Computers, Materials & Continua CMC, vol.62, no.1, pp.293-307, 2020

SSD Real-Time Illegal Parking Detection Based on Contextual Information Transmission

Huanrong Tang[1], Aoming Peng[1], Dongming Zhang[2], Tianming Liu[3] and Jianquan Ouyang[1, *]

Abstract: With the improvement of the national economic level, the number of vehicles is still increasing year by year. According to the statistics of National Bureau of Statics, the number is approximately up to 327 million in China by the end of 2018, which makes urban traffic pressure continues to rise so that the negative impact of urban traffic order is growing. Illegal parking-the common problem in the field of transportation security is urgent to be solved and traditional methods to address it are mainly based on ground loop and manual supervision, which may miss detection and cost much manpower. Due to the rapidly developing deep learning sweeping the world in recent years, object detection methods relying on background segmentation cannot meet the requirements of complex and various scenes on speed and precision. Thus, an improved Single Shot MultiBox Detector (SSD) based on deep learning is proposed in our study, we introduce attention mechanism by spatial transformer module which gives neural networks the ability to actively spatially transform feature maps and add contextual information transmission in specified layer. Finally, we found out the best connection layer in the detection model by repeated experiments especially for small objects and increased the precision by 1.5% than the baseline SSD without extra training cost. Meanwhile, we designed an illegal parking vehicle detection method by the improved SSD, reaching a high precision up to 97.3% and achieving a speed of 40FPS, superior to most of vehicle detection methods, will make contributions to relieving the negative impact of illegal parking.

Keywords: Contextual information transmission, illegal parking detection, spatial attention mechanism, deep learning.

1 Introduction

Illegal parking-a common problem in the field of transportation security is urgent to be solved. According to the statistics of National Bureau of Statics, the number of vehicles is up to 327 million in China by the of 2018, which makes urban traffic pressure continue to

[1] Key Laboratory of Intelligent Computing and Information Processing, Ministry of Education, College of Information Engineering, Xiangtan University, Xiangtan, 411105, China.

[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029, China.

[3] Department of Computer Science, the University of Georgia, Athens, Georgia, USA.

[*] Corresponding Author: Jianquan Ouyang. Email: oyjq@xtu.edu.cn.

rise and the impact of illegal parking on urban traffic order is growing. There are lots of unnecessary traffic accidents, such as a heavy-duty trailer which hits the crosswalk in a city, making the non-motor vehicle lanes unusable and thus unfortunately leading to the death of a pedestrian. It undoubtedly shows the public transportation safety problems caused by illegal parking need to be eagerly solved [Zheng (2014)]. In the light of above, it is of great significance to propose a method that can automatically detect the illegal parking in real-time, which is embodied in two aspects: on the one hand, the method could relieve and avoid the public transportation safety accidents caused by illegal parking, on the other hand, it could also provide law enforcement officers with more efficient monitoring and management tools.

Our study on detection of illegal parking belongs to the field of object detection in computer vision, the performance is mainly evaluated by speed, precision and robustness. Traditional methods are mainly based on ground loop and manual supervision, which may miss detection and cost much manpower, and then researchers do image analysis to supervise illegal parking. There are limited number of related studies on this topic, few methods have been proposed to establish such a system [Maddalena and Petrosino (2013); Mu, Ma and Zhang (2015); Wahyono, Filonenko and Jo (2015)], most of which are based on separation of foreground and background. Specifically, vehicles are first extracted from background and then are tracked. An alarm will be triggered if a vehicle was found to be stationary and it lasts over a preset time in the rectangular region of interest (ROI).

Maddalena et al. [Maddalena and Petrosino (2013)] utilized sophisticated background modeling strategies to extract the foreground objects, it does well in simple traffic environment except the crowded scenes. Mu et al. [Mu, Xing and Zhang (2015)] proposed a method that subtracting the background constructed by Gaussian Mixture Model to extract the foreground, and then a vehicle is recognized by detecting wheels, it effectively separates the foreground and background, but a vehicle cannot be detected when its wheels are occluded. Wahyono et al. [Wahyono, Filonenko and Jo (2015)] proposed to use background subtraction to get candidates of stationary regions and verify a vehicle by exacting scalable histogram of oriented gradient (SHOG) features followed by support vector machines (SVM) classification. It performs well when lighting changes, but the design of the SHOG features is hard and cannot treat well with complex weather conditions. Overall, most of above methods are easily affected by various environments, such as illumination changing, occlusion and weather.

In order to enhance the robustness and precision of the detection in complex environments, feature extraction plays an important role. The aim of the object detection is to locate all the objects and specify each object category on a given image or video [Meng, Rice, Wang et al. (2018)]. With the rapidly development of deep learning, the R-CNN [Girshick, Donahue, Darrell et al. (2014)], SPP-Net [He, Zhang, Ren et al. (2014)], Fast R-CNN [Ren, He, Girshick et al. (2015)], Faster R-CNN [Ren, He, Girshick et al. (2015)] algorithms who based on region proposal are gradually evolved and got better performance step by step. Finally, the Faster R-CNN has achieved an essentially end to end detection system with a high precision, but its speed still needs improving. After that, end to end object detection algorithms such as You Only Look Once (YOLO) [Redmon, Divvala, Girshick et al. (2016)] and SSD [Liu, Anguelov, Erhan et al. (2016)] which have

obtained a higher speed and precision, and the latter SSD have gotten a balanced performance on precision and speed. but its shortcoming is the detection of small objects. Particularly, to improve the detection about small objects, the Deconvolutional Single Shot Detector (DSSD) [Fu, Liu, Ranga et al. (2017)] introduces contextual information to get a better feature representation ability by replacing the basic VGG with ResNet [He, Zhang, Ren et al. (2016)] and does much better in small objects detection. What's more, Google DeepMind had proposed Spatial Transformer Networks (STN), whose differentiable module do not require redundant annotations and can adaptively learn the spatial transformation of different data. It can not only transform the input spatially, but also can be inserted into the arbitrary layer of the existing network as a network module to achieve the spatial transformation of different feature maps, it is essentially a spatial domain attention learning mechanism. So that we try to add STNs module to SSD to further improve the performance of overall model.

Based on the above, the main contributions made in this paper are described below:

a. Aiming at the problem that object detection lacks deep semantic feature information when using the feature information of shallow network in the SSD to predict object. This paper is inspired by DSSD and dedicated to making use of the deconvolution layer to achieve the contextual information fusion. Thus, the improved SSD model that can extract deeper and focused details is proposed, which obtained a higher precision on PASCAL VOC2007 than baseline SSD by 1.5%.

b. In view of the fact that the pooling layer in Convolutional Neural Network has gotten certain spatial robustness at the expense of very important location feature information, this paper introduces a spatial attention transformation to the SSD to relieve the problem.

c. Applying the improved model which combines contextual fusion and STNs to illegal parking detection in sophisticated scenes and training it with the public datasets BIT-Vehicle [Dong, Pei, He et al. (2014)] and vehicle images of the PASCAL VOC2007 together, the improved model finally achieved good results that a precision up to 97.3% and the real-time performance with a speed of 40FPS.

2 Related works

2.1 The single shot detector (SSD)

SSD is a general detector published by Wei Liu in ECCV 2016, which takes advantages of Faster R-CNN, YOLO and multiscale pyramids. Concretely, it discretizes the output space of bounding boxes into a set of default boxes, and every feature map owns different aspect ratios and scales. Additionally, it combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes.

Specifically, the framework of SSD is described by the following Fig. 1.

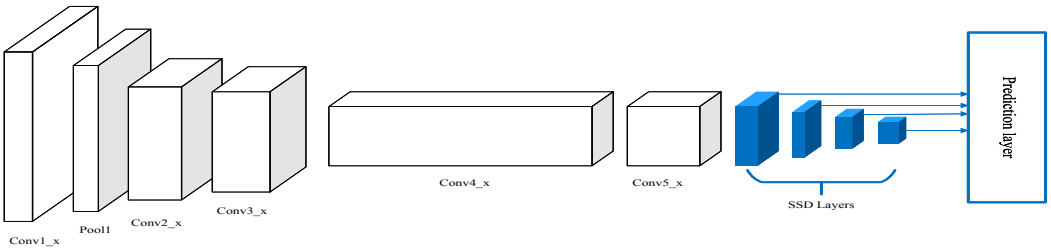

Figure 1: Framework of SSD

The loss function of the SSD is described below:

$$L(\mathrm{x},c,l,g)=\frac{1}{N}\left(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)\right) \tag{1}$$

where N is the number of matched default box for a ground truth. L_{conf} is confidence loss and L_{loc} is localization loss. $x_{ij}^{p}=\{1,0\}$ is an indicator for matching the i_{th} default box to the j_{th} ground truth box of category p. The localization loss is a Smooth L1 loss between the predicated box (l) and the ground truth box (g) parameters. The parameters of location and category are trained at the same time. The scale of the default boxes for each feature map is computed as:

$$\mathrm{s}_k=s_{\min}+\frac{s_{\max}-s_{\min}}{m-1}(k-1), k\in[1,m] \tag{2}$$

where $s_{\min}$ is 0.2 and $s_{\max}$ is 0.9, meaning the lowest layer has a scale ratio of 0.2 and the highest layer has a scale ratio of 0.9, and all layers in between are regularly spaced. In practice, one can also design a distribution of default boxes to best fit a specific dataset, thus in this paper, we made use of k-means to deal with our dataset to find proper ratio parameters. By combining predictions for all default boxes with different scales and aspect ratios from all locations of many feature maps, there are diverse set of predictions, covering various input object sizes and shapes.

2.2 The deconvolutional single shot detector (DSSD)

Convolutional neural networks have inherent problems in structure: the receptive fields of high-level networks are relatively large, and the semantic information representation ability is strong, but the resolution of feature maps is low, and the representation ability of geometric information is weak; the receptive fields of low-level networks are relatively small, the information representation ability is strong, although the resolution is high, the semantic information representation ability is weak. SSD used multi-scale feature maps to predict objects, high-level feature information with large receptive fields to predict large objects, and low-level feature information with small receptive fields to predict small targets. This brings up a problem: the classification performance of the SSD for the small objects will be poor when using the feature information of the low-level network to predict the small objects due to the lack of high-level semantic features. The idea to solve

this problem is to fuse high-level semantic information and low-level semantic information, which could enrich the prediction of the multi-scale feature map of the input and finally improves the accuracy.

Based on the above description, DSSD proposed a deconvolution and "width-narrow-wide" asymmetric "hourglass" structure to inspire us to address the above problem. DSSD is one of the improved SSD model, it replaces the basic VGG network with Resnet-101 and introduced the residual module before the classification and regression. After the auxiliary convolution layer added by the SSD, the deconvolution layer is added to form "wide-narrow-wide" hourglass structure. One of the biggest improvements in DSSD compared to SSD is that DSSD has been greatly improved in the detection of small objects. The final part of the paper also shows the detection performance of small objects. Even so, the detection speed is much slower than SSD because the Resnet-101 is too deep.

2.3 Spatial transformer networks

Spatial Transformer Networks (STNs), the spatial transformation network was released by Google's DeepMind. The classification network is simpler and more efficient by executing inverse spatial transformation on the data to eliminate the influence made by the deformation of the target images.

Its framework is pictured as the following Fig. 2, which could be divided into three parts: localisation network, grid generator and sampler, which can be inserted into the existing CNN model.

In the Fig. 2, U is the input feature map that passed to a localisation network which regresses the transformation parameters θ, the regular spatial grid G over V is transformed to the sampling gird $\Gamma_\theta(G)$, which is applied to U to produce the output feature map V. The combination of the localisation network and sampling mechanism defines a spatial transformer. What's more, the STNs can be seen as a generalization of differentiable attention to any spatial transformation.

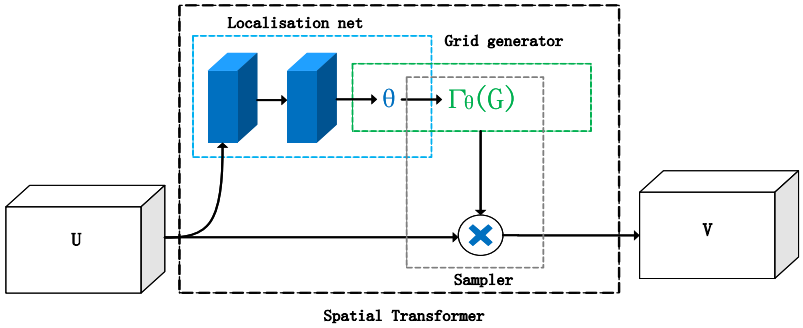

Figure 2: Framework of spatial transformer networks

2.3.1 Localisation network

The localisation network takes the input feature map $U \in R^{H \times W \times C}$ with width W, height H and C channels and outputs θ, the parameters of the transformation Γ_θ to be applied

to the feature map: $\theta = f_{loc}(U)$. The size of θ will be changed depending on the transformation type that is parameterized.

The localisation network function $f_{loc}()$ can take any form, such as a fully-connected network or a convolutional network, but should include a regression layer as the last layer to produce the transformation parameters θ.

2.3.2 Parameterized sampling grid

As the paper mentioned, by pixel the paper refers to an element of a generic feature map, not necessarily an image. In general, the output pixels are defined to lie on a regular grid $G = \{G_i\}$ of pixels $G_i = \left(x_i^s, y_i^s\right)$, forming an output feature map $V \in R^{H' \times W' \times C}$, where H' and W' are the height and width of the grid, and C is the number of channels, which is same in the input and output.

In the next expression, we show an case for 2D affine transformation A_θ.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \Gamma_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{3}$$

where $\left(x_i^t, y_i^t\right)$ are the target coordinates of the regular grid in the output feature map, are $\left(x_i^s, y_i^s\right)$ the source coordinates in the input feature map that define the sample points, and A_θ is the affine transformation matrix.

In Eq. (3), the transform allows cropping, translation, rotation, scale, and skew to be applied to the input feature map, and requires only 6 parameters (the 6 elements of A_θ) to be produced by the localisation network.

2.3.3 Differentiable image sampling

The next expression is based on the above, Each $\left(x_i^s, y_i^s\right)$ coordinate in $\Gamma_\theta\left(G\right)$ defines the spatial location in the input where a sampling kernel is applied to get the value at a particular pixel in the output V.

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k\left(x_i^s - m; \Phi_x\right) k\left(y_i^s - n; \Phi_y\right) \forall_i \in \left[1 \ldots H'W'\right] \forall_c \in \left[1 \ldots C\right] \tag{4}$$

where Φ_x and Φ_y are the parameters of a generic sampling kernel $k(\ )$ which defines the image interpolation (e.g., bilinear), U_{nm}^c is the value at location $\left(n, m\right)$ in channel of the input, and V_i^c is the output value for pixel i at location $\left(x_i^t, y_i^t\right)$ in channel c. In theory, any sampling kernel can be used, as long as (sub-)gradients can be defined

with respect to x_i^s and y_i^s. For example, using the integer sampling kernel reduces (4) to

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \delta\left(round\left(x_i^s + 0.5\right) - m\right) \delta\left(round\left(y_i^s\right) - n\right) \tag{5}$$

where $round(\ )$ means x to the nearest integer and $\delta(\ )$ is the Kronecker delta function. This sampling kernel equates to just copying the value at the nearest pixel to $\left(x_i^s, y_i^s\right)$ to the output location $\left(x_i^t, y_i^t\right)$. Alternatively, a bilinear sampling kernel can be used, giving it as below.

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max\left(0,1-|x_i^s - m|\right) \max\left(0,1-|y_i^s - n|\right) \tag{6}$$

To allow backpropagation of the loss through this sampling mechanism we can define the gradients with respect to U and G. For bilinear sampling (6) the partial derivatives are

$$\frac{\partial V_i^c}{\partial U_{nm}^c} = \sum_n^H \sum_m^W \max\left(0,1-|x_i^s - m|\right) \max\left(0,1-|y_i^s - n|\right) \tag{7}$$

$$\frac{\partial V_i^c}{\partial \mathrm{x}_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max\left(0,1-|y_i^s - n|\right) \begin{cases} 0 & if \quad |m - x_i^s| \geq 1 \\ 1 & if \quad m \geq x_i^s \\ -1 & if \quad m < x_i^s \end{cases} \tag{8}$$

the $\dfrac{\partial V_i^c}{\partial y_i^s}$ is similarity to (8).

3 Proposed method

As shown in Fig. 3 below, the work we have done is mainly divided into two pipelines. The first is the acquisition and processing of monitoring video data, which includes data cleaning and labeling, it is not the focus of this paper, but it clearly shows the overall work. The second pipeline includes the improved SSD of illegal parking detection method which will be described in the later section.

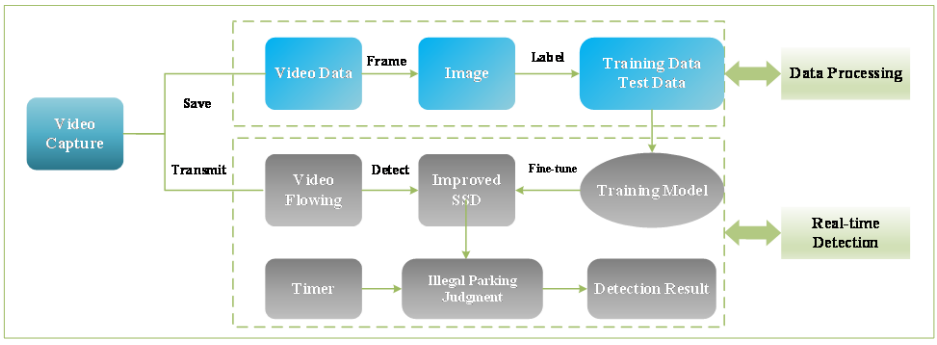

Figure 3: Overall workflow

3.1 Improved SSD based on contextual information and STNs

Based on the above, in the case that SSD does well in large objects, so that we pay more attention to improve the detection of small objects. In our study, we use deconvolution to feed deeper layer information back to the shallow layer and add STNs to our model, which introduces contextual information transmission and spatial transformer network to promote the performance of our improved model.

In addition, the STNs module is computationally very fast and does not impair the training speed so that there is little time overhead when used normally, what's more it will potentially accelerate the model by subsequent downsampling that applied to the output of the transformer.

The improved model architecture is shown in the following Fig. 4.

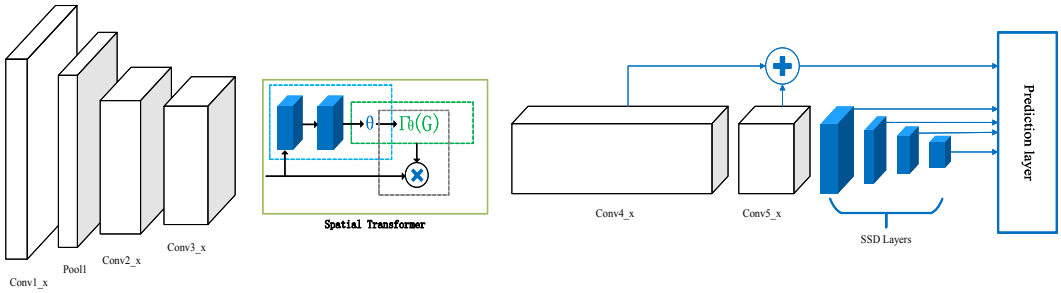

Figure 4: Improved SSD model architecture

As mentioned in the paper [Cao, Xie, Yang et al. (2017)], contextual information fusion will inevitably introduce unnecessary noise, so it is necessary to find suitable layers for connection. The connection module was inspired by DSSD and ResNet, its specific structure is shown in Fig. 5.

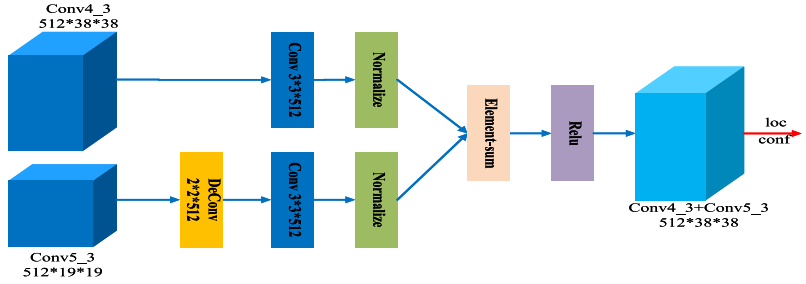

Figure 5: Connection module

In order to ensure that Conv5_3 layer and Conv4_3 layer get the same size to connect, the Conv5_3 layer is connected to a 2*2*512 bilinear deconvolution layer for upsampling. Then, the Conv4_3 and Conv5_3 layers are respectively connected to a 3*3*512 convolution layer to strenthen the ability of learning features. At the same time, the normalization layer is respectively connected to accelerate training speed and finally the feature information fusion is established after respectively connecting the ReLU function and the rule is adding up the corresponding feature value.

The best combination is tested by the dataset PASCAL VOC2007 and verified the

precision of Conv4_3, Conv4_3+Conv5_3, Conv4_3+fc6 and Conv3_3+Conv4_3+ Conv5_3. The results will be displayed in the section of 3.3, which indicates that Conv4_3+Conv5_3 makes more excellent performance.

3.2 Illegal parking detection

The illegal parking detection method proposed in this paper is different from the methods which execute background modeling with background segmentation, that is very susceptible to complex environmental factors.

3.2.1 Definition of illegal parking

The definition of illegal parking is pictured as following Fig. 6:

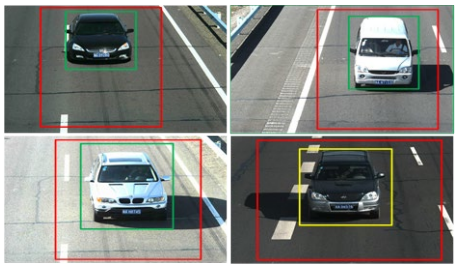

Figure 6: Examples of illegal parking (Images come from training dataset)

Within the fixed monitoring range of the camera, an alarm will be automatically triggered when there are vehicles staying over a preset time in the red alert area, the color of the bounding box will change from green to yellow.

3.2.2 Process of illegal parking

The specific flowchart is approximately shown in Fig. 7.

Inspired by Xie et al. [Xie, Wang, Chen et al. (2017)], we firstly optimize the flowchart to suit our method. Secondly, we set up the alert area and read video by frame to the improved SSD. Third, we calculate IOU between all the detected bounding boxes and alert area. IOU is a method used for calculating the proportion of overlapping parts for two bounding boxes. If it overlaps, the method will output all bounding boxes within illegal area. Finally, we determine whether there is an illegal parking according to following key aspects below:

a. Status Analysis

For each vehicle, we will firstly get two bounding boxes of the adjacent two frames and calculate respective centroid, then we calculate Euclidean distance of the two positions to determine whether it is driving, if the distance is greater than a given threshold, it is judged to be moving, otherwise, it is judged to be stationary.

b. Timer Strategy

For each vehicle, the timer is started as soon as the vehicle is moving, and the timer is cleared to zero once the vehicle is stationary.

c. New or Old Box

Since the SSD will always detect all the vehicles in an input image, so we need to match the newly detected boxes with the bounding boxes in last frame. Here we match two bounding boxes by calculating the IOU between them. If an IOU of two bounding boxes is greater than the given threshold, we think the two bounding boxes contain the same vehicle and the new box will inherit the timing information of the old box.

An alarm will be automatically triggered when there are vehicles staying over a preset time in the alert area.

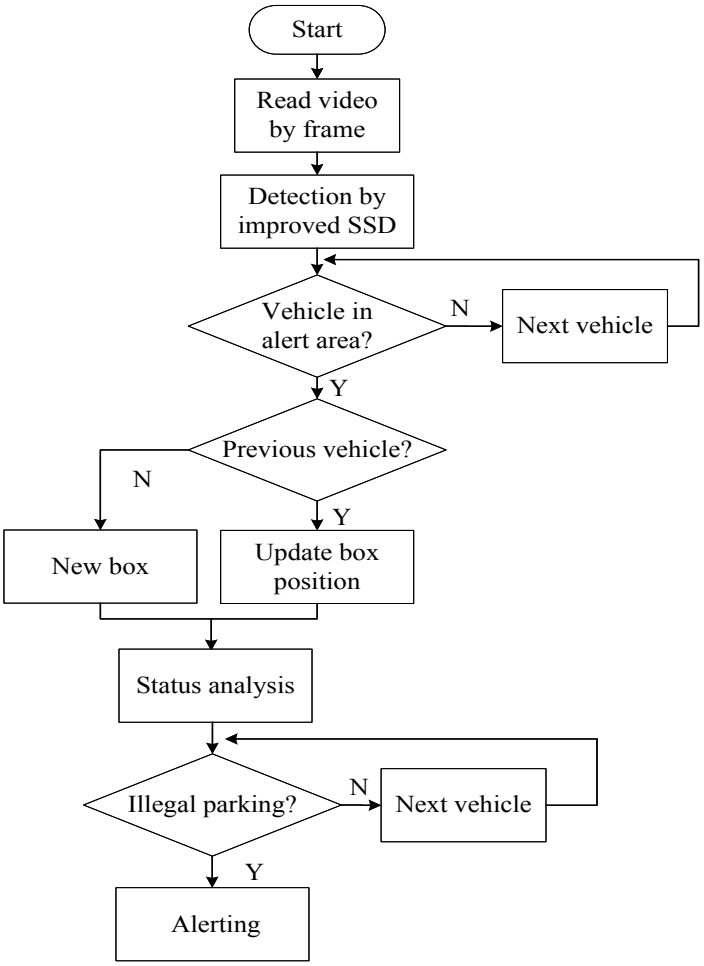

Figure 7: Flowchart of detection

3.2.3 Evaluation of illegal parking

We redefine precision and recall as the following formulas:

$$\Pr ecision = \frac{True_{illegal-parking}}{True_{illegal-parking} + False_{illegal-parking}} \tag{9}$$

$$\mathrm{Re}call = \frac{True_{illegal-parking}}{True_{illegal-parking} + False_{N-illegal-parking}} \tag{10}$$

$True_{illegal-parking}$: the times of correctly detecting illegal parking;

$False_{illegal-parking}$: the times of incorrectly detecting illegal parking;

$False_{N-illegal-parking}$: the times of missing detection of illegal parking.

4 Experiment results

4.1 Experiment configuration

OS: Ubuntu-Server 16.04(64 bit), CPU: 2.2 GHz Dual-Core CPU, Memory: 16 GB, GPU: Matrox Electronics Systems Ltd. MGA G200e [Pilot] Server Engines, Tensorflow Version: 1.4.0, CUDA Version: 8.0, CUDANN Version: 8.0.61.

4.2 Experiment dataset

The dataset mainly comes from the BIT-Vehicle and PASCAL VOC2007. The former has a total of 9850 vehicle images all over the China, it contains various color, illumination, car model, the latter is a classic dataset widely used in object detection.

Specifically, we use K-means to deal with the dataset to get proper ratio parameters in SSD model, the flow of K-means is generally as follows:

(1) Initialize the sample dataset.
(2) Select random K initial clustering centers for the analysis data.
(3) Calculate the distance between each point and the center point in the sample dataset and divide it into the category that the distance is closest.
(4) Calculate the average of all points in each cluster and set it as a new clustering center
(5) Repeat Step 3.
(6) Repeat Step 4 until met the preset iteration condition.

Then, the operation for our dataset is described below:

(1) Read the labelling files of dataset.
(2) Get the coordinates of ground truth bounding boxes and calculate length, width, finally generate the 2D coordinate.
(3) Start clustering with the coordinate as the feature until met the iteration condition. (Select K=4 in our dataset).

Finally, the K-means result is show in Fig. 8, we get the ratio parameters $a_{BIT} = \{1, 0.9, 1.2, 1.5, 2\}$ which made our experimental performance better.

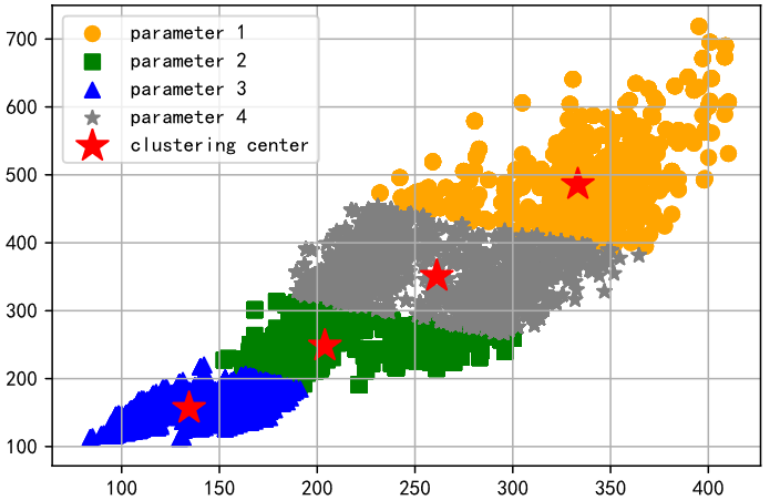

Figure 8: K-means result of our dataset

4.3 Experiment results of connection module

In this part, we display the test results of earlier Section 3.1, we choose the PASCAL VOC2007 dataset to train and use "xavier" method for weight initialization. The learning rate and loss curve in the experiment are shown in Fig. 9.

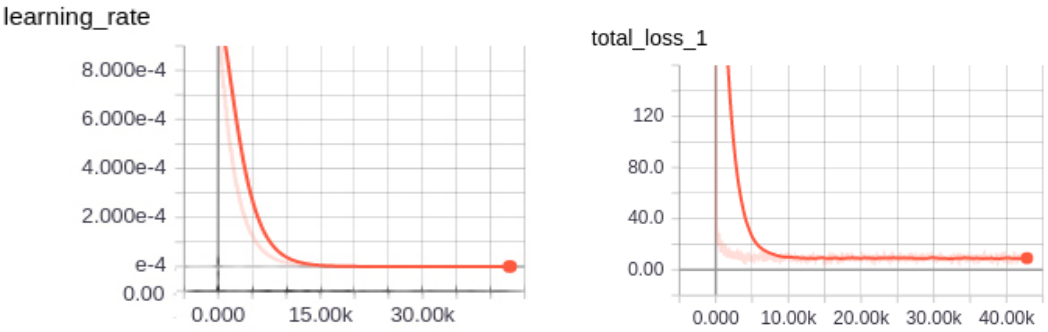

Figure 9: Learning rate and total loss

It shows that the learning rate and the loss are close to convergence when the training epoch is 15K times. The experiment result of Section 3.1 described in Fig. 10.

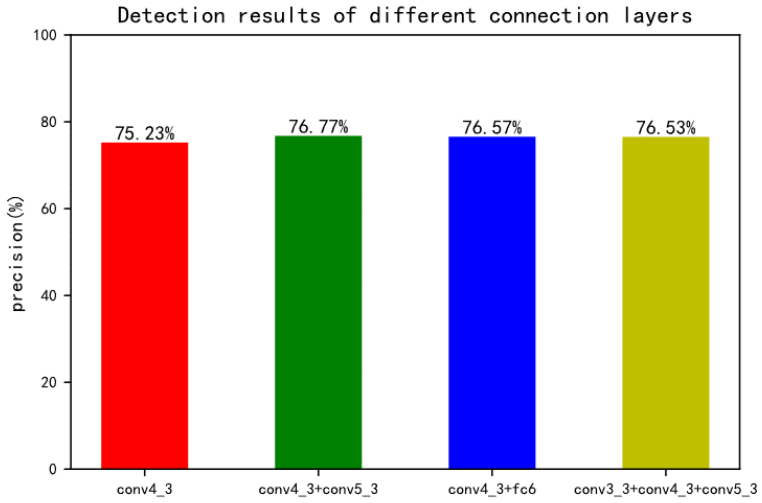

Figure 10: Detection results of different connection layers

In order to find out the most suitable combination of connection module, we repeat experiments with VOC 2007 to take the mean. As shown in Tab. 1, Conv4_3 + Conv5_3 makes the better performance that its precision is higher than Conv4_3 by 1.5%, while fc6 has a larger receptive field for small objects compared to conv5_3, which will introduce more background noise and make the performance slightly worse.

4.4 Result of detection of illegal parking

In order to verify the improved effect, we changed different detection model for comparative experiments. The experiment results are shown in Tab. 1 below.

Table 1: Detection results of different basic models

Basic Model	Precision (%)	Recall (%)	FPS
DPM	92.32	83.75	1/3
Faster R-CNN	93.45	87.72	5
Yolo-v2	94.22	95.42	38
Improved SSD	**97.30**	**96.84**	**40**

As shown in Tab. 1, It clearly shows that our method has better performance because of the strong ability to extract features, which gets a high precision up to 97.3% and a speed of 40 FPS, superior to the other methods for illegal parking detection with same conditions.

5 Conclusion

In our paper, we proposed a real-time illegal parking detection method to relieve the negative impact caused by illegal parking. We designed and verified an improved SSD

inspired by DSSD and STNs, which takes advantage of contextual information fusion and spatial transformer networks (spatial attention mechanism). Due to our improvement, it enhances the robustness of detection in various environments and improved the detection accuracy on small objects. With the improved SSD, we exactly achieved a high precision that up to 97.3% and a speed of 40FPS real-time detection, superior to the other methods for illegal parking detection with the same conditions.

In our future work, we will make further improvement on our detection method for workflow optimization and keep studying for new methods such as Nai et al. [Nai, Li, Li et al. (2018)], which proposes a novel local sparse representation based on tracking framework for visual tracking.

Acknowledgement: This research has been supported by NSFC (61672495), Scientific Research Fund of Hunan Provincial Education Department (16A208), Project of Hunan Provincial Science and Technology Department (2017SK2405), and in part by the construct program of the key discipline in Hunan Province and the CERNET Innovation Project (NGII20170715).

References

Cao, G. M.; Xie, X. M.; Yang, W. Z.; Liao, Q.; Shi, G. M. (2017): Feature-fused SSD: fast detection for small objects. *International Conference on Graphic and Image Processing.*

Dong, Z.; Pei, M. T.; He, Y.; Liu, T.; Dong, Y. M. et al. (2014): Vehicle type classification using unsupervised convolutional neural network. *IEEE International Conference on Pattern Recognition*, pp. 172-177.

Fu, C.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A. (2017): DSSD: deconvolutional single shot detector. *Computer Vision and Pattern Recognition*.

Girshick, R. (2015): Fast R-CNN. *IEEE International Conference on Computer Vision*, pp. 1440-1448.

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. (2014): Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587.

He, K. M.; Zhang, X. H.; Ren, S. Q.; Sun, J. (2016): Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. (2014): Spatial pyramid pooling in deep convolutional networks for visual recognition. *European Conference on Computer Vision*, pp. 346-361.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. et al. (2016): SSD: single shot multibox detector. *European Conference on Computer Vision*, pp. 21-37.

Meng, R. H.; Rice, S. G.; Wang, J.; Sun, X. M. (2018): A fusion steganographic Algorithm based on faster R-CNN. *Computers, Materials & Continua*, vol. 55, no. 1, pp. 1-16.

Maddalena, L.; Petrosino, A. (2013): Stopped object detection by learning foreground

model in videos. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 723-735.

Mu, C. Y.; Ma, X.; Zhang, P. P. (2015): Smart detection of vehicle in Illegal parking area by fusing of multi-features. *International Conference on Next Generation Mobile Applications, Services and Technologies*, pp. 388-392.

Nai, K.; Li, Z. Y.; Li, G. J.; Wang, S. Q. (2018): Robust object tracking via local sparse appearance model. *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 4958-4970.

Ren, S. Q.; He, K. M.; Girshick, R.; Sun, J. (2015): Faster R-CNN: towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, pp. 91-99.

Redmon, J.; Divvala, S.; Girshick, R; Farhadi, A. (2016): You only look once: unified, real time object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788.

Wahyono, W.; Filonenko, A.; Kang-Hyun, J. (2018): Illegally parked vehicle detection using adaptive dual background model. *Industrial Electronics Society*, pp. 25-28.

Xie, X. M.; Wang, C. Y.; Chen, S.; Shi, G. M.; Zhao, Z. F. (2017): Real-time illegal parking detection system based on deep learning. *International Conference on Deep Learning Technologies.*

Zheng, Y. (2014): *Research on Key Technologies of Illegal Parking Forensics (Ph.D. Thesis)*. Shanghai Jiao Tong University, China.