Fast Unconstrained Vehicle Type Recognition with Dual-Layer Classification

Xiao-Jun Hu[1], Bin Hu[1], Chun-Chao Guo[1], and Jian-Huang Lai[1,2](✉)

[1] School of Information Science and Technology,
Sun Yat-sen University, Guangzhou, China
{hux999m,chunchaoguo,huglenn232}@gmail.com

[2] Guangdong Key Laboratory of Information Security Technology,
Sun Yat-sen University, Guangzhou 510006, China
stsljh@mail.sysu.edu.cn

Abstract. This paper tackles the problem of vision-based vehicle type recognition, which aims at outputting a semantic label for the given vehicle. Most existing methods operate on a similar situation where vehicle viewpoints are not obviously changed and the foreground regions can be well segmented to extract texture, edge or length-width ratio. However, this underlying assumption faces severe challenges when the vehicle viewpoint varies apparently or the background is clutter. Thus we propose a dual-layer framework that can jointly handle the two challenges in a more natural way. In the training stage, each viewpoint of each type of vehicles is denoted as a sub-class, and we treat a pre-divided region of images as a sub-sub-class. In the first layer, we train a fast Exemplar-LDA classifier for each sub-sub-class. In the second layer of the training stage, all the Exemplar-LDA scores are concatenated for the consequent training of each sub-class. Due to introducing Exemplar-LDA, our approach is fast for both training and testing. Evaluations of the proposed dual-layer approach are conducted on challenging non-homologous multi-view images, and yield impressive performance.

Keywords: Vehicle type recognition · Exemplar-LDA · Dual-layer classification

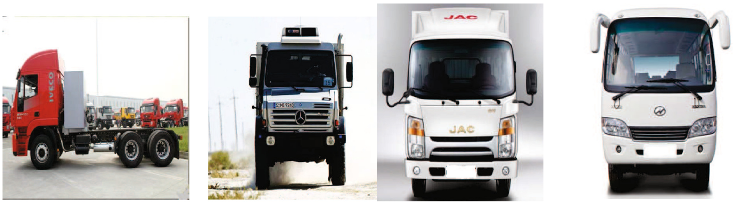
1 Introduction

Vehicle type recognition based on vision has received great attention for its broad range of applications, including autonomous driving, intelligent transport and visual surveillance. It is an issue that aims to assign semantic labels to a vehicle image or video, such as truck, bus, sedan, or motorcycle. A framework for vehicle type recognition mainly consists of two components, namely vehicle representation and feature classification. Representation of vehicles is rather difficult comparing with that of other objects, since the appearance of different vehicles within the same type can vary drastically. This can be seen in Fig. 1. Even the appearance of the same type can change apparently with the angle.

Y.-J. Zhang (Ed.): ICIG 2015, Part II, LNCS 9218, pp. 275–283, 2015.
DOI: 10.1007/978-3-319-21963-9_25

To alleviate the difficulty on vehicle representation and classification, many existing methods mainly focus on constrained situations, for instance, video sequences captured from a single view [2,10,12,14,15,17]. In those sequences, intra-class appearance variations are not drastic and foreground regions can be segmented to further extract edge, length-width ratio, or other discriminative cues. This leads to easy representation and classification, as well as brings about an implicit deficiency that the model trained from a specific dataset is difficult to be applied to another scenario. For classification, hierarchical classification has shown its strength in saliency detection [16], face verification [1], and so on. However, previous works on vehicle type recognition usually rely on a single-level classification framework, including common AdaBoost [17], SVM [6,14] and KNN [9,12], which tends to be easily influenced by appearance changes.

Inspired by the recent fine-grained object categorization [1], we propose a unified framework for vehicle type recognition. Vehicles are represented in a higher level and unconstrained situation, which leads to that our representation can tolerate arbitrary viewpoint changes and large intra-class appearance variations. In the training stage, each viewpoint of each kind of vehicles is denoted as subset, and we treat a region of images in a subset as a sub-class. Different from [1] that trained one-vs-one SVM for each pair of sub-classes, we train a fast Exemplar-LDA classifier [7] for each sub-class firstly, also called region-based one-vs-one features. This is mainly due to the fact that Exemplar-LDA is much faster than Exemplar-SVM as well as achieving competing performance. In the second layer of the training stage, all the Exemplar-LDA scores are concatenated as a new feature vector for the consequent training of each vehicle type. Given a test image, scores of the first level are obtained from all the trained Exemplar-LDA classifiers. Then they are concatenated and fed into a SVM classifier for the final vehicle type output. The two layers are closely related, which can well handle both representation and classification.



(a) (b)

Fig. 1. Challenges of vehicle type recognition. (a): Large intra-class appearance variation of 2 trucks caused by viewpoint change; (b): Type confusion caused by appearance similarity between a bus and a truck.

This paper aims to resolve two problems, including unconstrained vehicle recognition and fast vehicle recognition. The first problem is solved with the help of our dual-layer framework that generates a higher representation, and the second goal is achieved via the introduction of exemplar-LDA in the first layer. The main contributions of this work are three-fold. (1) We propose an approach to unconstraint vehicle type recognition, which jointly resolves larger intra-class variations within the same type as well as viewpoint-independent vehicle type recognition. (2) Our representation does not rely on any prior assumptions of a vehicle. Extraction of vehicle foreground regions is not deemed necessary, which allows our representation can be generalized to either video sequences or still images. (3) We leverage Exemplar-LDA to the first layer of our dual-layer classification, considering that the Exemplar-LDA is much faster than the conventional Exemplar-SVM and achieves competitive performacne. Thus our approach can be practicably implemented in real-world applications.

2 Related Work

Vehicle type recognition has drawn great attention in the past years. We will review closely related works following the two common components of this issue, namely vehicle feature representation and feature classification.

In terms of feature extraction, shape and texture features are more popular than color descriptors, since vehicle type is independent on color in most cases. An intuitive way to categorize vehicle types is to represent and classify the given vehicle with shape or edge cues. The contour and aspect ratio of vehicles are commonly exploited in [2,5,9,13] for representation. In [15], Tian et al. made use of color information and taxi symbols to detect taxi with the priori knowledge about the position of plate number. In [12,14], two new descriptors are introduced, named extract Square Mapped Gradients (SMG) and Locally Normalized Harris strengths (LNHS), respectively. However, both of them requires a frontal view for a car image. Reference [6,17] extracted SIFT and HOG to embed texture information of vehicle. Ma et al. [10] proposed modified-SIFT along with SIFT to demonstrate vehicles that made obvious performance improvements. It was extracted by first obtain SIFT on edge points, and then cluster them, which is followed by the final key point selection.

As for feature classification of vehicles, most common classification algorithms are employed in this field. SVM classifiers are ultilized in [6,14], while kNN is used in [9,12]. Also a Bayesian classifier was used by [10]. Moreover, Zhang [17] cascaded different strong classifier with Adaboost, which showed an obvious improvement.

Although extensive studies on vehicle type recognition have been presented, it remains a challenging problem for the reason that most existing approaches work in constrained scenarios. [2,10,12,13,17] are based on cameras whose positions are fixed, which means that their background is unchanged. [2,10,13] relies on foreground region extraction. It cannot achieve satisfactory performance if it confronts complicated background or the background is clutter. [2,10,12,14,15, 17] capture vehicles from an unchanged viewpoint. Hence they are sensitive to viewpoint variations and cannot work in realistic scenes, where diverse angles of vehicles occur frequently. [2,5,9,10,13] extract edge features and show impressive improvements on their testing data. However, their limitations emerge obviously when given images have low resolution or contain a clutter background.

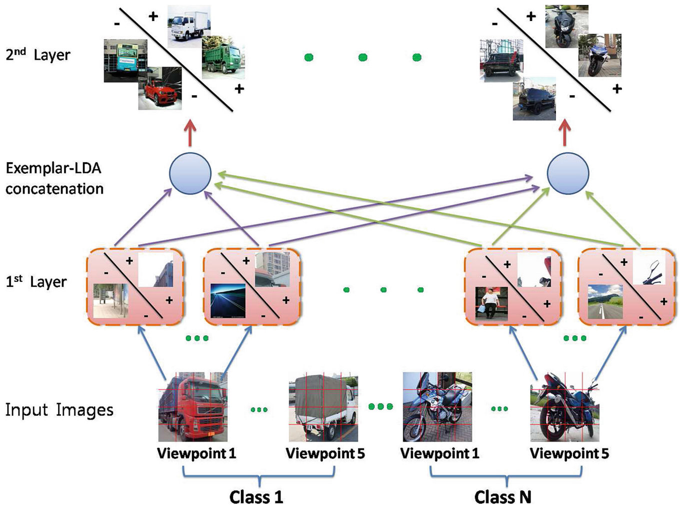


Fig. 2. Sketch of the framework. Best viewed in color.

3 The Dual-Layer Framework

3.1 Overview of the Framework

Figure 2 demonstrates the sketch of our framework. In the vehicle classification dataset, we have four different types of vehicles (bus, car, motorcycle, truck). The training images of each vehicle category are further divided into five subsets according to the viewpoints. This division in training stage is critical, considering that intra-class appearance variations are extremely drastic. During feature learning, we treat each subset of vehicle as a sub-class, so that we totally have 20 classes in the final classification. Inspired by part-based models in many fine-grain classification works, we divide images into $n \times m$ overlapping windows, each of which plays the role of a part model. We extract HOG descriptors from each window and train a linear classifier $P_{i,j}$. Here i denotes a sub-class ranging from 1 to 20 in our experiment, and j indicates a window in the image. Please see Sect. 3.3 for details of training part classifiers. In fact, the underlying meaning of each trained part classifier is that the part classifier describes the local appearance attribute of a vehicle. In the feature extraction stage, each part models are applied to correspond location of image, and the detection scores in different sub-classes and different parts are concatenated to form our final feature representation

$$\begin{aligned} Ftr(H) = &[\varPhi\left(H_1, P_{1,1}\right), ..., \emptyset\left(H_{m*n}, P_{1,m*n}\right), ..., \\ &\varPhi\left(H_1, P_{20,1}\right), ..., \emptyset\left(H_{m*n}, P_{20,m*n}\right)] \end{aligned} \tag{1}$$

where $H_i \in [1, m*n]$ are HOG descriptors in each window, and $\varPhi$ denotes the score function. Given this feature representation, we train 20 One-Vs-All classifiers with linear SVM. At testing stage, we obtain the predicted type of the vehicle by selecting the class with highest classification score ignoring the angle.

3.2 Training Exemplar-LDA for Each Part

Although POOF [1] has achieved significant performance improvement in fine-grained classification, it still shows some limitations, one of which is that training linear SVM for each part is time-expensive. The Exemplar-SVM method [11] in objet detection domain faces the same problem since it needs to train a SVM classifier for each training sample. In Exemplar-LDA [7], the authors replace the SVM with LDA which can be trained very efficiently. We can treat LDA model as a linear classifier with its weight given by $\omega = \Sigma^{-1}\left(\mu_1 - \mu_0\right)$. Here Σ is the class-independent covariance matrix and μ_i is corresponding to class-dependent mean. Exploiting the scale and translation invariance property of nature image [8], μ_0 and Σ in [7] are estimated offline, and then reused for all object categories.

Explicitly, given an image window of a fixed size, the HOG descriptor is a concatenation of gradient orientations histogram in each 8×8 cell. We denote x_{ij} as the feature vector of the cell in correspond location (i, j). We extract HOG descriptor from different scales and translations in training image, and compute the mean HOG feature $\mu = E\left[x_{ij}\right]$. Then we average the features over all locations and images. Exploiting the translation invariance property, for a category with size of N_0 cells, we can construct μ_0 by conveniently replicating μ over all N_0 cells.

Likewise, we can treat Σ as a block matrix with blocks $\Sigma_{(i,j),(l,k)} = E\left[x_{ij}x_{lk}^T\right]$. Under the assumption of translation invariance, we assume that $\Sigma_{(i,j),(l,k)}$ only depends on the relative offset $(i-k)$ and $(j-l)$, then we have

$$\Sigma_{(i,j),(l,k)} = \Gamma_{(i-l),(j-k)} = E\left[x_{uv}x_{(u+i-l),(v+j-k)}^T\right] \tag{2}$$

Instead of learning $dN_0 \times dN_0$ matrix Σ, we only have to learn $d \times d$ matrices $\Gamma_{(s,t)}$ for all possible offset (s, t). Here d indicates the dimension of HOG descriptor. We learn the μ and matrices $\Gamma_{(s,t)}$ from all subwindows extracted from PASCALVOC 2010 dataset [3], which contains 10000 natural images. Σ can be reconstructed from Γ using Eq. 2.

3.3 Accelerating with Exemplar-LDA

As mentioned in Sect. 3.1, we need to train a linear classifier for each part of each sub-class. Here it is unpractical to employ conventional SVMs, which involve the hard sample mining procedure that is time-consuming. Therefore, we utilize the powerful Exemplar-LDA to train our part-based models. We average HOG descriptors of all positive samples of each part in order to obtain μ_1. The mean background μ_0 and the covariance matrix Σ are calculated following Sect. 3.2. The final weight of a trained linear classifier can be computed using

$$\omega = \Sigma^{-1}(\mu_1 - \mu_0). \tag{3}$$

Table 1. Comparison results.

	Bus	Car	Motorcycle	Truck
BoW	66 %	67 %	81 %	64 %
HOG + RF	71 %	85 %	89 %	73 %
Ours	**91 %**	**93 %**	**99 %**	**92 %**

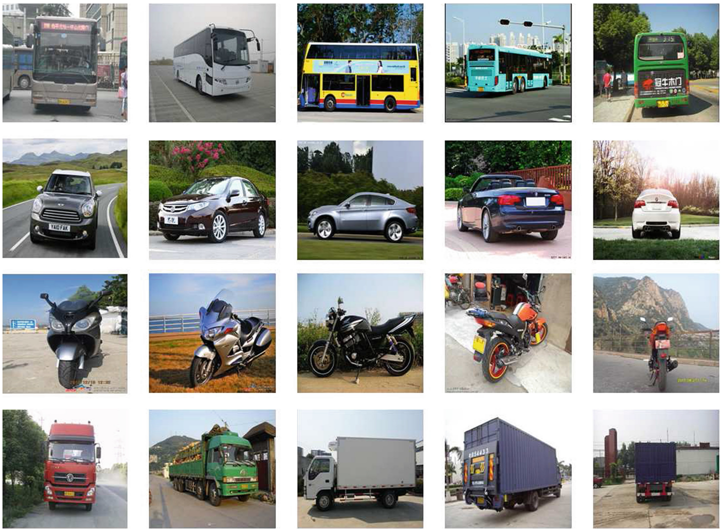
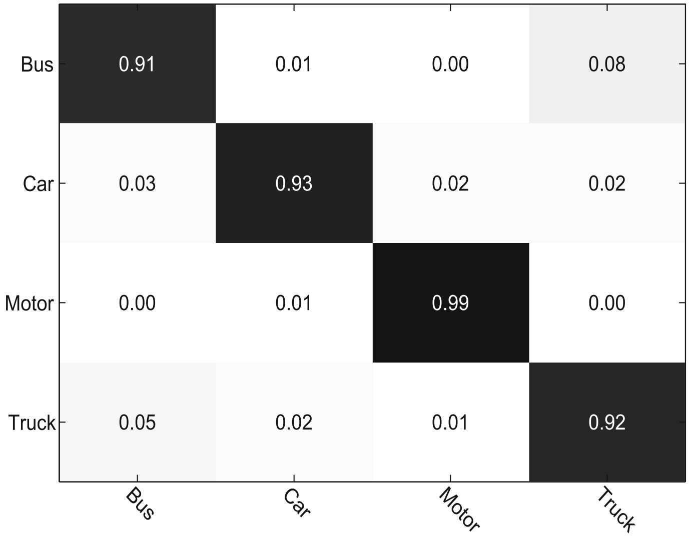


Fig. 3. Samples of the dataset.

4 Experimental Results

4.1 Dataset and Experiment Settings

There exists 4 vehicle types in our dataset, including car, bus, motorcycle and truck. 4000 images are contained in this dataset in total, with 5 different viewpoints for each type. Considering the symmetry of vehicles, only 5 viewpoints are collected, including 0 degree, 45 degree, 90 degree, 135 degree and 180 degree, which contains a circle from frontal view to rear view. 3600 images are used in the training stage, and the remaining 400 images are treated as the test set. This dataset is rather challenging for the reason that the backgrounds are clutter and the intra-class appearance variations are large. Samples of the dataset can be seen in Fig. 3. In our experiments, we denote each vehicle type as a class, each viewpoint of each class as a sub-class, and each region of each sub-class as a sub-sub-class. Here all the images are normalized into the same resolution with 224*224, and each image is divided into 4*4 regions. Consequently, there are 4 classes, 4*5 sub-classes, and 4*5*16 sub-sub-classes. In the first layer, 320 Exemplar-LDA classifiers are trained, while in the second layer, 20 SVMs are trained for all the sub-classes.



	Bus	Car	Motor	Truck
Bus	0.91	0.01	0.00	0.08
Car	0.03	0.93	0.02	0.02
Motor	0.00	0.01	0.99	0.00
Truck	0.05	0.02	0.01	0.92

Fig. 4. The confusion matrix of our classification results.

4.2 Quantitative Results and Analysis

As can be seen in Fig. 4, our dual-layer approach achieves an average accuracy of 93 %. For every type, our recognition rate is over 90 %, especially for the motorcycle, whose accuracy even reaches 99 %. Besides the strength of our method, another reason of the high performance is that the appearance of motorcycles is obviously different form other vehicles. From Fig. 4, we can see that the main threaten comes from the classification of the buses and the trucks. This is mainly due to the fact that the two types are similar in appearance in some specific views, for instance, in the 135 degree.

In our experiments, we set two widely used models as the benchmark comparison methods, including the bag of words model (BoW) [4] and the Random Forest model using HOG features (HOG + RF). Table 1 gives the quantitative accuracy, from which we can see that our approach obviously outperforms the two benchmark methods.

Note that, our approach is rather fast in both training and testing, and thus it is appropriate to be implanted in real-world scenarios. The training time of per Exemplar-LDA consumes 0.3283 s, whilst the time of classification for each image only costs 0.0105 s.

5 Conclusion

This paper has presented a dual-layer framework for fast unconstrained vehicle type recognition. Unlike many existing methods, this work does not rely on strong assumptions on vehicle appearance or viewpoints. We jointly handle viewpoint changes and large intra-class appearance variations, as well as throw away the dependence of foreground region extraction. Our approach has the ability to work in clutter videos or even only a single-shot image. In our framework, region-based one-vs-one features are firstly extracted and the corresponding Exemplar-LDA classifiers are trained. Secondly, all the classification scores are concatenated as a new feature for the consequent SVM classifier. Due to introducing Exemplar-LDA, our approach has much lighter computation load than the conventional method using one-vs-one SVM training. In the future, we will enrich our approach and extend it to more challenging scenarios, such as type recognition for occluded vehicles.

Acknowledgement. This work was supported by NSFC (No. 61173084) and the National Science & Technology Pillar Program (No. 2012BAK16B06).

References

1. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 955–962. IEEE (2013)
2. Chen, Z., Ellis, T., Velastin, S.A.: Vehicle type categorization: a comparison of classification schemes. In: 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 74–79. IEEE (2011)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
4. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 524–531. IEEE (2005)
5. Feris, R., Siddiquie, B., Zhai, Y., Petterson, J., Brown, L., Pankanti, S.: Attribute-based vehicle search in crowded surveillance videos. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 18. ACM (2011)
6. Gandhi, T., Trivedi, M.M.: Video based surround vehicle detection, classification and logging from moving platforms: issues and approaches. In: Intelligent Vehicles Symposium, pp. 1067–1071. IEEE (2007)
7. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 459–472. Springer, Heidelberg (2012)
8. Hyvrinen, A., Hurri, J., Hoyer, P.O.: Natural Image Statistics: A Probabilistic Approach to Early Computational Vision. Springer, London (2009)
9. Iqbal, U., Zamir, S., Shahid, M., Parwaiz, K., Yasin, M., Sarfraz, M.: Image based vehicle type identification. In: 2010 International Conference on Information and Emerging Technologies (ICIET), pp. 1–5. IEEE (2010)
10. Ma, X., Grimson, W.E.L.: Edge-based rich representation for vehicle classification. In: 10th IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1185–1192. IEEE (2005)

11. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)
12. Petrovic, V.S., Cootes, T.F.: Analysis of features for rigid structure vehicle type recognition. In: BMVC, pp. 1–10 (2004)
13. Rad, R., Jamzad, M.: Real time classification and tracking of multiple vehicles in highways. Pattern Recogn. Lett. **26**(10), 1597–1607 (2005)
14. Rahati, S., Moravejian, R., Mohamad, E., Mohamad, F.: Vehicle recognition using contourlet transform and SVM. In: 5th International Conference on Information Technology: New Generations. 2008 ITNG, pp. 894–898. IEEE (2008)
15. Tian, B., Li, B., Li, Y., Xiong, G., Zhu, F.: Taxi detection based on vehicle painting features for urban traffic scenes. In: 2013 IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp. 105–109. IEEE (2013)
16. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1155–1162. IEEE (2013)
17. Zhang, B.: Reliable classification of vehicle types based on cascade classifier ensembles. IEEE Trans. Intell. Transp. Syst. **14**(1), 322–332 (2013)