Sensors and Materials, Vol. 31, No. 2 (2019) 569–578
MYU Tokyo

S & M 1796

Fine-grained Vehicle Classification Technology Based on Fusion of Multi-convolutional Neural Networks

Wei Zhu,[1] Shaoyong Yu,[2] Xiaodong Zheng,[1*] and Yun Wu[3]

[1]School of Software Engineering, Xiamen University of Technology, Xiamen 361024, China
[2]College of Mathematics and Information Engineering, Longyan University, Longyan 364012, Fujian, China
[3]College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

(Received August 26, 2018; accepted November 22, 2018)

Keywords: vehicle detection, fine-grained classification, deep learning, CNN

With the development of cities, the rapid growth of vehicle ownership has given rise to traffic violations and traffic safety problems resulting in casualties. Therefore, with the rise of intelligent transportation in smart cities, intelligent traffic video monitoring systems have attracted considerable attention. The industry and academia have begun to consider the problem of increasing the intelligent functions of video monitoring systems. The intelligent functions of current intelligent traffic video monitoring systems focus on object detection and tracking, and abnormal situation alarms, for example. As the core functions of intelligent traffic video monitoring systems involve vehicle detection and classification of fine-grained problems, research in this area is very difficult. Moreover, no good, substantial products are available, especially with the fine-grained vehicle classification function, and no practical research has addressed this issue. In this paper, we propose an approach based on the fusion of convolutional neural networks (CNNs) to solve the problem of vehicle detection and fine-grained classification. In the fine-grained vehicle classification problem, the differences in a class are greater than the differences between classes. As a result, the classification accuracy is not sufficiently high to achieve efficient fine-grained vehicle classification.

1. Introduction

As an application field of target detection, vehicle detection has been widely used in many industries. After vehicle detection, fine-grained classification identification is required. Owing to the small difference between fine classification objects and the fact that the difference within a class is often greater than that between classes, it is difficult to study the problem of vehicle fine size classification. At present, there are no practical research results related to this challenging research topic.

*Corresponding author: e-mail: zxd@xmut.edu.cn
https://doi.org/10.18494/SAM.2019.2133

ISSN 0914-4935

Fine-grained classification is a subdomain of target identification, and its main purpose is to distinguish various subclasses under the same basic category. Different from the general coarse classification of objects, fine-grained classification distinguishes similar objects that are visually very similar. Take the fine-grained dogs classification as an example. Poodles and bichon are very similar in appearance. Fine-grained classification aims to accurately identify different dog species.

At present, research on fine-grained classification focuses on the expression of the image level to the expression of the mining semantic component level, and looks for clues in details. Training methods range from all end-to-end-oriented methods to the integration of a variety of methods, all of which make full use of prior knowledge in the training process. However, a large number of studies have focused on the problem areas of birds,[1–8] cats,[4] flowers,[4] aircraft,[8] dogs,[1,4,6,9] pedestrians,[10] and actions.[11] Relevant studies have also been carried out on the fine particle size classification of vehicles,[4,5,8,9] but there has been no specific research on the fine particle size classification of traffic. Few studies have addressed fine-grained vehicle classification owing to the lack of relevant standard data sets. Although in some early studies[12] vehicles were classified in a fine-grained way, those studies were mainly limited to the front and back views of vehicles. After the license plate was detected, previous studies extracted the region of interest (ROI) to generate a feature vector for classification. Stark *et al.*[13] also achieved good results in their fine-grained vehicle classification studies using the deformable part model (DPM). Prokaj and Medioni[14] used a 3D model of a vehicle to perform vehicle pose estimation, then projected it onto the 2D plane, and finally used the scale-invariant feature transform (SIFT) operator to compare different vehicles so as to classify vehicle fine-grain size. Their approach can solve the problem of inaccurate classification. Krause *et al.*[15] used a 3D CAD model to train shape classifiers to further improve the classification effect of Prokaj and Medioni.[14] Lin *et al.*[16] proposed the use of a 3D active shape model to capture vehicle trademark components so as to achieve fine-grained vehicle classification, and obtained better classification results than by other methods on their FG3DCar data set. Krause *et al.*[17] proposed the use of the convolutional neural network (CNN) method to study the distinctive components of a vehicle using these significant pieces to achieve fine-grained classification. All of the above works are based on the premise that all incoming vehicle images are "pure" images without complex backgrounds. Recently, Yang *et al.*[18] proposed the use of a convolution neural network for fine-grained vehicle classification and then regression of parameters. Krause *et al.*[19] also used the R-CNN method and combined the joint segmentation and automatic parts localization method to solve the problem of some components being unmarked.

In fact, vehicles are rigid and have their own characteristics, such as a unified structure. Each vehicle is composed of several fixed types of components, and the relative position between components is fixed. In addition, vehicles have symmetrical characteristics that can be applied to fine-grained classification models. Therefore, designing a fine-grained classification model applicable to vehicles by studying the existing fine-grained classification model and combining the characteristics of vehicles is theoretically and practically supported. So far, no fine-grained vehicle classification methods are applicable to practical transportation. Therefore, it is necessary to study this topic. We attempt to solve the problem of fine-grained vehicle classification through deep learning.

2. Analytic Structure

The definition of fine-grained vehicle classification is to classify vehicles by brand, car series, model, and year, such as identifying a vehicle as a "Buick-Weilang-Sedan-2018." Because classification is carried out within the categories of segmentation, the differences between objects are often very small and most of the time fine class differences within the larger than fine class differences. Thus, to achieve fine-grained classification, the core consideration is to find the significant characteristics of a small class, and then to identify the rich semantic components.[20–23] The problem of fine-grained vehicle classification is to identify the various components of the vehicle and distinguish them in accordance with their differences. Therefore, as long as the component detector of different components of the vehicle can be trained, the positions of the different components of the vehicle and the confidence of the vehicle model of each component can be detected using the input image, and then the result of each component detector can be combined to determine which fine-grained classification the vehicle image is most likely to belong to. Finally, the fine-grained vehicle classification problem with high similarity can be transformed into a vehicle component classification problem with a large difference.

2.1 Vehicle component detection model

A vehicle is a rigid body with a fixed structure that can be divided into 13 components: ceiling, headlight, inlet gate, hood, front windshield, tail lights, rear windshield, rearview mirror, front side door, back side door, trunk, front logo, and back logo. The 13 different components of the vehicle require 13 different components to be trained. When designing different component detection models, the improved Faster R-CNN model is still adopted. Taking the headlamp detection model as an example, it is processed in the following order: vehicle image, deep learning model of headlight detection, headlight subgraph, as seen in Fig. 1.

The model shown in Fig. 1 can detect the position of vehicle components from the vehicle image. Each part of the vehicle is represented by a rectangular area. Owing to the different sizes of each part, the proportion of the area in the whole vehicle image is different. Therefore, different components adopt different anchor sizes in the candidate area extraction network

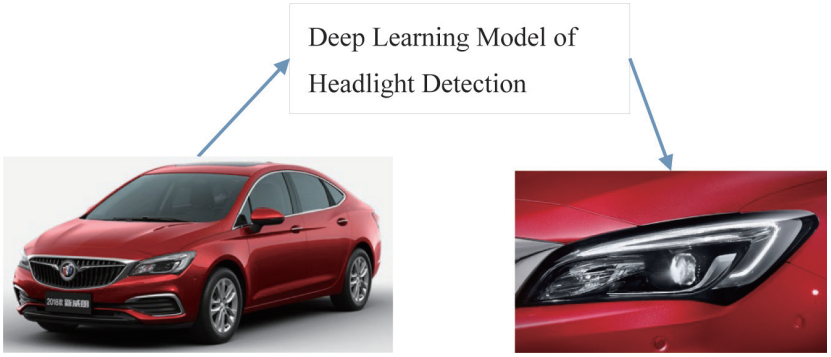


Fig. 1. (Color online) Flow chart of headlight detection model.

named the regional proposal network (RPN). Table 1 shows the average percentage of rectangular areas representing different components in the whole vehicle rectangular area and the ratio of length to width of rectangular areas representing each part. The purpose of the statistics on the average percentage of vehicle area and ratio of different components is to optimize the anchor settings in the RPN, so that the RPN can scan the whole image in accordance with these statistical data when extracting candidate areas, and reduce invalid candidate areas. Only in this way can the RPN achieve the best detection accuracy and speed.

After a component is detected by a component detector, it is necessary to judge to which part the vehicle component belongs. For example, when a headlight is detected, it is necessary to further determine the type of vehicle and its corresponding confidence level. When the production of polyhydroxyalkanoates (PHAs)[24] is used to compare the detected components with the component image in the component library of the dataset, it is possible to find the component image with the highest degree of similarity in the component library. The model category of the image of the corresponding component in the component library is the category of the model used to detect the component and the value of similarity is its confidence.

The steps of the algorithm are as follows.

(1) Reduce the image size to 8 × 8. The image has 64 pixels in total. The purpose of this is to remove the details of the image and retain only basic information, such as image structure/brightness, and eliminate the difference caused by image size or aspect ratio.
(2) Remove the color information in the image and retain only the grayscale level.
(3) Average the grayscale values of the 64 pixels.
(4) Compare the grayscale value of each pixel with the average value obtained in step (3). Those smaller than the average value are denoted as 0, and those greater than or equal to the average value are denoted as 1.
(5) Combine the results of the comparison in step (4) into a 64-bit string in order from top to bottom, and left to right; this 64-bit string is the fingerprint of the image.
(6) Calculate the "hamming distance" of two image fingerprints, i.e., the number of the same characters in the same position, and divide it into 64 to obtain the similarity of the two images.

Table 1
Average percentage and ratio of length to width.

Part name	Average percentage of vehicle area (%)	Ratio of length to width
Ceiling	8	4:1 3:1 2:1
Headlight	5	3:2 1:1
Intake gate	5	4:1 3:1 2:1
Hood	9	3:2 1:1
Front windshield	11	2:1 5:3
Tail light	3	3:2 1:1
Rear windshield	4	2:1 5:3
Rear view mirror	2	3:2 1:1
Front side door	6	3:2 2:1
Back side door	5	3:1 2:1
Trunk	3	2:1 5:3
Front logo	2	3:2 1:1
Back logo	2	3:2 1:1

The final results of the headlight test described in Fig. 1 indicate that the big lamp belongs to a Buick Weilang Sedan 2018, and the confidence level is 0.9.

2.2 Fine-grained vehicle classification model based on fusion of multi-CNNs

The model of each component adopts CNN training of a neural network that can detect the components with a total of 13 neural networks. After entering the vehicle image through these 13 networks, the classification and confidence of the 13 components of the vehicle are obtained, and then the classification of the vehicle is voted on. The confidence is the voting value. The sum of all voting confidence values obtained from each classification is the voting value of the classification. The classification with the highest voting value is taken as the classification of the vehicle image. The average of all voting confidence values of the classification is taken as the confidence of the classification.

$$\mathrm{B} = \frac{\max\left\{S_1, S_2, \ldots, S_j\right\}}{Maximum\,number\,of\;votes\;for\,the\,sum\,of\;confidence} \tag{1}$$

Note that Vi = {Cj, Bi} represents the ith vote and i = 1, 2, ..., 13, Cj, denotes the vote of the jth model, j = 1, 2, ..., N, N denotes the total class of vehicles, and B is the confidence of the ith vote, $0 \leq \mathrm{Bi} \leq 1$. Sj is the sum of the confidence in all voting obtained for the jth vehicle type. The vehicle model of the image is the jth type corresponding to the maximum of Sj, and the confidence level is B.

For the quick and efficient detection of 13 vehicle components using the CNN, the input image used an improved image of a Faster R-CNN, which represents the vast majority of vehicle after removing background graphics. A test image of a Faster R-CNN after detection is taken, as shown in Fig. 2. The whole vehicle image is input into 13 vehicle part detectors to obtain 13 confidence detection results. The confidence data obtained from the voting for these five vehicles is shown in Table 2.

Fig. 2. (Color online) Test picture.

Table 2
Confidence detection and voting results of 13 component detectors.

Component	Vehicle type				
	Toyota Corolla Sedan 2017 Confidence	Toyota Camry Sedan 2018 Confidence	Buick Regal Sedan 2017 Confidence	Buick Weilang Sedan 2015 Confidence	Porsche Macan SUV 2017 Confidence
Ceiling			0.7		
Headlight	0.9				
Intake gate	0.9				
Hood					0.8
Front windshield	0.7				
Tail light	0.7				
Rear windshield				0.7	
Rear view mirror		0.6			
Front side door	0.9				
Back side door	0.8				
Trunk				0.8	
Front logo	0.9				
Back logo	0.9				
Total votes	6.7	0.6	0.7	1.5	0.8

As can be seen from the voting results in Table 2, the total votes for the 5 types of vehicles are, respectively, 6.7, 0.6, 0.7, 1.5, and 0.8, and the maximum voting totals of 6.7 are taken as the fine-grained classification of the image. Therefore, in this case, the model image belonging to a fine-grained classification is a Toyota Corolla Sedan 2017 model. Using Eq. (1), the corresponding confidence level is found to be 6.7/8 = 0.8375.

3. Experiment and Improvement

3.1 Experiment

Because of the thousands of vehicles that have been seen so far, we need a huge amount of data to be able to identify the 13 vehicle components, and in this experiment, we apply a series of five models for training: Buick Regal Sedan 2017, Buick Weilang Sedan 2015, Toyota Corolla Sedan 2017, Toyota Camry Sedan 2018, and Porsche Macan SUV 2017.

For each of the components of the vehicle, about 2000 images were collected at an average of about 400 images per model, with 300 images being used to train the component detectors and 100 images used as verification sets. In the fine-grained classification test stage, 96 images of the above 5 models and 4 images of the Zhongtai SR9 were input for classification testing, and the accuracy was 68%. We identified the correct examples, as shown in Fig. 3.

As seen in Fig. 3, the fine-grained classification model successfully identified the Toyota Corolla and Camry and the Buick Regal and Weilang. During the experiment, some typical errors arose, as shown in Fig. 4.

In the results of the experiment shown in Fig. 4, the left image correctly classifies the Porsche Macan, but the right image mistakenly classifies the Zhongtai SR9 as a Porsche Macan. The Zhongtai SR9 imitates the appearance of the Porsche Macan, with only some minor details, and the resolution of the image is not sufficiently high. Therefore, these details are difficult to learn, thus leading to a system classification error.

(a) (b)

(c) (d)

Fig. 3. (Color online) Successful fine-grained vehicle classification. (a) Buick Regal Sedan 2017, (b) Buick Weilang Sedan 2015, (c) Toyota Corolla Sedan 2017, and (d) Toyota Camry Sedan 2018.

Fig. 4. (Color online) Errors in fine-grained vehicle classification of Porsche Macan SUV 2017.

3.2 Improvement

Although this method can classify vehicles with 68% accuracy, it is still far from being a practical application owing to the following problems.

(1) So far, the experiment has only been conducted for five models. The rate of change as new models are added has not yet been established.

(2) The categorization is too slow. The average speed of classifying 100 images is three times per second, which is far from the requirement of real-time performance. The main reason is that 13 independent CNNs are used in this study for component detection. The operations between different CNNs are not shared, and there are many repeated calculations.

Table 3
Results of fine-grained classification and CompCars classification.

	Top-1	Top-5
CompCars model	0.767	0.917
Model in this paper	0.791	0.941

Table 4
Classification accuracy before and after expansion of CompCars.

	Top-1		Top-5	
	Original	Extension	Original	Extension
CompCars model	0.767	0.757	0.917	0.908
Model in this paper	0.791	0.823	0.941	0.953

To further verify the classification performance of the fine-grained classification model, experiments were carried out using the CompCars[25] data set. Since the images of vehicle components in CompCars only include headlights, tail lights, and air intake gates, only a three-component detection network is allowed. The fine-grained classification results of the experiment are shown in Table 3.

It can be seen from Table 3 that the comparison of the Top-1 and Top-5 classification accuracies of the two models indicates that the fine-grained classification model adopted in this study is superior to the CompCars' classification model. It also expands upon CompCars. Since CompCars does not include the images of vehicle ceilings, hoods, front windshields, rear windshields, rearview mirrors, front side door, back side door, trunks, front logo, and back logo, we added 21,458 images to the CompCars data set, and then used the 13 component detectors in the fine-grained classification model herein to conduct our experiment. The classification results are shown in Table 4.

It can be seen from Table 4 that after the expansion of the CompCars' data set, the fine-grained classification effect of the CompCars' original classification model on the new data set changed minimally (both top-1 and top-5 have reduced accuracies). However, the model in this work has greatly improved classification accuracies of top-1 and top-5. The main reason is that the increased number of component detectors in the fine-grained classification method result in the comparison of more details.

4. Conclusions

In this work, we studied fine-grained vehicle classification technology based on the fusion of multi-CNNs. For vehicle images with complex backgrounds, the model first detects the vehicle area and then inputs the area into the fine-grained classification model for classification. This method filters the input of the fine-grained classification model, reduces the noise interference, and significantly improves the accuracy and speed of fine-grained classification.

We proved that the fusion of multi-CNNs can achieve fine-grained vehicle classification. The proposed method divides the vehicles into 13 components, trains one detector for each part, and then votes in accordance with the test results of the 13 components to classify the input image. The experimental results show that this method is effective, but the classification speed

has yet to be improved. It is hoped that this study will provide a reference for the application of the fine-grained vehicle classification technology based on the fusion of multi-CNNs.

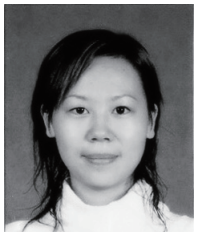
Acknowledgments

This work is supported by the Young Teachers' Education Scientific Research Project of Fujian province (Grant No. JT180455), Xiamen Science and Technology Foundation (Grant 3502Z20173035), Fujian Provincial Natural Science Foundation of China (Grant 2018J01570), and National Science Foundation of Fujian Province of China (No. 2018J01575).

References

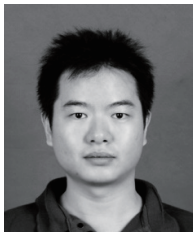
1 T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang: IEEE Conf. CVPR (2015) 842–850.
2 Z. Akata, S. Reed, W. Daniel, L. Honglak, and B. Schiele: IEEE Conf. CVPR (2015) 2927–2936.
3 N. Zhang, J. Donahue, R. Girshick, and T. Darrell: European Conf. CVS (2014) 834–849.
4 Q. Qi, J. Rong, Z. Shenghuo, and L. Yuanqing: IEEE Conf. CVPR (2014) 3716–3724.
5 K. Jonathan, J. Hailin, Y. Jianchao, and L. Fei-Fei: IEEE Conf. VPR (2015) 5546–5555.
6 L. Di, S. Xiaoyong, L. Cewu, and J. Jiaya: IEEE Conf. CVPR (2015) 1666–1674.
7 G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie: IEEE Conf. CVPR (2015) 1449–1457.
8 L. Tsung-Yu, A. RoyChowdhury, and S. Maji: IEEE Conf. Computer Vision (2016) 1449–1457.
9 X. Saining, Y. Tianbao, W. Xiaoyu, and L. Yuanqing: IEEE Conf. CVPR (2015) 2645–2654.
10 D. Hall and P. Perona: IEEE Conf. CVPR (2015) 5482–5491.
11 Z. Yang, N. Bingbing, H. Richang, W. Meng, and T. Qi: IEEE Conf. CVPR (2016) 3323–3331.
12 V. S. Petrovic and T. F. Cootes: British Machine Vision Conf. (2004) 587–596.
13 S. Michael, K. Jonathan, P. Bojan, M. David, J. J. Little, B. Schiele, and K. Daphne: British Machine Vision Conf. (2012) 228–236.
14 J. Prokaj and G. Medioni: IEEE Workshop ACV (2013) 1–7.
15 K. Jonathan, S. Michael, J. Deng, and L. Fei-Fei: IEEE Conf. ICCV (2013) 554–561.
16 L. Yen-Liang, V. I. Morariu, H. Winston, and L. S. Davis: European Conf. Computer Vision (2014) 466–480.
17 K. Jonathan, G. Timnit, D. Jia, L. Li-Jia, and L. Fei-Fei: 22nd Int. Conf. Pattern Recognition (2014) 26–33.
18 L. Yang, P. Luo, C. Change Loy, and X. Tang: IEEE Conf. CVPR (2015) 3973–3981.
19 K. Jonathan, J. Deng, S. Michael, and L. Fei-Fei: 2nd Workshop on Fine-Grained Visual Categorization (FGVC2) (2013) 1–2.
20 H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas: IEEE Conf. CVPR (2016) 1143–1152.
21 S. Huang, Z. Xu, D. Tao, and Y. Zhang: IEEE Conf. CVPR (2016) 1173–1182.
22 Y. Cui, F. Zhou, L. Yuanqing, and S. Belongie: IEEE Conf. CVPR (2015) 1153–1162.
23 W. Yaming, C. Jonghyun, V. Morariu, and L. S. Davis: IEEE Conf. CVPR (2016) 1163–1172.
24 L. Wai, W. Yujie, C. Pui Ling, C. Shun Wan, T. Yiu Fai, C. Hong, and P. Hoi Fu Yu: Environ. Technol. **38** (2017) 1779.
25 Y. Linjie, L. Ping, C. Change Loy, and T. Xiaoou: IEEE Conf. CVPR (2015) 3973–3981.

About the Authors



Wei Zhu is a college lecturer in School of Software Engineering at Xiamen University of Technology. Her research interests are in the areas of software development, information security, and artificial intelligence. (zhw@xmut.edu.cn)

Shaoyong Yu received his Ph.D. degree from Xiamen University in 2017. He is also a college lecturer in the School of Mathematics and Information Engineering, Longyan University. His research interests are in the areas of computer vision and deep learning. (syyu@xmut.edu.cn)



Xiaodong Zheng is a lecturer in School of Software Engineering at Xiamen University of Technology. His research interests are in the areas of software development, information security, and artificial intelligence. (zxd@xmut.edu.cn)

Yun Wu received her Ph.D. degree from Xiamen University in 2007. Her research interests are in the areas of artificial intelligence and big data. Her scientific contributions to AI are on soft computing and clustering algorithms.