Image Flame Detection Method Based on Improved YOLOv3

Ziteng Dai 1*
1School of Resources and Environmental Engineering, Wuhan University of Science and Technology, Wuhan 430081, China.
Email: 916485463@qq.com

Abstract. The general flame detection method is sensitive to environmental changes and has low robustness, so the detection rate is not high. In this paper, a deep learning flame detection algorithm based on the improved YOLOv3 algorithm is proposed to automatically extract flame characteristics; with the help of MobileNet's deep separable convolution to compress the deep network model, replace the YOLOv3 backbone network with MobileNet, decompose the standard convolution into deep convolution and point-wise convolution, which can effectively reduce the amount of model parameters, Use K-means clustering to generate Anchor Box at different scales, further enhance the detection effect of the model. This paper conducts data analysis experiments on the self-built flame data set. The results show that the optimized algorithm (M_YOLOv3) has achieved 97.34% average precision (AP) on the data set. The flame target detection algorithm proposed in this paper can achieve high-precision and rapid detection of multi-scale flame targets in flame images with different shapes and sizes.

1. Introduction
Fire, as a kind of disaster accident with high frequency, poses a huge threat to people's life and property safety, so the detection and early warning of fire is of great significance [1]. Traditional fire detection technology uses temperature sensors and smoke sensors to achieve fire detection. It has the disadvantages of limited detection distance, long response time, and low accuracy, which is far from meeting the needs of complex scenes. Thanks to the advancement of image processing technology and the development of computer vision technology, image-based fire detection technology has gradually been applied. It can detect the occurrence of fire by analyzing the characteristics of changes in the picture so as to overcome the defects of traditional detection technology in the application of large space.

In the field of flame detection, there are many flame image features. Traditional flame detection pays more attention to manually extracting features to build a detection model. For flames, features suitable for extraction include brightness features, center of gravity height, flame initial area change features, multifractal spectrum features, flame spike triangle sine value, edge jitter features[2], etc. Xie Linbai[3] and others first used the combination of RGB and HIS to extract the flame foreground, and then extracted the color and texture features of the flame foreground image. Dimitropoulos[4] and others proposed a detection method that uses flame color, stroboscopic and other features to model, and uses SVM classifier for classification. Only because dynamic texture analysis is used in the candidate area, the computational cost is relatively large. Geng Qingtian[5] and others used RGB model and YCbCr model to develop new color recognition rules to establish a flame detection model, reducing the interference caused by lighting background changes, and showing strong robustness under unfavorable lighting backgrounds. The accuracy of the manually designed feature detection model is highly dependent on the manually selected feature parameters, and there are certain limitations in application scenarios, detection accuracy and detection rate. In recent years, deep
learning theory has made breakthrough progress in target detection and other fields. Therefore, flame detection technology that utilizes the powerful characterization and modeling capabilities of feature learning algorithms has important research value and application prospects. At present, scholars have used deep learning technology to recognize flame images. Hui Tian[6] et al. proposed a flame recognition method based on Fast R-CNN, using a pre-trained convolutional neural network AlexNet model as a feature extraction network to automatically extract features, and realize classification of small samples through migration learning, which has good detection effects and strong anti-interference ability of the CNN model, but the CNN model has a higher feature dimension, which reduces the generalization ability of the classifier. Szegedy[7] et al. proposed a flame detection architecture based on the GoogleNet network, and used migration learning for fine-tuning. This method achieved a certain balance between detection accuracy and efficiency, but still has a high false alarm rate. With the further improvement of hardware computing power, more deep convolutional neural network models are applied to the field of target detection, which are mainly divided into two-step target detection algorithms based on R-CNN[8-9] and single-step target detection algorithms based on YOLO and SSD networks. Step target detection algorithm. The former uses the region suggestion method to generate a series of sample candidate frames, and classifies the samples through the convolutional neural network; the latter merges the two tasks of extracting candidate frames and classification into one network, and converts the border positioning problem into a regression problem for processing. The two methods have differences in the candidate frame generation strategy, which leads to differences in their performance. The former is superior in detection accuracy and positioning accuracy, and the latter is superior in detection rate. The YOLOv3 algorithm uses a network structure Darknet53 with better classification effect, and uses multi-scale features for prediction, which improves the recognition rate of small targets, maintains the speed advantage and improves detection accuracy.

Flame image detection has high requirements for timeliness. Therefore, under the premise of giving priority to the detection rate, this paper integrates the detection accuracy of each network model and proposes an improved flame detection model based on the YOLOv3 network. First obtain candidate frames through K-means clustering, and use multi-scale feature fusion to improve the detection accuracy of the model, and then replace the YOLOv3 backbone network with MobileNet according to the principle of deep network model compression. Under the premise of ensuring accuracy, effectively improve the large and complex calculations of the deep network model, and improve the execution speed of the algorithm.

2. Traditional YOLOv3 Target Detection Algorithm

2.1. Feature Extraction Network
The performance of target detection algorithm is closely related to the basic feature extraction network. YOLOv3 uses the skeleton network Darknet53 as the feature extraction network, which contains 53 convolutional layers. YOLOv2 changes the tensor size in the forward process through 5 maximum pooling, and in the basic network Darknet53 of YOLOv3, it is performed by increasing the step size of the convolution kernel 5 times. On the one hand, YOLOv3 adopts such a fully convolutional structure, and on the other hand, it introduces a residual neural network structure, which greatly reduces the difficulty of training the network, thereby improving the accuracy of network classification.

2.2. Inspection Process
YOLOv3 trains the network in a regression method, introduces the anchor mechanism to generate three sets of a priori boxes at three different scales, and the size of the subsequent bounding boxes will be fine-tuned based on the 9 a priori boxes. Input 416×416 image, perform feature extraction through the basic network, and input FPN structure. Finally, 3 scale feature maps are generated as predictions, which are 13×13, 26×26 and 52×52 respectively. Divide these feature maps into grid areas, predict three bounding boxes on each grid, and generate a total of (13×13+26×26+52×52)×3=10647 bounding boxes. Each bounding box predicts four coordinates: $t_x, t_y, t_w, t_h$. The distance from the target grid to
the upper left corner of the image is \((c_x, c_y)\) and its corresponding bounding box and height are \(p_w, p_h\). The corresponding prediction relationship is as follows:

\[
b_x = \sigma(t_x) + c_x, \quad b_y = \sigma(t_y) + c_y, \quad b_w = p_w e^{t_w}, \quad b_h = p_h e^{t_h}
\]

Each grid also predicts the probability of the object in the prediction box \(P_r(\text{Object})\), and scores the prediction box by formula

\[
\text{Conf}(\text{Object}) = P_r(\text{Object}) \times \text{IOU}_{\text{pred}}^{\text{truth}}
\]

Where \(\text{IOU}_{\text{pred}}^{\text{truth}}\) is the intersection ratio between the prediction box and ground truth, and \(\text{Conf}(\text{Object})\) is the confidence level. When there is a target in the grid, \(P_r(\text{Object})=1\), otherwise it is 0. The prediction box gets the final result after non-maximum suppression.

3. Improved YOLOv3 Algorithm

3.1. Dimensional Clustering Candidate Box Improvement

Flame, as a non-rigid body, varies greatly in its morphology during combustion and has different length-width ratios. And the Anchor a priori box of YOLOv3 is obtained based on the clustering of the VOC data set. Among the 20 types of targets, as large as bicycle, bus, as small as bird, cat, the size of the target varies greatly, and the obtained Anchor dimension is not universal. Therefore, it is necessary to select a set of a priori boxes suitable for the flame data set. Based on this, K-means clustering is used to cluster the width and height of the target frame in the FireDatasets dataset, with the number of clusters \(k\) as the number of a priori frames, and the width and height of the \(K\) clustering center frames as the a priori frame (Anchor) width and height. The average degree of overlap (Average IOU) is used as the measure of target cluster analysis to perform cluster analysis on the data set. The clustered Average IOU objective function \(f\) can be expressed as

\[
f = \arg \max \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{\text{IOU}(B, C)}{n}
\]

Where: \(B\) represents the sample, which is the target in the ground truth; \(C\) represents the center of the cluster; \(n\) represents the total number of samples; \(nk\) represents the number of samples in the \(k\)-th cluster center; \(\text{IOU}(B, C)\) represents the intersection of the center box of the cluster and the cluster box; \(i\) represents the sample number; \(j\) represents the sample number in the cluster center.

Select \(k=1\sim9\), and perform cluster analysis on the data set samples respectively, and the relationship between the number of a priori boxes \(K\) and Average IOU can be obtained as shown in Figure 1. With the increase of \(k\) value, the curve change tends to be stable, and the changing inflection point can be considered as the optimal number of anchor boxes. When \(k>6\), the curve tends to be stable, so the number of a priori boxes is selected as 6, which can speed up the convergence of the loss function and eliminate the error caused by the candidate boxes. The width and height of the cluster centers obtained in the Fire dataset are \((67,119), (135,164), (226,184), (385,287), (625,214), (824,607)\), respectively.
Figure 1. K-means dimensional clustering results.

3.2. Network Structure Improvement

This paper uses the characteristics of the MobileNet network to solve the standard convolution into deep convolution and point-wise convolution to reduce the amount of parameters, and replaces the YOLOv3 backbone network framework, using multi-scale fusion to form a MobileNet-based target detection algorithm.

3.2.1. MobileNet algorithms. MobileNet is a small and efficient CNN model proposed by Google. The basic unit of MobileNet is deep-level separable convolution, which can be decomposed into deep convolution and point-wise convolution, as shown in Figure 2.

Assuming the input feature map size is \(D_F \cdot D_F \cdot M\), the output feature map size is \(D_F \cdot D_F \cdot N\), where \(D_F\) is the width and height of the feature map. For the standard convolution \(D_K \cdot D_K\), the calculation amount is:

\[
D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F
\]

(4)

Compared with the standard convolution, the calculation amount of the depth separable convolution is reduced:

\[
\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2}
\]

(5)
3.2.2. Flame target detection model based on MobileNet. The basic components of the MobileNet network are deep-level separable convolutions, except for the fully connected layer, batchnorm is added after all layers, and the ReLU activation function is used, and finally the softmax classifier is used for classification. The entire network has 28 layers. In order to improve the target detection rate, this paper combines the convolutional neural network model compression method, combines MobileNet and YOLOv3 detection model, replaces the latter's backbone network with MobileNet, while retaining the YOLOv3 multi-scale prediction, and then obtains a lightweight detection model.

Figure 3. M_YOLOv3 network structure diagram.

4. Flame Target Detection Experiment

4.1. Experiment Environment
This paper uses the deep learning framework Keras to build an experimental operating environment, and combines the CUDA environment to perform GPU parallel acceleration operations.

4.2. Evaluation Index
This paper selects mean Average Precision (mAP) and single image detection time (ms) as the evaluation indicators. The average category accuracy can be obtained from the PR (precision-recall curve, PRC) curve. P represents Precision, which is accuracy, R Indicates Recall is the recall rate, and the calculation formula is as follows.

\[
P = \frac{TP}{TP + FP} \quad (6)
\]

\[
R = \frac{TP}{TP + FN} \quad (7)
\]

Among them, \(TP\) is a real case, \(FP\) is a false positive case, and \(FN\) is a negative positive case. P-R curve is drawn with precision and recall rate on Y axis and X axis.

4.3. Test Results and Analysis
In this paper, different target detection methods are trained and tested on the Fire dataset. The detection results are shown in Table 1. Based on the test results in Table 1, it can be seen that: 1. Compared with other typical target detection algorithms, the method M_YOLOv3 proposed in this paper adopts a multi-scale fusion strategy to improve the ability to express multi-scale target features, and achieves the best average detection accuracy under different IoU thresholds. The proposed method has a better detection effect for multi-scale targets; 2. The algorithm in this paper uses the principle of MobileNet's deep separable convolution to compress the network model, and decomposes the standard convolution into deep convolution and point-by-point convolution, which effectively reduces the amount of model parameters. The detection time of a single image of the improved model is 32.3ms, which has obvious speed advantages compared with other target detection algorithms.
Table 1. Comparison of results of different target detection methods.

| Detection method | Basic network | AP0.5/% | AP0.75/% | Detection time/ms |
|------------------|---------------|---------|----------|------------------|
| YOLOv3           | Darknet53     | 76.25   | 70.95    | 60.2             |
| Fast R-CNN       | VGG16         | 88.03   | 83.56    | 48.6             |
| SSD              | VGG16         | 80.70   | 57.36    | 41.3             |
| M_YOLOv3         | MobileNet     | 97.34   | 95.29    | 32.3             |

Figure 4. P-R curves of different models.

Figure 5. Comparison of detection results of different frameworks.

The PR curves of flame target detection results corresponding to different detection methods are shown in Figure 4. The detection framework based on MobileNet proposed in this paper is superior to other methods in terms of accuracy and recall rate. It can be seen from Figure 5 that compared to the SSD and Fast R-CNN target detection framework, the MobileNet-based M_YOLOv3 network reduces
the false detection rate of the target, and effectively improves the detection ability of small targets and multiple targets in the image.

5. Conclusion
In order to improve the real-time performance and robustness of flame detection, this paper uses the MobileNet network model to replace the YOLOv3 backbone network, which effectively reduces the amount of model parameters and improves the model detection rate. And use K-means clustering to generate Anchor Boxes at different scales, and use multi-scale feature fusion to fuse high-level and low-level feature semantic information to further improve the detection accuracy of the model. The experimental results show that the improved model has obvious advantages in detection accuracy and detection speed compared with other classic models, and achieves the optimal detection performance for multi-scale flame targets in different scenarios.

6. References
[1] Wang Bo. Research on flame image extraction and recognition technology based on machine learning[D]. Tianjin University, 2018.
[2] Wang Haizhen, Lian Zuozheng, Teng Yanping. A fire image recognition algorithm based on niche technology[J]. Computer Engineering and Science, 2010, 32(11): 75-78.
[3] Tan Yong, Xie Linbai, Feng Hongwei, Peng Li, Zhang Zhengdao. Image-based flame detection algorithm[J]. Progress in Laser and Optoelectronics, 2019, 56(16): 115-121.
[4] Dimitropoulos K, Gunay O, Kose K, et al. Video-based flame detection for the protection of Cultural Heritage[J]. International journal of heritage in the digital era, 2013, 2(1): 23-47.
[5] Geng Qingtian, Yu Fanhua, Zhao Hongwei, Wang Chuang. A new flame detection algorithm based on color features[J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(06): 1787-1792.
[6] Huitian, Halidan Abudureyimu, Du Han. Multi-type flame detection combined with Faster R-CNN[J]. Journal of Image and Graphics, 2019, 24(01): 73-83.
[7] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
[8] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.