An Effective and Robust Multi-view Vehicle Classification Method Based on Local and Structural Features

Z.Q. Xiang, X.L. Huang and Y.X. Zou*
School of Electronic and Computer Engineering
Peking University
Shenzhen, Guangdong 518055 China
*Email: zouyx@pkusz.edu.cn

***Abstract*—Big traffic data analysis for intelligent transportation is attracting more and more attention. Due to different designs of vehicles in the same class and the similarity of shape and textures between different classes, vehicle classification is remaining a challenge. In this paper, different from traditional methods that only classify vehicles to two or three types in one viewpoint, a novel method using local and structural features has been proposed for vehicle classification in real-time traffic system that has a good ability to categorize vehicles into more specific types and is robust to the changes in viewpoint. Specifically, local features are obtained using scale invariant feature transform (SIFT), and an efficient L2-norm sparse coding technique is used to reduce computational cost. Besides, vehicle building structures are extracted as structural features. Finally, linear support vector machine (SVM) is used as the classifier. The performance evaluations using real vehicle images extracted from surveillance videos in different viewpoints are carried out and five vehicle classes (SUV, truck, van, bus, car) are considered. Experimental results show that the proposed method can obtain an average accuracy of 95.95% in real-time, which validate the effectiveness of our method.**

I. Introduction

Intelligent transportation system (ITS) is advanced applications in which information and communication technologies are applied in the field of road transport, such as traffic management, surveillance and security. In recent years, with the increasing number of cameras deployed to traffic monitoring, the research and application of image-based vehicle detection and classification system attract more and more attention. For most traffic surveillance systems, vehicle detection, tracking and classification are three key stages, which are used to estimate desired traffic parameters. They are the foundations of traffic flow measurement, automatic incident detection, automatic road enforcement and criminal investigation [1]–[3]. For vehicle detection, most methods [4]–[6] assumed that the camera was fixed and then desired vehicle objects can be detected by background subtraction. Then, different tracking methods, like region-based tracking [5], [7], contour tracking [8], 3D model-based tracking [9]–[11], and feature-based tracking [12]–[14] were designed to track each vehicle object. After that, several vehicle features were extracted for vehicle classification [5], [14], [15]. Based on some simple geometric features, like shape, length, width, texture, etc., in [5], [16] only two categories were classified, i.e., cars and noncars. Song and Miao [17] classified vehicles into three categories by using a method based on spatial pyramid representation and BP neural network, an average accuracy of 76.52% was obtained. Peng et. al [18] classified vehicles into four categories by extracting a dense boosting binary feature computed with a boosted binary hash function and pooling the features in different resolutions. Nurhadiyatna et. al [19] developed a system to conduct classification of three vehicle types by using Gabor kernel for feature extraction, and the highest accuracy was 93.36% that was obtained using 18 features built by ten Gabor kernel combinations with random Forest classifier. It is hard to precisely classify vehicles into more types, due to the similarity of shape and textures between different classes. In addition, most of the previous studies mentioned above were only considering images captured from a stationary camera in one viewpoint, and did not consider the changes of viewpoint. The robustness to changes of viewpoint will make the erection of cameras more flexible.

In this paper, we focus on precisely classifying vehicle images obtained from traffic surveillance videos which are captured from different viewpoints into more specific types. At the beginning, background substraction with gaussian mixture model (GMM) is implemented for vehicle detection. Then, for local feature representation, dense SIFT is applied in consideration of the variant rotation, scale and illumination. Different from classical bag-of-words model, an efficient L2-norm sparse coding and multi-scale spatial max pooling are adopted to make the features more discriminative and representative. Besides, we observe that vehicles from different classes consist of different major parts. Considering different vehicle building structures between classes, color intensity values along vertical axis are extracted as structural features. After that, features fusion is adopted and support vector machine (SVM) is used for vehicle classification. Experimental results show that the proposed method is effective in classifying vehicles into specific types in real-time.

The rest of this paper is organized as follows. In the section II, framework of the proposed method and techniques are

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.58

described. Section III experimental results are presented and discussed in details. Finally, a conclusion is presented in Section IV.

II. Framework of the Proposed Method

In our study, we design and implement a surveillance video based vehicle classification method. It can be used for detecting and classifying vehicles from surveillance video sequences. Fig.1 shows the flowchart of the proposed method. The proposed method includes three parts: object detection, feature extraction and classification. In object detection phase, considering the detection accuracy and speed, background subtraction by gaussian mixture model is implemented for vehicle detection. The feature extraction and classification, which are the main issue we study in this paper, are described in detail.

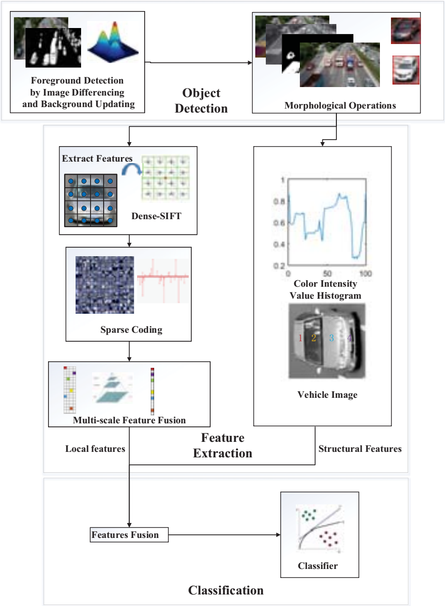

Fig. 1. The flowchart of proposed method.

A. Feature Extraction

1) The Local Feature Extraction: It is well known that proper feature extraction is core to image object representation. Several vehicle images in our database are presented in Fig.2. From Fig.2, we can observe that there are full of variance in rotation, scale and illumination. Since SIFT features are invariant to image scale and rotation, and also robust to change in illumination, noise, and minor changes in viewpoint, we adopt SIFT descriptor to extract local features. It is shown that SIFT descriptors outperform many other local descriptors in object recognition. In addition, considering the reduction of clutter after vehicle detection and the good performance of densely sampled SIFT descriptors in object recognition [20], we utilize a dense regular grid instead of commonly adopted interest points to extract SIFT features, which can capture more discriminative information about vehicle objects. For vehicle image $\boldsymbol{I}_i$, all the SIFT feature vectors extracted from $\boldsymbol{I}_i$ constitute a matrix $\boldsymbol{Y}_i$.

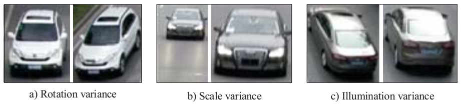

Fig. 2. Vehicle samples captured by different view points.

2) Sparse Encoded SIFT: In recent years, several authors have reported very good recognition results by means of encoding techniques [21]–[24]. In these encoding techniques, sparse coding (SC) was widely used in many state-of-the-art works, which can be formulated as the following:

$$\min_{\boldsymbol{A},\boldsymbol{V}} \sum_{i=1}^{M} \|\boldsymbol{p}_i - \boldsymbol{V}\boldsymbol{a}_i\|_2^2 + \lambda\|\boldsymbol{a}_i\|_1$$
$$s.t.\|\boldsymbol{v}_j\| \leq 1, \forall j = 1, 2, \cdots, L \qquad (1)$$

where a set of patches $\boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_M] \in \mathcal{R}^{128\times M}$ are extracted from a large number of images, M is the number of patches. $\boldsymbol{V} \in \mathcal{R}^{128\times L}$ is the codebook, and $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_M] \in \mathcal{R}^{L\times M}$ is the corresponding coding matrix for $\boldsymbol{P}$. The number of the basis vector of codebook $\boldsymbol{V}$ is denoted as L. In general, the codebook $\boldsymbol{V}$ is an overcomplete basis set.

It is noted that the process of solving the L1-norm constraint optimization problem in Eq.1 is computationally demanding when doing the online vector sparse coding, which reduces the practical value of the sparse coding. In this paper, we use an efficient coding algorithm to reduce the computational cost. We relax the L1-norm constraint by a weaker sparsity L2-norm constraint, as shown in Eq.2.

$$\hat{\boldsymbol{a}}_i = \arg\min_{\boldsymbol{a}_i} \|\boldsymbol{p}_i - \boldsymbol{V}\boldsymbol{a}_i\|_2^2 + \lambda\|\boldsymbol{a}_i\|_2^2$$
$$s.t.\|\boldsymbol{v}_j\| \leq 1, \forall j = 1, 2, \cdots, L \qquad (2)$$

The Eq.2 is a regularized least square problem. To solve this problem, we just need to use partial differential on the equations, then Eq.3 can be obtained

$$\hat{\boldsymbol{a}_i} = \boldsymbol{P}\boldsymbol{y}_i$$
$$\boldsymbol{P} = (\lambda\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T \qquad (3)$$

where $I \in \mathcal{R}^{L\times L}$ denotes a unit matrix. Eq.3 shows that there is only one parameter $\boldsymbol{V}$, and the codebook $\boldsymbol{V}$ can be obtained in dictionary learning phase. Because Eq.2 has analytical solution, $\boldsymbol{P}$ can be pre-calculated by Eq.3, the cost of computation on online is reduced vastly. Although the coding solution obtained by the L2-norm regularization is not rigorously sparse, the solution still has the property that it is discriminative and distinguishable.

3) Local Feature Fusion by Multi-Scale Spatial Max Pooling: After the local feature encoding phase, the sparse vector set $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_M] \in \mathcal{R}^{L\times M}$ can be obtained on each vehicle image. However, the encoded SIFT descriptor only represents the local property and ignores the global property. Based on the observation of vehicle images captured in detection phase, the images have complex backgrounds, so global and salient properties are crucial to robust classification. We intend to represent an image in a multi-scale feature fusion. The main procedure includes two individual parts: spatial pyramid matching (SPM) and max pooling. The absolute sparse codes is formulated as

$$\begin{gathered} \boldsymbol{z}_i = \max\{|\boldsymbol{a}_{i1}|, |\boldsymbol{a}_{i2}|, \cdots, |\boldsymbol{a}_{ik}|\} \\ \forall i, i = 1, \cdots, L \end{gathered} \tag{4}$$

where k is the number of local descriptors in the region, $\boldsymbol{a}_{ij}$ is the sparse coding on the local descriptors and $\boldsymbol{z}_i$ is the ith code which represents the region after max pooling. In this study, we divide the image into 1,4,16 parts respectively, then operate the max pooling on each part and concatenate the 21 parts directly to form the image representation feature vector $\boldsymbol{l}$. Fig.3 shows the procedure.

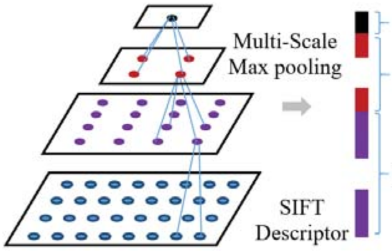

Fig. 3. The illustration of local feature fusion.

B. *Structural Feature Extraction*

Considering the similarity in the shape and textures of the parts from different class of vehicles, only using local features of vehicles can be hard to classify some vehicle types, like car, van and SUV. In this paper, we also consider the structural features of the vehicles that detect the major building parts of the vehicles to help classify vehicles, because we observe that vehicles from different class consist of different major parts. For example, van consists of three major parts, namely the roof, the rear window and the trunk while car consists of four to five parts, namely the hood, the windshield, the roof, the rear window and the trunk. The major building parts of each vehicle image are represented by the color intensity values along the vertical axis of that images. Fig.4 shows the color intensity value histogram of van and car examples respectively. From Fig.4, we can clearly see that the histograms indicate the build part of the vehicles.

Specifically, for a vehicle image, we convert the RGB image to gray image and resize it to 100×100. Then, we use an average filter of size 3×3 in order to reduce image noise due to illumination on the vehicle surface and to reduce detail due to variations in vehicle surface texture. Finally, structural feature vector $\boldsymbol{s} = [s_1, s_2, \cdots, s_{100}]$ is calculated as follows:

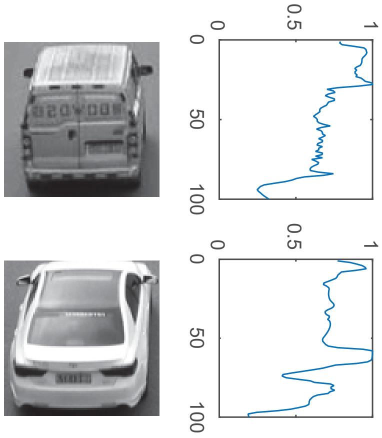

Fig. 4. The color intensity value histogram along vertical axis of van and car examples.

$$\begin{gathered} s_i = \sum_{j=1}^{100} I(i,j) \\ \forall i, i = 1, \cdots, 100 \end{gathered} \tag{5}$$

where i is the vertical index of image and j is the horizontal index of image. $I(i,j)$ is the color intensity of point (i,j).

C. *Classification*

In classification phase, we concatenate local feature $\boldsymbol{l}$ and structural feature $\boldsymbol{s}$ directly to form the final image representation $\boldsymbol{c}$. The classical multi-class linear support vector machine is taken as the classifier, which is considered as a popular and effective supervised machine learning technique.

III. Experiments and Analysis

A. *Database*

In order to analyze the performance of the proposed method, five sequences named as C01-C05 were used. All the sequences are acquired from boulevards with a fixed camera. Sequences are captured from different viewpoints, as show in Fig.5. Foreground detection and image processing method are used to implement the vehicle detection. One example of the results of the vehicle detection is shown in Fig.6. The resolution of videos is 1920×1080 , and the frame rate is 25 frames per second. In our study, the vehicles in the database will be classified into five categories: bus, truck, SUV, van and car. In total, 45103 vehicle pictures are obtained in predefined ROIs and tagged manually. The number of each vehicle type in different sequences is showed in Table I. All experiments are carried out using Matlab R2015a on a 4.0GHz with Intel Core i7 4790K CPU and 16G RAM. The operation system is the 64-bit windows 10.

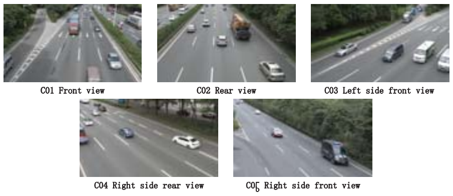
C01 Front view
C02 Rear view
C03 Left side front view
C04 Right side rear view
C05 Right side front view

Fig. 5. Examples in sequences C01-C05.

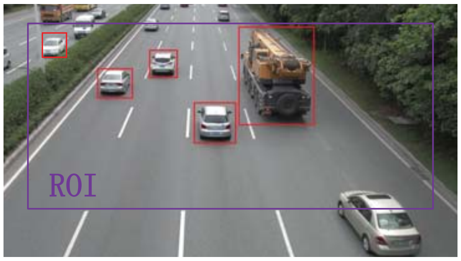

Fig. 6. One example result of the vehicle detection.

TABLE I
THE NUMBER OF VEHICLE IMAGES.

	SUV	Track	Van	BUS	Car	Total
C01	1664	858	925	912	3979	8338
C02	1286	1143	1007	1044	2063	6543
C03	1801	1620	1288	1004	4450	10163
C04	3079	1905	1205	942	5656	12787
C05	924	809	1060	956	3523	7272
Sum	8754	6335	5485	4858	19671	45103

B. Experimental Results

In this study, several experiments have been carried out. The experimental setting are as follows: 1) the SIFT descriptors for every 16×16 patches on a grid of step size 8 is employed in the local feature extraction phase. 2) the parameter λ in Eq.1 and Eq.2 is fixed to be 0.15 and the maximum iteration to be 50. The size of the codebook $\boldsymbol{V}$ is set as 128×1024. 3) 5 fold cross-validation scheme is used in the classification experiments. Our following experiments will verify the superiority of the proposed method in terms of classification accuracy, running time and robustness to different viewpoints.

1) Classification only use local features: Firstly, for each video sequence, we select 1000 vehicle objects (200 vehicle objects for each vehicle type) as testing data and the rest as training data. Only local feature $\boldsymbol{l}$ is used in this experiment. In order to demonstrate the proposed method is effective, we compare our proposed method with two algorithms, mainly focusing on feature extraction. One is Gandhi's method [25] applied the extraction of HOG features followed by SVM classifier, the other is LBP-SVM applied the extraction of LBP features followed by SVM classifier. The experimental results are shown in Table II. In Table II, we can clearly see that our method achieves highest classification accuracy among three algorithms. The lowest classification accuracy is 94.20% for C05, the right side front view. The highest classification accuracy is 96.50% for C04, the right side rear view.

TABLE II
CLASSIFICATION ACCURACY OF EACH SEQUENCE THAT USE ONLY LOCAL FEATURES.

	C01	C02	C03	C04	C05
Our method	94.80%	95.10%	95.40%	96.50%	94.20%
HOG-SVM [25]	90.20%	91.10%	93.30%	92.40%	92.60%
LBP-SVM	89.00%	88.70%	89.40%	90.20%	86.30%

Then, we mix vehicle object images from different sequences together to consider the classification performance in multi-view condition. We randomly pick 8500 samples as the testing data and the rest as the training data and also only local feature $\boldsymbol{l}$ is used. The per-class accuracy is computed and the final average classification accuracy is obtained by the mean of each process. The experimental results are shown in Table III.

TABLE III
CONFUSION MATRIX ACROSS DIFFERENT VEHICLE TYPES THAT USE ONLY LOCAL FEATURES.

	SUV	Truck	Van	Bus	Car	Total	Accuracy
SUV	1311	9	38	1	141	1500	87.40%
Truck	9	1410	6	23	52	1500	94.00%
Van	109	23	1329	7	32	1500	88.60%
Bus	2	17	2	978	1	1000	97.80%
Car	103	18	4	2	2873	3000	95.77%
Average						8500	92.95%

Form Table III, it clearly shows that the bus category has the highest classification accuracy of 97.80% while the SUV category has the lowest classification accuracy which is 87.40%. The average classification accuracy is 92.95%, which validates the effectiveness of our proposed system in multi-view condition.

2) Classification use local and structural features: In this experiment, experimental settings are same as above, but we use both local and structural features to represent the vehicle images. The experimental results are shown in Table IV and Table V.

TABLE IV
CLASSIFICATION ACCURACY OF EACH SEQUENCE THAT USE LOCAL AND STRUCTURAL FEATURES.

	C01	C02	C03	C04	C05
Accuracy	96.20%	96.30%	97.10%	97.50%	95.30%

TABLE V
CONFUSION MATRIX ACROSS DIFFERENT VEHICLE TYPES THAT USE LOCAL AND STRUCTURAL FEATURES.

	SUV	Truck	Van	Bus	Car	Total	Accuracy
SUV	1421	7	23	1	48	1500	94.73%
Truck	7	1417	10	19	47	1500	94.47%
Van	47	23	1394	7	29	1500	92.93%
Bus	1	13	2	983	1	1000	98.30%
Car	43	10	4	2	2941	3000	98.03%
Average						8500	95.95%

Table IV and Table V show that using structural feature can improve the classification accuracy of vehicles, which indicates that structural feature is complementary to local feature. We can observe that the easily confused vehicle types like SUV, car and van in Table III can be classified well by adding structural features, which is shown in Table V.

From above experimental results, we can conclude that the proposed method is effective to vehicle classification and robust to the changes in viewpoint.

3) The real-time performance of our proposed method: To testify the time performance of our proposed method, we test the running time in each phase. From table VI, it shows that the phase of feature extraction costs 41.31ms and classification costs 1.13ms. The overall time for a vehicle object is 42.44ms. As the result, we are confident that our method can meet the real-time acquirement in real world applications.

TABLE VI
THE AVERAGE RUNNING TIME FOR PER VEHICLE.

	Feature Extraction	Classification	Total
Time	41.31ms	1.13ms	42.44ms

IV. CONCLUSION

In this paper, an effective and robust vehicle classification method has been proposed. By using local and structural features, the proposed method categorizes vehicles into more specific types in multi-view. Sparse coding and multi-scale spatial max pooling are adopted to make the features more discriminative and representative. Experiments show that our proposed method is promising, not only obtains the high classification accuracies, but also works well in different viewpoints. Moreover, the classification speed is fast for real-time application. So, We are confident that our proposed method can be used as a working solution for vehicle classification in practical applications.

ACKNOWLEDGMENT

This work was partially supported by the Shenzhen Science & Technology Fundamental Research Program (No: JCYJ20150430162332418).

REFERENCES

[1] S. Murayama and M. Haseyama, "A note on traffic flow measurement for traffic surveillance video : Reduction of performance degradation in various environments," *Ite Technical Report*, vol. 109, pp. 175–178, 2010.

[2] B. Liu, J. Zhang, and X. Liao, "Projection kernel regression for image registration and fusion in video-based criminal investigation," in *Multimedia and Signal Processing (CMSP), 2011 International Conference on*, 2011, pp. 348–352.

[3] Y. Jiang, *Highway Traffic Automatic Detection System Based on Video and Image Processing*. Springer Berlin Heidelberg, 2013.

[4] P. Mclauchlan, D. Beymer, B. Coifman, and J. Mali, "A real-time computer vision system for measuring traffic parameters," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 1997, pp. 495–501.

[5] S. Gupte, O. Masoud, R. F. K. Martin, and N. P. Papanikolopoulos, "Detection and classification of vehicles," *Intelligent Transportation Systems IEEE Transactions on*, vol. 3, no. 1, pp. 37–47, 2002.

[6] W. L. Hsu, H. Y. M. Liao, B. S. Jeng, and K. C. Fan, "Real-time traffic parameter extraction using entropy," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 151, no. 3, pp. 194–202, 2004.

[7] J. C. Lai, S. S. Huang, and C. C. Tseng, "Image-based vehicle tracking and classification on the highway," in *Green Circuits and Systems (ICGCS), 2010 International Conference on*, 2010, pp. 666–670.

[8] R. Rad and M. Jamzad, "Real time classification and tracking of multiple vehicles in highways," *Pattern Recognition Letters*, vol. 26, no. 10, p. 15971607, 2005.

[9] N. H. C. Yung and A. H. S. Lai, *Detection of vehicle occlusion using a generalized deformable model*. IEEE, 1998.

[10] Z. W. Kim and J. Malik, "Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 524–531 vol.1.

[11] F. Bardet, T. Chateau, and D. Ramadasan, "Unifying real-time multi-vehicle tracking and categorization," in *Intelligent Vehicles Symposium, 2009 IEEE*, 2009, pp. 197–202.

[12] X. Ma and W. E. L. Grimson, "Edge-based rich representation for vehicle classification," in *Proceedings / IEEE International Conference on Computer Vision. IEEE International Conference on Computer Vision*, 2005, pp. 1185–1192 Vol. 2.

[13] N. K. Kanhere, S. T. Birchfield, and W. A. Sarasua, "Vehicle segmentation and tracking in the presence of occlusions," *Journal of the Transportation Research Board*, vol. 1944, no. 1, 2005.

[14] J. W. Hsieh, S. H. Yu, Y. S. Chen, and W. F. Hu, "Automatic traffic surveillance system for vehicle tracking and classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 175–187, 2006.

[15] O. Hasegawa and T. Kanade, "Type classification, color estimation, and specific target detection of moving targets on public streets," *Machine Vision Applications*, vol. 16, no. 2, pp. 116–121, 2005.

[16] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, 1998, pp. 129–136.

[17] S. Song and Z. Miao, "Research on vehicle type classification based on spatial pyramid representation and bp neural network," in *Image and Graphics*, ser. Lecture Notes in Computer Science, Y.-J. Zhang, Ed. Springer International Publishing, 2015, vol. 9219, pp. 188–196.

[18] Y. Peng, Y. Yan, W. Zhu, and J. Zhao, "Binary coding-based vehicle image classification," in *Signal Processing (ICSP), 2014 12th International Conference on*, 2014, pp. 918–921.

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

[20] F. F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories." *Cvpr*, vol. 2, pp. 524–531, 2005.

[21] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[22] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*, 2003, p. 1470.

[23] J. C. V. Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, *Kernel Codebooks for Scene Categorization*. Springer Berlin Heidelberg, 2008.

[24] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and*

Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.

[25] T. Gandhi and M. M. Trivedi, “Video based surround vehicle detection, classification and logging from moving platforms: Issues and approaches,” in *Intelligent Vehicles Symposium, 2007 IEEE*, 2007, pp. 1067–1071.