Small Face Detection Based On Feature Fusion

Zili Hao1*, Hangqi Shi2

1Software engineering, Beijing University of Technology, BeiJing, 100124, China
2Software engineering, Beijing University of Technology, BeiJing, 100124, China
*Corresponding author’s e-mail: haozl@emails.bjut.edu.cn

Abstract. While great progress has been made in face detection, one of the remaining challenges is to detect the faces of small targets in images. The small target face we define, mainly includes the absolute size of the face is not greater than 32×32 pixels and the relative proportion of the face size is not greater than one tenth of the image size. To meet this challenge, we propose an effective face detector, called SmallFaceBoxes, which has a better performance in both speed and accuracy. Specifically, our method has a lightweight but powerful network structure consisting of small target face detection layers and multi-scale convolution layers. The design of small target face detection layers enables detector to increase the representation ability of small and medium-sized faces in the image and improve the detection accuracy of small faces. The multi-scale convolution layers aim at enriching the receptive fields and discretizing anchors over different layers to handle faces of various scales. In addition, according to the size of receptive field, we propose a strategy of setting Anchor in multiple feature layers, thus improving the recall rate of small face. As a consequence, our algorithm performs well on the wideface dataset.

1. Introduction

Face detection is a very important task in the field of computer vision. Its development has gone through different stages. Many studies related to face detection have had very good performance in detection performance, they improve the accuracy and speed of detection algorithm from different angles. some works[1] pay attention to the design of network, and some data augmentation methods [2] and new loss designs [3]are presented. Besides, a few works [4] begin to redesign the matching strategy and label assignment process.

However, there are still some difficulties in the current face detection research, mainly including small target face detection, occluded face detection and multi-pose face detection. This paper mainly studies the small target face detection.

There are two main difficulties in small target face detection. First, small face of pixels are relatively small, and contains fewer information. Convolutional neural network can extract high-dimensional features of images, but some features of images will be lost in the process of calculation. Therefore, the high dimensional features lose the small face information seriously so that the algorithm is less effective in detecting small faces. Secondly, the size of the face in the image has a great change, not only does the algorithm have to improve the detection accuracy of small face, but also to ensure the effect of multi-scale face detection. Finally, while ensuring the detection accuracy of the algorithm, the detection speed of the algorithm should also be improved.

The main contributions of this paper are summarized as follows:
In view of the problem that the depth features of neural network cannot well represent the features of small faces, a Full Convolution Layers (FCL) is designed to reduce the loss of features in the process of convolution calculation.

In order to further improve the characterization ability of shallow features, this paper designed a Partial Feature Integration Layers (PFIL) to integrate high-level features and low-level features, and improve the accuracy of small face target detection.

In addition, this paper adopts the strategy of setting the anchors on the multi-level feature map. According to the feature map from different depth corresponding to the different size of the receptive field, the reasonable size of the anchors are set to improve the recall rate and accuracy of the detection algorithm.

The network structure is shown in Figure 1:

![Figure 1. network structure. (a)fully convolution layers (b)partial feature integration layers (c)Multiple Scale Convolutional Layers](image)

2. Related work

Thanks to the rapid development of convolutional neural network, the vast majority of current target detection and face detection algorithms use neural network as the main structure, which makes the speed and accuracy of detection algorithms have a qualitative leap compared with traditional algorithms.

2.1. Traditional detection algorithms

Previous face detection algorithms mainly use hand-craft features, such as Haar[5], in order to recognize and detect faces. However, face is affected by age, posture, illumination, occlusion, blur and other factors, and it is difficult to cover all conditions with hand-craft extracted features, so the detection effect of traditional algorithms is not enough to be applied in daily life.

2.2. Detection algorithm based on convolutional neural network

Different from traditional face detection algorithms, researchers use convolutional neural networks to extract image features and make inferential calculations. In order to improve the detection effect of the algorithm, people put forward different methods from many angles after continuous research. On the basis of SSD, S3FD extended the anchor-associated layer to the C3 stage, and proposed an anchor matching strategy to cover the small faces, Pyramidbox[1] emphasize the importance of data anchor sampling and enhancing context.

To sum up, methods presented in face detection almost cover every part of deep learning training from data processing to loss designs. It is obvious that all of these methods focus on the challenge of
hard faces. Based on the above work, this paper proposes a SmallFaceBoxes, a detector for tiny faces and small faces.

3. SmallFaceBoxes
This section introduces our specific work: Fully Convolution Layers, Partial Feature Integration Layers and Anchor Setting Strategy.

3.1. Fully Convolution Layers
Since the small face in the image contains less information, in order to reduce the loss of small face features in the calculation process, we only use convolution calculation to extract features in the first part of the network, thus forming the fully convolutional layers. The structure is shown in Figure. 1(a).

In order to effectively extract small face features, the full convolution module has a total of four layers. The convolution kernel of 5×5 is used in the first layer, and the convolution kernel of 3×3 is used in Conv1_2, Conv1_3 and Conv1_4. The step size of each convolution layer is 2. After the convolution of the first four layers, the total step size is 16, and the theoretical receiving field of the feature map is 33×33. At the same time, for the purpose of reducing the loss of image information, the number of convolution cores increases twice with the increase of network depth. The number of the first layer of convolution cores is set as 32. After the calculation of the first four layers of convolution, the number of convolution cores grows to 256.

3.2. Partial Feature Integration Layers
In order to solve the multi-scale problem of detected targets, some works extracted features from different layers of the network for prediction, such as SSD, but the SSD did not use enough low-level features, which are very helpful for detecting small objects. Later, some algorithms integrate the features of low resolution and high semantic information with those of high resolution and low semantic information, so that fusion features have rich semantic information, such as FPN. Because of network degradation, by using the method of feature fusion of all layers, the detection effect of the small face target is not significantly improved, but the computing overhead of the network is greatly increased. In the process of convolution calculation, the actual receptive field presents a Gaussian distribution in the theoretical receptive field. Thus, the current feature map can meet the detection requirements of tiny faces, which with a pixel value of about 16x16. Therefore, this paper proposes part of the feature fusion module. As shown in Fig. 1(b)

The partial feature fusion module contains the same three parts, each of which is composed of residual structure. When the features are enough to represent small faces, in order to prevent network degradation, this paper adopts three ResNet modules. As shown in Fig. 2(a). In the module, 1×1 convolution is firstly used to reduce the feature map depth and calculation cost. Then, a 3×3 convolution is used to extract the feature, and finally, 1×1 convolution is used to restore the feature map depth. The feature map size remains unchanged through the residual module. The features obtained from the three ResNet modules and the original features were added bit-by-bit to obtain the fusion features. Finally, 1×1 convolution was performed to reduce the impact of overlap effect.

3.3. Multiple Scale Convolutional Layers
In order to learn the face features of different scales, output features of the anchor-associated layers should respond to the sensory field of different scales. For this reason, Inception modules are adopted in this paper. Inception Modules consists of multiple convolution branches, each containing a different convolution layer, As shown in Fig. 2(b).

3.4. Anchor Set Strategies
In this paper, four feature maps are selected to set Anchor. Each feature map corresponds to a different receptive field size. In this paper, anchors matching the receptive field size are set. The scale of anchor
for the Conv1_5 layer is 16 and 32 pixels, for the Inception3 is 64 and 128 pixels, for the Conv3_3 layer and Conv4_2 layer are 256 and 512 pixels, respectively.

4. Experiment
This subsection introduces the implementation details including training dataset, data augmentation, matching strategy, loss function, etc.

4.1. Dataset
Our model is trained on the WIDER FACE training subsets.

4.2. Implementation Details

4.2.1. Data augmentation. Each training image is processed by the following data augmentation strategies in turn:
- Color distortion: Applying some photo-metric distortions similar to [6].
- Random clipping: We randomly clipped five square patches from the original image, one of which is the largest square patch, and the other sizes are all between [0.3, 1] of the short size of the original image. We then randomly select a patch for subsequent operations.
- Scale transformation: After cropping randomly, adjust the size of the selected square patch to 1024x1024
- Horizontal flipping: The resized image flips horizontally with a probability of 0.5.

4.2.2. Matching strategy. During training, we first matched each face to the anchor with the best Jaccard overlap, and then matched the anchor to any face with Jaccard overlap higher than the threshold (i.e., 0.35).

4.2.3. Loss function. We adopt a 2-class softmax loss for classification and the smooth L1 loss for regression.

4.2.4. Other implementation details. All parameters are randomly initialized using the "Xavier" method. We train model using SGD with 0.9 momentum, 0.0005 weight decay and batch size 8. The maximum number of iterations is 160k. The learning rate of the first 100k iteration is 0.001, then we
continue training for 30k iterations with 0.0001 and 0.00001, respectively. Our method is implemented in the Pytorch library.

4.3. Model analysis

We carried out extensive ablation experiments on the wideface val dataset to analyze our model. According to statistics, the proportion of small faces in the wideface val dataset is 93%. Therefore, it is reasonable to use this dataset to verify the effectiveness of our algorithm. For all the experiments, we use the same settings, except for specified changes to the components.

To better understand SmallFaceBoxes, we ablate each component one by one to examine how each proposed component affects the final performance. 1) First, we use the combined model of convolution and pooling to replace the full convolutional layer. In the convolution calculation, the size of the convolution kernel is 7×7 and 5×5, and the step size is 4 and 2, respectively. 2) Then, we replace bottleneck module with three convolutional layers, which all have 3×3 kernel size, at the same time, the operation of feature fusion is not used. 3) Finally, we take the place of Inception module with three convolutional layers in MSCL.

The ablative results are listed in Tab. 1 and some promising conclusions can be summed up as follows:

| Contribute   | FCL | PFIL | MSCL |
|--------------|-----|------|------|
| SmallFaceboxes |     |      |      |
| Accuracy (mAP) | 85.5  | 82.7  | 80.1  | 83.4 |

4.3.1. FCL is better. By comparing the data in columns 1 and 2 of Tab. 1, We found that replacing the FCL with the combination of convolution and pooling would result in the mAP of the algorithm dropping from 85.5 to 82.7.

4.3.2. PFIL is crucial. As can be seen from the results of Column 1 and 3 in Tab.1, the mAP decreases sharply from 85.5 to 80.1 without the strategy of feature fusion, which indicates the effectiveness of feature fusion strategy.

4.3.3. MSCL is efficient. According to the Tab. 1, we replaced the MSCL structure with a single convolution operation, and the mAP was reduced from 85.5 to 83.4. This structure is mainly aimed at multi-scale face detection.

5. Conclusion

In this paper, we point out that one of the difficulties in small target face detection is that the features extracted by the neural network cannot properly represent the face information. To solve this problem, we proposed the full convolution module and feature fusion module, as well as the strategy of setting reasonably sized Anchor in multiple feature layers. The experiment finds that methods proposed in this paper effectively improve the detection effect of small face.

References

[1] Xu Tang et al. “Pyramidbox: A context-assisted single shot face detector”. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 797–813.
[2] Zhihang Li et al. “Pyramidbox++: High performance detector for finding tiny face”. In: arXiv preprint arXiv:1904.00386 (2019).
[3] Jian Li et al. “DSFD: Dual Shot Face Detector”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2019.
[4] Yang Liu et al. “HAMBox: Delving Into Mining HighQuality Anchors on Face Detection”. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2020.

[5] Paul Viola and Michael Jones. “Robust real-time face detection”. In: null. IEEE. 2001, p. 747.

[6] A. G. Howard. Some improvements on deep convolutional neural network based image classification. arXiv preprint arXiv:1312.5402, 2013.