Vehicle Type Classification via Adaptive Feature Clustering for Traffic Surveillance Video

Shu Wang[1], Feng Liu[1,2], Zongliang Gan[1,2], Ziguan Cui[1,2]
1. Jiangsu Province Key Lab on Image Processing & Image Communications, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Key Lab of Broadband Wireless Communication and Sensor Network Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
Corresponding author: liuf@njupt.edu.cn

***Abstract*—Vehicle type classification has become an important part of intelligent traffic. However traditional methods can not deal with the varying situations in the reality. In this paper, a novel method is proposed to handle this task in the real road traffic surveillance video. In order to distinguish different vehicles, we categorize vehicles into three types: compact cars, mid-size cars, and heavy-duty vehicles. For a certain video, our method has four steps. First, a deep convolutional neural network is used to detect vehicles in the candidate region and a data set would be generated. Second, the main features of vehicles can be extracted using a fully-connected network. Also, for the sake of higher accuracy, weak labels given by pre-trained extreme learning machine (ELM) are fused into the final features, adding prior information proportionally. Third, K-means is implemented to learn three vehicle-type cluster centers adaptively. Finally, vehicle type will be recognized according to the closest distance principal. Experimental results show that the recognition rate outperforms other traditional methods, verifying the feasibility and effectiveness of the proposed method.**

***Keywords*—vehicle type classification; adaptive clustering; feature learning; deep learning; extreme learning machine.**

I. Introduction

Vehicle detection and classification has become a significant task in machine learning because of its potential applications, such as intelligent traffic systems [1] and autonomous driving systems. Besides, few vehicle types such as truck will not be expected to appear in some particular locations, so vehicle type classification is of value in surveillance videos which can reduce the security risks in our daily lives. However, it is challenging to tackle this problem because of the complexity of surveillance videos. The difficulties of vehicle type classification come from several aspects. First, in low-resolution surveillance videos, the vehicle features will be hard to extract. Because the limited size and low quality of videos always make vehicle images textureless. Second, different road conditions and varying illumination conditions make the problem more complicated [2]. And surveillance cameras sometimes have different viewpoints, increasing the difficulties in distinguishing vehicle types. Third, vehicles belong to the same type always have different appearances due to their different colors, postures and manufacturers. Meanwhile, some vehicles belong to different types are similar in appearance.

In recent years, numerous algorithms have been proposed to tackle this task. Chen [3] introduced a feature extraction method based on sparse learning and trained a linear SVM for vehicle classification. Karaimer [4] also used SVM for classification, but they employed shape-based and HOG features of vehicle images. Dynamic Bayesian Network (DBN) [5], [6] is another effective way to classify vehicles in traffic videos, because DBN has the ability to visualize the relationship of random variables. Nurhadiyatna [7] proposed a real-time vehicle classification framework using Gabor filters to extract features. With the remarkable success of deep learning, Zhang [8] used deep convolutional neural networks (CNN) to classify vehicle types, which do not require the finely cropped vehicle images. Qian [9] combined the high-layer features of deep network and some traditional features such as local binary patterns, and achieved high accuracy in their experiments.

Naturally, according to the vehicle size, vehicles can be divided into three categories: compact cars (sedans, taxis), mid-size cars (vans), and heavy-duty vehicles (buses, trucks), which are shown in Fig. 1. Unfortunately, traditional methods fail to adapt to different situations in real-world applications. In this paper, we propose a more smart method for vehicle type classification, which allows the algorithm to adaptively deal with the varying situations in the real surveillance. When it comes to a particular surveillance video, we firstly detect vehicles from a region of interest (ROI) by using a deep network based on Fast-RCNN [10]. And we can obtain a small data set for vehicle images if the detection process is applied to the front part of the video. Then we introduce a new method

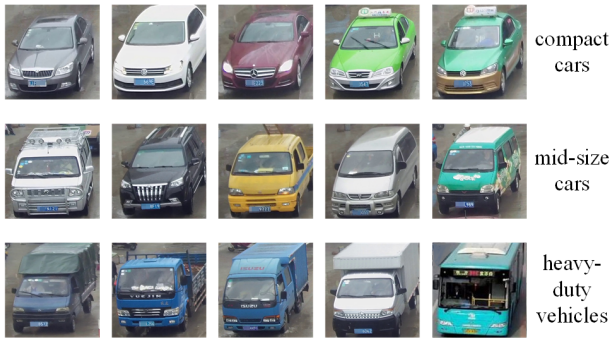


Fig. 1: Three different types of vehicles.

978-1-5090-2860-3/16/$31.00 ©2016 IEEE

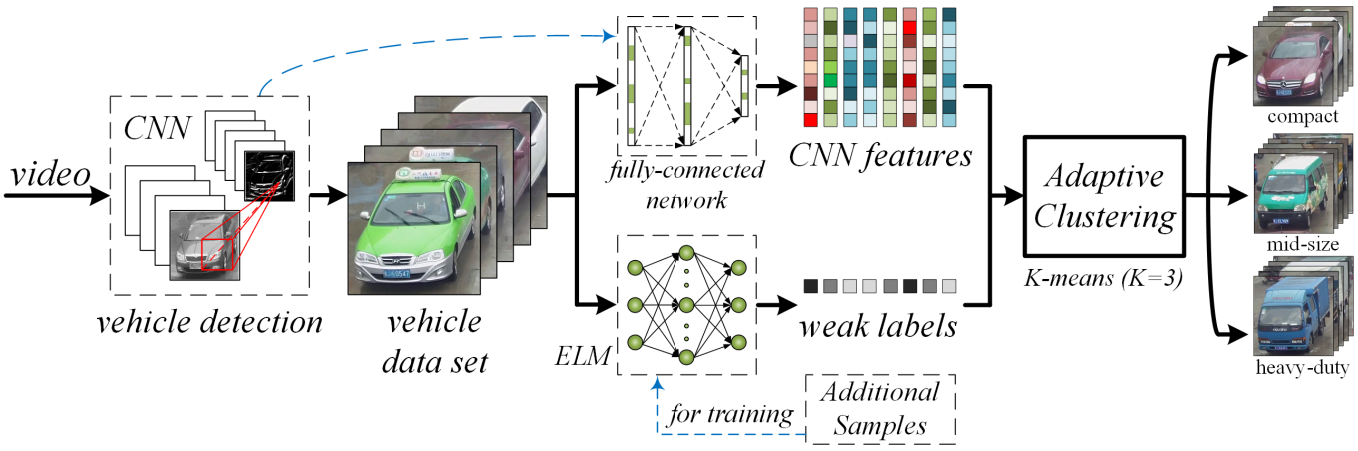


Fig. 2: Flowchart of the proposed method.

based on neural network to extract vehicle image features, combining a label from a pre-trained ELM [11] vehicle-type classifier. Finally, the feature vectors are clustered into three clusters via k-means method. It is shown that the clustering results can largely distinguish these three types of vehicles. As a result, we can determine the type of a vehicle in the latter part of the video by comparing the distances between a vehicle feature vector and three different cluster centers. Our experimental results show that the vehicle detection rate is over 98%, and the vehicle type recognition rate is about 2.76% higher than some other traditional methods.

II. Proposed Method

Our proposed method can effectively classify three kinds of vehicles in surveillance videos. The flowchart of the method is illustrated in Fig. 2. And the algorithm has four steps. First, vehicles can be detected from a selected ROI in a video sequence and a small data set of vehicle images is obtained. Second, the main vehicle-type features will be extracted from deep network, and the weak labels obtained from a pre-trained ELM classifier will be fused with the main features. Third, K-means is used for unsupervised learning in high-dimensional feature space. Finally, the last step is the identification step.

A. Vehicle Detection and Dataset Generation

Firstly, we set a fixed region of interests in a particular video sequence. Setting ROI has several advantages such as limiting the vehicle images to a suitable size range. Also, by setting proper candidate regions, the detection area will be decreased and the integrity of vehicles can be guaranteed.

Then, a pre-trained Fast-RCNN network which can output accurate vehicle coordinates has been implemented to detect vehicles in the selected region. And the vehicle images will be obtained from a video frame. If the detection process is applied to the front part of the video, a vehicle data set for the particular video scene would be generated in order to complete the following steps.

B. Feature Extraction

The features in this paper consist of two distinct parts: the features excavated by deep network, and the priori features as a form of weak labels, which we can see in Fig. 2.

(1) Features from Deep Network

From the most intuitive point of view, the main features of vehicle types can be excavated from the global information of vehicle images, rather than the local information. As a result, after obtaining the accurate coordinates of vehicle images using Fast-RCNN, we establish a fully connected neural network to classify three different vehicle types. The fact is that, in the fully-connected networks, the penultimate layer outputs can largely reflect the abstract characteristics of vehicle types, and can be used for predicting which type a vehicle is. Therefore, we use the weights in the front fully-connected layers as the filters to extract features.

$$feature_1 = g(\boldsymbol{W} \cdot \boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

where $\boldsymbol{W}$ denotes the weight matrix between the input layer and the penultimate layer. $\boldsymbol{b}$ is the threshold vector. $\boldsymbol{x}$ and $feature_1$ denote the input vector and the output features respectively. $g(\boldsymbol{x}) = (1 + exp(-\boldsymbol{x}))^{-1}$ is the sigmoidal activation function which normalize the data to 0-1 range.

(2) Priori Features

In order to better recognize vehicle types, a pre-trained ELM classifier is introduced to provide weak labels as priori features. As we can see in Fig. 3, we use cropped vehicle images as the training samples, and extract the CNN features [12] as the ELM input vectors.

The ELM classifier consists of three hidden layers with the node number 512-512-2000. In the first two hidden layers, two ELM auto-encoders are performed for unsupervised feature representation. it is known that the auto-encoder aims to learn a function $h_\theta(\boldsymbol{x}) \simeq \boldsymbol{x}$, which acts as some sort of feature extractor in multi-layer learning framework [13], [14]. And in the last hidden layer, random projection and supervised feature classification are used for learning the final classifier.

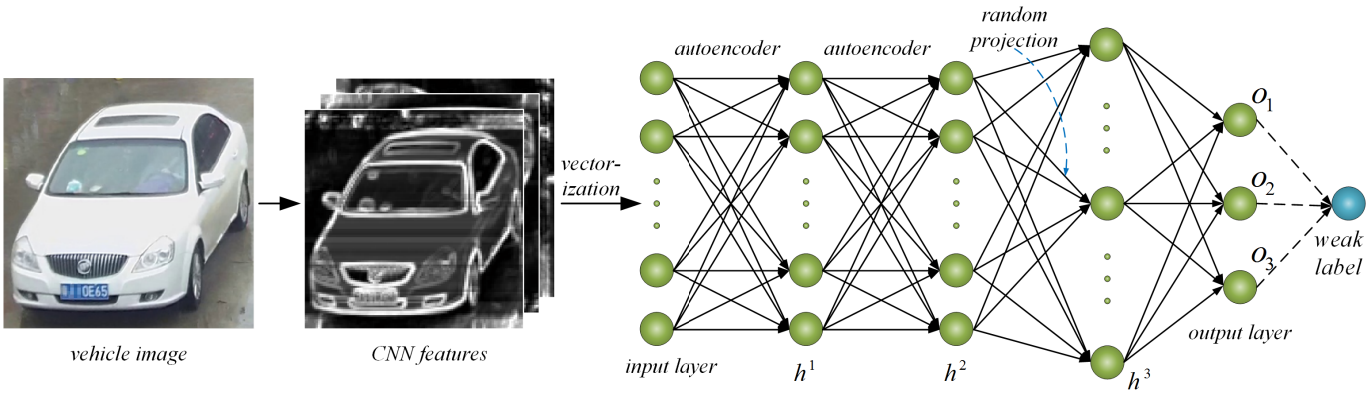


Fig. 3: Feature extraction and learning using ELM.

As the ELM network has three output nodes which indicate three different types of vehicles, we translate the ELM outputs into a weak label $feature_2$, which can be denoted as the following equation:

$$feature_2 = \underset{j}{\operatorname{argmax}}(o_j) - 2, \quad j = 1, 2, 3 \tag{2}$$

o_j denotes the j-th ELM output node. As a result, the weak label can be three values: -1, 0 or +1, implying the priori features of three vehicle types.

(3) Feature Fusion

After obtaining the features from deep network and the priori features, we can merge them together according to the queue model.

$$f^{(i)} = [feature_1^{(i)}, \lambda \cdot feature_2^{(i)}] \tag{3}$$

$f^{(i)}$ is the features of the i-th vehicle image in dataset, and $feature_1^{(i)}$ and $feature_2^{(i)}$ are the corresponding high-layer features from deep network and the corresponding weak label respectively. λ is a positive parameter which can adjust the weights between two kinds of features.

C. Unsupervised Learning

Theoretically, the fused features have the ability to make a distinction between different vehicle types, that is to say, the features of different vehicle types can be easily distinguished in the high-dimensional data space. Consequently, machine can get the vehicle type difference by unsupervised learning with vehicle features.

Our aim is to partition the vehicle images into 3 clusters in which all the images in one cluster belong to the same vehicle type. K-means clustering algorithm is a feasible and effective method to accomplish this task. Through iterative refinement, we can obtain 3 clusters, which correspond to three different vehicle types (compact cars, mid-size cars, heavy-duty vehicles). For each cluster, there is a cluster center which is denoted as c_k, $k = 1, 2, 3$.

We find that the clustering results can largely distinguish three types of vehicles. That is because adaptive feature learning can avoid the effects come from the outside environment.

D. Recognition and Classification

When it comes to the recognition process in the latter part of the video, we could compare the distance between the fused features obtained from the detection results and three vehicle-type cluster centers obtained from unsupervised learning. According to the shortest distance principle, if an extracted feature is very close to certain cluster center in the high-dimensional feature space, the vehicle with this feature would have a close relationship with the label corresponding to this cluster center. So, we can get the index of the closest cluster center to the vehicle feature f, which is denoted as r.

$$r = \underset{k}{\operatorname{argmax}} ||f - c_k||, \quad k = 1, 2, 3 \tag{4}$$

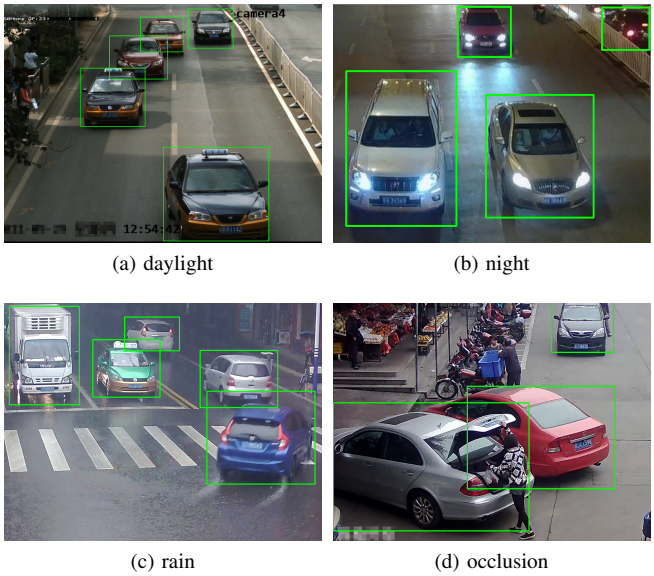
Then, the recognition result of a certain vehicle in the latter part of the video would be the label which binds to the index r.

III. Experimental Results

The Implementation of vehicle detection and classification algorithms on all the video sequences are carried out in Visual Studio 2013 and Matlab R2014a environment running in Core i5, 3.2GHZ CPU with 8-GB RAM. The training process of Fast-RCNN and fully-connected network are implemented in Sugon I450, with NVIDIA Tesla K20C for GPU parallel computing. For deep network based vehicle detection, 11,296 road surveillance images with 26,665 vehicle instances are carefully selected for deep learning. Also, in case of the overfitting problem in detection, 10,000 irrelevant images are served as negative training samples. For ELM training process in vehicle type classification, 13,860 vehicle images with 4,620 images for each vehicle type are used for learning the final ELM network weights. Therefore, our experimental results are based on the training with a great deal of image data.

A. Vehicle Detection

In our experiment, the model of deep neural network is *VGG_CNN_M_1024* which has a great capability to achieve high detection rate [15]. Meanwhile, compared with *VGG16* [16] or *OverFeat* [17], this model also has lower computational



(a) daylight (b) night

(c) rain (d) occlusion

Fig. 4: Detection results under different situations.

complexity, making it possible for the implementation in the real traffic surveillance. Also, for pre-processing, we use selective search [18], [19] to generate object proposals. The confidence threshold is set to 0.8, and the threshold of non-maximum suppression is set to 0.3.

The detection rate of our system is 98.63% in the real traffic videos, and the error detection rate is less than 0.1%. It takes about 12 hours for the training process with 100,000 iterations. And in the testing process (for 1080p videos), it takes 1.5s/frame with CPU or 150ms/frame with GPU. Moreover, due to the generalization ability of neural networks [20], our system performs well in various situations, such as daylight, rain, night and occlusion, as shown in Fig. 4.

B. Vehicle Type Classification

In our experiment, the fully connected network has 3 layers, with 128×128 nodes in the input layer and 128 neural nodes in the hidden layer. For the weak labels, the ELM network takes the CNN features with 512 dimensions as inputs. Two auto-encoders both have 512 nodes and the random projection layer has 2000 nodes. All the node numbers are the optimal values obtained from several experiments.

So, $feature_1$ is a 128-dimensional vector with 0-1 range, and $feature_2$ is a numerical variable with value 0, +1 or -1. The parameter of feature fusion λ is set to a low value, such as 4 or 5. As shown in Table I, the proposed framework can get 85.56% accuracy in real traffic surveillance, better than other methods with the same testing datasets.

TABLE I: THE VEHICLE TYPE RECOGNITION ACCURACY OF SEVERAL METHODS

Methods	Accuracy (%)
HOG+SVM [4]	81.42
HOG+ELM	82.80
The proposed method	85.56

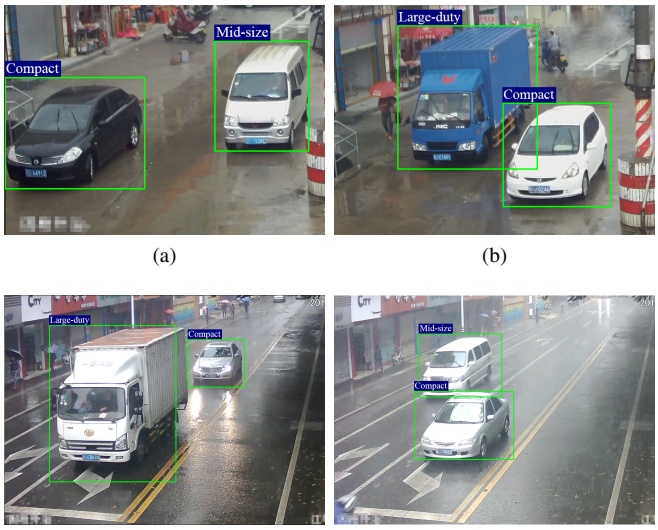
TABLE II: THE CONFUSION TABLE BETWEEN DIFFERENT VEHICLE TYPES

Actual \ Predict	Compact	Mid-size	Heavy-duty
Compact	**88.26**	9.61	2.14
Mid-size	6.60	**85.85**	7.55
Heavy-duty	16.33	12.24	**71.43**

Table II further shows the confusion matrix between three different vehicle types. As we can see, for compact vehicle which is the most common vehicle type, the recognition accuracy is higher than other two types. That is because the training samples of compact cars are abundant in our sample library, which enable the proposed model fit well with the experimental data. However, the accuracy on heavy-duty vehicle is relatively lower than other types, because heavy-duty vehicles have more variations, which lead to the difficulties of recognition in the real surveillance.

Moreover, from Fig. 5 and Fig. 6, we can see the vehicle type classification results of the proposed method. Among them, Fig. 5 is the results under the scene of urban street, while Fig. 6 demonstrates the results for the scene in highway. Also, our experimental results are robust to the different viewpoints of surveillance cameras. All these video sequences come from traffic surveillance in the reality.

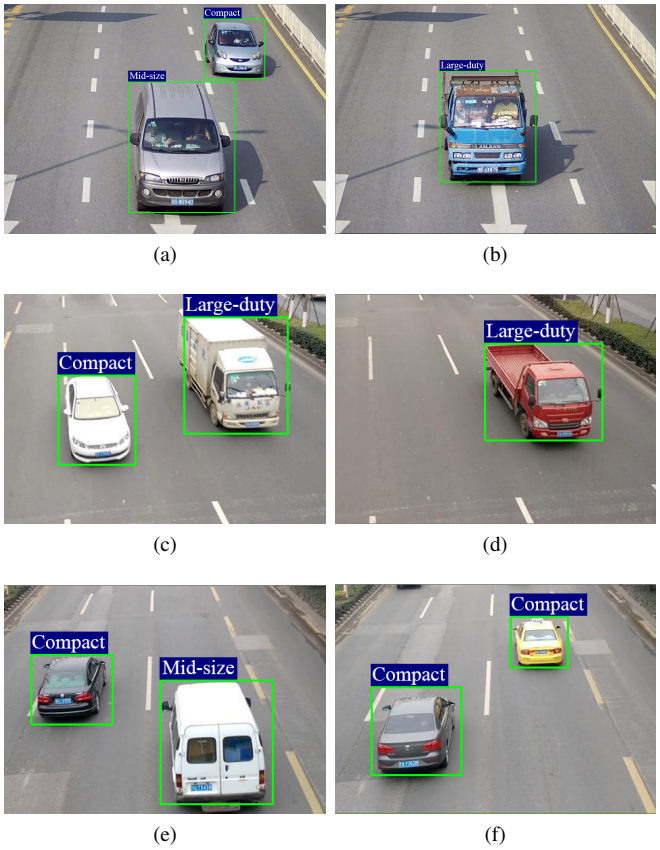
Compared with other traditional methods which can not adaptively deal with the various situations, the proposed method can perform well in the low-resolution traffic surveillance under different scenes.



(a) (b)

(c) (d)

Fig. 5: Vehicle type classification results for street scene.



(a) (b) (c) (d) (e) (f)

Fig. 6: Vehicle type classification results for highway scene.

IV. Conclusion

In this paper, a novel method is proposed for vehicle type classification. Experimental results show that the proposed method outperforms other traditional methods, and has a very wide application prospects. It is because this framework can avoid the uncontrolled effects of environment, such as hash image conditions, illumination conditions and various road situations. Also, our method can be extended to an online learning framework if the processing power of servers is great enough. From the above, we truly believe that the proposed method can be widely implemented in the real surveillance systems and other classification problems.

Acknowledgment

This work is supported by National Natural Science Foundation of China (NSFC) (61501260, 61471201), Natural Science Foundation of Jiangsu Province (BK20130867), Jiangsu Province Higher Education Institutions Natural Science Research Key Grant Project (13KJA510004), The peak of six talents in Jiangsu Province (2014-DZXX-008), Natural Science Foundation of NUPT (NY214031), and "1311 Talent Program" of NUPT.

References

[1] B. F. Momin and T. M. Mujawar, "Vehicle detection and attribute based search of vehicles in video surveillance system," in *Circuit, Power and Computing Technologies (ICCPCT), 2015 International Conference on*, March 2015, pp. 1–4.

[2] A. Mukhtar, L. Xia, and T. B. Tang, "Vehicle detection techniques for collision avoidance systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2318–2338, Oct 2015.

[3] Z. Chen, N. Pears, M. Freeman, and J. Austin, "Road vehicle classification using support vector machines," in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, vol. 4, Nov 2009, pp. 214–218.

[4] H. C. Karaimer, I. Cinaroglu, and Y. Bastanlar, "Combining shape-based and gradient-based classifiers for vehicle classification," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, Sept 2015, pp. 800–805.

[5] Y. Liu and K. Wang, "Vehicle classification system based on dynamic bayesian network," in *Service Operations and Logistics, and Informatics (SOLI), 2014 IEEE International Conference on*, Oct 2014, pp. 22–26.

[6] M. Kafai and B. Bhanu, "Dynamic bayesian networks for vehicle classification in video," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 100–109, Feb 2012.

[7] A. Nurhadiyatna, A. L. Latifah, and D. Fryantoni, "Gabor filtering for feature extraction in real time vehicle classification system," in *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Sept 2015, pp. 19–24.

[8] F. Zhang, X. Xu, and Y. Qiao, "Deep classification of vehicle makers and models: The effectiveness of pre-training and data enhancement," in *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2015, pp. 231–236.

[9] H. Qian, Y. Zhang, and C. Liu, "Vehicle classification based on the fusion of deep network features and traditional features," in *Advanced Computational Intelligence (ICACI), 2015 Seventh International Conference on*, March 2015, pp. 257–262.

[10] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1440–1448.

[11] J. Xu, H. Zhou, and G. B. Huang, "Extreme learning machine based fast object recognition," in *Information Fusion (FUSION), 2012 15th International Conference on*, July 2012, pp. 1490–1496.

[12] M. Elmikaty and T. Stathaki, "Car detection in high-resolution urban scenes using multiple image descriptors," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, Aug 2014, pp. 4299–4304.

[13] J. Tang, C. Deng, and G. B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 809–821, April 2016.

[14] E. Cambria, G. B. Huang, L. L. C. Kasun, and et al, "Extreme learning machines," *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 30–59, Nov 2013.

[15] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3973–3981.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–14, 2016.

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013.

[18] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[19] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1879–1886.

[20] L. Yang, J. Liu, and X. Tang, *Object Detection and Viewpoint Estimation with Auto-masking Neural Network*. Cham: Springer International Publishing, 2014, pp. 441–455.