Vehicle Classification using Transferable Deep Neural Network Features

Yiren Zhou, Ngai-Man Cheung
yiren_zhou@mymail.sutd.edu.sg, ngaiman_cheung@sutd.edu.sg

arXiv:1601.01145v1 [cs.CV] 6 Jan 2016

***Abstract*—We address vehicle detection on rear view vehicle images captured from a distance along multi-lane highways, and vehicle classification using transferable features from Deep Neural Network. We address the following problems that are specific to our application: how to utilize dash lane markings to assist vehicle detection, what features are useful for classification on vehicle categories, and how to utilize Deep Neural Network when the size of the labelled data is limited. Experiment results suggest our approach outperforms other state-of-the-art.**

***Index Terms*—Vehicle Classification, Deep Neural Network**

I. Introduction

Vehicle detection and classification are important parts of Intelligent Transportation Systems. They aid traffic monitoring, surveillance, and traffic counting, which are necessary for tracking the performance of traffic operations. Existing methods use various types of information for vehicle detection and classification, including acoustic signature, radar signal, frequency signal, and image and video representation. The evolution of image processing techniques, together with wide deployment of road cameras, facilitate image-based vehicle detection and classification.

Various approaches to image-based vehicle detection and classification have been proposed recently. Sivaraman and Trivedi [1] use active learning to learn from front part and rear part vehicle images, and achieves 88.5% and 90.2% precision on front and rear part vehicle detection. Chen et al. [2] use a Measurement Based Feature (MBF) and intensity pyramid-based HOG (IPHOG) combined feature set for vehicle classification on front view road images. A rear view vehicle classification approach is proposed by Kafai and Bhanu [3]. They define a feature set including tail light and plate position information, then pass it into hybrid dynamic Bayesian network for classification.

The focus of this work is on vehicle classification based on rear view vehicle images. Given a rear view image captured by a static road camera from a distance along a multi-lane highway (Fig. 1), our goal is to localize the vehicles in the image and subsequently classify the vehicles into passenger vehicles and non-passenger vehicles ("passenger" category and "other" category). Less efforts have been devoted in rear view vehicle classification [3]. Rear view vehicle classification is an important problem as many road cameras capture rear view images. It is challenging as rear views are less discriminative (compared to side view, for instance). Furthermore, it is more challenging for images captured from a distance along multi-lane highways. Partial occlusions between vehicles or by roadside objects complicate detection and classification. Vehicle motion may smear the vehicle details. In addition, both classes (passenger vehicle and other) in our dataset have considerable in-class variances, while difference between the two classes are not very distinctive. It is a challenging task even for a human without proper training.

This paper makes contributions to both vehicle detection and classification. For detection, we propose to make use of the lane markings to assist localization of vehicles. In particular, the dash lane markings provide dimension information to assist rejection of partially occluded vehicles. For classification, we propose to use Deep Neural Network (DNN). Many image classification methods based on Deep Neural Network (DNN) have been proposed recently [4]. These methods achieve state-of-the-art on various datasets [5]. However, direct application of DNN is not possible in our case, as the size of our labelled dataset is too small compared to number of parameters inside DNN architecture. Directly training on DNN would result in overfitting and reduce classification accuracy. On the other hand, it is extremely laborious to construct a properly labelled vehicle dataset to train a DNN.

In this work, we use another approach to take advantage of a DNN architecture. We use a very recent result: the higher layers of a DNN trained on a specific large labelled dataset could be general enough for another distant dataset / image classification task [6]. Thus we extract the features from a specific layer inside a properly-trained DNN, and transfer them to our specific classification task. We apply dimensional reduction on extracted features and train a SVM for classification. We demonstrate that this feature transfer approach is effective in a vehicle classification problem. Furthermore, we analyse the layer activation. We find that in our case the DNN architecture provides a way to learn rich mid-level vehicle features and semantic representations that are specifically related to vehicle perception. This enables high classification accuracy in our vehicle classification. We use Alexnet [4] as our DNN trained on ImageNet. An efficient Deep Learning framework called *Caffe* [7] makes it feasible to run Alexnet on normal computer. There are also other deep learning framework available [5], we choose *Caffe* for

its popularity and efficiency. Experimental results show that our vehicle detection method achieves good precision. Our vehicle classification method outperforms state-of-the-art.

Remaining sections are organized as follows. Section II describes the details of our proposed approach. Section III shows experimental results. Section IV concludes the paper.

Fig. 1: Example road image in the dataset.

II. Methodology

Fig. 2 depicts our approach with two main steps: vehicle detection and vehicle classification.

A. Vehicle detection from road image

Our dataset images are taken from a static camera along an express way. These images contain rear view of vehicles on multiple lanes (Fig 1). Given that these images have same background, we use several steps to extract potential vehicle regions: (i) compute the background using temporal median filter. (ii) apply background subtraction to obtain difference images between each image and background. (iii) use median filter to remove noise in the difference images. (iv) apply Otsu's method to determine foreground against background. The foreground generated from previous steps contains several connected regions. Each connected region could represent: a whole vehicle, a partial vehicle, multiple vehicles that overlap with each other, or objects outside road region. We want to keep only the regions that contain whole vehicles.

In our approach, we localize the positions of dash lane markings inside background image, and use this positions to determine road region in images. The regions contain all the lanes are considered as road regions. We also extract each single lane to provide distance information for later process.

Here we use Connected Component Analysis (CCA) to discard regions with invalid size, aspect ratio, or location on the image. We also make use of a measure called *normalized width*. Each connected region is associated to one single lane based on its centroid position. In order to make use of the lane information in the road image, we normalize the width of each connected region based on the associated lane. Normalized width is the width of connected region divided by the width of lane at the centroid of the connected region. With the normalized width, we can fairly compare the size of vehicles at different distance from the camera.

Remaining regions are considered as valid vehicle regions that contain whole vehicles. We further crop vehicle regions by their bounding boxes, and output the cropped images.

B. Vehicle classification using Alexnet features

We use the rear view vehicle images obtained from Section II-A as input images for classification. Vehicle images will be classified into two classes: *passenger* class and *other* class. Passenger vehicle class includes sedan, SUV, and MPV, other vehicle class includes van, truck, and other types of vehicle. Both classes have large in-class variance. Also the difference between passenger vehicles and other vehicles is not distinctive. These make it difficult to distinguish between these two classes. Fig. 3 shows examples for both vehicle classes. As we can see from the sample images, Fig. 3(a) is MPV, and Fig. 3(b) is taxi. They are both passenger vehicles but different in shape, color, and size. Fig. 3(a) is MPV, and Fig. 3(c) is van. They are in different classes, but similar in shape, color, and size. The classification between passenger vehicles and other vehicles has semantic meanings included, so only low-level vision features are not enough for classification. We need high-level vision features for semantic representations [8].

(a) (b) (c) (d)

Fig. 3: Vehicle image examples for both classes. (a) passenger. (b) passenger. (c) other. (d) other.

We apply a Deep Convolutional Neural Networks (DCNN) approach, extract features from Alexnet [4]. Here we use Alexnet model in a widely-adopted open source deep learning framework called *Caffe* [7]. The model is applied on a popular dataset called ImageNet LSVRC-2012 (ILSVRC-2012). For each vehicle image detected from Section II-A, we resize it to 256 $\times$ 256, make it valid Alexnet input. Then the resized image is passed into Alexnet. Fig. 4 shows the structure of Alexnet. Alexnet has 5 convolutional layers (named as conv1 to conv5) and 3 fully-connected layers (named as fc6, fc7, fc8). Each convolutional layer contains multiple kernels, and each kernel represents a 3-D filter connected to the outputs of the previous layer. For fully-connected layers, each layer contains multiple neurons. Each neuron contains a positive value, and it is connected to all the neurons in previous layer.

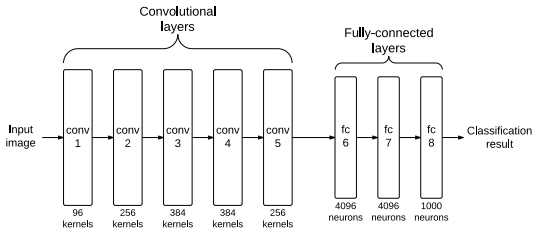

Fig. 4: Structure of Alexnet.

Here we extract the third last and second last fully connected layer (i.e. layer fc6 and fc7) in Alexnet as the generic image representation (to be justified later). Each image representation is a 4096-dimension vector, obtained from the 4096 neurons in layer fc6 (or fc7). Here we consider the extracted layer as a feature vector $f = [f_1, f_2, ..., f_{4096}]$. This is a transfer learning approach (details to be discussed). After we obtain the image representations, Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) is used for dimensional reduction, each feature vector is transformed and reduced to a vector $f' = [f'_1, f'_2, ..., f'_m]$, m is the reduced dimensionality.

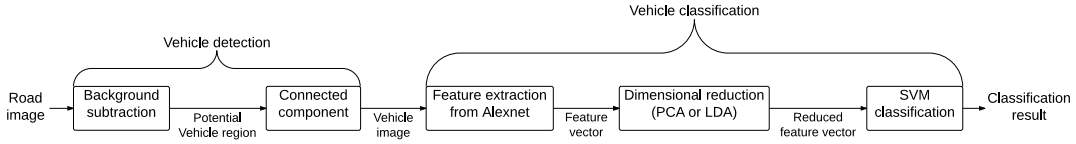

Fig. 2: Pipeline of vehicle detection and classification process.

Then Support Vector Machine (SVM) with linear kernel is used for classification.

Different layers in a Deep Neural Network (DNN) are often considered to have different level of features. The first few layers contain general features that resemble Gabor filters or blob features. The higher layers contain specific features, each representing particular class in dataset [6]. Thus features in higher layers are considered to have higher level vision information compared to general features in base layers. To understand this in our particular problem, Fig. 5 shows several average images we obtained from vehicle images. Given a specific feature f_i we extracted from Alexnet, we sort all the vehicle images based on value of f_i. The images that have highest values on this feature are chosen. Then we calculate the average image of these images. The 4 images in Fig. 5 represents average images for 4 different features (i.e. $f_{i_1}, ..., f_{i_4}$, here $i_1, .., i_4 \in \{1, ..., 4096\}$). We can recognize specific types of vehicles from these average images. Fig. 5(a) represents a specific type of normal sedan. Fig. 5(b) is taxi. Fig. 5(c) is van. And Fig. 5(d) represents truck. Human can easily associate these average images to certain types of vehicles, meaning that the features related to these images contain high-level visualization information related to semantic meanings, and could be very helpful for our vehicle classification.

Alexnet model is trained on ILSVRC-2012 with 1.2 million images in 1000 categories (including general kinds of natural and man-made images), and we use this model to classify our dataset with 400 vehicle images. For transfer learning using Alexnet, we need to consider two main factors: size of the new dataset, and the similarity between the original and new datasets [6]. Since the size of the new dataset is very small compare to original dataset (10^2 vs 10^6), it is not a good idea to fine-tune Alexnet for our dataset due to overfitting concerns. So we use Alexnet as fixed feature extractor instead. Another concern is the similarity between new dataset and original dataset. If the two datasets are very similar, we consider higher-level features in Alexnet are also relevant to new dataset. If the two dataset is not similar, we consider lower-level features in Alexnet are more useful because these features contain more general information. For our dataset (vehicle images) and ILSVRC dataset (natural and man-made images), our dataset (vehicle images) can be considered as a subset of ILSVRC dataset (vehicle images and other images). ILSVRC dataset contains specific kinds of vehicles (race car, ambulance, fire truck, etc.), while our problem has a more general categorization (”passenger” and ”other” category) for our dataset. So we expect to use higher-level features, but not exactly the top layer (which is too class-specific).

In experiment section, we compare the results of using fc6 and fc7 for feature extraction, and PCA and LDA for dimensional reduction.


(a) (b) (c) (d)

Fig. 5: Average image of the vehicles with high values on a specific feature.

III. Experimental results

In this section the experimental results of the proposed vehicle detection and vehicle classification method are presented. The size of road images in our dataset is 4184 $\times$ 3108. For the vehicle detection process, the road images are resized to 1046 $\times$ 777 using Bicubic interpolation. After vehicle detection, we map the vehicle regions back to original road images, and crop vehicle image in original resolution. Typical resolution of vehicle images is around 500 $\times$ 500. For vehicle classification step, all vehicle images are resized to 256 $\times$ 256 to pass into Alexnet. The vehicle detection method is implemented in MATLAB. For vehicle classification method, the feature extraction from Alexnet is under Caffe framework [7], dimensional reduction methods (PCA and LDA) and SVM are implemented in MATLAB.

A. Vehicle detection experiment

Our dataset contains 300 road images with same background. 983 vehicle images are cropped from road images by our method. Among 983 vehicle images, 940 are valid images, i.e., each image contains a whole vehicle. Among 43 invalid images, most of them contain multiple vehicles that overlap heavily. Vehicle detection precision is 95.6%.

B. Vehicle classification experiment

1) Experiment on public dataset: To compare our method with state-of-the-art vehicle classification methods, we perform our method on a public dataset provided in [9]. We use same experiment setting in [10] to perform fair comparison. There are three types of vehicles in this dataset: sedans, vans, and taxis. Following [10], three experiments are performed: *cars* vs *vans*, *sedans* vs *taxis*, and *sedans* vs *vans* vs *taxis*. Note that sedans and taxis are all regarded as cars.

From each of the vehicle images, we extract two feature vectors (from layer fc6 and fc7) with 4096 dimensions using Alexnet. Then, PCA (or LDA) is used to reduce vector dimension (for PCA, dimension for both Alexnet vectors are reduced to 50; for LDA, dimension for different vectors are reduced to 1 or 2 based on number of vehicle class). Then, SVM with linear kernel is applied for classification.

Table I shows accuracy comparison among Alexnet-based methods and other state-of-the-art methods. For Alexnet-based

methods, given the same extracted feature (fc6 or fc7), LDA gives better result than PCA. Given the same dimensional reduction method (PCA or LDA), Alexnet-fc6 feature is better than Alexnet-fc7 feature. Compare with other state-of-the-art methods, Alexnet-fc6-LDA achieves best results on *cars* vs *vans*, and *sedans* vs *vans* vs *taxis* classification, and second-best result on *sedans* vs *taxis* classification. Note that *cars* vs *vans* classification is more similar to our problem (*passenger* vs *other*). *Sedans* vs *taxis* and *sedans* vs *vans* vs *taxis* classification is more fine-grained than our classification.

Accuracy (%)	Cars vs vans	Sedans vs taxis	Sedans vs vans vs taxis
PCA+DFVS [10]	98.50	97.57	95.85
PCA+DIVS [10]	99.25	89.69	94.15
PCA+DFVS+DIVS [10]			96.42
Constellation model [9]	98.50	95.86	
Alexnet-fc6-LDA	99.75	97.27	97.74
Alexnet-fc7-LDA	99.25	96.97	97.36
Alexnet-fc6-PCA	99.50	96.97	95.66
Alexnet-fc7-PCA	98.50	95.15	91.70

TABLE I: Accuracy comparison among different approaches on public dataset. Reported results from [10] are used.

2) Experiment on our dataset: From previous vehicle detection step we get 940 vehicle crop images. 714 and 226 images are manually labelled as passenger vehicles and other vehicles, respectively. To make the size of two class comparable, we randomly select 200 sample images from each class. These images can be downloaded from [11].

Here we compare the performance of the proposed vehicle classification method with one state-of-the-art image classification method: Fisher vector with SIFT descriptor [12][1]. From each vehicle image, we extract a feature vector (fc6 or fc7) with 4096 dimensions using Alexnet. Similarly, from each vehicle image, we first compute SIFT descriptors of the image, each SIFT descriptor represents a 128-dimension vector. Then PCA is applied for dimensional reduction, we choose first half of the 128-dimension vectors, so 64 dimensions are reserved. Finally, fisher encoding with 32 Gaussian distributions is used to generate a 4096-dimension fisher vector. Therefore, for the same vehicle image, we get three 4096-dimension vectors: two extracted from Alexnet (from layer fc6 and fc7), one from fisher encoding. Note that fisher vector and Alexnet feature vector have the same dimension.

We apply the same classification approach using all vectors. First, PCA (or LDA) is used to reduce vector dimension (for PCA, dimension for all Alexnet vectors and fisher vector are reduced to 50; for LDA, dimension for different vectors are reduced to 1). Then, SVM with linear kernel is applied for classification. We do not separate the dataset into training and testing set, so 10-folder cross validation is applied during classification. We label the methods as 4 Alexnet-based methods (Alexnet-fc6-LDA, Alexnet-fc7-LDA, Alexnet-fc6-PCA, Alexnet-fc7-PCA), and 2 SIFT-FV-based methods (SIFT-FV-LDA, SIFT-FV-PCA). We run each method for 5 times to get 5 accuracy results, and report the mean accuracy.

[1]We are unable to run the code from [10], thus we do not include their methods.

Accuracy (%)	LDA	PCA
Alexnet-fc6	97.00	96.45
Alexnet-fc7	96.80	96.10
SIFT-FV [12]	92.30	91.30

TABLE II: Accuracy comparison on our dataset

Table II shows the accuracy comparison among Alexnet-based methods and SIFT-FV-based methods. Given the same extracted feature, LDA gives better result than PCA. Given the same dimensional reduction method, Alexnet approaches give better results than SIFT-FV, and Alexnet-fc6 is slightly better than Alexnet-fc7. Alexnet-fc6-LDA achieves the best result 97%, it is about 5% higher than SIFT-FV-LDA.

Experiment results on both datasets show that using LDA for dimensional reduction gives better results. Using fc6 layer for feature extraction is better than using fc7, which is consistent with our analysis: we want to use high-level features that is not too class-specific.

IV. Conclusion

We have investigated a vehicle detection and classification method. From the multi-lane road images, we have proposed to normalize vehicle size based on lane information, in order to detect vehicle precisely. Because of the limited size of the labelled dataset, our method extracts features from a DNN trained on another dataset. We have analysed the transferred features from DNN. We found that these features capture mid-level vehicle information and can be very helpful for our classification problem. Our approach has achieved some of the best classification accuracy.

References

[1] S. Sivaraman and M. M. Trivedi, "Real-time vehicle detection using parts at intersections," in *15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2012, pp. 1519–1524.

[2] Z. Chen, T. Ellis, and S. Velastin, "Vehicle detection, tracking and classification in urban traffic," in *15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2012, pp. 951–956.

[3] 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, 2012.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.

[6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[8] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *arXiv preprint arXiv:1411.7766*, 2014.

[9] X. Ma and W. E. L. Grimson, "Edge-based rich representation for vehicle classification," in *Tenth IEEE International Conference on Computer Vision (ICCV)*, vol. 2. IEEE, 2005, pp. 1185–1192.

[10] A. Ambardekar, M. Nicolescu, G. Bebis, and M. Nicolescu, "Vehicle classification framework: a comparative study," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 29, 2014.

[11] "Our dataset," https://goo.gl/nPSMOu, 2015.

[12] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.