Intelligent Recognition of Vehicle Information in Surveillance Video

Jia He*, Weiwei Meng and Tingting You

North China Municipal Engineering Design & Research Institute Co.,Ltd. Tianjin, China

*Corresponding author: hejia20@cemi.com.cn

Abstract. In this paper, a method of real-time vehicle speed and vehicle type recognition is proposed. The vehicle type recognition is based on residual network to increase the convergence speed and improve the feature expression ability, and attach the centre loss to improve recognition accuracy of similar vehicles. The speed recognition is based on moving object detection. Experimental results show that the average error of vehicle speed is no more than 5%, and the average precision of vehicle type recognition is 85%, towards minibus and cars, the precision reaches 98.7%, which is superior to the traditional recognition method.

Keywords: Intelligent Transport System, residual network, intelligent recognition.

1. Introduction

The Intelligent Transport System has been continuously developed [1], so higher requirements are imposed on the acquisition of traffic information. Speed and type of vehicles are important traffic information, they play an important role in vehicle load research, traffic flow statistics, and illegal vehicle tracking.

Video-based vehicle detection can use the monitoring equipment to achieve identification without damaging the pavement structure, which overcomes the problems of high detection cost and difficult maintenance. Currently, vision-based vehicle recognition mainly adopts matching methods, which are as follows: 1) Feature matching, using the features of artificial designed. These features cannot express the high-order information of the image due to the lack of expressiveness, and the robustness is poor. 2) Model matching, modelling the vehicle image and then using the similarity criterion to judge the vehicle type, this method cannot effectively capture the edge information of the vehicle, so it is easy to miss detection.

With the development of deep learning, convolutional neural networks have shown great advantages in object recognition and image classification [2-3]. First of all, compared with the traditional method, convolution neural network can automatically extract the features of the image by using the original images as the input data and the large-scale data set as the driver, which overcomes the problem that the traditional feature extraction method has poor adaptability and performance in complex environment. In addition, convolution neural network has a strong resistance to image translation, rotation, scaling and other deformation, which greatly improves the accuracy of object recognition [4-6]. Residual network (ResNet) is a new type of deep convolutional network proposed by Kaiming He in 2015. It won
the ILSVRC challenge and performed very well in image detection, location and classification [7-8]. Compared with the previous convolutional neural network, the residual unit solves the problem of degradation caused by the increase of network depth, which makes the network depth get a great extension, and has a strong feature expression ability.

2. Identification of vehicle speed

2.1. Recognition of vehicles
In the video, the speed of vehicle is fast, and the frame rate of the video is large, which puts high demands on the real-time detection. In addition, the camera required for detection is fixed, which means the scene of the video cannot move globally, so the background subtraction is used for vehicle detection. The background subtraction is a common method for moving target detection. The principle is to use the difference between current frame of the image and the background model to detect the motion region, and its performance mainly depends on the background modelling technology.

2.1.1. Background modeling. In this paper, multi-frame averaging method is used to model the background. The basic idea is to take serial frames in the video image and calculate the average value of the gray values to replace that of target vehicle appearance area. This modeling method has the advantages of simple algorithm, fast calculation speed and can eliminate the influence of background on vehicle detection.

2.1.2. Threshold selection. In order to make the system have good adaptability, the method of threshold segmentation should be used, which can automatically determine the size of the threshold. Ostu algorithm is based on the grey histogram of the image, according to the maximum distance between classes to determine the segmentation threshold, the algorithm is simple and accurate, described as follows:

Suppose the image has L grey levels, \( f_i \) represents the number of pixels with a grey value of \( i \), and the total number of pixels is \( N = \sum f_i \), the probability of occurrence of each gray value is \( P_i = f_i / N \), \( t \) represents the threshold, Pixels are divided into two categories according to grey scale: background class \( B \) and foreground class \( F \), the probability that they will show up is:

\[
P_B = \sum_{i=0}^{t} P_i, \quad P_F = \sum_{i=t+1}^{L-1} P_i
\]

The grey mean values of \( B \) and \( F \) are:

\[
\omega_B = \sum_{i=0}^{t} \frac{iP_i}{P_B}, \quad \omega_F = \sum_{i=t+1}^{L-1} \frac{iP_i}{P_B}
\]

The total grey mean of the image is:

\[
\omega_0 = P_B \omega_B + P_F \omega_F
\]

From this, the variance of \( B \) and \( F \) can be obtained:

\[
(t) = P_B (\omega_B - \omega_0)^2 + P_F (\omega_F - \omega_0)^2
\]

The \( t \) that makes the variance the largest is the optimal threshold, that is:

\[
t = \text{Arg max}_{0 \leq t \leq L} \{ \sigma^2(t) \}
\]
2.1.3. Extraction of moving objects. To identify the target, we need to extract the foreground of the image data, the steps are as follows:

(i) Get the difference between the current frame and the background frame of video.
(ii) Find out the foreground target and perform binarization.
(iii) Use the mathematical morphology processing to smooth the edge and filter the noise.
(iv) Search the connected target. After binarization, the area connected by white points forms the foreground target of the current frame.

2.2. Vehicle speed recognition

The virtual coil method is used to identify the vehicle speed. This method performs the detection by calculating the time required for the vehicle to pass a fixed distance. The working principle is to set a plurality of virtual coils in the lane of the video image, and starting timing when the vehicle enters the virtual coil, at the same time the virtual coil begins to activate. When the vehicle leaves the virtual coil, the coil state is reset and the timing ends.

The virtual coil is set according to the known coordinate points of the actual road, and the arrangement is shown in Fig 1. The state of the virtual coil is mainly judged by the gray change. When the gray change of the image in the coil area exceeds threshold, the coil is activated and it is determined that the vehicle enters the virtual coil. Specific steps are as follows:

(i) Calculate the gray of foreground image named $F_I(x, y)$ and the gray of background image named $B_I(x, y)$ in the current frame virtual coil range.
(ii) Difference between foreground image grayscale and background image grayscale to obtain $D_I(x, y)$.
(iii) If the gray change is greater than the threshold value, it is considered that there is a vehicle passing through, and if the gray change is less than the threshold value, it is considered that no vehicle has passed or has left the coil.
(iv) Calculate the time difference between the activation of two virtual coils.

![Figure 1. Virtual coil arrangement and processing.](image)

3. Vehicle type recognition based on deep residual network

3.1. Residual network

Based on the traditional residual network and combining the characteristics of vehicle pictures, this paper attempts to improve the residual structure, reduce the number of network layers and improve the calculation speed and recognition accuracy. Finally, an improved deep residual network structure is proposed. The specific parameters are shown in Table 1. This network inputs $192 \times 192 \times 3$ RGB images. In the first layer, it adopts the original ResNet strategy that 64 convolution kernels of $7 \times 7$ is used, and the stride is 2. Then connected with the maximum pooling layer of $3 \times 3$, where overlapping pooling is used to enhance the robustness of target deformation. After the maximum pooling layer, 8 residual units are connected, and then connected with the average pooling layer, the output is $1 \times 1 \times 2048$. Finally, it acts on the full connection layer of 5 neurons, and the total number of layers is 26.
Table 1. Structure of the network in this paper

| Name       | Output size | Filter size/stride,pad |
|------------|-------------|------------------------|
| Input      | 192×192×3   |                        |
| Conv1      | 96×96×64    | 7×7/2,3,3×3/2          |
| MAX Pool   | 48×48×64    | 3×3/2                  |
| Conv2      | 24×24×256   | (1×1/1,0;3×3/2,1;1×1/1,0)×2 |
| Conv3      | 12×12×512   | {1×1/1,0;3×3/2,1;1×1/1,0}×2 |
| Conv4      | 6×6×1024    | (1×1/1,0;3×3/2,1;1×1/1,0)×2 |
| Conv5      | 6×6×2048    | {1×1/1,0;3×3/1,1;1×1/1,0}×2 |
| Avg-pooling| 1×1×2048    | 6×6/1                  |

The residual unit is as shown in the fig 2, the three convolutional layers are 1×1, 3×3, 1×1, where the 1×1 convolution layer act to change the dimension, and the 3×3 convolution layer inherits the performance of the VGG network. In addition, the unit is pre-activated, and all convolutional layers use ReLU as the activation function, that is, using batch normalization before the activation function and convolution. This structure greatly reduces the parameters and accelerates the training speed.

3.2. Centre loss function

The feature extraction ability of convolution neural network is generally evaluated by Softmax loss function, and the network parameters are optimized by feedback loss value. In the process of network training, with the increase of the number of iterations, the loss value decreases gradually, and the types of samples are separated.

The Softmax loss function is defined as follows:

\[
\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_i^T x_i + b_i}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}
\]

In the process of recognition, because of the complexity of data distribution, the distance of the same kind of data is often larger than that of different classes in the data space, which makes the Softmax loss function cannot guide the network to carry out classification learning. In order to solve the problem, this paper adds centre loss to constrain [9], the way is to control the offset of each type of sample and the centre of this kind of sample, so that the same kind of samples can be aggregated together.

The Centre loss function is defined as follows:

\[
\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^{m} \| x_i - c_{yi} \|^2
\]

Where \( c_{yi} \) represents the centre of features in the depth of i, \( x_i \) represents the feature before the fully connected layer, and m represents the size of the mini-batch.

Combine the central loss with the softmax loss, and use \( \lambda \) to control the proportion, defined as follows:
\[ \mathcal{L}_C = - \sum_{i=1}^{m} \log \frac{e^{w_i^T x_i + b_i}}{\sum_{j=1}^{n} e^{w_j^T x_i + b_j}} + \frac{1}{2} \sum_{i=1}^{m} \| x_i - c_i \|_2^2 \]  

(8)

4. Experiment and results

4.1. Datasets

The data in this paper comes from the road monitoring video under real conditions, and the images applied to vehicle recognition are extracted from six different scenes. According to the characteristics of vehicle load, this experiment classifies the vehicles into five categories, namely car, minibus, bus, truck, and large trailer, as shown in fig 3.

![Figure 3. The data sets of various types of vehicles](image)

4.2. Experimental results

The experiment is carried out on a lane of road through monitoring video. The video is 10 minutes 20 seconds long and the frame rate is 25 frames per second. In vehicle speed recognition, the real speed is obtained by outdoor reporting. In addition, cars, minibuses, buses, trucks and large trailers are recorded as 0, 1, 2, 3, 4. Finally, the experimental result is shown in Table 2, and the prediction result is shown in fig 4.

| Vehicle serial number | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|-----------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| Measured speed (km/h) | 78.2 | 66.7 | 56.7 | 92.1 | 82.2 | 85.4 | 82.1 | 83.4 | 90.1 | 60.0 | 84.2 | 96.3 |
| Actual speed (km/h)  | 75.3 | 68.2 | 52.1 | 95.4 | 80.2 | 83.6 | 83.1 | 85.7 | 94.2 | 63.7 | 88.2 | 93.1 |
| Identified type      | 0  | 0  | 3  | 0  | 0  | 0  | 1  | 1  | 0  | 4  | 2  | 1  |
| Actual type          | 0  | 0  | 3  | 0  | 0  | 0  | 1  | 1  | 0  | 3  | 2  | 1  |
| Vehicle serial number| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| Measured speed (km/h) | 54.4 | 77.1 | 75.7 | 62.8 | 63.6 | 84.2 | 91.0 | 83.3 | 57.9 | 62.8 | 75.5 | 78.6 |
| Actual speed (km/h)  | 57.7 | 73.2 | 71.4 | 65.4 | 65.7 | 86.2 | 93.5 | 81.3 | 59.8 | 60.2 | 71.5 | 73.4 |
| Identified type      | 4  | 2  | 0  | 0  | 0  | 0  | 1  | 4  | 1  | 1  | 2  |
| Actual type          | 4  | 2  | 0  | 0  | 0  | 0  | 1  | 3  | 1  | 1  | 2  |
4.3. Results analysis

4.3.1. Accuracy analysis of vehicle speed recognition. The experiment deals with the images of vehicle entering and leaving the virtual coil. Results show that the proposed method can process video in real time at 25 f/s and detect the speed of the vehicle. The detectable vehicle speed range is 0<S<200 km/h, and the average error is less than 6%.

4.3.2. Algorithm analysis of vehicle recognition. Fig 5 shows the visualization results of the feature map extracted by the network in this paper, where (a) is the input image, (b) represents the conv1 layer, (c) represents the conv2 layer, (d) represents the overall feature map in conv1, and (e) represents the overall feature map in conv2. We can find that the shallow network has higher resolution in feature map and contains more features, but it mainly extracts texture and detail features; the deep network extracts higher-order features, such as contours, headlights, license plates, the corresponding feature maps have lower resolution and the extracted features are more representative.

4.3.3. Model performance comparison. In order to test the performance, this model is compared with GoogLeNet VGG16 and AlexNet. At last, using the same test set to test the precision and recall of different models for five types of vehicles, and the results are shown in fig 6. From the comparison, it can be seen that the model in this paper has the best performance. For each type of vehicle, it has the highest precision, and the average precision is 85%. Aimed at minibus and cars, the precision reaches 98.7%, which is much better than that of GoogLeNet, VGG16 and AlexNet models.

5. Conclusions
In this paper, a real-time vehicle speed and type intelligent recognition system is established. In speed recognition, the technology of moving object detection with less parameters are applied, which takes up
fewer computing resources and meets the real-time needs. In vehicle type recognition, the advanced residual structure is used to establish the 26-layer network, and the high-order features are extracted by increasing the depth of the network, which enhances the robustness of the system in the real complex environment. In addition, the advantage of residual learning is used to reduce the network parameters, thus speeding up the convergence speed. Finally, combined with centre loss and softmax loss, the network learning is effectively guided to have more discrimination characteristics, and the accuracy of similar vehicle recognition is improved.

References
[1] Z. Li, K. Zhang, B. Chen, et al, Driver identification in intelligent vehicle systems using machine learning algorithms, IET Intelligent Transport Systems, 2019, 13 (1), 40–47.
[2] S. Natarajan, A. K. Annamraju, C.S. Baradkar, Traffic sign recognition using weighted multi-convolutional neural network, IET Intelligent Transport Systems, 2018, 12(10), 1396-1405.
[3] H. Hiary, H. Saadeh, M. Saadeh, et al, Flower classification using deep convolutional neural networks, IET Computer Vision, 12(6), 855-862.
[4] D. Liu, Z. Wang, Deep convolution neural networks for vehicle classification, Application Research of Computers, 2016, 33 (3), 930－932.
[5] L. Shi, Y. Wang, Y. Cao, Vehicle Type Recognition Based on Deep Convolution Neural Network, Computer Science, 2018, 45 (05), 280-284.
[6] Y. Yu, Q. Jin, Y. Fu, et al. Fine-grained Classification of Car Models Using Fg-CarNet Convolutional Neural Network, Acta Automatica Sinica, 2018, 44(10), 1864-1875.
[7] H. Pham, L. Khoudou, A. Crouzil, Learning and recognizing human action from skeleton movement with deep residual neural networks. 8th International Conference of Pattern Recognition Systems (ICPRS 2017), Madrid, Spain, July 2017.
[8] L. Yu, J. Chen, Y. Zhang, Deep Spectrum Prediction in High Frequency Communication Based on Temporal-Spectral Residual Network, China Communications, 2018, 15(09), 25-34.
[9] Y. Wen, K. Zhang, Z. Li, et al. A Discriminative Feature Learning Approach for Deep Face Recognition, in Leibe B., Matas J., Sebe N., Welling M. (Ed.): 'Computer Vision – ECCV 2016' (Springer Cham Press, 2016), pp.499-515