International Journal of Computational Intelligence Systems
Vol. **13(1)**, 2020, *pp.* 1153–1160
DOI: https://doi.org/10.2991/ijcis.d.200805.002; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

YOLOv3: Face Detection in Complex Environments

Lin Zheng Chun[1,2], Li Dian[1,*], Jiang Yun Zhi[1,*], Wang Jing[1,*], Chao Zhang[1]

[1]*School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong 510665, China*
[2]*School of Computer Science, Guangdong University of Technology, Guangzhou, Guangdong 510665, China*

ARTICLE INFO

Article History
Received 20 Jun 2020
Accepted 28 Jul 2020

Keywords
Face detection
Complex environment
Priori box
Multiple score values

ABSTRACT

Face detection has been well studied for many years. However, the problem of face detection in complex environments is still being studied. In complex environments, faces is often blocked and blurred. This article proposes applying YOLOv3 to face detection problems in complex environments. First, we will re-cluster the data set in order to find the most suitable a priori box. Then we set multiple score values to make it possible to predict the results of multiple sets of images and find the optimal score value. Experimental results show that after adjustment, the model has more advantages in face detection than the original model in complex environments. The average accuracy is more than 10% higher than that of aggregate channel feature (ACF), Tow-stage convolutional neural network (CNN) and multi-scale Cascade CNN in face detection benchmarks WIDER FACE. Our code is available in: git@github.com:Mrtake/-complex–scenes-faceYOLOv3.git

1. INTRODUCTION

Face detection is a fundamental and critical task in various face technical. The early Viola-Jones [1] detector utilizes AdaBoost algorithm and Haar-like features to train. Since that, lots of subsequent work focuses on improving the performance of the algorithm. Subsequently, deformable part models (DMPs) [2–4] is introduced into face detection tasks by modeling the relationship of deformable facial parts. However, these methods rely on the designed features which are less representable and trained by separated steps.

With the continuous development of the convolutional neural network (CNN), a lot of progress for face detection has been made in recent years due to utilizing modern CNN-based object detectors, including R-CNN [5–8], SSD [9], YOLO [10–12], feature pyramid network (FPN) [13]. Benefiting from the powerful deep learning approach to extract image features. Compared with designed feature method, the CNN-based face detector achieves better performance, which provides a new foundation for future methods.

Recent, the anchor-based detection framework mainly used to face in complex environments such as WIDER FACE. Face Faster R-CNN [14] detected each region by region of interest (ROI), however, although this method has a high precision, it has a large amount of computation and a slow detection speed. Face R-FCN [15] combines the full convolution network with the region-based network module to dectect face and eliminates the effect of nonuniformed contribution in each facial part using a position-sensitive average pooling, but the detection speed is still not ideal. FacenessNet [16] designed the face features, taking into account hair, eyes, nose, mouth and beard, but face detection accuracy is not high in complex environment.

The above work points out that the face detector with better performance is usually very slow, and if the image feature extraction and generalization are not sufficient, the face cannot be effectively attention when there is blur and occlusion. Considering the above two problems, the paper proposes to apply YOLOv3 [12] to face detection in complex environments.

Firstly, Darknet-53 network has very good detection speed, which is significant for face detection. Because the environment is complex and changeable in the actual application, we need to locate the face quickly to meet the practical application.

Secondly, the features extracted by Darknet-53 network have strong generalization ability. This ensures that the detector is adaptable to different environments. At the same time, the idea of FPN is adopted in the network, which is effective for face detection at different scales.

For clarity, the main contributions of this work can be summarized as two fold:

1. The prior box in our network has been adjusted several times. Four groups of prior frames were set up for the experiment, and the prior frame with the best experimental effect was selected.
2. For face detection in complex environments, it is very difficult to choose the score value setting. But the setting of the score value will directly affect the final result. Currently, there is no unified standard for setting this value. Therefore, we add

*Corresponding authors. Email: jiangyunzhi2008@163.com, 862729693@qq.com, wj_adr@163.com

a step of detecting multiple values in the final detection process, which can detect the results under multiple score values at the same time.

3. We achieved better performance through adjustments to YOLOv3 on the common face detection benchmarks WIDER FACE.

The rest of the paper is organized as follows: Section 2 provides an overview of the related works. Section 3 introduces the framework and adjustments of YOLOv3. Section 4 presents the experiments and Section 5 concludes the paper.

2. RELATED WORK

Anchor-based face detectors. Faster R-CNN [7] was the first algorithm to propose Anchor, and then this idea was widely used in two-stage and one single shot object detectors. In recent years, anchor-based detectors [9,10] have made great progress. S3FD [17] proposed anchor matching strategy, which designs scales of anchors to ensure that the detector can handle various scales of faces well. FaceBoxes [18] introduces anchor densification to make different types of anchors on the image have the same density.

Multi-scale detector. To improve the performance of object detector to handle object of different scale, Many state-of-the-art researches [17,19,20] construct different structures in the same framework to detect objects of different sizes, design high-level features to detect larger objects and low-level features to detect smaller objects, FPN [13] proposed a top-down architecture to use high-level semantic feature maps at all scales. Detectors [21] proposes to add feedback and recursive repeat stacking to the structure of FPN, so that it can pay attention to the features that have not been noticed before. Recently, Tang [20] introduces a low-level feature pyramid network layer (LFPN), a top-down structure starting from the middle layer instead of the top layer. Because not all high-level features are necessarily helpful for detecting smaller faces.

3. YOLOv3

This section introduces the object detector YOLOv3, and how to adjust it for face detection in complex scenes. We first briefly introduce the network architecture in Section 3.1. Then we perform multiple clustering adjustments on the prior box for the data set, hoping to find the most suitable prior box for this data set in Section 3.2. Finally, Section 3.3 presents to set multiple score values to prediction.

3.1. Network Architecture

YOLOv3 is an improved version of YOLO and YOLOv2. The main change in its network structure is the introduction of residual blocks, which ensures that even if the YOLOv3 network becomes deeper, the model can still converge quickly. In order to better deal with the problem of overlap, the loss function uses binary cross-entropy loss; The multi-scale fusion method is adopted to merge the high-level semantics with the low-level, which improves the sensitivity to small targets. These improvement measures mainly improve the detection accuracy. In the COCO data set, the AP_{50} (average precision) standard, and the accuracy reaches 57.9%.

Table 1 Comparison of backbones. Accuracy, billions of operations, billion floating point operations per second and FPS for various networks.

Backbone	Top-1	Top-5	Billion Floating-Point Operations (BFLOP/s)	Frames Per Second (FPS)
Darknet-19 [11]	74.1	91.8	1246	**171**
Darknet-53	77.2	**93.8**	**1457**	78
ResNet-101 [22]	77.1	93.7	1039	53
ResNet-152 [22]	**77.6**	**93.8**	1090	37

Network architecture is shown in Figure 1. Three feature maps of different scales in the image are extracted through the convolutional network, and the feature maps is divided into S × S grid; perform nonmaximum suppression (NMS) processing on the predicted bounding box and the prior box for each grid and calculate the confidence of the output bounding box, it is effectively filter out the useless bounding box.

Darknet-53 network. YOLOv3 uses the Darknet-53 network as the overall framework, with a total of 53 convolutional layers. The residual block structure of the ResNet [22] network is introduced into the network, and each residual block is composed of two layers of convolution, and the connection method adopts a jump connection. This reduces the complexity of the model, while the number of required training parameters has not increased significantly. YOLOv3 proposed a speed comparison experiment on the backbone network in the paper, and the results are quoted in Table 1. Each network is trained under the same configuration and tested with 256×256 image to obtain the test accuracy of single-size image. The running time is measured by processing 256 × 256 image on Titan X. All tests are performed on ImageNet.

FPN network. Feature Pyramid essentially analyzes image information on multiple scales, because the image includes many object features of different sizes, any single-scale analysis will cause the omission of object features. FPNs [13] is developed on the feature image pyramid. YOLOv3 draws on its ideas, and its architecture is shown in Figure 2.

3.2. Multiple Clustering Adjustments on the Prior Box

The YOLOv3 algorithm utilizes K-means to perform dimensional clustering on the data set COCO, but for the WiderFace data set used in this article, this priori box is not the best choice. We must re-cluster the prior boxes based on the data. In YOLOv3, to prevent the emergence of "big box advantage," introduce intersection over union (IOU) to measure the similarity between the candidate box and the actual box. Thus for our distance metric we use

$$d(box, centrodi) = 1 - IOU(box, centroid) \quad (1)$$

Through the cluster analysis of the data set, the results are shown in Table 2. With the different K value of the number of clusters, the average_IOU is also changing, with it continues to rise, which means that the a priori boxes obtained by our clustering are better representative of different scales. However, the more a priori boxes are selected, the better the detection performed will be? This question cannot be answered here. According to the clustering results

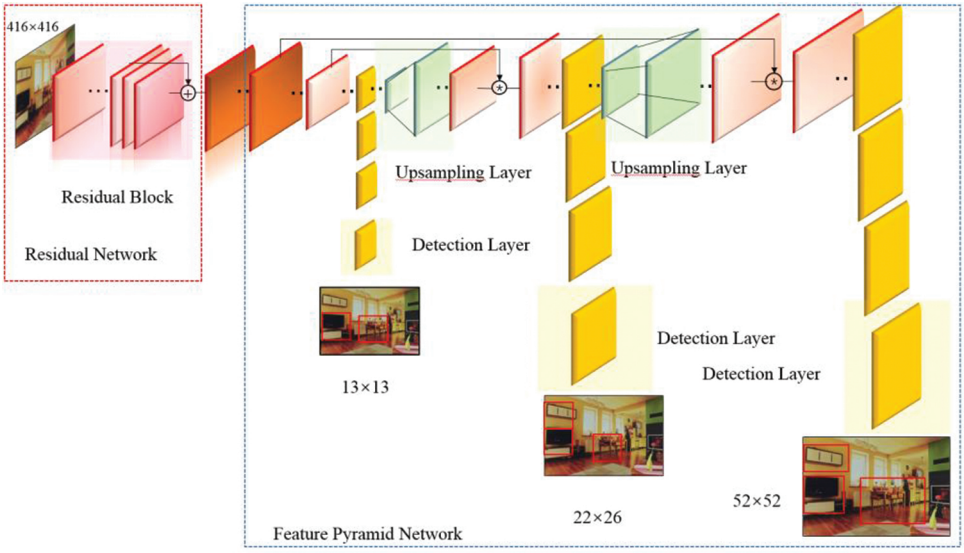


Figure 1 | Architecture of YOLOv3. It consists of **Residual Network, Feature Pyramid Networks (FPN).**

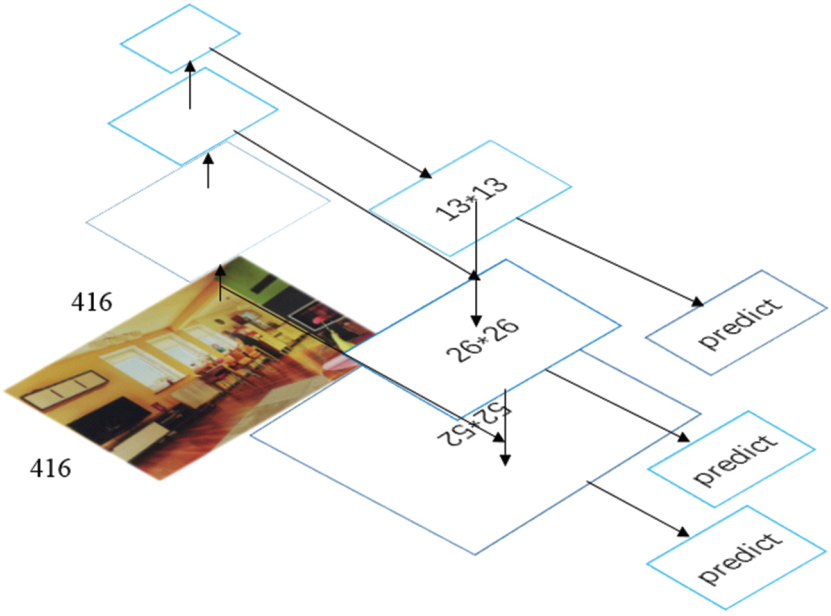


Figure 2 | Feature pyramid network (FPN). Two feature accumulations were performed on the last layer of feature maps.

Table 2 | Average_intersection over union (IOU) changes.

k	3	5	7	9	11	13	15	17	19	21	24
Avg_IOU	0.602	0.697	0.747	0.784	0.798	0.821	0.821	0.836	0.848	0.847	0.856

given in Table 2, we will select the four sets of values k = 9, 15, 21, 24 for experimental. The selected 4 groups priori box is shown in Table 3.

3.3. Multiple Score Value Prediction

YOLOv3 introduce the NMS [23] algorithm to extract the most likely object and its corresponding frame in the result. It is usually taken IOU $\geq$ 0.5, however, there is no standard for the choice of confidence, and it should be adjusted according to the tested data set. The confidence set in YOLOv3 is usually relatively large; such a setting is obtained after adjustment on the coco data set. So we cannot use such a high confidence because there are many small faces in the data set we use. A high confidence will miss a large number of face, while a low confidence will produce too many negative cases. In order to find a reasonable score, multiple score values are set in

Table 3 | Clustering results with different k values.

K	Clustering Result
9	(6 × 6), (10 × 13), (15 × 19), (20 × 26), (28 × 35), (39 × 50), (58 × 74), (98 × 131), (228 × 304)
15	(3 × 3), (4 × 5), (5 × 7), (7 × 9), (8 × 13), (10 × 12), (12 × 14), (13 × 18), (14 × 16), (16 × 20), (20 × 26), (28 × 36), (42 × 53), (73 × 96), (168 × 224)
21	(5 × 5), (6 × 8), (7 × 10), (9 × 11), (11 × 13), (13 × 16), (15 × 21), (16 × 18), (17 × 27), (18 × 23), (18 × 21), (20 × 25), (21 × 30), (25 × 29), (29 × 36), (34 × 44), (43 × 54), (56 × 72), (79 × 104), (124 × 166), (268 × 355)
24	(5 × 5), (6 × 8), (7 × 9), (8 × 12), (8 × 9), (9 × 10), (10 × 12), (10 × 14), (12 × 15), (13 × 18), (14 × 14), (15 × 19), (16 × 24), (17 × 17), (17 × 20), (18 × 26), (20 × 22), (23 × 28), (27 × 35), (35 × 44), (47 × 59), (68 × 90), (112 × 150), (250 × 332)

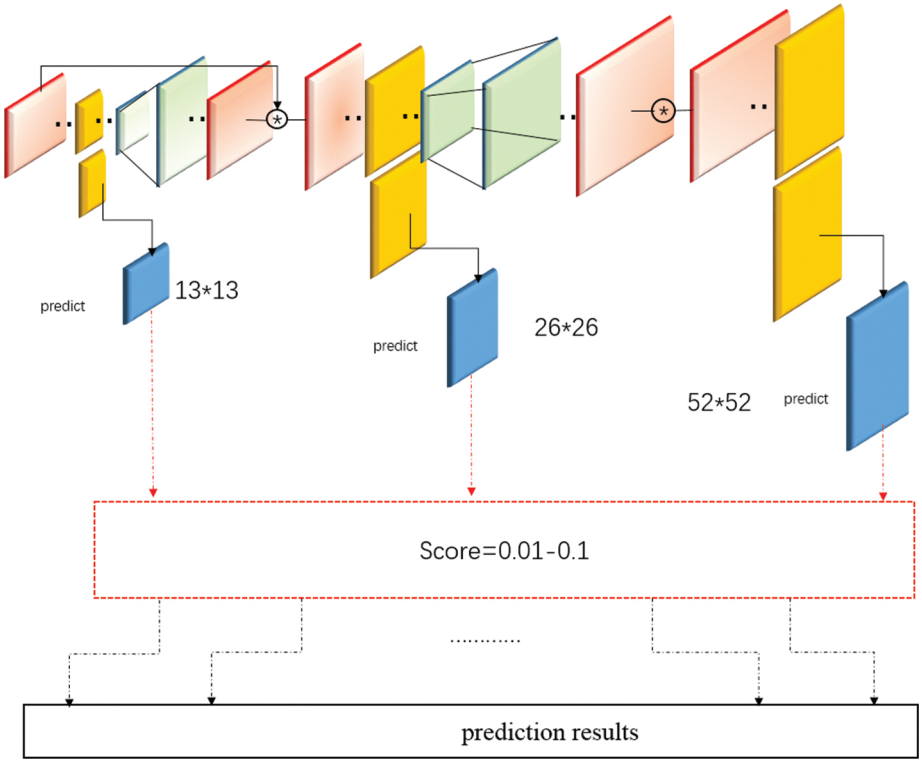


Figure 3 | **Multi-score value prediction.** We set multiple score values. Values from 0.01 to 0.1. Values above 0.1 are too large and the detection accuracy is too low.

the output layer of the network, and multiple predictions are made for each detection target. This avoids manually searching for confidence.

$$\mathbf{predicted}\left(\mathbf{score},\mathbf{image}\right) = \mathbf{f_{ij}} \tag{2}$$

i represents the number of image; j represents the number of predicted object in each image; score represents multiple confidence. f_{ij} represents the predicted result under different confidence. The multi-score value prediction is shown in Figure 3.

4. EXPERIMENTS

In this section, we first train YOLOv3 models under 4 groups of different k values. The training set is WIDER FACE. Then, the prediction results under the 4 sets of k values are tested on the WIDER FACE verification set benchmark, and select the most suitable priori box. Finally, through multiple score predictions under the appropriate a priori box, find the most suitable score threshold for this data set.

Table 4 | K value change comparison. The score value is 0.09 and 0.045.

K	Easy	Medium	Hard	Score
9	0.797	0.747	0.493	0.09
9	0.801	0.758	0.535	0.045
15	0.782	0.755	0.481	0.09
15	0.787	0.766	0.518	0.045
21	0.770	0.706	0.441	0.09
21	0.776	0.721	0.486	0.045
24	0.719	0.683	0.458	0.09
24	0.728	0.700	0.5	0.045

Sets of model analysis under different k values. To select the appropriate k value, the paper selects 4 groups of different k values for control experiments; in the case of the same network settings, calculate the average accuracy rate (mAP) under three different difficulties in the WIDER FACE verification set for comparison. The results are shown in Table 4.

Table 5 The impact of score changes on accuracy. There are also 3 sets of intersection over union (IOU) values set in the table. It can be observed that the change of IOU will indeed affect the accuracy to a certain extent. But, this effect gradually becomes weaker as the score value continues to increase. In the following 4 sets of score values, the change of IOU does not affect the change of accuracy.

Score	Easy	Medium	Hard	IOU
0.01	0.804	0.764	0.558	0.45
0.01	0.790	0.769	0.550	0.50
0.01	0.805	0.762	0.552	0.55
0.02	0.803	0.762	0.552	0.45
0.02	0.804	0.769	0.550	0.50
0.02	0.804	0.761	0.546	0.55
0.03	0.803	0.760	0.546	0.45
0.03	0.803	0.760	0.544	0.50
0.03	0.803	0.759	0.540	0.55
0.045	0.801	0.758	0.535	0.45
0.045	0.802	0.758	0.533	0.50
0.045	0.802	0.756	0.529	0.55
0.06	0.800	0.754	0.522	0.45
0.06	0.800	0.754	0.522	0.50
0.06	0.800	0.753	0.516	0.55
0.075	0.799	0.751	0.507	0.40
0.075	0.799	0.751	0.507	0.45
0.09	0.797	0.747	0.493	0.40
0.09	0.797	0.747	0.493	0.45
0.1	0.797	0.744	0.484	0.40
0.1	0.796	0.744	0.484	0.45

Table 6 Comparison of the two models in WIDER FACE validation sets.

Model	Easy	Medium	Hard
Adj_YOLOV3	0.805	0.762	0.552
Orig_YOLOV3	0.601	0.509	0.238

According to Table 4, it can be seen that the increase of the a priori frame does not improve the accuracy, but has decreased. The reason for the drop caused by label rewriting; in the training set, the face distribution is very dense and the length and width of the face are very close, so there is more serious label rewriting [24]. At the same time, the a priori box is a group of 3, which are used to detect large-scale faces, medium-scale faces and small-scale faces. However, in the WIDER FACE data set, not all faces of all scales exist; in the training set, there are a large number of small-scale and medium-scale faces. If we continue to train, small-scale and medium-scale faces are forced to be assigned to different layers for prediction, which is very unreasonable.

Multi-score value prediction. In the previous section, we decided to introduce 9 prior boxes after experimentation to alleviate the impact of label rewriting. But, the choice of score value still needs to be adjusted. Set multiple sets of score values to make multiple predictions for multiple pictures at the same time. The results are shown in Table 5.

According to the table, we can know that as the score continues to increase, the accuracy is constantly decreasing, and the overall trend is downward. You can choose score = 0.01 or 0.02 as the confidence threshold.

Finally, we compare the adjusted model with the unadjusted model. The results are shown in Table 6.

In order to show the effect of the picture, we selected the detection results in 20 scenes. The test results are shown in Figure A1 in the Appendix.

4.1. Evaluation on Benchmark

We evaluate our YOLOv3 on the most popular face detection benchmarks, including Face Detection Data Set and WIDER FACE [25].

WIDER FACE Dataset. It contains 32203 images and 393703 annotated faces with a high degree of variability in scale, pose and occlusion. The database is split into training (40%), validation (10%) and testing (50%) set, where both validation and test set are divided into "easy," "medium" and "hard" subsets, regarding the difficulties of the detection. Our model is trained only on the training set and evaluated on both validation set with the state-of-the-art face detectors, such as []. Figure 4 presents the precision-recall curves and mAP values. Our detector is not perform good on three subsets, i.e., 0.805 (easy), 0.762 (medium), 0.552 (hard) for validation set.

5. CONCLUSION

Through the adjustment of YOLOv3, it is applied to the problem of face detection in complex environments. But according to the experimental results, the performance is not optimal. The main reason is that YOLOv3 has label rewriting in the face-intensive data set. Since small-sized faces and medium-sized faces are the majority in the data set, it is very unreasonable to simply allocate them to three layers for training. Therefore, YOLOv3 needs to make reasonable improvements to specific data sets to further improve the accuracy of face detection in complex environments. However, the constant adjustment of the parameters in this paper has greatly improved its detection accuracy compared with the original network, so it has certain application value.

ACKNOWLEDGMENTS

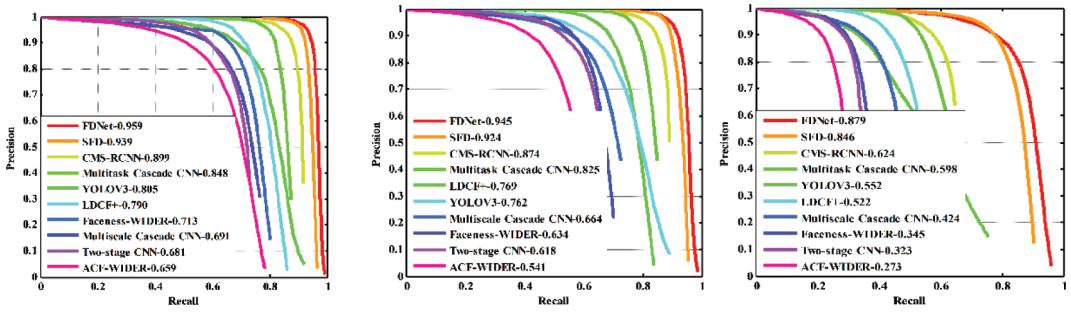
First of all, I would like to thank Mr. Lin very much for providing important comments during the process of writing the paper and making serious revisions to the paper; at the same time, I would like to thank Mr. Jiang for providing us with experimental equipment and experimental environment.Finally, I am very grateful to the other members of the group for providing materials and technical support for the paper, and Guangdong Technical Normal University for funding the paper.

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interest to this work.We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

AUTHOR CONTRIBUTIONS

Lin Zheng Chun is the first author. He put forward the idea of the paper and provided guidance in the experiment. Li Dian is the second author. He conducted experiments and wrote the thesis. Jiang Yun Zhi is the corresponding author and provided the



(a) Val: Easy (b) Val: Medium (c) Val: Hard

Figure 4 | Precision-recall curves on WIDER FACE validation sets.

experimental environment. Wang Jing and Chao Zhang are the third authors and provided reference materials during the writing of the paper.

FUNDING STATEMENT

National Natural Science Foundation of China (Youth) (61702118); Guangdong Provincial Department of Education's Young Innovative Talent Project (Natural Science) (2016KQNCX089); Guangdong Province General University Characteristic Innovation Project (Natural Science) (2017KTSCX113)

REFERENCES

[1] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2004), 137–154.
[2] M. Mathias, R. Benenson, M. Pedersoli, L.V. Gool, Face detection without bells and whistles, in European Conference on Computer Vision, Zurich, Switzerland, 2014.
[3] J. Yan, Z. Lei, L. Wen, S.Z. Li, The fastest deformable part model for object detection, in Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014.
[4] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012.
[5] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014.
[6] R. Girshick, Fast R-CNN, in IEEE International Conference on Computer Vision, Santiago, Chile, 2015, pp. 1440–1448.
[7] S. Ren, K. He, R. Girshick, *et al.*, Faster R-CNN: towards real-time object detection with region proposal networks, in Advances in Neural Information Processing Systems, Held in Montreal, Canada in 2015, pp. 91–99.
[8] K. He, G. Gkioxari, P. Dollár, *et al.*, Mask R-CNN, in IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 2961–2969.
[9] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.Y. Fu, A.C. Berg, SSD: single shot multibox detector, in European Conference on Computer Vision, Amsterdam, The Netherlands, 2016.
[10] J. Redmon, S. Divvala, R. Girshick, *et al.*, You only look once: unified, real time object detection, in IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 779–788.
[11] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 7263–7271.
[12] J. Redmon, A. Farhadi, Yolov3: an incremental improvement, in IEEE Conference on Computer Vision and Pattern Recognition, Held at the Hawaii Convention Center in Honolulu, Hawaii. 2017, pp. 6517–6525. arXiv:1804.02767
[13] T.Y. Lin, P. Dollár, R. Girshick, *et al.*, Feature pyramid networks for object detection, in IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 2117–2125.
[14] H. Jiang, E. Learned-Miller, Face detection with the faster R-CNN, in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, Washington, DC, USA, 2017, pp. 650–657.
[15] J. Dai, Y. Li, K. He, *et al.*, R-FCN: object detection via region-based fully convolutional networks, in Advances in Neural Information Processing Systems, Held in Barcelona 2016, pp. 379–387.arXiv:1605.06409
[16] S. Yang, P. Luo, C.C. Loy, *et al.*, Faceness-Net: face detection through deep facial part responses, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018), 1845–1859.
[17] S. Zhang, X. Zhu, X. Lei, H. Shi, X. Wang, S.Z. Li, S3FD: single shot scale-invariant face detector, in International Conference on Computer Vision, Venice, Italy, 2017.
[18] S. Zhang, X. Zhu, X. Lei, H. Shi, X. Wang, S.Z. Li, Faceboxes: a CPU real-time face detector with high accuracy, arXiv preprint arXiv: 1708.05234, 2017.
[19] M. Najibi, P. Samangouei, R. Chellappa, L.S. Davis, SSH: single stage headless face detector, in International Conference on Computer Vision, Venice, Italy, 2017.
[20] X. Tang, D.K. Du, Z. He, *et al.*, PyramidBox: a context-assisted single shot face detector, in European Conference on Computer Vision, Munich, Germany, 2018, pp. 812–828.
[21] S. Qiao, L.-C. Chen, A. Yuille, DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution, arXiv preprint arXiv: 2006.02334, 2020.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hong Kong, China, 2016, pp. 770–778.

[23] A. Neubeck, L. Van Gool, Efficient non-maximum suppression, in 18th International Conference on Pattern Recognition (ICPR'06), IEEE, Hong Kong, China, 2006, vol. 3, pp. 850–855.

[24] P. Hurtik, V. Molek, J. Hula, *et al.*, Poly-YOLO: higher speed, more precise detection and instance segmentation for YOLOv3, arXiv preprint arXiv: 2005.13243, 2020.

[25] S. Yang, P. Luo, C.C. Loy, *et al.*, WIDER FACE: a face detection benchmark,in Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 5525–5533.

APPENDIX

Figure A1 | Part of test on validation sets. Only 20 differnent scenarios in here.