A Comprehensive Study of the Effect of Spatial Resolution and Color of Digital Images on Vehicle Classification

Khaled F. Hussain, Mahmoud Afifi, and Ghada Moussa

***Abstract*— Vehicle-type classification is considered a core module for many intelligent transportation applications, such as speed monitoring, smart parking systems, and traffic analysis. In this paper, many vision-based classification techniques were presented relying only on a digital camera without the need for any extra hardware components. Dimension and color are two important characteristics of any digital image that affect the cost of the digital camera used in the image acquisition. In this paper, we present a comprehensive study of the effect of these two characteristics on the vehicle classification process in terms of accuracy and performance. We apply a set of different state-of-the-art image classifiers to the BIT-Vehicle and LabelMe data sets. Each data set is downscaled into different scales to generate a variety of spatial resolutions of each data set. Besides, we examine the effect of color by converting each color version to a gray-scale one. At last, we draw a valid conclusion in regards to the impact of these two characteristics (i.e., dimension and color) on the classification accuracy and performance of the image classification methods using more than 46 000 individual experiments. Experimental results show that there is no significant influence of both color and spatial resolutions of the vehicle images on the classification results obtained by most state-of-the-art image classification methods. However, there is a correlation between the spatial resolution and the processing time required by most image classification methods. Our findings can play an important role in saving not only money, but also time for vehicle-type classification systems.**

***Index Terms*— Vehicle classification, vision-based classification, computer vision, deep learning.**

I. Introduction

TRAFFIC applications are becoming increasingly ubiquitous due to the remarkable growth in the number of vehicles. Vehicle classification is considered a substantial module for many traffic applications, such as traffic analysis, speed vehicle monitoring, smart parking systems, fuel type determination (i.e., diesel or petrol), and behavior analysis of drivers. Many methods were developed to improve the accuracy of vehicle classification relying on either specific hardware components (i.e., sensors) or ordinary digital cameras. Nowadays, the accuracy of existing vision-based methods is considered extremely high. However, it is still a challenging task. This category of vehicle classification methods is considered the most efficient way to classify vehicles without the need for extra hardware equipment besides digital cameras. One of the key factors that impacts the cost of these methods, from an economic perspective, is the specifications within the digital camera such as: sensor type, color supported, bit depth, dynamic range, digital and optical zoom, focal length, ISO/IR sensitivity, shutter speed, frame rate, manufacturer, and spatial resolution. Despite the numerous specifications of digital cameras, the spatial resolution is considered one of the main factors affecting their cost [1], [2]. Where, the cost of a digital camera is directly proportional to its spatial resolution [3]. Generally, color digital cameras are more expensive than the gray-scale ones. Moreover, color images with high resolution require more processing time and storage space compared to low resolution gray-scale images.

In this paper, we present a wide study on the effect of both spatial resolution and color of digital images on the accuracy and performance of vision-based vehicle classification methods. We aim to provide an answer for the question: what are the best values for these two properties of digital cameras to use in vision-based vehicle classification systems? In other words, how many pixels are sufficient to build an accurate vision-based vehicle classification system? In order to draw a coherent relation between the classification accuracy, and the spatial resolution and color of vehicle images, we use the BIT-Vehicle [4] and LabelMe [5] datasets. We apply many vision-based classification methods to various versions of each dataset with a different spatial resolution and color. The answer of the mentioned question helps to understand the effect of these two properties on vehicle classification methods. Consequently, that helps to decide the most appropriate digital camera to use in vehicle classification systems.

The rest of this paper is organized as follows. In Section II, we present a review of the existing vehicle classification methods. In Section III, we present the details of how this study is conducted and a brief description of the vision-based classifiers used in this study. Next, the experimental results

Manuscript received May 19, 2017; revised December 24, 2017 and April 2, 2018; accepted May 11, 2018. The Associate Editor for this paper was J. Zhang. *(Corresponding author: Khaled F. Hussain.)*

K. F. Hussain is with the Computer Science Department, Faculty of Computers and Information, Assiut University, Asyut 71515, Egypt (e-mail: khussain@aun.edu.eg).

M. Afifi is with the Electrical Engineering and Computer Science Department, Lassonde School of Engineering, York University, Toronto, ON M3J 1P3, Canada.

G. Moussa is with the Civil Engineering Department, Faculty of Engineering, Assiut University, Asyut 71515, Egypt.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2838117

TABLE I

A SUMMARY OF DESCRIPTORS USED IN THIS STUDY AS AN EMBEDDED STEP IN THE HANDCRAFTED FEATURE-BASED METHODS (i.e., BoVW [13], VLAD [14], AND FV [15])

Descriptor	Brief description
SIFT [16], [17]	Difference of Gaussian filtering (DoG) is used to detect robust features (salient points). Then, HOG is used to describe these features by producing a 128-dimensional feature vector.
DSIFT [18]	Dense SIFT (DSIFT) is similar to SIFT, however, it does not detect the salient points, SIFT descriptors are computed in densely sampled locations instead.
Opp-SIFT	Opponent-SIFT applies SIFT descriptor with taking into account the color information of the image. The opponent-color channels are used instead of the RGB color space. Each pixel color is represented using the following equation: $\begin{bmatrix}\hat{E_1}\\ \hat{E_2}\\ \hat{E_3}\end{bmatrix} = \begin{bmatrix}0.06 & 0.63 & 0.27\\ 0.3 & 0.04 & -0.35\\ 0.34 & -0.6 & 0.17\end{bmatrix}\begin{bmatrix}R\\ G\\ B\end{bmatrix}$, where $\hat{E_1}$ and $\hat{E_2}$ are the color information and $\hat{E_3}$ represents the intensity information [19].
Opp-Hist	The opponent-histogram is based on the opponent-like color system. Where, it is based on constructing a combined histogram containing three one-dimensional histograms of $\hat{E_1}$, $\hat{E_2}$, and $\hat{E_3}$ [19]. The opponent-histogram is considered a specular invariant descriptor by using the opponent angle Θ given by: $\Theta = tan^{-1}(\frac{\hat{E_1}'}{\hat{E_2}'})$, where $\hat{E_1}'$ and $\hat{E_2}'$ represent the first derivative of $\hat{E_1}$ and $\hat{E_2}$, respectively. The opponent-histogram is represented using 36 bins.
HSV-SIFT [20]	By applying SIFT to the three components of the HSV color space separately, the HSV-SIFT descriptor is created. This descriptor is represented by a 384-dimensional feature vector.
HUE-Hist [21]	To avoid the lack of precision of HUE values in the case of low saturated colors, Van de Weijer *et al.* [21] suggested that by weighting each HUE sample individually with respect to its saturation, the problem of the HUE instability can be fixed. HUE-histogram describes the given image using 36 bins.
HUE-SIFT [21]	To increase the accuracy of discrimination of SIFT descriptors, color features are integrated using the weighted hue histogram.
C-SIFT [22]	Colored SIFT (C-SIFT) constructs the SIFT descriptors in a Color invariant (C-invariant) space ($\frac{\hat{E_1}}{\hat{E_3}}$, $\frac{\hat{E_2}}{\hat{E_3}}$).
*rg*SIFT	SIFT descriptor is applied on the image in a normalized color space rg that is computed by: $r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$. Since the r and g values are calculated from a linear mapping of the RGB components, *rg*SIFT is invariant to linear light intensity changes in addition to being scale/shift invariant descriptor [19].
RGB-SIFT	This descriptor is applied on the image in a normalized color space, which is generated using the following equation: $\begin{pmatrix}R_{norm}\\ G_{norm}\\ B_{norm}\end{pmatrix} = \begin{pmatrix}\frac{R-\mu_R}{\sigma_R}\\ \frac{G-\mu_G}{\sigma_G}\\ \frac{B-\mu_B}{\sigma_B}\end{pmatrix}$. Like *rg*SIFT, RGB-SIFT descriptors are created for each normalized color channel [19].
RGB-Hist	The RGB-Hist descriptor is based on the RGB color space without any consideration for invariant properties.

are elaborated and discussed in Section IV. Lastly, we present conclusions and recommendations drawn from this empirical study in Section V.

II. EXISTING VEHICLE CLASSIFICATION METHODS

The existing vehicle classification methods can be grouped into two main categories: (1) hardware-based and (2) vision-based methods. In this section, we briefly review the related research methods of each category.

A. Hardware-Based Methods

There are many techniques developed using specific hardware components rather than, or in addition to, cameras. A range sensor was used in [6] to obtain vehicle features (i.e., length, width, height, and speed). The classification was performed using a feed forward neural network. Global Positioning System (GPS) was used to estimate vehicle trajectories as a distinctive feature to classify vehicles [7]. Ma *et al.* [8] achieved a high classification accuracy using wireless accelerometer and magnetometer sensors to calculate vehicle axles. Although these methods obtain high classification accuracy, specific hardware equipment and settings are required.

B. Vision-Based Methods

Many vision-based methods were developed, solely based on regular digital images, for vehicle classification. Some of them focus on the geometrical characteristics of vehicles, while other were designed to detect and determine vehicle logos. On the other hand, other techniques rely on visual features, which is considered a robust mechanism to get a neat vision-based vehicle classification system.

1) Dimension-Based Methods: By fitting a vehicle image with a deformable vehicle model, Lai *et al.* [9] estimated vehicles' dimensions to classify vehicles according to their 3D dimensions. As this method obtains inadequate classification accuracy, Zhang *et al.* [10] used a generic 3D vehicle model containing twelve shape parameters for vehicle recognition. The 3D model is projected into the vehicle's image by estimating the twelve shape parameters and vehicle's pose using a local gradient-based method.

2) Logo-Based Methods: Recognizing the vehicle's logo is another way for vehicle classification, where the vehicle's logo and model indicate its category. The Region Of Interest (ROI) is determined using a sliding window, and recognized using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM) by Llorca *et al.* [11]. Peng *et al.* [12]

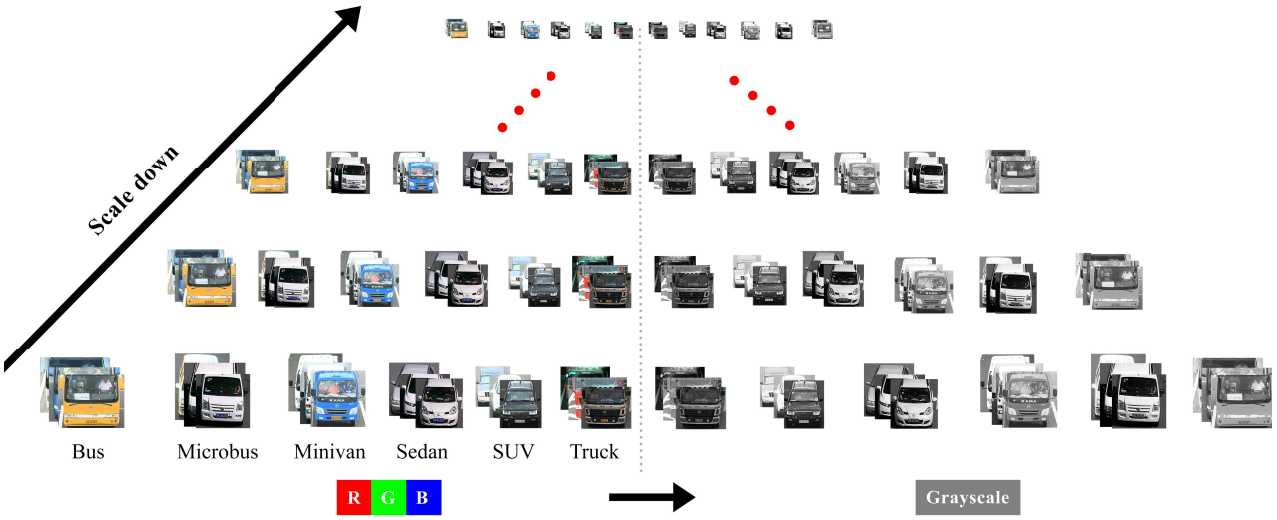

Fig. 1. An overview of the pre-processing of the BIT-Vehicle dataset [4]. Each category in the dataset is down-sampled 9 times by using bicubic interpolation. In the right side of the hierarchy, each category is converted to a new gray-scale version by using a weighted sum of the color channels. This process is applied to the LabelMe dataset [5] as well.

presented a technique based upon the statistical random sparse distribution to represent vehicle logos in frontal-view poor quality vehicle images.

3) Feature-Based Methods: Feature-based methods deal with extracting visual handcrafted features from the vehicles' images using local or global feature representation. In the literature, there are many handcrafted feature descriptors developed to detect and describe visual features in digital images, see Table I. In the vehicle classification field, visual features are computed using extracted edge points [23], Scale Invariant Feature Transform (SIFT) features [24], Principal Component Analysis (PCA) [25], Gabor and log-Gabor filters [26], [27], Haar-like features [28], [29], HOG and Pyramid Histogram of Oriented Gradients (PHOG) [30], [31], or HOG with simple shape-based features [32].

Machine learning techniques, such as k-nearest neighbor, SVMs, and deep learning techniques, were used for vehicle classification [4], [6], [28]. Recently, deep learning is considered the state-of-the-art to solve many problems, such as speech recognition, document retrieval, and object recognition [33]. In image classification challenges, deep learning has received a great attention because of its remarkable accuracy. Dong *et al.* [4] presented a semi-supervised Convolutional Neural Network (CNN) achieving 88.11% accuracy on the BIT-Vehicle dataset [4]. Zhou *et al.* [34] proposed two learning networks to detect the vehicle and extract its attributes (i.e., viewpoint, color, and type) from traffic surveillance videos.

In general, many other approaches used to solve image classification problems, such as Bag-of-Visual Words (BoVW) [13], Vector of Locally Aggregated Descriptors (VLAD) [14], and Fisher Vector (FV) [15].

III. Datasets and Methods

To conduct this study, we apply several image classification techniques to two different datasets. The first one, the BIT-Vehicle dataset, is a complex dataset of frontal-view vehicle images collected under different conditions. The second one, the LabelMe dataset, is used to study the effect of the spatial resolution and color on the vehicle detection problem —vehicle/non-vehicle classification. In this section, we concisely describe the datasets and the classification methods used in this study.

TABLE II
Statistics of the Original Spatial Resolution of the BIT-Vehicle Dataset [4]

Category	Width (pixel)			Height (pixel)		
	Min.	Max.	Mean	Min.	Max.	Mean
Bus	441	1,087	875.0	381	1,029	705.1
Microbus	252	903	528.4	256	1,135	664.4
Minivan	298	1,129	660.0	320	1,135	684.1
Sedan	250	770	509.9	225	867	519.0
SUV	257	819	561.9	264	942	652.5
Truck	430	1,268	890.7	282	1,200	758.5
Total	**321.3**	**996**	**671.0**	**288**	**1,051.3**	**663.9**

A. Preparation of the Datasets

Dong *et al.* [4] presented the BIT-Vehicle dataset which has many challenges, such as bad weather and lighting conditions, background confusion, and variety of vehicle colors. This dataset covers six different categories of vehicles (Sedan, Microbus, SUV, Bus, Minivan, and Truck). The dataset contains 9,850 high-quality frontal-view images whose spatial resolutions are either 1,920 × 1,080 pixels or 1,600 × 1,200 pixels. Each image contains one or more vehicles. There are 558 Bus images, 883 Microbus images, 1,392 SUV images, 5,919 Sedan images, 476 Minivan images, and 823 Truck images. Table II shows basic statistics of the original spatial resolution of each category in the dataset.

Another dataset presented by MIT known as LabelMe dataset [5], that contains a large collection of annotated

real scene images for many different objects. The dataset is continually growing by uploading new images or adding annotations to existing images. In this study, a dataset of vehicles and non-vehicle images from the LabelMe dataset is used (1,810 images for each class). The average spatial resolutions of vehicle and non-vehicle images are 1903 $\times$ 1412 and 1557 $\times$ 1122 pixels, respectively.

In this study, we aim to study the effect of spatial resolution on the accuracy and performance of image classification methods. By down-sampling the original dataset D_1 nine times (10% for each time), we generate new datasets $D_{0.9}, D_{0.8}, ...D_{0.1}$ with gradual scales. For each image in the original dataset, we apply bicubic resampling to generate a downscaled image by $\mathbb{S}$ scale factor, where $\mathbb{S} \in \{0.9, 0.8, ..., 0.1\}$. In order to study the color effect on the vehicle classification process, we convert each dataset to a gray-scale version using the weighted summation of color images' RGB values, see Fig. 1. These processes were applied to both the BIT-Vehicle and LabelMe datasets.

B. Classification Methods

In order to draw a robust conclusion, we use four different classification methods, which are: (1) BoVW [13], (2) VLAD [14], (3) FV [15], and (4) CNNs [35]–[39] Fig. 2 summarizes the handcrafted image classification methods used in this study.

1) Bag-of-Visual Words (BoVW): BoVW is based on building a distribution of extracted visual features from a given image. These features are clustered using K-means clustering algorithm [40] to create a set of clusters. The centers of these clusters are called visual words or codewords. These clusters construct a vocabulary (also called a codebook). Thus, each image can be represented as a vector of histogram of codewords which is called BoVW. In the training step, a set of labeled images is used to create a pool of vectors. The decision boundaries among the vectors are determined by an SVM classifier to categorize further images. Csurka *et al.* [13] used Harris affine detector [41] to detect the salient points to specify local patches of images. The local patches are described by SIFT in order to create histogram of features. In this study, we test different descriptors in addition to the SIFT descriptor, see Table I.

2) Vector of Locally Aggregated Descriptors (VLAD): VLAD is considered an extension of the BoVW. It uses the K-means to cluster the descriptors of all the training images to construct the visual vocabulary. For each image, a vector quantization is applied to each descriptor by finding its approximate nearest neighbor visual words. Rather than computing the histogram of visual words only, the concatenation of the sum of vector difference between image descriptors and visual words is computed. Then, the component-wise mass normalization is applied. For each training image, the VLAD descriptors are calculated, and used with the associated labels to train an SVM classifier.

3) Fisher Vector (FV): FV is considered as an alternative to the BoVW. FV uses Gaussian Mixture Models (GMMs) instead of the K-means to generate the visual vocabulary.

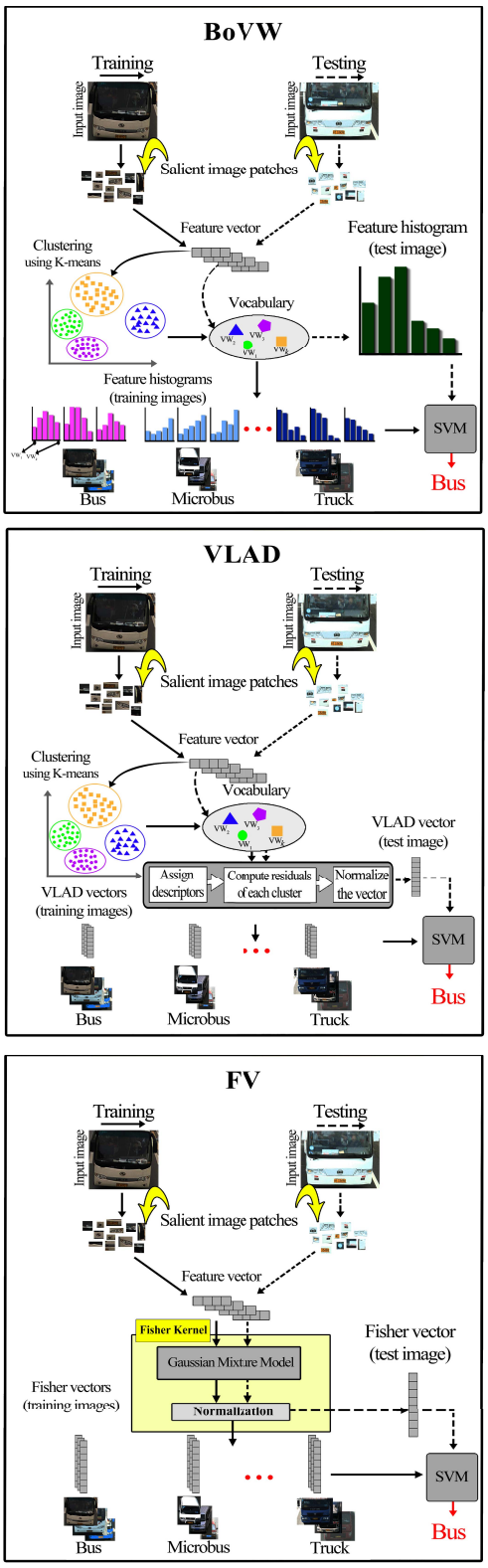

Fig. 2. A summary of the handcrafted image classification methods used in this study (BoVW [13], VLAD [14], and FV [15]).

It computes the second-order information for visual words in the visual vocabulary. For each image, the concatenation of the mean and covariance deviation vectors of the descriptors, for each GMM, is computed. Lastly, the vector is normalized.

TABLE III

BEST ACCURACIES (%) ACHIEVED USING BOVW, VLAD, FV, AND CNN METHODS FOR COLOR AND GRAY VERSIONS OF **THE BIT-VEHICLE DATASET** [4]. THE LAYOUT $L2$ REFERS TO CONSTRUCTING TWO IMAGE PYRAMID LEVELS FROM THE VEHICLE IMAGE

Scale	**Classification method**							
	BoVW		**VLAD**		**FV**		**CNN/ResNet**	
	Gray DSIFT ($L2$)	**Color RGB-SIFT** ($L2$)	**Gray DSIFT** ($L2$)	**Color RGB-SIFT** ($L2$)	**Gray DSIFT** ($L2$)	**Color RGB-SIFT** ($L2$)	**Gray**	**Color**
0.1	83.75	36.25	91.32	86.21	92.43	86.75	90.39	90.71
0.2	87.43	76.88	93.47	89.42	93.83	90.04	93.51	93.90
0.3	89.50	77.83	94.24	89.75	94.82	92.21	**94.90**	**95.12**
0.4	88.17	80.13	94.03	90.25	94.85	92.17	-	-
0.5	88.50	81.92	94.48	90.88	94.63	92.79	-	-
0.6	88.71	81.96	94.63	90.33	94.66	91.54	-	-
0.7	88.42	81.33	94.25	90.38	94.75	92.08	-	-
0.8	88.50	81.83	94.14	90.08	94.52	93.08	-	-
0.9	88.67	82.67	94.08	90.58	94.43	91.68	-	-
1.0	88.25	82.75	94.27	90.58	94.71	92.29	-	-

4) Deep Learning for Image Recognition: The CNNs use weight-sharing to exploit the nature that comparable structures happen in within various areas in an image. The CNNs consist of different types of layers, such as Convolutional, Pooling, and Fully-Connected (FC) layers. Each layer transforms the 3D input volume to a 3D output volume of neuron activations. There are different CNN architectures presented in the last few years. For example, AlexNet [35] consists of 60 million parameters and 650,000 neurons. It has five Convolutional layers with Max-Pooling on the first, second, and fifth layers. There are three FC layers added to the end of the network. VGGNet [37] is another powerful network that is deeper than AlexNet. The main contribution of the VGGNet is showing that deeper network is more effective. A drawback of the VGGNet is that it is more costly to train due to the large number of parameters used compared to AlexNet. GoogLeNet [38] was presented to increase the depth and width of the network while decreasing the number of parameters in the network by the development of Inception Module and replacing the FC layer with Average Pooling layer. Another powerful architecture is ResNet [39]. The two main contributions of the ResNet are: (1) the use of a ResNetblock to learn residual functions with reference to the ResNetblock inputs instead of learning unreferenced functions, and (2) the use of deeper network (a depth of up to 152 layers). CNNs, such as AlexNet, VGGNet, GoogLeNet, and ResNet, are trained using large numbers of images. From these images, CNNs learn multiple levels of rich feature representations from an extensive variety of images. The learned features are often better than handcrafted features, such as SIFT and HOG. It is faster to fine-tune pre-trained CNNs instead of training a new CNN from scratch to solve a new problem. This process is implemented by using an existing pre-trained CNN as an initial point in the training phase.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the experimental results of applying the aforementioned classification methods to the datasets discussed in Section III. We performed 46,800 individual experiments using the LabelMe and BIT-Vehicle datasets (1,800 experiments for each color version and 540 experiments for each gray-scale version of each dataset) using the Matlab implementations of BoVW, VLAD, FV, and CNN architectures presented by Vedaldi and Fulkerson [42] and Vedaldi and Lenc [43]. The experiments were done on an Intel® core™ 2 Quad CPU Q9550 @ 2.83GHz machine with 12 GB RAM and NVIDIA® GeForce® GTX 960 graphics card.

We followed the cross-validation approach used by Dong *et al.* [4]. For each category in the dataset $D_{\mathbb{S}}$, we randomly picked 200 images for training and 200 images for testing; the training and testing sets have no images in common. For the image classification methods that utilize color-based descriptors, the experiments were carried out using only the color versions of the datasets. As described in Section III, we have tested two groups of image classification methods (i.e., handcrafted and CNN methods). For the handcrafted methods, we have used 1,024 and 64 clusters to construct

TABLE IV

CONFUSION MATRIX OF ONE OF THE EXPERIMENTS FOR THE BEST ACCURACY ACHIEVED USING RESNET [39] ON THE **BIT-VEHICLE DATASET** [4]

	Confusion matrix obtained using $D_{0.3}$					
	Bus	**Microbus**	**Minivan**	**Sedan**	**SUV**	**Truck**
Bus	**1.000**	0.000	0.000	0.000	0.000	0.000
Microbus	0.005	**0.955**	0.000	0.015	0.025	0.000
Minivan	0.000	0.025	**0.920**	0.000	0.010	0.045
Sedan	0.000	0.025	0.000	**0.930**	0.045	0.000
SUV	0.000	0.025	0.005	0.035	**0.935**	0.000
Truck	0.000	0.005	0.030	0.000	0.000	**0.965**

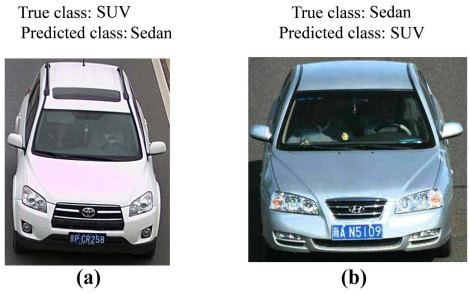

Fig. 3. Examples of the visual similarities between vehicle images that belong to different categories. (a) a true-color SUV image is misclassified by ResNet [39] as a Sedan vehicle image. (b) a true-color Sedan image is misclassified by ResNet as an SUV vehicle.

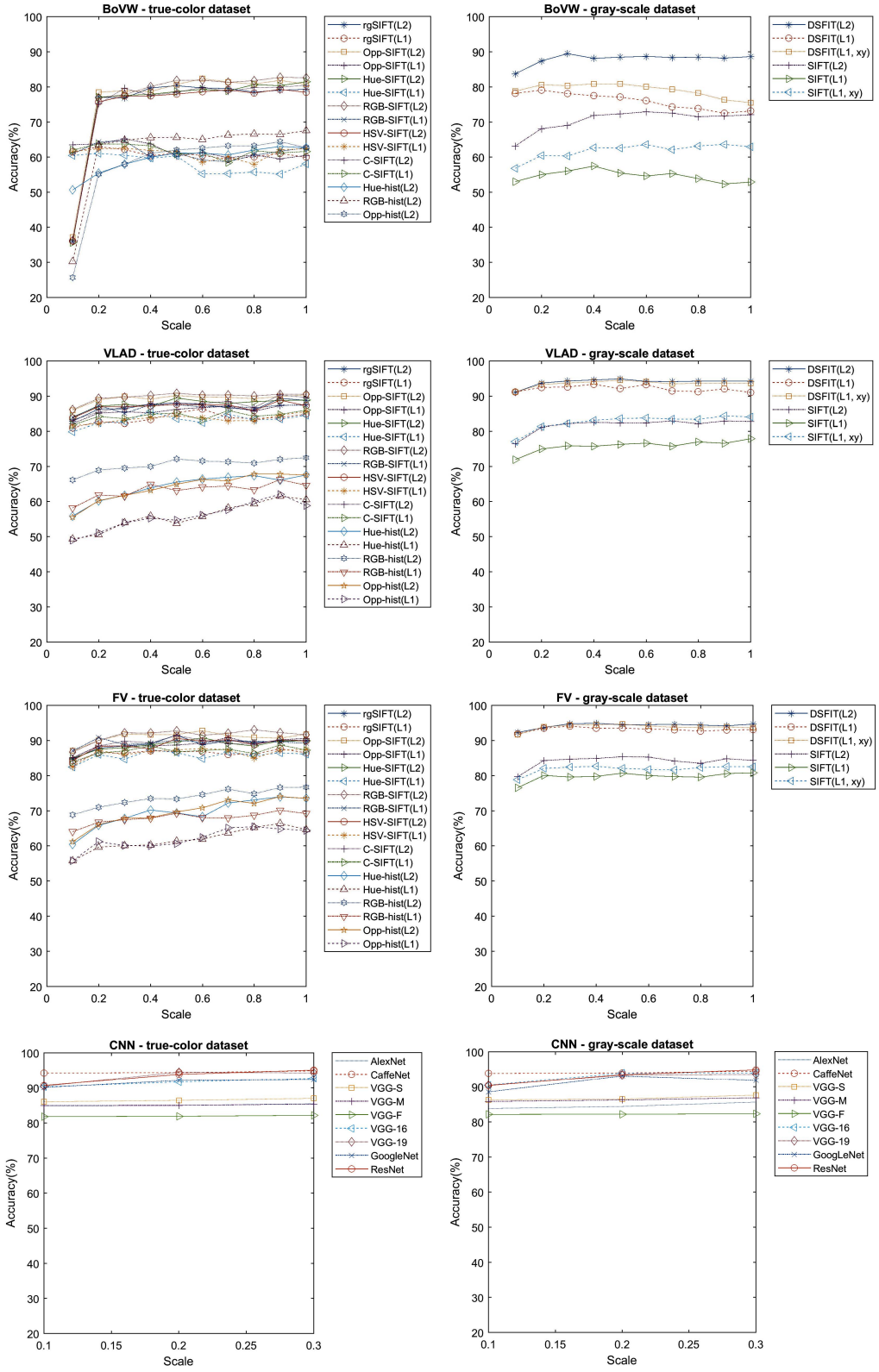

Fig. 4. The effect of down-sampling on recognition rate using the **BIT-Vehicle dataset** [4]. The methods whose average recognition rate over all datasets is less than 50% were excluded.

the visual vocabulary of the BoVW, and VLAD and FV, respectively. CNNs were trained using stochastic gradient descent with momentum for 200 epochs, or until there was no improvement of the classification accuracy. The learning rates of CNNs were $\lambda = 10^{-3}$ for the last FC layer of each model and $\lambda = 10^{-4}$ for the other layers.

TABLE V

BEST ACCURACIES (%) ACHIEVED USING BOVW, VLAD, FV, AND CNN METHODS FOR COLOR AND GRAY VERSIONS OF THE **LABLEME DATASET** [5]. THE LAYOUT $L2$ REFERS TO CONSTRUCTING TWO IMAGE PYRAMID LEVELS FROM THE VEHICLE IMAGE

Scale	Classification method							
	BoVW		VLAD		FV		CNN/ResNet	
	Gray DSIFT ($L2$)	Color HUE-SIFT ($L2$)	Gray DSIFT ($L2$)	Color HUE-SIFT ($L2$)	Gray DSIFT ($L2$)	Color HUE-SIFT ($L2$)	Gray	Color
0.1	94.24	94.24	94.70	91.54	95.50	93.09	98.87	97.52
0.2	95.08	93.03	96.59	95.74	96.00	96.24	98.79	97.43
0.3	94.58	91.87	97.09	94.55	95.56	97.06	**99.34**	**98.95**
0.4	94.07	94.06	95.70	95.24	98.02	94.22	-	-
0.5	96.59	93.77	98.00	96.00	97.20	95.50	-	-
0.6	96.21	94.29	96.04	96.56	96.77	97.24	-	-
0.7	93.72	93.34	98.07	95.26	96.54	96.50	-	-
0.8	93.27	94.06	95.51	97.57	95.74	97.03	-	-
0.9	94.75	93.39	94.75	96.72	95.58	96.23	-	-
1.0	94.71	94.26	96.75	96.06	98.58	96.21	-	-

For each method described in Section III, we reported the average of both testing time and recognition rates of the 20 experiments for each dataset. BoVW, VLAD, and FV methods were applied using all descriptors in Table I. In this stage, we have used two different layouts to represent input images: (1) using the whole input image ($L1$) and (2) constructing two spatial pyramid levels ($L2$) to increase the accuracy. The first level of the pyramid contains the whole input image. The second level is constructed by dividing the input image into four sub-images. Moreover, the impact of feature locations (xy) are considered for each descriptor.

We fine-tuned nine pre-trained CNNs: (1) AlexNet [35], (2) CaffeNet [36], (3) VGG-S, (4) VGG-M, (5) VGG-F, (6) VGG-16, (7) VGG-19 [37], (8) GoogLeNet [38] and (9) ResNet [39]. All the mentioned CNN architectures require a specific image dimension (224 $\times$ 224 pixels, except alexNet which requires 227 $\times$ 227 pixels). Consequently, the experiments were carried out using the dataset versions that are below the required resolution of the CNN architectures, namely the datasets $D_{0.3}$, $D_{0.2}$, and $D_{0.1}$.

A. A Comparative Study of Accuracy

Table III shows best accuracies achieved by the classification methods (i.e., BoVW, VLAD, FV, and CNNs) on the BIT-Vehicle dataset, using testing images. Where, the CNNs/ResNet attains the highest accuracy (95.12%), which outperforms the accuracy reported by Dong *et al.* [4] (88.11%). Using training images, all classification methods achieve 100% accuracy except the BoVW that attains 93.78% on average. Table IV demonstrates the confusion matrix for the best accuracy achieved using ResNet. We justify that the misclassification occurred due to the high similarity of some vehicle images that belong to different categories, as shown in Fig. 3.

The DSIFT descriptor can be considered as the best descriptor for the BoVW, VLAD, and FV methods, as it achieves better results than other descriptors for most cases, as shown in Fig. 4. For all descriptors, features' locations do not improve the classification accuracy except in the case of SIFT and DSIFT descriptors with the layout $L1$. Moreover, the $L2$ layout is better than $L1$ for all descriptors.

TABLE VI

PRECISION, RECALL, AND F1 SCORE FOR THE BEST ACCURACY ACHIEVED USING BOVW, VLAD, FV, AND CNN METHODS ON THE **LABELME DATASET** [5]

	Classification method			
	BoVW DSIFT ($L2$)	VLAD DSIFT ($L2$)	FV DSIFT ($L2$)	CNN ResNet
Precision	0.995	0.995	0.995	0.995
Recall	0.939	0.966	0.975	**0.995**
F1 score	0.966	0.980	0.985	**0.995**

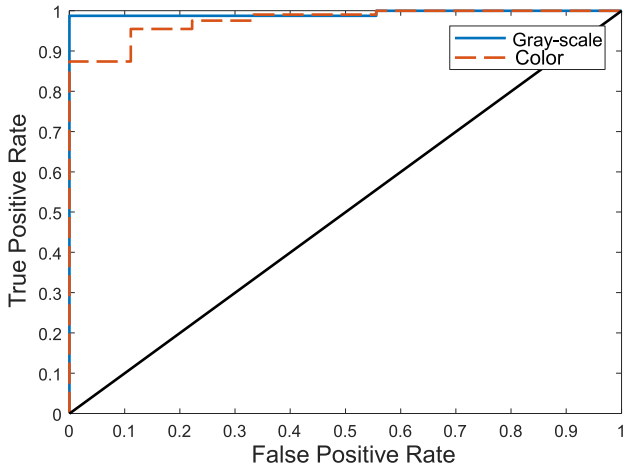

Fig. 5. ROC curve for the best accuracy achieved using ResNet for color and gray versions of the **LabelMe dataset** [5].

For vehicle/non-vehicle images from the LabelMe dataset, best accuracies achieved by the classification methods (i.e., BoVW, VLAD, FV, and CNNs), using testing images, are presented in Table V. The ResNet outperforms other methods not only in the classification accuracy but also in the recall and F1 score, as shown in Table VI. Fig. 5 shows the ROC curve for the best accuracy obtained by the ResNet on the LabelMe dataset. Using training images, a 100% accuracy was attained by all classification methods.

B. Resolution/Accuracy Analysis

Classification methods with average recognition rates, over all datasets, less than 50% were excluded. Fig. 4 illustrates the effect of down-sampling on recognition rates using the

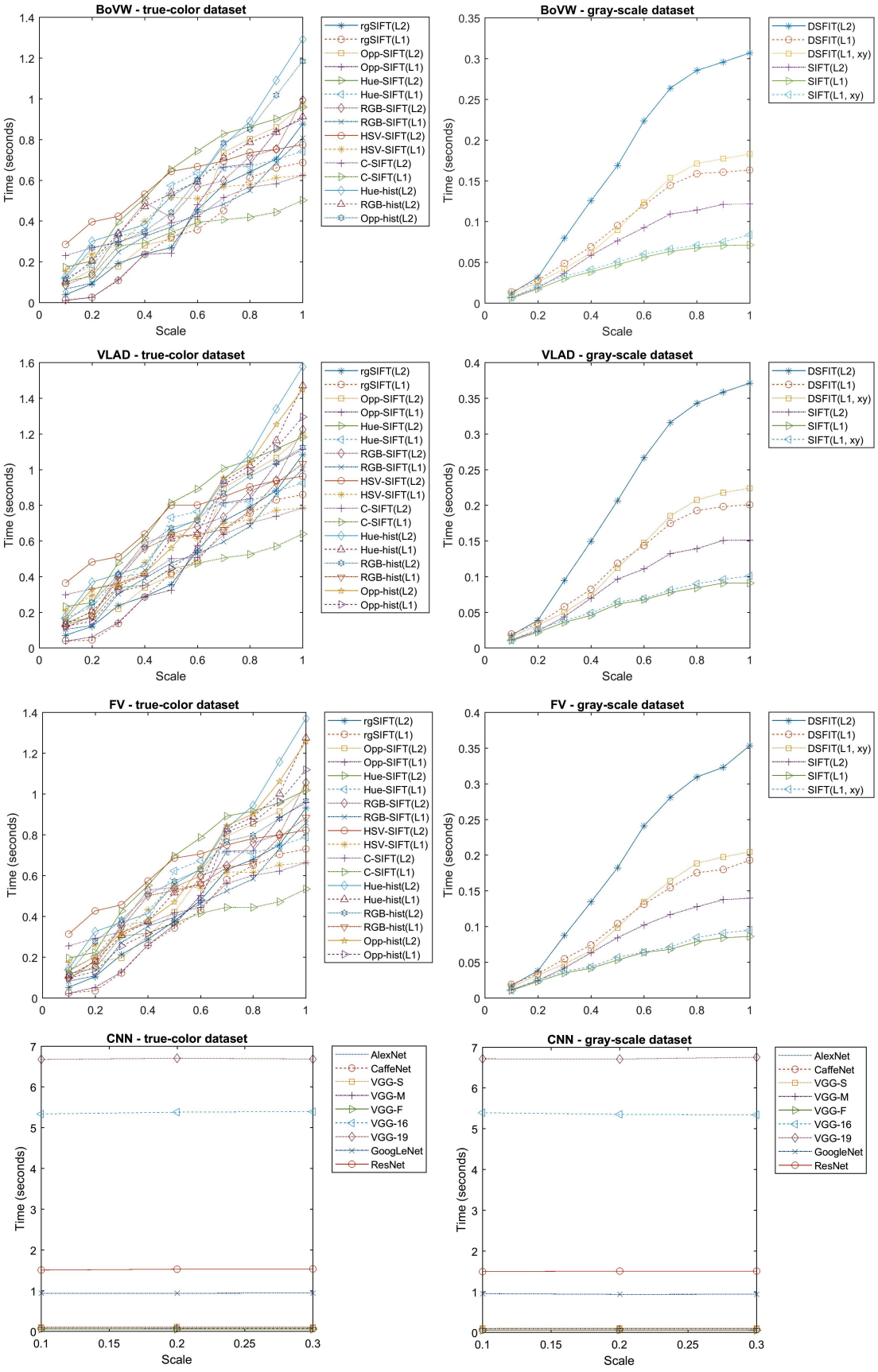

Fig. 6. The effect of down-sampling on the performance (testing time). The time shown is the average time required per image during the testing process on CPU. The classification methods whose average accuracy over all datasets is, roughly, equal or less than 50% were excluded.

BIT-vehicle dataset. Generally, the recognition rates of almost all classification methods are stable against changes in spatial resolution and color of the datasets. Where reducing the size of an image is equivalent to keeping low-frequency components and trimming the high-frequency components. The majority of image perceptual information is stored in the low-frequency components. Thus, reducing the size of an image has no significant influence on the vehicle classification accuracy.

It is worth noting the following observations. First, the accuracy goes down when we downscale the dataset to 10% of the original size, especially with the BoVW. Second, the recognition rates using gray-scale descriptors are, in most cases, better than those with color descriptors. This indicates that the color information may mislead the classifier rather than be an improving factor, for applications which focus on general classification of vehicle types. Eventually, we can note that the same findings are applicable to the vehicle/non-vehicle classification, see Table V that shows the results obtained using the LableMe dataset.

C. Resolution/Time Analysis

For a fair comparison between handcrafted and CNN-based methods, we have tested the CNN architectures on CPU. The training time is not an issue, since the process is performed off-line before deploying the classification method. The resolution/time curves in Fig. 6 illustrate the relationship between the spatial resolution of images and the time. As the image size becomes smaller, the time required to classify the given image is reduced for BoVW, VLAD, and FV classification methods. In the case of using CNN, the time is approximately constant because the CNN architectures require specific image dimensions that all images are resized into before being given as input to the CNN. We can notice that VGG-19 is considered the worst choice in terms of inference time.

Finally, Fig. 6 shows that the color also has an impact on the performance. The time required to classify gray-scale images is less than or equal to the time required for color images. The reason is that the computational cost for one gray-scale channel is less than the computational cost for three color channels.

V. Conclusions and Recommendations

In this paper, we have presented a comprehensive study of the effect of both image spatial resolution and color on the vehicle classification accuracy. We have studied the relationship between image resolutions and the time required for the classification process. We have attempted to answer the questions: how much spatial resolution is sufficient to build an accurate vision-based vehicle classification system? Do image colors have an obvious effect on vehicle classification accuracy? That leads to more understanding of the influence of the camera characteristics (i.e., spatial resolution and color) on the vehicle classification accuracy obtained by different methods. Moreover, this paper has presented a comparative study of the effectiveness of different methods (i.e., BoVW, VLAD, FV, and CNNs) in the vehicle classification task. We have found that there is no significant influence of both color and spatial resolutions of the vehicle images on the vehicle classification accuracy for most cases. The ResNet architecture is considered the top classification method with an accuracy rate of 95.12% using the BIT-Vehicle dataset and 99.34% using the LabelMe dataset. For BoVW, VLAD, and FV methods, it is recommended to use the DSIFT descriptor. Where, DSIFT outperforms SIFT, RGB-SIFT, CSIFT, HSV-SIFT, hue-SIFT, opponent-SIFT, *rg*SIFT, opponent-histogram, hue-histogram, and RGB-histogram in most cases. Eventually, results have shown that the color information is not a significant parameter for general vehicle type classification applications. To sum up, two recommendations are extracted from the experiments for designing fully automated vehicle classification system:

1) This study shows that images with low spatial resolution achieve approximately the same recognition rates as images with high spatial resolution for vehicle classification. That means that higher spatial resolution images have too many unnecessary details for vehicle classification and low resolution images filter out these unnecessary details. Thus, there is no need for a camera with a high spatial resolution —a camera with low spatial resolution is adequate for the vehicle classification task. Specifically, the vehicle pixels (i.e., the ROI) in the image should be equal or larger than $\sim 150 \times 150$ pixels on averages for the BoVW. For the VLAD, FV, and CNN-based methods, this condition can be omitted, where they are stable even using $\sim 90 \times 90$ pixels on average. Furthermore, the processing time for images with high spatial resolution is greater than the processing time for images with low spatial resolution.
2) There is no need for true-color cameras; a monochrome camera is sufficient, especially if the main goal of the system is general vehicle type classification. Color information does not improve the classification accuracy for most image classification frameworks.

References

[1] W. S. Warner and B. R. Slaattelid, "Multiplotting with images from the Kodak DCS420 digital camera," *Photogramm. Rec.*, vol. 15, no. 89, pp. 665–672, 1997.
[2] J. Xiao, "Technological advances in digital cameras: Welfare analysis on easy-to-use characteristics," *Marketing Lett.*, vol. 19, no. 2, pp. 171–181, 2008.
[3] R. J. Poulo, V. X. Nguyen, V. S. Bostrom, and W. L. Gorman, "Creating high resolution images," U.S. Patent 6 535 650, Mar. 18, 2003.
[4] Z. Dong, Y. Wu, M. Pei, and Y. Jia, "Vehicle type classification using a semisupervised convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2247–2256, Aug. 2015.
[5] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.
[6] K. F. Hussain and G. S. Moussa, "Automatic vehicle classification system using range sensor," in *Proc. IEEE Conf. Inf. Technol., Coding Comput. (ITCC)*, vol. 2, Apr. 2005, pp. 107–112.
[7] Z. Sun and X. Ban, "Vehicle classification using GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 37, pp. 102–117, Dec. 2013.
[8] W. Ma *et al.*, "A wireless accelerometer-based automatic vehicle classification prototype system," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 104–111, Feb. 2014.
[9] A. H. Lai, G. S. K. Fung, and N. H. C. Yung, "Vehicle type classification from visual-based dimension estimation," in *Proc. IEEE Intell. Transp. Syst.*, Aug. 2001, pp. 201–206.

[10] Z. Zhang, T. Tan, K. Huang, and Y. Wang, "Three-dimensional deformable-model-based localization and recognition of road vehicles," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 1–13, Jan. 2012.

[11] D. F. Llorca, R. Arroyo, and M. A. Sotelo, "Vehicle logo recognition in traffic images using HOG features and SVM," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 2229–2234.

[12] H. Peng, X. Wang, H. Wang, and W. Yang, "Recognition of low-resolution logos in vehicle images based on statistical random sparse distribution," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 681–691, Apr. 2015.

[13] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis. (ECCV)*, Prague, Czech Republic, vol. 1, nos. 1–22, 2004, pp. 1–2.

[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3304–3311.

[15] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[17] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[18] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 524–531.

[19] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[20] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[21] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, Jan. 2006.

[22] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1978–1983.

[23] X. Ma and W. E. L. Grimson, "Edge-based rich representation for vehicle classification," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1185–1192.

[24] X. Zhang, N. Zheng, Y. He, and F. Wang, "Vehicle detection using an extended hidden random field model," in *Proc. 4th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1555–1559.

[25] Y. Peng, J. S. Jin, S. Luo, M. Xu, and Y. Cui, "Vehicle type classification using PCA with self-clustering," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2012, pp. 384–389.

[26] A. Nurhadiyatna, A. L. Latifah, and D. Fryantoni, "Gabor filtering for feature extraction in real time vehicle classification system," in *Proc. 9th Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2015, pp. 19–24.

[27] J. Arróspide and L. Salgado, "Log-Gabor filters for image-based vehicle verification," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2286–2295, Jun. 2013.

[28] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 508–517, Mar. 2015.

[29] X. Wen, L. Shao, Y. Xue, and W. Fang, "A rapid learning algorithm for vehicle classification," *Inf. Sci.*, vol. 295, pp. 395–406, Feb. 2015.

[30] M. Cheon, W. Lee, C. Yoon, and M. Park, "Vision-based vehicle detection system with consideration of the detecting location," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1243–1252, Sep. 2012.

[31] B. Zhang, "Reliable classification of vehicle types based on cascade classifier ensembles," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 322–332, Mar. 2013.

[32] H. C. Karaimer, I. Baris, and Y. Bastanlar, "Detection and classification of vehicles from omnidirectional videos using multiple silhouettes," *Pattern Anal. Appl.*, vol. 20, no. 3, pp. 893–905, 2017.

[33] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[34] Y. Zhou, L. Liu, L. Shao, and M. Mellor, "DAVE: A unified framework for fast vehicle detection and annotation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 278–293.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[36] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[37] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–11.

[38] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[41] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. Eur. Conf. Comput. Vis.* Copenhagen, Denmark: Springer, 2002, pp. 128–142.

[42] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.

[43] A. Vedaldi and K. Lenc, "Matconvnet—Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

Khaled F. Hussain received the B.S. and M.S. degrees in electrical engineering from Assiut University, Assiut, Egypt, in 1994 and 1996, respectively, and the Ph.D. degree in computer science from University of Central Florida, Orlando, FL, USA, in 2001. From 2002 to 2006, he was a Visiting Assistant Professor with University of Central Florida. Since 2007, he has been with the Computer Science Department, Faculty of Computers and Information, Assiut University, where he is currently an Associate Professor, the Executive Director of the Multimedia Laboratory, and a Vice Dean. His major research interests include computer vision, computer graphics, augmented reality, and computer animation.

Mahmoud Afifi received the B.S. and M.S. degrees in information technology from the Faculty of Computers and Information, Assiut University, Egypt, in 2009 and 2015, respectively. He was with Assiut University from 2011 to 2016. He is currently with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Canada. His research interests include image processing, computer vision, and computational photography.

Ghada Moussa received the B.S. degree (Hons.) in civil engineering from Assiut University, Assiut, Egypt, in 2000, and the M.S. degree in structural engineering and the Ph.D. degree in transportation engineering from University of Central Florida, Orlando, FL, USA, in 2003 and 2006, respectively. She is currently an Associate Professor in traffic and highway engineering with the Civil Engineering Department, Faculty of Engineering, Assiut University, and the Executive Director of the Highways Laboratory, Assiut University. Her major research interests include traffic data analysis, applying machine learning into traffic engineering, traffic operation, and highway safety.