参赛队员姓名：**杨学霖**

中学：**华南师范大学附属中学**

省份：**广东省**

国家/地区：**中国**

指导教师姓名：**杨晓安**

报告标题：**Faster R-CNN over Attention: Shared Bikes Detection in Surveillance Video**

Faster R-CNN over Attention: Shared Bikes Detection in Surveillance Video

Yang Xuelin
the Affiliated High School of South China Normal University

Abstract

With the rapid development of bicycle-sharing system since 2016, illegal parking has become one of the most significant social issues. While the existing management methods cannot meet the demand, object detection technology provides a possibility. The Video Surveillance System can be utilized as a rich source of information. However, there is technical challenge to handle complex scenarios. Faster R-CNN, chosen because of its excellent overall performance, cannot specify the proposals of shared bikes especially in distant view. Processing the consecutive images in videos is neither necessary nor efficient. The original results are limited by interference of moving bikes and blocking effects.

In this paper, I introduce a *Faster R-CNN over Attention* (FoA) that combines the from-coarse-to-fine visual mechanism with deep neural networks. FoA realizes fine-grained detection of shared bikes in complex circumstances. It consists of an *Attention Region Extraction* (ARE) and an optimized Faster R-CNN. In ARE, I introduce the concept of *Attention Region* and define it as video background, then apply Gaussian Mixture Model. ARE processes the surveillance video into discrete regions, allowing concentration on certain areas and efficiency in detection. In Faster R-CNN, I put forward an *anchor box optimization* in regards of the remarkable region-based characteristic in R-CNN. The optimization applies k-means to refine the anchor boxes and generate more appropriate region proposals.

I construct a categorized shared bike data set of 4,291 images and 12,697 labeled objects for FoA training and testing. FoA performs excellently with the detection and recognition ratios of 94.19% and 92.73% respectively. The optimization is proved to make a difference.

The paper not only provides a new direction for solving illegal parking issue and managing bicycle-sharing system, but also generalizes FoA into a new detection framework of *Attention Region Extraction plus Region-Based Convolutional Neural Network*, which can be applied to certain object detection in various scenes. The mechanism that combines ARE and R-CNN organically allows the towards-real-time detection to possess both practicality and universality. FoA is one of its robust implements.

Keywords: Shared bikes, object detection, Attention Region, Gaussian Mixture Model, Faster R-CNN, clustering

Contents

1 Research Background

1.1 Bicycle-Sharing System

Bicycle-sharing is a new form of sharing economy[1] originated in 2016. By applying mobile Internet technology, the dockless shared bikes provides an easier, cheaper and more convenient way of transportation. It soon became the focus of the public and investors nationwide and grew exponentially in 2017. According to some statistics, until July, 2017, there are about 70 bike-sharing companies, more than 16 million bikes, 130 million registers and 1.5 billion service use in total[2]. Users of mobike and ofo are in highest proportion. It is predicted that bike-sharing will continue gaining momentum in the next 3 years[3].

The gigantic scale, numerous users and discrete bike-sharing platforms cause big management concern. Illegal parking is one of the most severe problems above all. It limits the development of bicycle-sharing. Neither the rules nor the parking areas have been set in most of the cities in China. The shared bikes are often placed randomly, clogging bus stations, subway entrances, emergency exits, pavements and even motorways, which brings tremendous potential danger. Some users, on the other hand, park the bikes in small streets, communities or close to their house. Others throw away the faulted bikes deliberately. The discrete distribution causes difficulties in managing and converging shared bikes for companies and governments.

Figure 1: Some scenes of illegal parking. **Left:** Encirclement of shared bikes in Beijing Bawangfen bus station (Visual China). **Right:** Misplaced shared bikes on streets (Wuhan Evening Paper).

As for this problem, the existing management methods are still limited.The common one is human inspection. For example, the shared bike operators hire people to go around on the roads every day, which requires at least 50 people every 10 thousand bikes. Some set up a “credit system” to encourage user tip-offs. Some citizens act as “bike hunters” spontaneously to maintain the order of bike-sharing system by reporting troubles and correcting the misplaced bikes. However, searching is a continuous process lack of destination. Human inspection has a heavy workload. And it is not easy to manage or find the discrete bikes away from the main roads and public areas either.[4~6]

Electronic fence is another pilot technology. In the official publication of ten departments in August, 2017 in China, it is encouraged to apply new technology such as electronic fence to provide convenience. As a result, a number of cities and companies are speeding up the study and application of electronic fence. The technology can be utilized in two ways[7]. The first way, pilot in Beijing, is based on Global Position System (GPS). However, it is limited by the precision of GPS and power consumption of communication. The second way, pilot in

Shanghai etc., is based on bluetooth technology. It requires parking points and shared bikes to install bluetooth devices and provide associated network and power support. The method increases modification and maintenance cost, and may lead to municipal or construction drawbacks.

To summarize, introducing information technologies to the management of shared bikes is not only a significant supplement of current measures but also a trend in the future. Facing the problem of illegal parking, it is essential and urgent to develop a new feasible method and solution.

1.2 Research Status Review

1.2.1 Convolutional Neural Network

The study on Convolutional Neural Network (CNN) has been conducted in depth with the rapid development of deep learning. CNN, as one of the efficient detection method, is a structure of neural network applied in various fields including pattern recognition, image classification, etc. It has become of the attention in recent years.

In 1960s, Hubel and Wiesel came up with the concept of "Receptive Field[10]" based on the study of the cerebral cortex of organisms, and discovered the mechanism of information classification. Inspired by these works, Fukushima put forward "Neocognition[11]" in 1980, which could be considered the prototype of CNN. It was the first artificial neural network based on the hierarchical structure and local connectivity of neurons. Since that, researchers have begun using an artificial neural network called "Multilayer Perception[12]" to take the place of manual feature extraction and training the network by gradient descend method. They introduced Back Propagation (BP) to calculate error gradients, which was proved efficient later.

In 1990, Lecun constructed the classical model of CNN[13]. He then optimized CNN and introduced LeNet-5[14] as a multilayer neural network for handwritten digit recognition. LeNet-5 has the same structure as classical networks, using BP[15] for training. By simulating cerebral cortex and extracting typical features, it requires no image preprocessing and is superior to other algorithms at that time. During this period, additionally, Zhang came up with Shift-Invariant Artificial Neural Network (SIANN)[16] for digit recognition. However, due to the limitation of computing power and the lack of labeled data for network training, CNN did not give a excellent performance in some problems which are more complicated, losing the continuous focus of the researchers.

The amount of data and images increased exponentially with the development of Internet, and, at the same time, various public data sets got hand-classified. Network training was sped up by GPUs, giving the possibility of deeper networks.

In large scale image classification, Krizhevsky[17] developed AlexNet and got the first place in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by about 10% of recognition error lower than the second team which use traditional method. AlexNet is similar to LeNet-5, but with deeper layers. The success of AlexNet gave rises to more effective CNN models, such as ZFNet[18], VGGNet[19], GoogleNet[20] and ResNet[21]. They were getting deeper and deeper. For example, ResNet, the winner of ILSVRC 2015, has the error rate of 3.57%, lower than the rate of human eyes of 5.10%. The number of layers in ResNet is 152, about 7 times larger than in VGGNet and 20 times in AlexNet. Deeper networks can learn the

nonlinear relationship between input and output better and get more robust features. However, accompanied problems of difficult convergence and over-fitting need to be carefully considered and solved.

CNN also performs well in problems of computer vision. Nowadays, it saw heavy use in research areas including natural language processing, image recognition, etc.

1.2.2 Object Detection

Object detection refers to the process of locating objects and classifying a type of specific object from others. The traditional detection methods mainly use manual features, such as HOG[22], DPM,[23] SIFT[24], etc.

In 2014, Ross Girshick came up with Region-based Convolutional Neural Network (R-CNN). It firstly uses Selective Search, a segmentation algorithm, to extract region where objects possibly exist, then gets the features and classifies the image by Support Vector Machine (SVM). It works well on public detection data set PASCAL VOC. R-CNN is the beginning of the series. In this series, various methods such as SSD[25], SPP-Net[26], YOLO[27], etc. was developed in succession, improving the detection results.

In 2015, Ross Girshick introduced Fast R-CNN[28-29] to speed up the slow R-CNN. It only extract features once to access to all feature vectors and performances better in PASCAL VOC. However, due to the use of Selective Search based on CPU, computing region proposals occupies high proportion of total time, which influences its applicability to some extent.

In 2016, Faster R-CNN[30] was developed. It uses Region Proposal Network (RPN) instead of the original Selective Search when computing proposals, and gives an excellent overall performance in speed, efficiency and precision. Details of the algorithm are described in Section 3.3.

In the study of R-CNN series, the main direction is to improve the detection results by changing network construction or training methods.

As for the problem in this paper, technological gaps due to the complexity of scenarios still exist. The previous methods have strict requirements on scenes, such as high resolution and simple backgrounds, while the situation is quite complicated in reality. Due to impact factors (including lightning, weather, time, etc.) and interference of pedestrians, trees and passing cars, it is technically challenged to detect shared bikes under such complex circumstances.

1.3 Applications of Image Processing in City Management

In the process of constructing Safe City and Smart City recent years, governments have built up a strong protection network to secure the whole cities, spreading security cameras over almost every roads and streets. The videos have rich information, high quality and availability. Take Guangdong province as an example. About 2.8 million collection points of public surveillance video, which cover main public areas and industries, have been installed from 2011 to 2015. The resolution, frame rate, online rate and online integrity of the 24 hours' surveillance videos are over 1080P, 25 fps, 95% and 98% respectively[8,9]. It is planned that the coverage of main public areas and industries will reach 100%[3] by 2050. So will the proportion of new high-definition cameras.

The Video Surveillance System provides an effective and scientific way of city

management. Image processing technology is used for extracting valid information, depending on different requirements and situations. Face recognition, vehicle recognition and video tag are several successful representatives.

To solve the problem of illegal parking, detecting shared bikes is the primary and pivotal step. Applying image processing technology for the automatic detection of illegal parking can greatly reduce the workload of human and video inspection. The Video Surveillance System can be utilized as a rich source of information. So I tried to conduct my study at this entry point.

Applications of image processing in city management are often general object detection, in which bikes are considered as a whole type instead of being classified into different brands. On the other hand those detection methods have strict requirements on scenes, while the situation here is quite complicated in reality. To summarize, there has been no study on fine-grained shared bike detection under complex circumstances based on CNN.

2 Purpose

This paper is to achieve the following 3 goals:

(1) Provide a new application of Video Surveillance System. Expand the field of resource utility.

(2) Design a new method for illegal parking problem. Apply computer vision technology to automatic detection of shared bikes.

(3) Take shared bike detection as an example to discuss a generalization of specific object detection.

3 Shared Bikes Detection in Surveillance Video

3.1 Framework

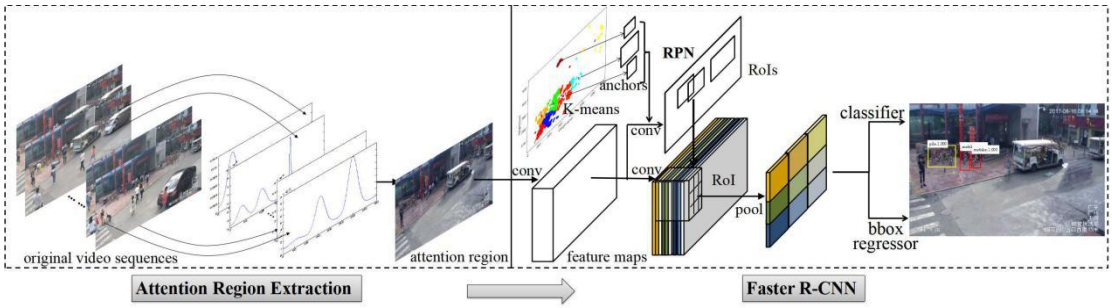


Figure 3-1: Faster R-CNN over Attention (FoA). In this paper, I introduce a FoA that combines the from-coarse-to-fine visual mechanism with deep learning neural networks. I introduce the concept of *Attention Region*. I also put forward an *anchor box optimization* using k-means in regards of the characteristics of them in reality. Based on gradual region extraction, this algorithm realizes the fine-grained detection of shared bikes under complex circumstances.

It consists of an *Attention Region Extraction* (ARE) and an optimized Faster R-CNN. Firstly, in ARE, using video frame as input, the area where objects possibly appear, defined as Attention Region, is computed by Gaussian Mixture Model iteratively.

Secondly, in Faster R-CNN, it generates feature maps (from multiple convolution and

pooling layers) and region proposals (from RPN) afterwards. Classification and border regression are achieved by a detection network in Faster R-CNN. The algorithm applies k-means to cluster ratio and scale indices as an optimization. The network ultimately access to the categories and location information of shared bikes.

FoA can also be generalized for detection of certain object in various scenes, discussed in Chapter 5.

3.2 Attention Region Extraction

Although Faster R-CNN has excellent overall performance in general object detection, it cannot be simply applied and fulfill the goal of specific object detection, shared bike detection in this paper.

The input is surveillance video, which has mainly two ways to process. The first way is to use tracking technology in video processing, which tracks the movements and status of objects. The second way is to transform the videos into images, then apply the detection technology. This way is less complicated and more efficient. Moreover, the images are discrete, which means that the operations afterwards do not need to concentrate on the relations between two images. It can save the time and calculation resource.

Using images directly from screenshot as the input of neural network holds limitations and disadvantages. For example, the shared bikes may be blocked by moving objects. The moving bikes would also be detected. The complexity of scenarios gives rise to improvements on the previous detection methods.

Attention Region. Based on Attention Model (AM) and the visual mechanism, this paper introduces the concept of Attention Region. As the detected object, shared bikes are rested relatively and their positions and postures are unchanged in a period of time. The Attention Region is considered the background of the video.

Attention Model (AM) in deep learning is a resource distribution introduced from AM of human brain in cognitive psychology. It can be described that humans have a global view at a moment but will concentrate on a special part of their view, known as focus. Specifically in the videos, shared bikes are of attention.

In visual mechanism, humans will find the objects after locating the attention region. For example, when searching for a cup in a room, we will probably look at the table first because the cup may probably be there. The mechanism is from coarse to fine. When it comes to shared bikes, we would firstly search the areas such as roads and pavements instead of observing moving objects such as pedestrians or cars running by.

Gaussian Mixture Model. I use robust Gaussian Mixture Model (GMM) to extract the Attention Region, of which main steps are matching and renewing.

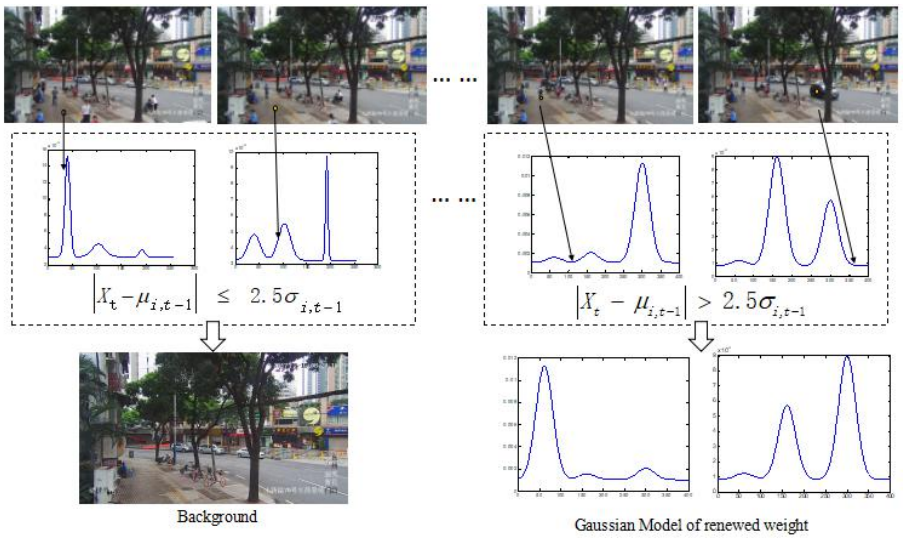


Figure 3-2: GMM. The key operations in GMM are matching and renewing. The following is the detailed process:

(I) Assume X_t is the value of a pixel at moment t, which satisfies Formula (3-1) of probability density in a period:

$$p(x_t) = \sum_{i=1}^{k} \omega_{i,t} \times \mathcal{N}(x_t, \mu_{i,t}, \sigma^2{}_{i,t}) \tag{3-1}$$

k is the number of Gaussian Distribution. $\mathcal{N}$ is the i[th] Gaussian Distribution at moment t, $\mu_{i,t}$ is its mean and $\sigma^2_{i,t}$ is the variance. $\omega_{i,t}$ is the weight of the i[th] Gaussian Distribution.

(II) For every images in the video frame, compare X_t to k weighted models. If their relation satisfies Formula (3-2), they are considered matched and the distribution index will be modified by Formula (3-3) to (3-5). Here $M_{i,t} = 1$. α is the renewal rate.

$$\left|X_t - \mu_{i,t-1}\right| \leq 2.5\sigma_{i,t-1} \tag{3-2}$$

$$\omega_{i,t} = (1-\alpha) \times \omega_{i,t-1} + \alpha \times M_{i,t} \tag{3-3}$$

$$\mu_{i,t} = (1-\rho) \times \mu_{i,t-1} + \rho \times X_t \tag{3-4}$$

$$\sigma^2_{i,t} = (1-\rho) \times \sigma^2_{i,t-1} + \rho \times \left(X_t - \mu_{i,t}\right)^2 \tag{3-5}$$

(III) If none of the present models satisfies Formula (3-2), the model of the least weight will be replaced, assigned a new mean, greater variance and smaller weight. Other models keep their means and variances. Their weights are computed by Formula (3-3). Here $M_{i,t} = 0$.

(IV) Every models then are arranged in a descending order of $\omega_{i,t} / \sigma_{i,t}$. If the top B models satisfy Formula (3-6), they are considered the background. T is the proportion of the

background.

$$B=\arg_b\min(\sum_{i=1}^{b}\omega_i > T) \tag{3-6}$$

3.3 Faster R-CNN

3.3.1 Introduction

Faster R-CNN[30] is a detection framework based on deep convolutional neural network. It was introduced in 2016 by Ross Girshick. It consists of a Region Proposal Network (RPN) and a detection network. In Faster R-CNN, the four fundamental steps (region proposal generation, feature extraction, classification and location refinement) are integrated in the network. The two networks share convolution layers and all calculations are completed in GPU without repetitions, which greatly improves the speed of detection.

3.3.2 Convolution, Pooling and Fully-Connected Layers

As a network based on CNN, Faster R-CNN contains networks of shared index, which can import original images and do not need preprocessing. The basic construction of Faster R-CNN includes convolution layer (*conv*), pooling layer (*pool*) and fully-connected layer (*fc*).

Convolution Layer. *Conv* layers are used for feature extraction, learning feature expressions of input images. Define x_i^l as the i[th] feature map in the l[th] layer, k_{ij}^l as the *conv* kernel between x_i^{l-1} and x_j^l, b_j^l as the bias term of x_j^l. x_i^l is calculated by Formula (3-7) below:

$$x_j^l = f(\sum_i x_i^{l-1} * k_{ij}^l + b_j^l) \tag{3-7}$$

$f(\bullet)$ is the activation function. Rectified Linear Unit (ReLU)[31] is chosen in this paper:

$$f(x) = \max(0, x) \tag{3-8}$$

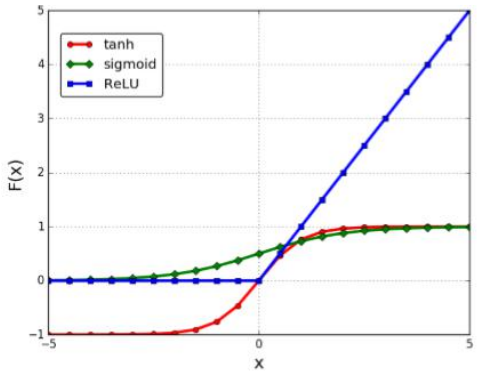


Figure 3-3: Activation functions[32]. Shown in Fig. 3-3, ReLU is chosen for certain reasons compared to others: It is more biologically plausible than widely used logistic sigmoid or hyperbolic tangent (tanh). While sigmoid and tanh have values near 0 varying gently, ReLU will not cause gradient dissipation as sigmoid does, and will be able to increase the speed of convergence. Its operations, including comparison, multiplication and addition only, reduce

the computation and do not require pre-training. It controls sparse outputs by providing zeros, which lead to non-overfitting CNN.

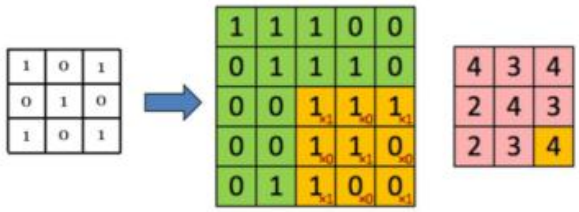


Figure 3-4: Conv operation[32]. Various kernels are for different features, shown in Fig. 3-3. Sliding the kernels on a image then adding the products up until all pixels are scanned. The output of conv is feature maps. The indices of kernels are pending. They can be trained to adapt to specific tasks.

Pooling layer. Pooling layers are often set after *conv* layers, reducing computations and indices while keeping prominent and effective information (or dominant features). They improve the generalization ability of the model and possess invariance. For each feature map x_k^l, the output of its pooling layer is:

$$x_k^{l+1} = pool(x_k^l) \tag{3-9}$$

$pool(\bullet)$ is the pooling function such as average-pooling[33] or max-pooling[34]. Max-pooling is chosen here. It computes the features in areas of certain scale by keeping the maximal number. Max-pooling has translation invariance.

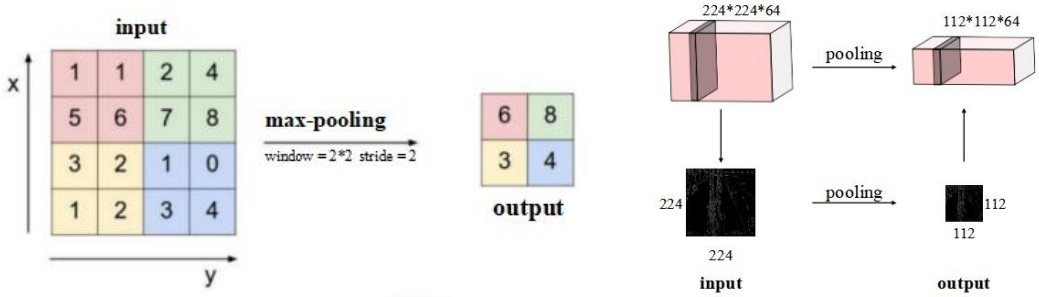


Figure 3-5: Max pooling[32]. **Left:** Operation. **Right:** Result.

Fully-Connected Layer. *Fc* layers are used for combining features and computing outputs. Formula (3-10) shows the process of connecting every neurons in *fc* to all neurons in the input layer. The symbols are of the same meanings as defined before.

$$x_j^l = f(\sum_i w_{ij}^l x_i^{l-1} + b_j^l) \tag{3-10}$$

When *fc* is the last layer, it acts as a classifier. Softmax is used as classifier in this paper, taking feature vectors from the previous layer as input and probabilities of every categories as output. The category of the highest probability is the predicted result of the network.

3.3.3 Region Proposal Network

In RPN, a sliding window is used to scan the feature maps. It is fully connected to a spatial window in feature map, and mapped to a low-dimensional vector. The vector is the input of classification (*cls*) and regression (*reg*) layers afterwards. Region proposals are generated by anchor boxes every step. The operation process has translation invariance. Regions are rectangles, the same as other object detection methods[35-37].

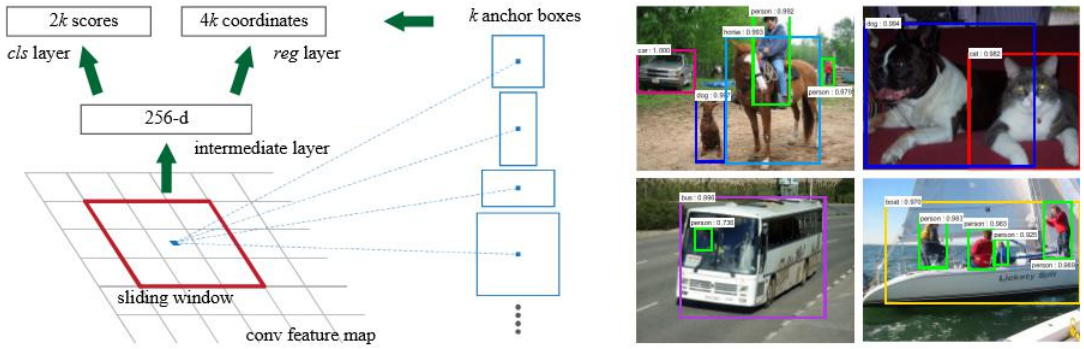


Figure 3-6: Left: Region Proposal Network (RPN). **Right:** Examples of general object detection using RPN proposals on PASCAL VOC 2007 test[30]. A sliding window, with k anchor boxes, predicts k region proposals at one time, transformed to a vector of 256 dimensions. The *cls* layer generates 4k coordinates of k region proposals while the *reg* layer generates 2k scores of whether the k proposals are object or background.

Anchor boxes and their operations holds translation invariance. They have scales and aspect ratios. Originally in Girshick's paper, k equals to 9. In regards of the features of shared bikes, k is set to 12 specifically. An optimization is embedded, discussed in Section 3.4.

3.3.4 Detection Network

Minimize the loss function following multi-task loss in Fast R-CNN. The loss function is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{3-11}$$

$$L_{reg}(t_i, t_i^*) = \sum_{i\in\{x,y,w,h\}} smooth(t_i - t_i^*) \tag{3-12}$$

$$smooth(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, |x| \geq 1 \end{cases} \tag{3-13}$$

$L_{cls}(p_i, p_i^*)$ is the classification loss and $L_{reg}(t_i, t_i^*)$ is the regression loss. L_{cls} is log loss over two classes (object and non-object). i is the index of an anchor. p_i is the predicted probability of anchor i being an object. The ground-truth label p_i^* =1 when the anchor has positive label; p_i^* =0 when it is negative. t_i is a vector representing 4

parameterized coordinates of the predicted bourndng box. t_i^* is a ground-truth box with positive label. The coordinates are parameterized following [38]:

$$t_x = (x - x_a)/w_a,\ t_y = (y - y_a)/h_a,\ t_w = \log(w/w_a),\ t_h = \log(h/h_a).$$

$$t_x^* = (x^* - x_a)/w_a,\ t_y^* = (y^* - y_a)/h_a,\ t_w^* = \log(w^*/w_a),\ t_h^* = \log(h^*/h_a).$$

It can be considered the bounding-box regression from anchor box to the ground-truth box nearby. (x, y) is the coordinates of the center. w is the width and h is the height. x, x_a and x^* are for the predicted box, anchor box and ground-truth box respectively, which is the same as y, w and h. N_{cls}, N_{reg} and λ are for normalization.

The output of RPN is transferred to RoI pooling, *cls* and *reg* layers, where the discrete probability of being a object, $p = (p_0, p_1, \cdots p_k)$, and the location regression, $t^k = (t_x^k, t_y^k, t_w^k, t_h^k)$ are obtained.

The loss function of RPN is defined as:

$$L(p, u, t^u, v) = \sum_i L_{cls}(p, u) + \lambda[u \geq 1] \sum_i L_{reg}(t^u, v) \tag{3-14}$$

$$L_{cls}(p, u) = -\log p_u \tag{3-15}$$

$$L_{reg}(t^u, v) = \sum_{i \in \{x,y,w,h\}} smooth(t^u - v) \tag{3-16}$$

$L_{cls}(p, u)$ is the *cls* loss of category u. $L_{reg}(t^u, v)$ is the *reg* loss of category u. $v = (v_x, v_y, v_w, v_h)$ represents the labeled location and $t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$ represents the predicted location. When $u \geq 1$, $[u \geq 1] = 1$; otherwise $[u \geq 1] = 0$. $Smooth(x)$ is defined in Formula (3-13).

3.3.5 Network Training

When training RPN, a binary class label (of being an object or not) is assigned to each anchor. Two types of anchors are given positive label:

(I) Anchor or anchors of the highest Intersection of Union (IoU) with a ground-truth box.

(II) Anchor of an IoU higher than 0.7 with ground-truth box.

IoU is a common assessment function of contact ratio between detected result and ground-truth box, defined in Formula (3-16):

$$IoU(GT, DR) = \frac{S_{GT \cap DR}}{S_{GT \cup DR}} \tag{3-17}$$

Notice that a ground-truth can have many positive anchors. It is proved that rule II can basically satisfy the generation of enough positive anchors. However, as for some extreme

cases, such as the case that none of the IoUs between anchor and ground-truth box are higher than 0.7, rule I is adopted.

On the other hand, negative anchors are labeled. Those anchors do not have IoU higher than 0.3 with any ground-truth box. Discard the anchors which are useless for training. They may have neither positive nor negative labels, or may be crossing the border of images.

For the whole network, Girshick's alternating training is adopted in order to ensure the shared *conv* layers:

(I) RPN training. RPN is initialized with an ImageNet-pre-trained model and fine-tuned end-to-end for region proposal task.

(II) Detection network training (using proposals generated in step I RPN). The network is also initialized with an ImageNet-pre-trained model. RPN and the detection network do not share *conv* layers at this point.

(III) Fix the shared *conv* layers and fine-tune RPN. The layers unique to RPN are revised. RPN is initialized with the detection network. Now the two networks share layers.

(IV) Fix the shared *conv* layers and fine-tune the detection network. The *fc* layers are revised.

3.4 Anchor Box Optimization

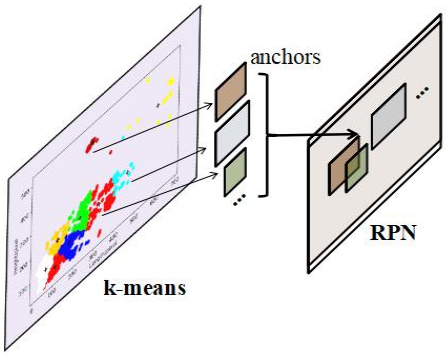


Figure 3-7: Anchor Box Optimization using k-means in RPN.

Faster R-CNN is mainly for general object detection, locating objects in various sizes in an image. When it comes to specific object detection (shared bikes, in this paper), considering too many sizes when detecting leads to errors. On the other hand, calculating the sizes in which objects would not likely to appear is invalidly and time-consuming.

In order to avoid the situation, I think about optimizing the algorithm in regards of its region-based mechanism, the most remarkable characteristic of this algorithm. Region proposals are of significance in the algorithm, generated by RPN from feature maps and would be used directly in detection network. The optimization about region proposals is reasonable and seems likely to make a difference.

In the generation of region proposals, anchor boxes matter. Better anchor boxes generate region proposals which are more appropriate to specific object. Proper proposals can increase IoUs between detected objects and ground-truth boxes, resulting in decrease of the possibility of detecting two adjacent objects as one.

As for specific object detection, although the postures and sizes of the object varies, they

can be classified into limited types approximately, which gives rise to the idea of extracting the features of postures and sizes. However, detecting the postures and sizes separately are not much efficient. Regular detection of postures has difficulties. There is no need for pointing out the features of size in detection. A simpler method has to be adopted to figure out the features of them, then generate a set of better indices of anchor boxes in RPN. It has to be non-referenced because the features are unknown, and, efficient as well.

I adopt k-means[39] to refine the sizes and scales of anchor boxes. It is proved to improve the detection and recognition results in experiments. K-means is an unsupervised learning algorithm for clustering. It does not need repetitions on calculating after the anchor boxes are set. The detailed process is explained below:

Assume that there are m objects of n dimensions to be computed. The ith $(i = 1, 2, \cdots, m)$ object's coordinate vector $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}\}$.

Firstly pick k objects μ_1 , μ_2 ,…, $\mu_k \in \mathbb{R}^n$ from all m objects randomly as original cluster center. Classify them into k categories depending on how similar they are to the cluster center, using Euclidean Distance as measurement specifically in this paper. The category of the ith object $c^{(i)}$ is defined as:

$$c^{(i)} = \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2, \quad j = 1, 2, \cdots, k \tag{3-18}$$

Then renew the coordinates of k cluster centers. Here, they are computed by the means of the coordinates in the same cluster:

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}, \quad j = 1, 2, \cdots, k \tag{3-19}$$

$1\{c^{(i)} = j\}$ is defined as:

$$1\{\mathrm{c}^{(i)}=\mathrm{j}\} = \begin{cases} 1, \mathrm{c}^{(i)} = j \\ 0, \mathrm{c}^{(i)} \neq j \end{cases} \tag{3-20}$$

Repeat the process until standard assessment function gets to the minimum value or the times of repetition reaches the threshold. The standard assessment function is defined in Formula (3-20) as the sum of distances from every object to the center in k clusters.

$$f(x) = \min \sum_{i=1}^{k} \sum_{x \in \mu_i} dist(x, \mu_i) \tag{3-21}$$

The number k should be set, which is not easy to estimate in reality. Let k=9 because there are 9 anchor boxes originally. Notice that the number of anchor boxes is set to 12.

4 Experiments

4.1 Settings

Experimental Environment. The experiments are conducted on a PC with 32GB RAM and 3.6GHz CPU. GEFORCE GTX GPU is associated. The OS is Windows 7 (x64). Caffe is installed as the deep learning framework.

Data Set. I construct a categorized shared bike data set of 4,291 images and 12,697 labeled objects is used for algorithm training and testing. The number of images in training set and test set is 3,861 and 430 respectively.

Figure 4-1: Samples in the data set. The data is collected from the Video Surveillance System of a high angle. Typical scenes such as the entrances of subway station, sideways, bus stations, hospitals, and so on are included in the data set. It also covers a variety of conditions such as the weather (sunny, rainy, foggy, etc.), time periods (daytime and night), levels of illumination and postures of shared bikes, which allows the data set to hold certain representative value.

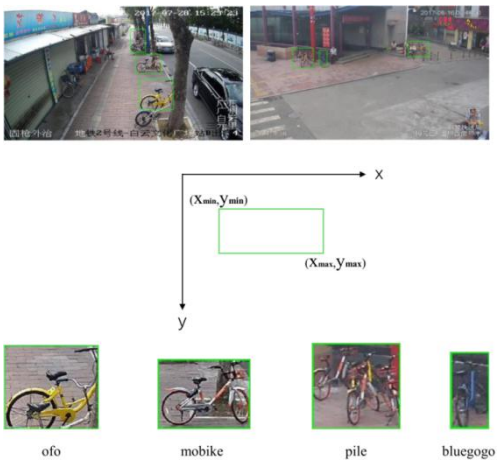


Figure 4-2: Shared bikes labeling. 12,697 objects are manually labeled in total, classified into 4 categories: ofo, mobike, bluegogo and pile (bikes overlapped). The number of objects in the test set is 1,427.

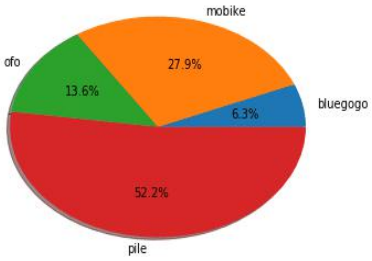


Figure 4-3: Distribution of objects in different categories in the data set.

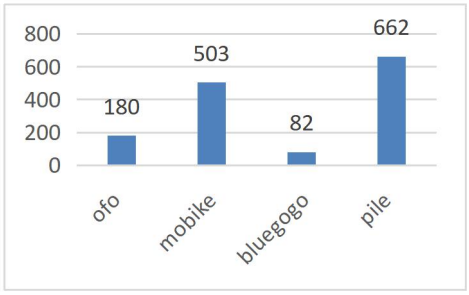


Figure 4-4: Diagram of objects in different categories in the test set.

As shown in the figures, the number of piles is of abundance. The numbers of objects in each category differ from 800 to 1,600, corresponding with the unbalanced distribution of shared bikes in reality.

Evaluation. Detection is judged by IoU (defined in Formula (3-17) in Section 3.3.5). The accuracy of detection (representing whether the bikes have been figured out) is defined as:

$$acc(i) = \frac{N_i(IoU \geq T)}{N_i} \tag{4-1}$$

T is the threshold of judging whether the object is successfully detected. $N_i(IoU \geq T)$ is the number of objects detected correctly in the i[th] category, while N_i is the total number of objects in it.

The precision of recognition (representing whether the bikes have been classified successfully) is defined as:

$$prec(i) = \frac{TP_i}{TP_i + FP_i} \tag{4-2}$$

As for the i[th] category, TP_i is the number of objects successfully recognized, while FP_i is the number of objects recognized as non-i[th] category objects.

Initialization. FoA is initialized with the original indices of Faster R-CNN in [30], while optimized FoA is optimized by k-means. Indices of anchor boxes are refined and generated

from the labeled shared bikes in the data set. Indices of anchors in FoA is: scales = [8,6,32], ratios = [0.5,1,2].

Table 4-1: Clustering results.

Length / pixel	Height / pixel
231.3131313	189.496633
463.5333333	201.6044444
62.95156125	104.4059247
114.2843735	76.16564993
608.3362256	366.3210412
286.4364641	458.441989
115.733614	217.6512327
324.9880192	175.4816294
178.6647666	126.347177

Table 4-2: Indices of anchors.

No.	Anchor index in FoA / pixel				Anchor index in optimized FoA / pixel			
1	-83	-39	100	56	-54	-30	71	47
2	-175	-87	192	104	-96	-56	113	73
3	-359	-183	376	200	-201	-121	218	138
4	-55	-55	72	72	-306	-186	323	203
5	-119	-119	136	136	-30	-51	47	68
6	-247	-247	264	264	-56	-91	73	108
7	-35	-79	52	96	-121	-191	138	208
8	-79	-167	96	184	-186	-291	203	308
9	-167	-343	184	360	-24	-66	41	83
10	/	/	/	/	-46	-116	63	133
11	/	/	/	/	-101	-241	118	258
12	/	/	/	/	-156	-366	173	383

4.2 Results and Analysis

The results show that FoA gives excellent performance in detection and recognition. It can adapt the cases and reduce the interference of moving objects effectively. The anchor box optimization is proved to make a difference in comparison.

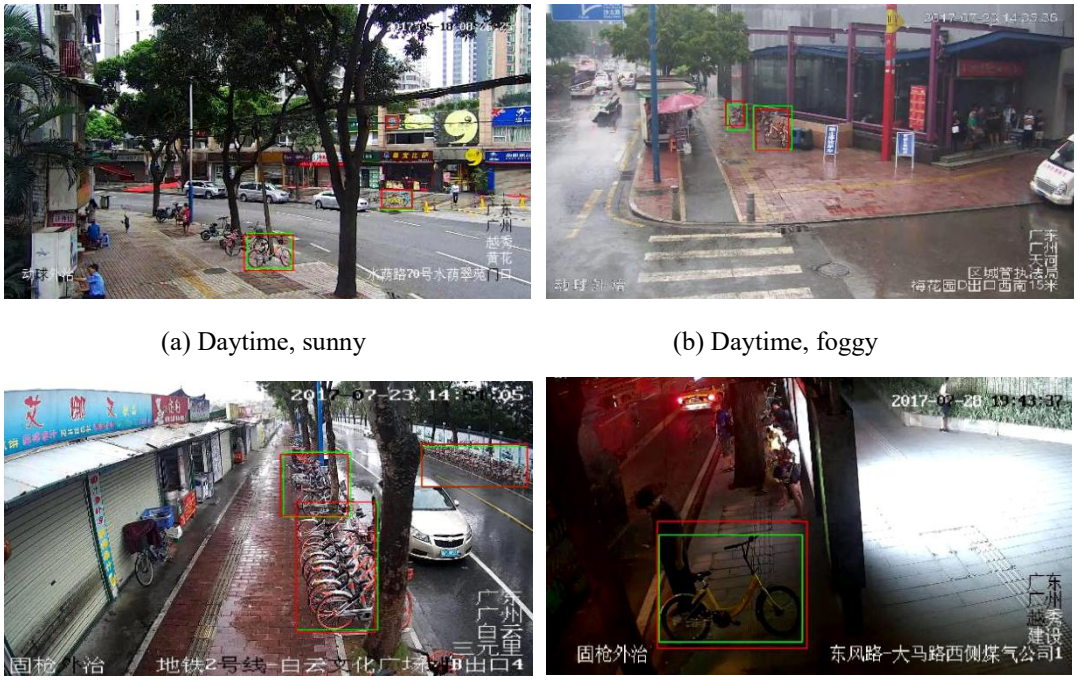
Detection Results

Table 4-3: Detection rates: % (IoU=0.7).

	ofo	mobike	bluegogo	pile	average
FoA	77.3251	90.1064	100.0000	98.1021	91.3834
Optimized FoA	86.2716	91.6934	100.0000	98.7856	94.1877

The overall detection rate is increased after *anchor box optimization*. The detection rate of ofo in optimized FoA is increased by 9.0% while others do not differ much. The increase of

detection rate influences in the improvement recognition rate.



(a) Daytime, sunny (b) Daytime, foggy

(c) Daytime, rainy (d) Night

Figure 4-6: Samples of detection results. As shown in the figures, green boxes are locations labeled artificially and red ones are computed by the algorithm. Basically they overlap with each other very much. And at the same time, even in some bad conditions such as at night, objects can still be detected accurately.

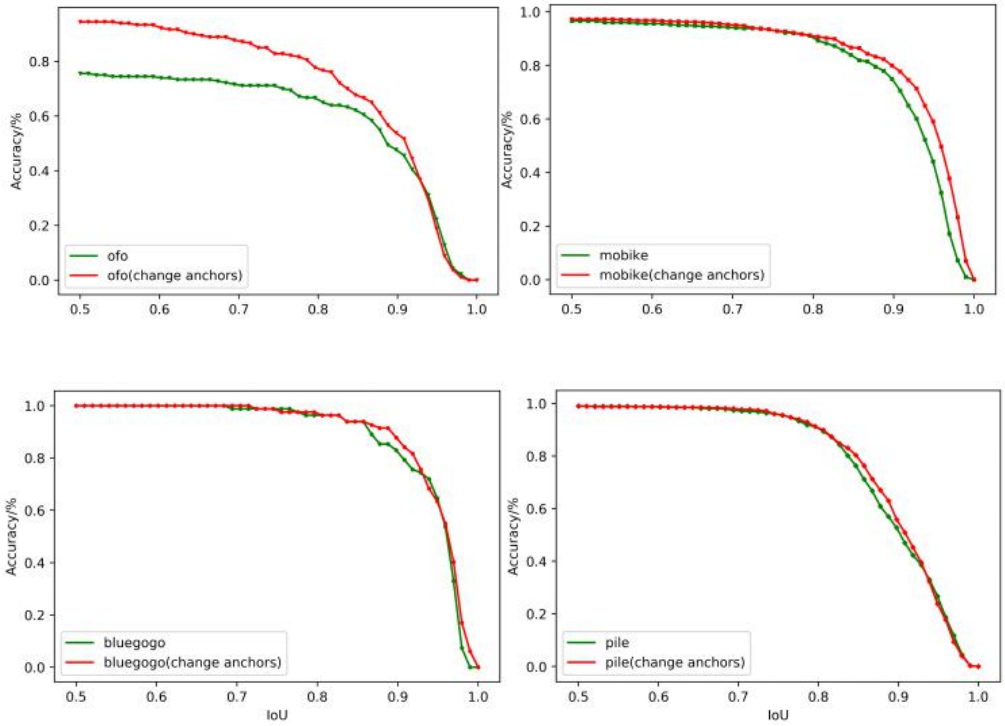


Figure 4-7: Comparisons of detection rates before and after changing anchors. At the same IoU, detection results are better if the accuracy is higher; while at the same accuracy,

detection results are better while the IoU is better. The accuracy of all 4 categories is improved under all IoU. Ofo's accuracy increases obviously when IoU is below 0.9.

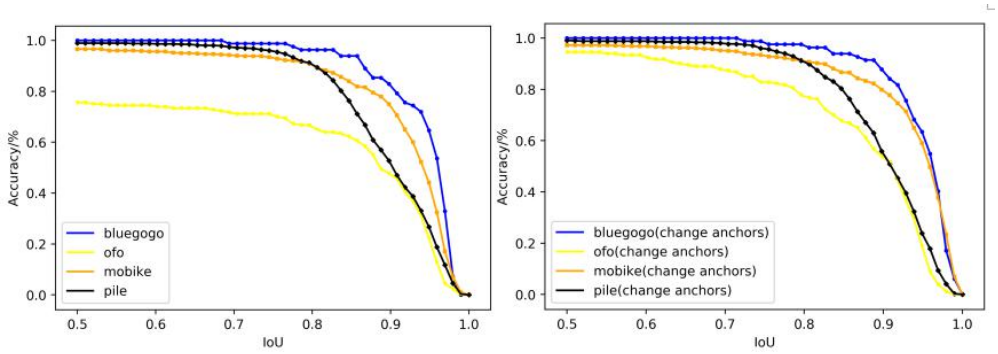


Figure 4-8: Comparisons of detection rates. **Left:** Detection rates of FoA. **Right:** Detection rates of optimized FoA. The permutation of accuracy of each category differs under different IoUs. In object detection, it is considered successful detected if IoU > 0.5. In optimized FoA, more than 90% objects in the test set can be successfully detected when IoU=0.5, which represents good detection results.

In order to balance the detection and recognition results, IoU is assigned to a value from 0.5 to 0.8. The decrease of detection rates becomes rapid when IoU > 0.7. So a successful detection is considered to have a IoU > 0.7.

Recognition Results

Table 4-4: Detection rates: % (IoU=0.7).

	ofo	mobike	bluegogo	pile	average
FoA	72.2295	90.8895	99.4574	90.3351	88.2279
Optimized FoA	90.5216	90.9091	99.5669	89.9224	92.7300

The overall recognition rate is increased after *anchor box optimization*. The recognition rate of ofo in optimized FoA is increased by 18.3% while others do not differs much.

Errors in recognition can be classified into 2 types: false-detected and undetected errors. In the following figures, the left one is computed by FoA while the right one is by optimized FoA. As shown in the samples, optimized FoA can avoid some mistakes of FoA and improve the recognition rate.

Figure 4-9: Samples of false-detected error. **Left:** Pedestrian detected as ofo in FoA. **Right:** Detected correctly in FoA.

Figure 4-10: Samples of false-detected error. **Left:** Bluegogo detected as mobike in FoA. **Right:** Bluegogo detected correctly in optimized FoA.

Figure 4-11: Samples of undetected error. **Right:** Undetected bluegogo in FoA. **Left:** Detected bluegogo in optimized FoA.

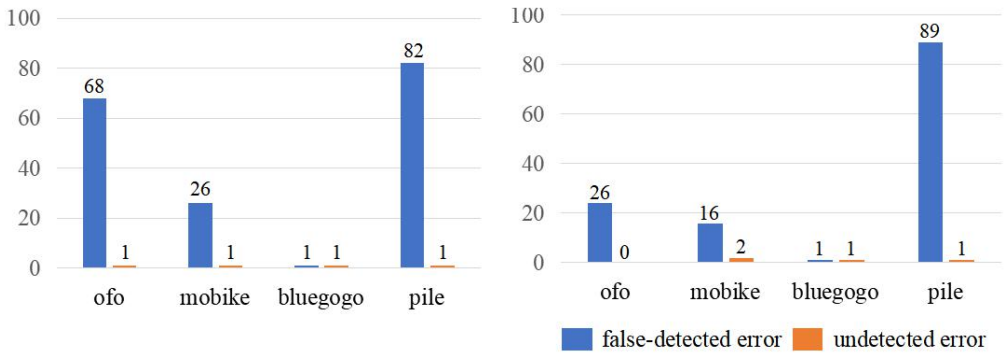


Figure 4-12: Errors Statistics. **Left:** Errors of FoA. **Right:** Errors of optimized FoA. The results are better for optimized FoA than FoA. The majority of errors comes from detecting the character area in the background as pile. A number of ofos are undetected in FoA. It is because the scales of anchors are too large, which leads to inappropriate proposals and makes it easy to detect several bikes as one. After refining the anchor boxes, the number of undetected errors reduces obviously; so does the total number of errors.

5 Prototype of a New Detection Framework

In the study of R-CNN series, the most modern networks nowadays have given excellent performances. Improvements in structures and training methods are made to get better results or faster speeds, applying to general object detection. As for specific object detection, the complexity of scenarios leads to drawbacks not only on application but also on the development of more effective and practical neural networks.

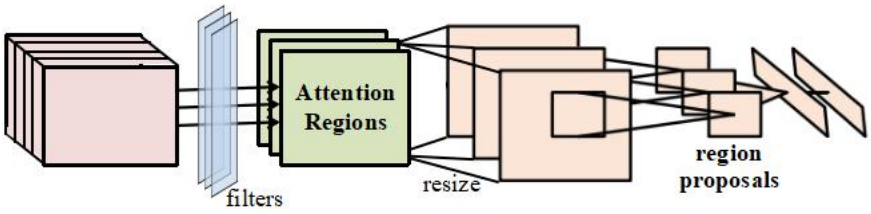


Figure 5-1: ***Attention Region Extraction plus Region-Based Convolutional Neural Network*** (ARE plus R-CNN). This paper puts forward a new prototype of framework for specific object detection. FoA is one of its robust improvements and holds its properties.

The ARE part corresponds with the region-based mechanism of R-CNN series, providing a more robust and feasible way to locate certain objects. An attention region can be defined as the background, foreground, or other areas the object possibly exists - but no need to be defined too accurately.

There are several filters for ARE, basically 3 filters. For example, a GMM and 2 unit filters in FoA. Then resize the region to feed the neural network. Notice that the filters are not narrowly defined. They represent different function modules or operations which compute the results from previous and present data. The filters simplify the input by wiping out part of interference.

The resize operation allows attention regions to have different sizes and scales. It links the two major parts of the framework and avoids complex and extra calculations. It holds, and refines the characteristics of translation invariance in R-CNN.

Table 5-1: Implements of ARE plus R-CNN

	Object	Attention Region	Filter α	Filter β	Filter θ
Motion-sensitive	Shared bike	Background	GMM	Unit	Unit
	Vehicle	Background	GMM	Unit	Unit
		Foreground	GMM	Tracking	Unit
Color-sensitive	Rivers and lakes (in map)	Light area	Low-pass Filter	Unit	Unit
Multi-sensitive	Wild tigers (in infrared video)	Dark area, foreground	High-pass Filter	GMM	Tracking

* A unit filter passes the original values of pixels.

ARE plus R-CNN can be generalized for certain object detection in various scenes. The framework is an organic combination and improvement of traditional image processing technology and modern deep neural networks. It helps allocate the computational resource appropriately. By paying attention on the joins of tradition techniques to deep learning, it provides possibility and a new concern to break over the drawbacks on detection. By coincident, this idea is also mentioned in the latest paper of Non-local Neural Network by Kaiming He on arXiv[40].

6 Conclusion and Future Work

The followings are conclusion of the paper:

(1) A review on the status of bike-sharing system and object detection is conducted. Computer vision technology is applied to help solve illegal parking problem. It is a new application to detect shared bikes automatically by image processing techniques.

(2) A complete algorithm called *Faster R-CNN over Attention* is put forward for shared bike detection in surveillance video. The concept of Attention Region is introduced based on Attention model and human visual mechanism. And FoA is optimized by refining anchor boxes using k-means.

(3) A shared bike data set is constructed for experiments. The algorithm is proved to possess both practicality and universality, while the optimization makes a difference. Distinguished and robust features of shared bikes are learned in the experiments.

(4) FoA is generalized into a prototype of a new detection framework. Its feasibility, characteristics and implements are discussed.

The followings are the future work:

(1) Attention Region Extraction or network training can be improved. For example, remove the character areas to reduce errors on piles, or detect components of shared bikes to reduce the blocking effect.

(2) The data set can be enriched. For example, add some posture samples and try to detect the postures.

(3) The new framework for certain object detection can be applied in various scenes, such as vehicle detection, creature detection, etc.

References

[1] 交通运输部, 中央宣传部等.关于鼓励和规范互联网租赁自行车发展的指导意见[Z].2017-08-01

[2] 公安部道路交通安全研究中心.共享单车发展对交通管理的影响分析及对策建议研究报告[Z]. 2017

[3] 公安部道路交通安全研究中心.在共享经济视角下推进共享单车道路交通管理的思考及建议[Z]. 2017

[4] 广州市交通委员会. 广州市中心城区城市道路自行车停放区设置技术导则[Z].2017-03-21

[5] 成都市公安局交通管理局. 成都市中心城区公共区域非机动车停放区位技术导则[Z].2016-12-26

[6] 厦门市市政园林局. 厦门市自行车停放区设置指引(试行) [Z].2017-02

[7] 廖应成.共享单车电子围栏技术使用调研报告[Z]. 2017-08-08

[8] GAT 669-2008, 城市监控报警联网系统技术标准[S].

[9] GAT 367-2001, 视频安防监控系统技术要求[S].

[10] Hubel D H, Wiesel T N. Receptive fields and functional architecture of monkey striate cortex[J]. Journal of Physiology, 1968, 195(1):215.

[11] K. Fukushima. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological cybernetics, 1980.

[12] Ruck D W, Rogers S K, Kabisky M． Feature selection using a multilayer perception[J]. Journal of Neural Network Computing, 1990, 2(2):40-48.
[13] Lecun Y, Boser B, Denker J S, et al. Handwritten digit recognition with a back-propagation network[C]. Advances in Neural Information Processing Systems. Morgan Kaufmann Publishers Inc. 1990:396-404.
[14] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
[15] Hecht-Nielsen. Theory of the backpropagation neural network[C]. International Joint Conference on Neural Networks. IEEE, 1989:593-605 vol.1.
[16] Zhang W, Itoh K, Tanida J, et al. Parallel distributed processing model with local space-invariant interconnections and its optical architecture[J]. Applied Optics, 1990, 29(29):4790-4797.
[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
[18] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]. European conference on computer vision. Springer International Publishing, 2014: 818-833.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[20] 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.
[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[22] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. computer vision and pattern recognition, 2005: 886-893.
[23] Felzenszwalb P F, Girshick R, Mcallester D, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.
[24] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
[25] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. european conference on computer vision, 2015: 21-37.
[26] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[C]. european conference on computer vision, 2014: 346-361.
[27] Redmon J, Divvala S K, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. computer vision and pattern recognition, 2015: 779-788.
[28] Girshick R. Fast R-CNN[J]. Computer Science, 2015.
[29] Girshick R. Fast R-CNN[C]. international conference on computer vision, 2015: 1440-1448.
[30] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016:1-1.
[31] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines.

In ICML, 2010.
[32] Minxian Yuan. Fine-grained Vehicle Model Recognition Based on Image Understanding [D]. Guangdong: Sun Yat-sun University, 2017.
[33] Wang T, Wu D J, Coates A, et al. End-to-end text recognition with convolutional neural networks[C]. Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012: 3304-3308.
[34] Boureau Y L, Ponce J, LeCun Y. A theoretical analysis of feature pooling in visual recognition[C]. Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 111-118.
[35] Zitnick C L, Dollar P. Edge Boxes: Locating Object Proposals from Edges[C]. european conference on computer vision, 2014: 391-405.
[36] Szegedy C, Reed S, Erhan D, et al. Scalable, High-Quality Object Detection[J]., 2014.
[37] Uijlings J, Sande K E, Gevers T, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.
[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
[39] Macqueen J. Some Methods for Classification and Analysis of MultiVariate Observations[C]// Proc. of, Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.
[40] X. Wang, K. He, R. Girshick, A. Gupta. Non-local Neural Networks[R]. arXiv: 1711.07971.

Acknowledgement

I would like to give my sincere gratitude to my tutors, Prof. Xiying Li, Prof. Nong Xiao, and Prof. Zibin Zheng in Sun Yat-sen University. Thanks to the postgraduate students in Intelligent Transportation Lab and Mobile Internet and Finance Big Data Lab. They helped a lot by advising me patiently, inspiring me, and providing academic resources and materials. They are always willing to answer my questions and make considerate comments on my outline.

My heartfelt thanks also go to teachers, Mr. Xiao'an Yang and Mr. Tao Xia, in the Affiliated High School of South China Normal University who encouraged me to conduct my study and gave me useful advice.

I am pleased to acknowledge the officers and experts from Guangdong Public Security Department. They gave me experiment materials. I learned lots of management experience and ideas. Their professional knowledge and sense of responsibility influenced me much.

Finally, in particular, I would like to express my gratitude to my friends and parents for their support.