Autonomous garbage detection for intelligent urban management

Ying Wang[1] and Xu Zhang [1,a]

[1]*School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China*

Abstract. With the development of smart city in major cities at home and abroad, especially the management of smart city, how to improve the intelligence level of urban environment monitoring and evaluation has become an important research topic. It is of great value to rapidly and accurately detect garbage from urban images in the application of intelligent urban management. This paper aims to adopt a deep learning strategy for automatic garbage detection. By training a Faster R-CNN open source framework with region proposal network and ResNet network algorithm, we look over garbage detection results on garbage images. In addition, to improve the accuracy of the method, a data fusion and augmentation strategy is proposed. As a result, experiments show that the method has favorable generalization ability and high-precision detection function.

1 Introduction

Monitoring and cleanliness assessment of garbage area in urban scenes mainly rely on manual inspection and photographic record, which makes it a difficult and time consuming task [1]. During the inspection process, human intervention and cumbersome problems often happen. The quality of sanitation work has been affected. Different from pedestrians, vehicles and other objects, garbage have no relatively clear definition. Due to the judgment of garbage always has certain subjectivity, in different situations, it will produce different judgment results. Since the diversity of scenes where garbage appears, accuracy of test results will be affected. With the development of smart city, we expect to provide an automatic detection method of urban garbage to help alleviate urban garbage problems.

2 Motivation

Before the development of deep neural networks, features were manually designed [2], then followed by a classifier. Some research focused on the classification and recycling of garbage a few years ago [3]. For example, Sudha S *et al*. proposed a model for classifying objects as biodegradable and non-biodegradable [4]. Although the traditional object detection already has some mature techniques, due to the morphological diversity, illumination diversity, background diversity and other factors of the target object, the detection precision for the unfixed form objects such as urban garbage is still a tough problem to solve.

The past decade has witnessed a rapid development of massive data and high-performance computing systems such as graphics processing units (GPUs).Now region-based CNN detection methods have dominated many tasks of computer vision. It is such an exciting area that can extract the high-level features and the hierarchical feature representations of the objects [5]. Girshick *et al*. introduced a region-based CNN (RCNN) for object detection [6], from 2014 to now, R-CNN, Fast R-CNN, Faster R-CNN, ION, HyperNet, SDP-CRC, YOLO, G-CNN, SSD and other increasingly fast and accurate object detection methods have emerged.

There is very little research on garbage detection at home and abroad now. In the first, Mittal proposed a GarbNet network to complete the classification of garbage and non-spam images, achieving a mean accuracy of 87.69% on the GINI dataset [7]. In [8], they developed GoogLeNet [9] as its classification architecture, which was similar to OverFeat model, to detect the target objects such as cigarette butts and leaves from a height of several meters.

The major contributions of this paper are presented as follows.

1) We develop a Faster R-CNN open source framework with region proposal network and ResNet network algorithm, using ResNet network to replace the previous VGG network as the basic convolution layers.
2) To optimize the performance of the model, we collect urban scene images containing garbage and urban scene images without garbage. By using fine-tuning strategy, we apply the pre-training model parameters which has been trained in coco dataset to our network.
3) We propose a dataset fusion strategy, which integrates the garbage dataset with several other datasets of typical categories in urban scenes.

In summary, the method has near-real time and generalization capabilities. Through experiments, we observe that the false detection rate of the garbage area has been significantly improved, and the recognition accuracy

[a] Corresponding author: xuzhang@shu.edu.cn

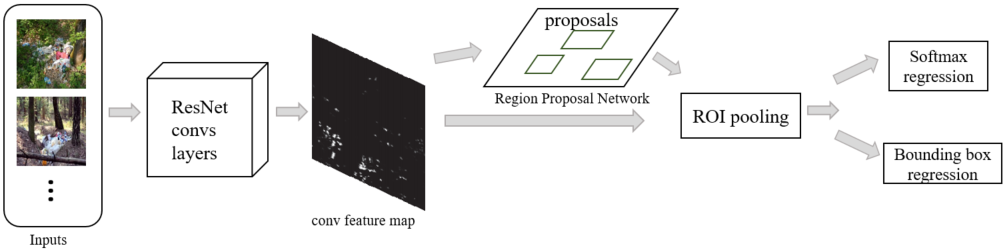

Figure 1. Overall network architecture.

of small areas has been improved.

3 Detection network architecture

Here, we briefly introduce our deep learning strategy.We firstly use the ResNet network as the basic network and apply it to Faster R-CNN framework. The ResNet network proposed by He Kaiming et al. in [10], which can avoid the simple stacking of convolutional neural network gradient disappearance or explosion and precision degradation. Here, ResNet network is used as the conv layers to extract features from the original image to obtain the feature map. Faster R-CNN solves the speed bottleneck of R-CNN and Fast R-CNN [11], further improves network performance. Faster R-CNN object detection method includes four basic steps, including region proposal generation, feature extraction, classification and location optimization.

The specific implementation process of the network is shown in Figure 1. During training, we input the dataset image firstly, through the shared convolution layer of ResNet, generating the feature map. Then RPN layer gets the output and generate a large number of region proposals.

3.1 Region Proposal Networks

Region Proposal Network (RPN) is modeled with a fully convolutional network, which takes images as input and finally outputs a set of bounding box proposals. To achieve these proposals, at each sliding-window location, we predict k proposal boxes at the same time, so the reg layer has 4k outputs, that is, the coordinate encoding of k boxes [12]. The cls layer outputs 2k scores, which is the estimated probability that each proposal box is a target or a non-target. The k reference boxes are parameterized by the corresponding k boxes called anchors. Each anchor is centered on the center of the current sliding window and corresponds to specific scale and aspect ratio. Following the default setting of the network, we use 3 scales and 3 aspect ratios, so that there are 9 anchors in each sliding position.

IoU is defined as the degree of coincidence between the bounding box predicted by the system and the ground-truth box marked in the original image. The anchor is associated with the highest IoU value whether it is a positive or negative label. When the probability of the garbage region marked in the experiment is higher than 0.7 or overlaps with a ground truth (GT) bounding box with the highest IoU value(possibly less than 0.7), we assign positive labels to them [10]. If the IoU value is less than 0.3 for all GT bounding boxes, then we identify them as negative labels. After setting the label, RPN calculates the loss for the labeled region detection. Following the multi-task loss definition [12], the loss function for an image in Faster R-CNN is defined as,

$$L(\{p_i\},\{t_i\})=\frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*) \quad (1)$$

Where i is an index of the anchor in the mini-batch, and p_i is the predicted probability of anchor i being an object. If the anchor is positive, the GT label p_i^* is 1, otherwise is negative, p_i^* is set to 0. t_i is a vector representing the four parameterized coordinates of the predicted bounding box, and t_i^* is the coordinate vector of the GT bounding box corresponding to the positive anchor. Reference [13], L_{cls} and L_{reg} represent classification losses and regression losses respectively. They are normalized by N_{cls} and N_{reg} and weighted by a balancing parameter λ.

The ROI pooling layer has two inputs, one is the feature map obtained through the conv layers, the other is an RPN output that represents a matrix of N*5 of candidate rectangular frames. Where N is the number of ROIs, the first column represents the image index [14], and the remaining four columns represent the remaining top left and bottom right coordinates. Map the coordinates in the ROI to the feature map, and divide the mapped area into sections of the same size (the number of sections is the same as the output dimension). After obtaining the coordinates on the feature map, we use a pooling layer to obtain a uniform output. The biggest benefit of ROI pooling is that it greatly increases processing speed.

3.2 Fine-tuning strategy

The mainstream machine learning method in the current vision field is deep learning methods based on high performance parallel computing and massive training data. However, due to the high complexity of the deep learning model, it is very easy to over-fitting, we need collect massive data and run on the GPU for several weeks to solve the problem [12]. These conditions are difficult for most of us to achieve. In order to be able to train with a smaller dataset in a more practical way, we use a technique called transfer learning or fine-tuning to assign a trained network to the new datasets.

We can transfer weights from a pre-trained architecture to fine-tune our network architecture. In this paper, we initial weights of the model by ResNet pre-trained results on the coco dataset to solve the problem of insufficient training datasets.

3.3 Data fusion and augmentation

Data collection is the most important preparation for object detection. Our experiment essentially solves a two-class problem, that is, whether the area contains garbage. Considering the complexity and diversity of urban scenes, we collect non-garbage urban scene classes that contain buildings, neat roads and neat lawn.

On the other hand, garbage does not have a certain shape, when using computer detection strategy, some object classes in the urban scene can match the garbage to similar attributes easily, which causes error detection and reduces the accuracy of the detection. In order to improve the generalization ability of the model, we consider using data fusion strategy to increase the background classes. We set pedestrians, vehicles, buildings, neat roads and neat lawn as background class, so that the algorithm can more effectively distinguish between garbage and other object classes.

4 Experiments and analysis

In this paper, we constructed a garbage dataset using a large amount of urban garbage images dataset obtained from the sanitation department. Afterwards we merged with background classes such as non-garbage urban scene images datasets and other categories object datasets as shown in Figure 2. In the test results, the background classes are not displayed, only the garbage class is displayed, thereby improving the robustness of the algorithm.

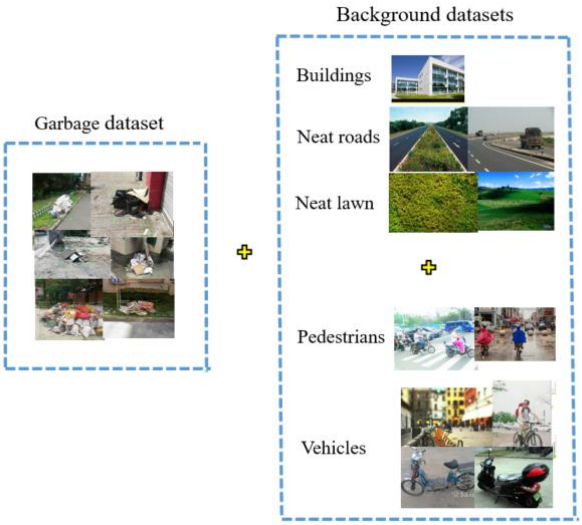

Figure 2. Data fusion and augmentation.

4.1 Description of the dataset

We unified all image format into jpg and named them in consecutive numbers. Both the training picture and the test picture are placed in a named directory. The garbage dataset of this experiment contained 816 images, 596 among them were train images and 220 were test images. These images were photographed from different scenes and different directions, which could detect the diversity of images effectively. In addition, we used the data fusion strategy and per background class consists of at least 50 images.

4.2 Framework Settings

We train the garbage detection model based on a pre-trained COCO model, we run the model 7200 iterations with a batch size of 5, the weight attenuation of 0.0005 and a learning rate is 0.0001. The size of images was not fixed in the model. During the training, we visualized the change of the loss function in the tensorboard via a browser, as shown in Figure 3. Machine learning library "Tensorflow" [15] was used to create the network.

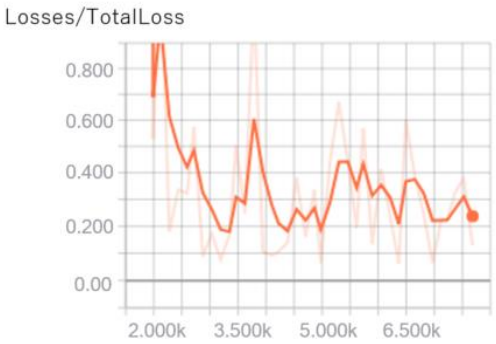

Figure 3. Loss function.

To verify the detection accuracy of the model, IoU is set to 0.7, that is, the area below the threshold in the image is not marked. In the statistics of experimental data, to avoid over-fitting, the threshold cannot be set too high, so after testing, a good experimental result can be obtained when the threshold is set to 0.7. For garbage region in the image, the number on the green box indicates the probability of the target object. The higher the score is, the more the system can determine that the target region is the garbage object.

4.3 Evaluation of the detection

We employ the evaluation criteria that the object detection needs to involve location information of the object in the image and classify the object. AP (Average Precision) is the average accuracy rate, which is a performance indicator that reflects the overall situation. The algorithm of this paper achieves the speed of 6 seconds per image on GTX750ti GPU.The detection accuracy of the method used in this paper reaches 0.89, and the accuracy is defined as

$$AP=\frac{TP}{TP+FP} \quad (2)$$

Where TP indicates the number of positive labels, FP indicates the total number of non-garbage object that was detected as garbage and missed detection.

4.4 Qualitative assessment

Urban garbage has extremely diversity in color texture and geometric form. For computer recognition, many things in urban scenes have characteristics similar to garbage, so misdetection happens. It is also observed that when these objects and garbage appear at the same time, false detection often occurs. For example, misdetections of pedestrian and vehicles are shown in Figure 4.

(a)

(b)

Figure 4. Examples of misdetection phenomenon in the experiment. (a) Error detection of pedestrian. (b) Error detection of vehicles.

Next, we propose an image fusion strategy, and Figure 5 shows the test results after the improvement. We visualize the test results, mark the objects of the corresponding category with a rectangular box and the specific parameters.

Figure 5. Test results after using image fusion strategy.

4.5 Comparison of experiments

In order to evaluate the effectiveness of the proposed algorithm, some urban garbage scene images are collected to compare and evaluate the detection results of this paper and the existing urban garbage detection methods. The comparison test is an algorithm proposed by Mittal, and the test data sets all use images that are not sent to the training. In the three groups of Figure 6, the left half is the result images detected the algorithm proposed by Mittal [7]. This method has limited optimization for the target detection frame. From the result picture, it can be seen that the detected target region is limited, and the small region target object is missed. By contrast, the right half of Figure 6 shows result using our algorithm for detection, which could detect small region objects.

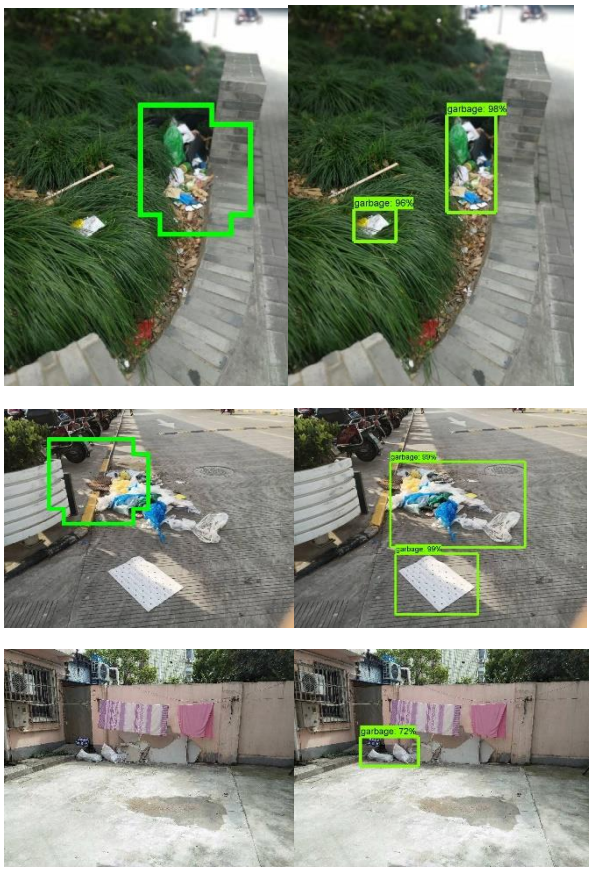

Figure 6. Comparisons of object detection results using Mittal's algorithm and our algorithm for garbage detection.

5 Conclusion

Based on the Faster R-CNN object detection framework, we present a way of using the ResNet network algorithm as the convolutions layers, which improves the accuracy of object detection and location. We achieve the experiment results as expected, the network demonstrates its efficient generalization ability when the small region objects occur. Our data fusion strategy overcomes region misdetection problem. Finally, the near-real time and high-precision detection of garbage in urban scenes is realized, which has high practical value. It remains an open challenge to further reduce the detection time with the aim of rapidly and high precision detection.

Acknowledgment

This research was partially supported by the National Nature Science Foundation of China (Grant no. 51575332 and no. 61673252).The authors are grateful to the participants helped to provide urban scenes datasets and annotate the images.

References

1. Hoornweg, Daniel, BhadaTata, et al. Environment: Waste production must peak this century, Nature, **502(7473)**, 615-7 (2013)
2. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, et al. Backpropagation applied to handwritten zip code recognition. Neural computation, **vol.1**,**no.4**, pp. 541–551 (1989).
3. Brinez L J C, Rengifo A, Escobar M. Automatic waste classification using computer vision as an application in colombian high schools, Networked and Electronic Media. IET, **10 (5)-10 (5)** (2017)
4. Sudha S, Vidhyalakshmi M, Pavithra K, et al. An automatic classification method for environment: Friendly waste segregation using deep learning, Technological Innovations in ICT for Agriculture and Rural Development. IEEE, 65-70 (2016)
5. Erhan D, Szegedy C, Toshev A, et al. Scalable Object Detection Using Deep Neural Networks, **3(4)**, 2155-2162 (2013)
6. R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, In *CVPR*, pages 580–587, (2014)
7. Mittal G, Yagnik K B, Garg M, et al. SpotGarbage:smartphone app to detect garbage using deep learning, ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 940-945 (2016)
8. Rad M S, Kaenel A V, Droux A, et al. A Computer Vision System to Localize and Classify Wastes on the Streets, International Conference on Computer Vision Systems. Springer, Cham, 195-204 (2017)
9. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions, 1-9 (2014)
10. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition, 770-778 (2015)
11. Girshick, Ross. Fast R-CNN. *Computer Science* (2015)
12. Ren, Shaoqing, et al. "Faster R-CNN: towards real-time object detection with region proposal networks." *International Conference on Neural Information Processing Systems* MIT Press, 91-99 (2015)
13. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 248-255 (2009)
14. Sermanet P, Eigen D, Zhang X, et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, Eprint Arxiv, (2013)
15. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning, (2016)