.273	0.273
Two-stage CNN [2]	0.681	0.681	0.618	0.618	0.323	**0.324**
Faceness [15]	0.713	0.713	0.634	0.634	0.345	**0.350**
Multiscale Cascade[15]	0.691	0.691	0.664	0.664	0.424	**0.433**
LDCF [2]	0.790	0.790	0.769	**0.773**	0.522	**0.545**
Multitask Cascade [12]	0.848	**0.849**	0.825	**0.826**	0.598	**0.614**
CMS-RCNN [10]	0.899	0.899	0.874	**0.876**	0.624	**0.635**
Scale Face [14]	0.868	**0.870**	0.867	**0.868**	0.772	**0.782**
HR[8]	0.925	0.925	0.910	**0.912**	0.806	**0.813**

TABLE II. RESULT ON WIDER FACE HARD SET

Method	NMS	Soft-NMS	SNMS
Hard set	0.806	0.810	0.813

3.3 Experiments on FIC

We make a 'harder' dataset -- Faces in the crowd (FIC) directly. FIC is a crowd face detection dataset, in which images are selected from face dataset WIDER FACE, FDDB, AFW and web. We choose 30 images and label 6474 faces. This dataset includes 10 grayscale images and 20 color images, the largest number of faces in one image is 868 and the smallest number of faces in one image is 32.

We retrain HR detector on FIC. As is shown in Fig.7, Precision-Recall curve shows that coexistence prior has higher mAP both on original HR (from 0.098 to 0.104) and the retained HR (from 0.224 to 0.243). As is shown in Fig.8, it is obvious that the proposed approach finds more true faces, these representative results demonstrate that coexistence prior makes sense to crowd challenge in face detection.

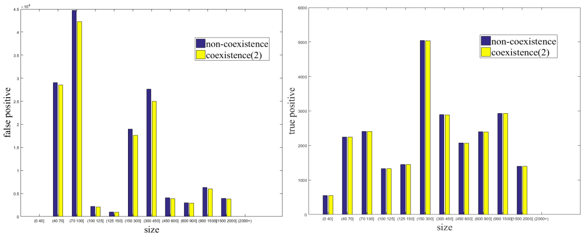

Fig. 6. Difference of false positive and true positive.

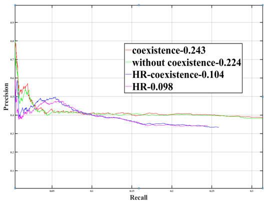

Fig. 7. Precision-Recall curve on FIC.

4. CONCLUSION

We propose a two-step general approach according to context. SNMS provides a refine consideration of the score to avoid arbitrary discard or preserve the bounding box. Coexistence prior makes sense to detect multi-scale and low resolution faces in crowd challenge. The proposed method does not require any extra training and is simple to implement. In future research, we will employ more coexistence strategies into face detection in the crowd.

5. ACKNOWLEDGEMENTS

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).

6. REFERENCES

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

[2] Shuo Yang, Ping Luo, Chen-Change Loy, Xiaoou Tang. "Wider face: A face detection benchmark." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[3] N. Bodla ,B. Singh , R. Chellappa, L.S. Davis. "Soft-NMS -- Improving Object Detection With One Line of Code." arXiv preprint arXiv:1704.04503 ,2017.8.

[4] Tychsen-Smith, Lachlan, and L. Petersson. "Improving Object Localization with Fitness NMS and Bounded IoU Loss."arXiv preprint arXiv:1711.00164 ,2017.10.

[5] Hosang, Jan, R. Benenson, and B. Schiele. "A Convnet for Non-maximum Suppression." German Conference on Pattern Recognition Springer International Publishing, 2016:192-204.

[6] Neubeck, Alexander, and L. V. Gool. "Efficient Non-Maximum Suppression." International Conference on Pattern Recognition IEEE Computer Society, 2006:850-855.

[7] Rothe, Rasmus, Matthieu Guillaumin, and Luc Van Gool. "Non-maximum suppression for object detection by passing messages between windows." Asian Conference on Computer Vision. Springer, Cham, 2014.

[8] P. Hu, and D. Ramanan. "Finding Tiny Faces." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2017:1522-1530.

[9] Dalal, Navneet, and B. Triggs. "Histograms of Oriented Gradients for Human Detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on IEEE, 2005:886-893.

[10] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 1271–1278.

[11] K, He, X. Zhang, S. Ren, J. Sun. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[12] K. Zhang, Z. Zhang, Z. Li, Y. Qiao. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." IEEE Signal Processing Letters 23.10(2016):1499-1503.

[13] B. Yang, J. Yan, Z. Lei, SZ. Li. "Aggregate channel features for multi-view face detection." IEEE International Joint Conference on Biometrics IEEE, 2014:1-8.

[14] Yang, Shuo, et al. "Face Detection through Scale-Friendly Deep Convolutional Networks." arXiv preprint arXiv:1706.02863 (2017).

[15] S. Yang, P Luo, CC. Loy, X. Tang. "From Facial Parts Responses to Face Detection: A Deep Learning Approach." IEEE International Conference on Computer Vision IEEE Computer Society, 2015:3676-3684.

[16] Torralba. Antonio，Murphy. Kevin P，Freeman. William T，Rubin. Mark A. "Context-based vision system for place and object recognition." IEEE International Conference on Computer Vision, 2003. Proceedings IEEE, 2003:273-280 vol.1.

[17] W. Xiang, DQ. Zhang,V. Arhitsos, H. Yu. "Context-aware Single-Shot Detector." arXiv preprint arXiv:1707.08682 ,2017.7

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

[19] Zhu, Chenchen, et al. "CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection." Deep Learning for Biometrics. Springer, Cham, 2017. 57-79.

[20] Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C., Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology, l4. 1982.

[21] A. Rosenfeld and M. Thurston. Edge and curve detection for visual scene analysis. IEEE Transactions on computers, 100(5):562–569, 1971.