Detection of Cars in High-Resolution Aerial Images of Complex Urban Environments

Mohamed ElMikaty, *Member, IEEE*, and Tania Stathaki, *Member, IEEE*

***Abstract*— Detection of small targets, more specifically cars, in aerial images of urban scenes, has various applications in several domains, such as surveillance, military, remote sensing, and others. This is a tremendously challenging problem, mainly because of the significant interclass similarity among objects in urban environments, e.g., cars and certain types of nontarget objects, such as buildings' roofs and windows. These nontarget objects often possess very similar visual appearance to that of cars making it hard to separate the car and the noncar classes. Accordingly, most past works experienced low precision rates at high recall rates. In this paper, a novel framework is introduced that achieves a higher precision rate at a given recall than the state of the art. The proposed framework adopts a sliding-window approach and it consists of four stages, namely, window evaluation, extraction and encoding of features, classification, and postprocessing. This paper introduces a new way to derive descriptors that encode the local distributions of gradients, colors, and texture. Image descriptors characterize the aforementioned cues using adaptive cell distributions, wherein the distribution of cells within a detection window is a function of its dominant orientation, and hence, neither the rotation of the patch under examination nor the computation of descriptors at different orientations is required. The performance of the proposed framework has been evaluated on the challenging Vaihingen and Overhead Imagery Research data sets. Results demonstrate the superiority of the proposed framework to the state of the art.**

***Index Terms*— Airborne imagery, automatic target recognition, car detection.**

NOTATIONS

X, x	Scalar.
$\mathbf{x}$	Column vector.
$\mathbf{X}$	Matrix.
$y(x)$	Function.
$(.)^T$	Transpose operator.
$(.)^H$	Hermitian transpose operator.
$*$	Convolution operator.
$\#\mathbf{x}$	Cardinality of $\mathbf{x}$.
IE{.}	Expectation operator.
$\mathrm{IR}_{\mathbf{xx}}$	Covariance matrix across $\mathbf{x}$.
d	Eigenvalue.
D	Dimensionality.
λ	Wavelength.
θ	Orientation of the normal to the strips of Gabor function.
ψ	Phase offset.
σ_g	Standard deviation of a Gaussian envelope.
γ	Spatial aspect ratio.

Manuscript received November 3, 2016; revised March 9, 2017 and May 2, 2017; accepted June 3, 2017. This work was supported by the Engineering and Physical Sciences Research Council Industrial Cooperative Awards in Science and Technology in collaboration with BAE Systems Military Air and Information, U.K., under Award 12220127. *(Corresponding author: Mohamed ElMikaty.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: m.elmikaty@imperial.ac.uk; t.stathaki@imperial.ac.uk).

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/TGRS.2017.2716984

I. INTRODUCTION

ROBUST detection and localization of small targets in airborne imagery have attracted many researchers in the fields of image analysis and computer vision over the past years. They have numerous useful applications in security, surveillance, defense, remote sensing, and other domains. Small targets, e.g., cars, shipping containers, aeroplanes, and military installations, are usually depicted within small areas that are represented by a small number of pixels in high-resolution aerial images of low ground sampling distance (GSD). Aerial images are commonly captured by sensors mounted on unmanned aerial vehicles, aeroplanes, and earth-observation satellites. In this paper, cars are considered the small targets of interest and the task is to automatically detect and locate cars relative to the rest of a given image of a typical urban scene. This is a tremendously challenging task. First, because of the huge intraclass variability that stems from the significant differences in the visual appearance among cars, as far as various features are concerned, such as colors, size, shape, contours, and boundaries. Furthermore, it is obvious that cars' appearance explicitly depends on the viewing angle. In order to overcome this problem of variability among cars, many researchers have focused on the task of the detection of cars in airborne imagery, wherein a collection of images are taken from a vertical or a slightly oblique viewpoint. This constrained viewpoint facilitates the detection process by eliminating that strong dependence on the viewing angle. In addition, most of the cars would possess almost similar sizes. Second, there is a significant interclass similarity between cars and other objects that arises as a result of the existence of various types of noncar objects in urban scenes, which have similar visual appearance. For instance, similar edge maps or color distributions to that of cars, such as those of buildings' windows, patios, and balconies, can be found.

In this paper, a novel framework is proposed that has the ability to carry out the robust detection and localization of cars in high-resolution airborne imagery of complex urban scenes. Original contributions of this paper include: 1) the design of a novel method that evaluates the likelihood of the detection windows to contain a target in order to reduce

0196-2892 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

IEEE Proof

search areas; 2) a new method to estimate accurately the dominant orientation of detection windows; 3) the introduction of adaptive distributions of cells over which image descriptors can be computed in a new fashion without rotating the original patch as it is a computationally expensive process and it alters the relative distribution of pixels; and 4) a novel encoding of the local distribution of complementary image cues for cars as well as the associated principal components into the feature vector. Points of strength of the proposed framework are: 1) robust performance in complex scenes of urban environments represented by the achieved higher precision and lower false-positive rate and 2) efficient encoding of image cues exploiting lower dimensionality than the state of the art.

This paper is organized as follows. In Section II, a brief discussion of previous works is presented. The proposed framework is introduced in Section III, followed by the description of the Vaihingen data set that was used for training and testing and the Overhead Imagery Research data set (OIRDS) that was only used for testing in Section IV. Results are demonstrated and discussed in Section III-B1. The conclusions are drawn in Section VI.

II. Related Work

Several methodologies have been proposed to solve the problem of the detection of cars in airborne imagery from a perpendicular or a slightly oblique viewpoint [1] and several types of images have been investigated, such as color images, synthetic aperture radar, and others [2], [3]. Nevertheless, our focus in this paper is on the detection of cars in color images and notable related past works will be discussed in this section.

Previous frameworks mainly adopted flat explicit or implicit models, i.e., models that do not use deep architectures. They can be broadly categorized into three main groups: the first group uses explicit models, the second group uses implicit models, and the third group exploits road maps alongside either an implicit or an explicit model.

A. Detection of Cars Using Explicit Models

Early works tried to solve the problem of the detection of cars using explicit models, wherein 2-D and 3-D templates that resemble image cues of cars are designed. Cars are claimed to be found, when the cost of matching between image descriptors of the patch under examination and those of the predesigned template is lower than a predefined threshold. For example, proposed works in [4]–[6] exploited a 3-D template that resembled the geometric structure of cars' contours and the distribution of the colors of cars' bodies, windscreens, and shadows. In a similar fashion, Choi and Yang [7] adopted a "bottom–up" approach, wherein blobs of pixels that possess similar geometric and radiometric properties to that of cars were formed using a mean-shift clustering algorithm [8].

Problems of such template-based models usually arise as a result of the high computational cost that is incurred by the matching process and their performance usually degrades, when they are applied in urban scenes with heavily cluttered backgrounds. This results in low precision rates because of the significant interclass similarity.

B. Detection of Cars Using Implicit Models

Unlike explicit models, implicit models do not exploit the predesigned templates. They use different image descriptors that are engineered to capture and encode different cues of cars. A classifier is trained on the extracted features and is then used to decide whether the patch under examination belongs to the car subspace or the noncar subspace. Several descriptors have been used. Nevertheless, there are no solid theoretical foundations about the choice of descriptors in the open literature, which can accommodate any possible image processing scenario. However, some guidelines can certainly be found. An example of the implicit models is the work that was conducted by Nguyen *et al.* [9], who used Haar wavelets [10], histogram of oriented gradients (HOG) [11], local binary patterns (LBPs) [12], and a mean-shift clustering algorithm [8] for postprocessing. Kembhavi *et al.* [13] followed a similar approach, i.e., they exploited color probability maps, pairs of pixels, and HOG to depict color and geometric structural properties. Likewise, Shao *et al.* [14] used HOG, LBP, and opponent histograms in their vehicle detector and used an intersection kernel support vector machines (SVMs) classifier. Following the basic idea of implicit models but in a different way, Razakarivony and Jurie [15] built two separate models: one for the HOG descriptors of vehicles and the other for those of the background using manifold learning. ElMikaty and Stathaki [16] addressed the problem using HOG, Fourier, and truncated pyramid color self-similarity jointly with a linear SVM classifier. Madhogaria *et al.* [17] used only HOG to detect cars in a two-stage classification process. In their first stage, HOG descriptors were computed over rotated images alongside a linear SVM classifier and a causal Markov random field (MRF) classifier [18]. Using superpixel segmentation, Chen *et al.* [19] extracted meaningful patches. Then, patches were rotated so that HOG descriptors could be computed. A sparse-coding-based dictionary was learnt and used to detect cars.

Although acceptable results of implicit models have been reported, their performance is significantly affected by the choice of features, the implementation, and the dimensionality of the descriptors used and classifiers. Moreover, it can be observed that in order to overcome the problems of intraclass variability and the interclass similarity, either a high-dimensional feature descriptor was used in conjunction with linear classification or a low-dimensional feature descriptor was used alongside nonlinear or multistage classification.

C. Detection of Cars Using Road Maps

In order to restrict search areas to regions that are more likely to contain cars, several previous works have exploited road maps to extract roads and parking areas as an initial step in order to decrease the false-positive rate. A representative example is the work conducted by Moon *et al.* [20], [21], wherein the authors made use of the locations of roads and buildings and utilized the derivative of the double-exponential filter [22] to extract gradients. Similarly, Zheng *et al.* [23] extracted roads and parking areas and used morphological structuring elements [24] and a classifier to distinguish actual car pixels. In a similar manner, Eikvil *et al.* [25] proposed a detection approach that consisted of a segmentation stage to

IEEE Proof

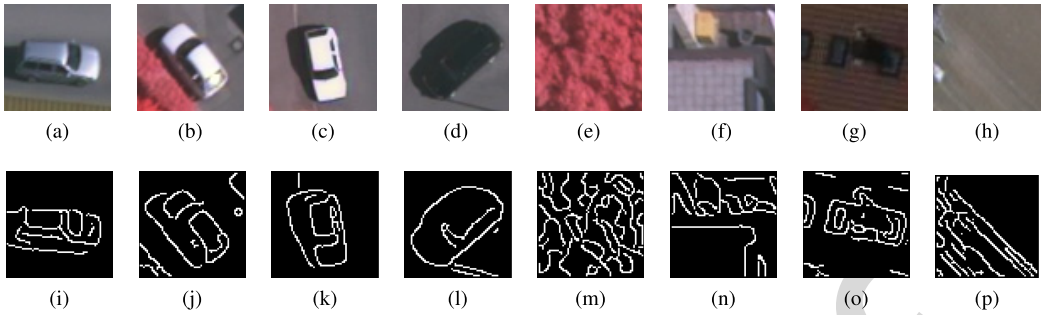


Fig. 1. Extraction of edges in target-class and non-target-class patches using Canny edge detectors [28] with an associated Gaussian filter of standard deviation of $\sqrt{2}$. Target-class patches are shown in (a)–(d), and their corresponding edges in (i)–(l), respectively. Non-target-class patches are shown in (e)–(h), and their corresponding edges in (m)–(p), respectively. Shown images were scaled for viewing.

separate regions that contain cars, followed by a two-stage object-classification framework exploiting road networks. Following the same idea, Tuermer *et al.* [26] proposed the use of road databases, disparity maps, and an HOG object detector.

Whilst exploiting road maps provides better performance in terms of true-positive and false-positive rates, it is based on the prior knowledge of accurate road maps, a precise map-projection method, which is computationally expensive, and the use of extra hardware, such as a global positioning system.

Unlike flat models, deep models for the detection of cars can hardly be found in the literature as they require vast training data sets, which are challenging to acquire. To the best of the our knowledge, only one work was proposed by Chen *et al.* [27] to solve the problem of car detection using deep models. In this paper, a hierarchal deep model was used in a hybrid fashion. Although high accuracy has been achieved, their testing data set compromised a set of images of vast parking areas containing a large number of cars and lacked the existence of complex shapes that characterize urban scenes.

III. Proposed Framework

Regarding a car in a given aerial image, its boundaries are expected to have almost rectilinear structures that correspond to the edges of the body and front and rear windscreens as shown in Fig. 1. Although these rectilinear structures form together a considerable number of crucial points of interest, e.g., high-curvature points and corners, they are incapable of discriminating cars from other objects. This is because in some cases, other noncar objects possess similar edge maps to those of cars, such as buildings' roofs and windows, and they can be easily misidentified as cars. Based on this observation, implicit models are seen to be superior to explicit models as they capture fine details and a simple linear classifier can learn from the training samples how to distinguish well-encoded combined image descriptors that correspond to the cues related to cars.

In this paper, three image descriptors are computed over a sliding detection window to characterize the distribution of gradients, colors, and texture. These image cues were chosen among other cues as cars possess: 1) almost rectilinear structures as aforementioned; 2) distinctive color distribution; and 3) mostly smooth surfaces with low amount of texture. The proposed framework adopts a gray-scale fixed-size window of 64×64 pixels with an associated stride length of four pixels as in a set of aerial images captured at a given GSD, cars have almost similar sizes. Square patches were chosen instead of rectangular patches as it was found after exhaustive experimental validations that square patches result in better performance. They incorporate some of the contextual information and they facilitate the computation of image descriptors at different orientations. The proposed framework comprises four stages, namely, window evaluation, extraction of features and encoding, classification, and postprocessing as shown in Fig. 2. Details of the implementation of each stage are discussed in Sections III-A–III-D.

Choice of Features and Parameters: All values of the parameters used in the implementation of the proposed framework were set empirically by testing excessively the properties of a large number of image samples from the training data set that will be described later in Section IV-A using cross validation.

A. Window Evaluation

In order to reduce search areas, it is necessary to investigate detection windows that are highly unlikely to contain cars at an early stage. This reduces the false-positive rate. In addition, it helps shorten the processing time by discarding a large number of windows, as it can be seen in Fig. 3. To achieve this, a novel three-step filtering process is proposed to measure the objectness [29] of the detection window and its likelihood to contain a target at an early stage based on several cues, including the local distribution of colors, texture, and gradients.

The first step is to investigate the variation in the distribution of gray levels across a detection window. This is examined by the computation of two 32-bin gray-level histograms $\mathbf{h}_w$ and $\mathbf{h}_c$, one for the whole detection window and the other for the 32×32 central square area. If the similarity between the two histograms is more than 50%, the window is discarded as this highly indicates that it belongs to a flat area with a constant or a slowly varying gray-level distribution, such as empty roads and vegetation areas. The Manhattan distance $r_{\text{Manhattan}}$ between these two color histograms is computed to measure their similarity with a threshold value of 2048 as the size of the detection window is 64×64 pixels

$$r_{\text{Manhattan}} = \sum_{k=1}^{32} |\mathbf{h}_w(k) - \mathbf{h}_c(k)|. \quad (1)$$

IEEE Proof

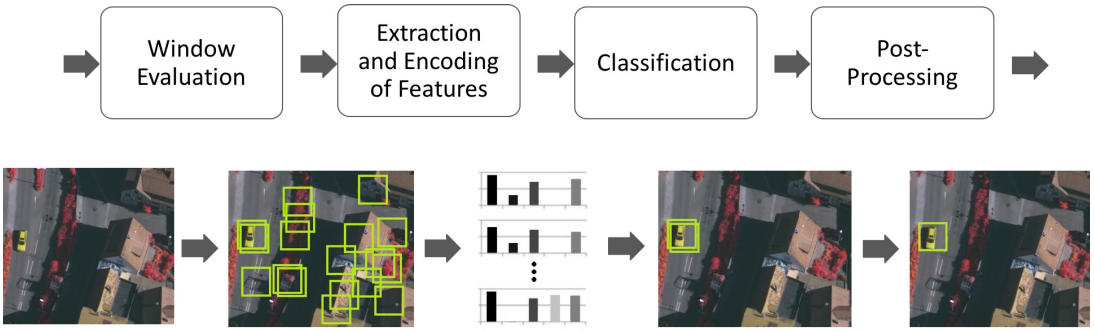


Fig. 2. Flow diagram of the adopted four-stage methodology. (Top) Block diagram shows the sequence of the four main stages of the adopted general methodology. (Bottom) Original image, processed image using the window-evaluation stage, computed image descriptors for each successful candidate window, output of the classification stage, and final detection after postprocessing (from Left to Right). Shown images were scaled for viewing.



(a) (b)

Fig. 3. Application of the window-evaluation stage on a sample region of the Vaihingen data set. (a) Original image. (b) Corresponding processed image. Areas that have been eliminated by the window-evaluation stage as they were highly unlikely to contain cars are marked with black. Shown images were scaled for viewing.

The second step is to measure the texture of the detection window. Pixels of each local cell of 4×4 pixels are divided into two groups according to whether the pixel value is larger than the average value of the cell or smaller. Then, the difference between the averages of the two groups is calculated. If the standard deviation σ_{texture} of the computed differences is larger than a predefined threshold value of 32, the detection window will not be considered for further processing. The intuition behind this is that a high standard deviation σ_{texture} of the mean values of the differences among average intensities (more than one eighth of the intensity range [0,255]) indicates a highly textured surface that cannot be a car as cars tend to have smooth surfaces.

The third step is to examine the local distribution of gradients. For a given gray image $\mathbf{I}$, magnitudes of intensity gradients $\mathbf{M}_g(x, y)$ at each pixel and their orientations $\mathbf{O}_g(x, y)$ are calculated over the detection window. This is done by the computation of the first-order discrete derivatives $(\partial\mathbf{I}/\partial x)$ and $(\partial\mathbf{I}/\partial y)$ of $\mathbf{I}$ using a nonsmoothing filter of $[-101]$ in both the vertical and horizontal directions

$$\mathbf{M}_g(x, y) = \sqrt{\left(\frac{\partial\mathbf{I}(x, y)}{\partial x}\right)^2 + \left(\frac{\partial\mathbf{I}(x, y)}{\partial y}\right)^2} \tag{2}$$

$$\mathbf{O}_g(x, y) = \arctan\left(\frac{\partial\mathbf{I}(x, y)}{\partial y} \Big/ \frac{\partial\mathbf{I}(x, y)}{\partial x}\right). \tag{3}$$

Then, the magnitudes of gradients of local cells of 4×4 pixels are accumulated independently in 9-bin histograms $\mathbf{g}_{\text{cell}}$ according to their unsigned orientation

$$\mathbf{g}_{\text{cell}}(k) = \sum_{x,y} \mathbf{M}_g(x, y) \mid \mathbf{O}_g(x, y) \in \mathbf{o}_k, \quad k \in \{1, 2, \ldots, 9\} \tag{4}$$

where $\mathbf{o}_k$ is the kth bin of orientation interval of width 20°, followed by computing the covariance matrix $\text{IR}_{\mathbf{gg}} = \text{IE}\{\mathbf{g}\mathbf{g}^H\}$ of the $L2$-normalized responses $\mathbf{g}_{\text{cell}} \leftarrow \mathbf{g}_{\text{cell}}/(\|\mathbf{g}_{\text{cell}}\|_2^2 + \epsilon^2)^{1/2}$ from local cells, where ϵ is a small constant and is equal to 0.01. For a typical detection window containing a car, there should be a dominant orientation of the gradients that corresponds to the direction, which is perpendicular to the orientation of the car in that particular window. Therefore, if the covariance matrix $\text{IR}_{\mathbf{gg}}$ is eigendecomposed, the maximum eigenvalue $d_{\max}$ must exceed a predefined threshold value of 0.09, otherwise the window is discarded. The intuition behind this is that the second-order statistics given by the covariance matrix $\text{IR}_{\mathbf{gg}}$ can be used to determine the number of the dominant orientations of gradients and their directions using the rank of $\text{IR}_{\mathbf{gg}}$ and the order of its significant eigenvalues d_k, respectively

$$[d_{1,\text{cells}}, \ldots, d_{9,\text{cells}}]^T = \text{eig}(\text{IR}_{\mathbf{gg},\text{cells}}). \tag{5}$$

Estimation of the Orientations of Cars: The orientation of cars O_{car} is discretized into nine orientations at 10°, 30°, 50°, 70°, 90°, 110°, 130°, 150°, and 170°. It is estimated to be in a perpendicular orientation to the bin orientation corresponding to the maximum eigenvalue $d_{\max}$ of the covariance matrix $\text{IR}_{\mathbf{gg}}$ of locally accumulated gradients computed over 4×4 cells.

IEEE Proof

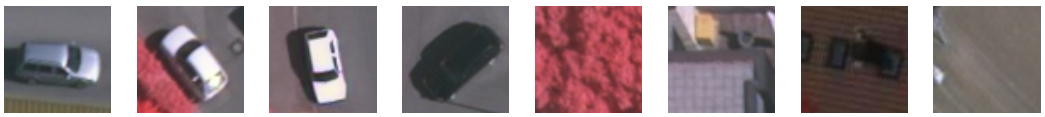


Fig. 4. Estimation of the dominant orientation in target-class and non-target-class patches. Target-class patches are shown in (a)–(d). Non-target-class patches are shown in (e)–(h). Estimated dominant orientations were (a) $O_{\text{dominant}} = 170°$, (b) $O_{\text{dominant}} = 130°$, (c) $O_{\text{dominant}} = 110°$, (d) $O_{\text{dominant}} = 50°$, (e) $O_{\text{dominant}} = 110°$, (f) $O_{\text{dominant}} = 170°$, (g) $O_{\text{dominant}} = 170°$, and (h) $O_{\text{dominant}} = 130°$. Shown images were scaled for viewing.

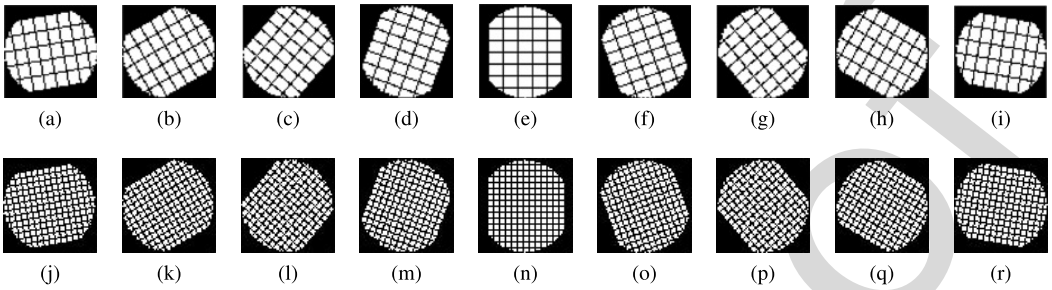


Fig. 5. Cells of original size of 8×10 pixels are shown in (a)–(i): (a) 10°. (b) 30°. (c) 50°. (d) 70°. (e) 90°. (f) 110°. (g) 130°. (h) 150°. (i) 170°. Cells of original size of 4×4 pixels are shown in (j)–(r): (j) 10°. (k) 30°. (l) 50°. (m) 70°. (n) 90°. (o) 110°. (p) 130°. (q) 150°. (r) 170°. Shown images were scaled for viewing.

Examples of estimating the dominant orientation of detection windows are shown in Fig. 4

$$O_{d_{\max}} = (10 + 20(k-1))° | d_k = \max_{k=1,\ldots,9} (\text{eig}(\text{IR}_{\mathbf{gg}})). \quad (6)$$

B. Extraction and Encoding of Features

Detection windows are traditionally divided into square cells of equal size over which image descriptors are computed locally. For rotation-variant features, detection windows are rotated in most cases so that they would have a specific orientation. However, this rotation process is time-consuming and it affects the relative distribution of the original pixels due to the nonlinear transformation between coordinates. To overcome this problem, adaptive cell distributions are proposed, wherein the shape of cells is no longer fixed-size squares. Two sets of adaptive cell distributions are used in the proposed framework and they are shown in Fig. 5. In the first set, distributions of cells are originated from the rotation of a mask oriented horizontally of size 64×50 pixels and divided into ordered cells of size 8×10 pixels to the previously mentioned discretized orientations, whereas in the second set, distributions of cells are originated from the rotation of a mask oriented horizontally of size 64×52 pixels and divided into ordered cells of size 4×4 pixels. Corners of each mask are rounded by eliminating pixels that are located at a distance more than 31 pixels from the center of the window.

Image descriptors are computed over each candidate detection window after being divided into one of cell distributions according to the estimated dominant orientation of the detection window as follows.

1) Gradient Descriptors: Cars have rectilinear structures and a dominant orientation as previously mentioned. To encode this feature, gradient descriptors are exploited. They are among the well-established image descriptors and they are known as being powerful single-cue descriptors [30]. In this paper, gradient descriptors are computed over the first set of the cell distributions, shown in Fig. 5, based on the estimated dominant orientation of the detection window (described in Section III-A). As already mentioned, magnitudes of gradients $\mathbf{M}_g$ of local cells are accumulated in 9-bin histograms $\mathbf{g}_{\text{cell}}$ according to their orientation $\mathbf{O}_g$ independently without the trilinear interpolation process recommended by Dalal and Triggs [11] as it is a computationally expensive process and it has a minimal effect on the overall performance. The response for each cell $\mathbf{g}_{\text{cell}}$ is normalized using $L2-$normalization so that illumination-invariance properties can be achieved. The final gradient descriptor is formed from the concatenation of the $L2$-normalized eigenvalues $d_{k,\text{block}}$ of the covariance matrices $\text{IR}_{\mathbf{gg},\text{block}}$ of the cell responses of groups of four adjacent cells. It has a dimensionality of 216 as there are 24 blocks of cells, each has nine eigenvalues. This differs from the original implementation of [11] in the way local cells are distributed and that the final descriptor consists of the eigenvalues of the local covariance matrices rather than the absolute cell responses to the histogram of gradients.

2) Color Descriptors: Cars exhibit a distinctive distribution of colors: the color of windscreens is always dark in aerial images, and the color of the car boot is usually the same as the color of the car bonnet. Therefore, if a local descriptor that encodes the local similarity and variation of colors across pixels is used, it will facilitate the classification process. A new modified version of the Color Self-Similarity descriptor proposed by ElMikaty and Stathaki [16] is suggested. It incorporates the principal components of the covariance matrices of the local cell responses. The new color descriptor $\mathbf{c}_{\text{cell}}$ calculates the similarity among adjacent cells by computing the intersection between corresponding local-color-histogram bins

$$\mathbf{c}_{\text{cell}} = [\mathbf{h}_c \cap \mathbf{h}_n] = \left[\left[\sum_k \min(\mathbf{h}_c(k), \mathbf{h}_n(k))\right]\right] \quad (7)$$

IEEE Proof

where $\mathbf{h}_c$ and $\mathbf{h}_n$ are the two gray-level histograms of the central cell and a neighboring cell, respectively, each has 16 bins, $\mathbf{h}_c(k)$ and $\mathbf{h}_n(k)$ are the scores of the kth bins of the two histograms, and $\lceil . \rceil$ indicates that the result is truncated at the size of an unsigned integer of eight bits as recommended by ElMikaty and Stathaki [16]. Local descriptors are computed on two levels in a pyramid approach using the two sets of cell distributions shown in Fig. 5. In the first level, descriptors are computed for 24 cells on a neighborhood of four cells, whereas in the second level, descriptors are calculated for 130 cells with a neighborhood of four cells, two cells with a neighborhood of three cells, and four cells with a neighborhood of two cells as small cells around the edges of the masks were not included. The final color descriptor has 3190 features and is formed by concatenating the $L2$-normalized local responses from all cells into a single vector and the $L2$-normalized eigenvalues of the covariance matrices $\mathrm{IR}_{\mathbf{cc},\mathrm{block}}$ of the local responses for each group of cells in each level of the pyramid.

3) Texture Descriptors: Texture is a common image cue that has been used, among other cues, to solve the problem of object detection [31]. Many researchers have used Gabor filters to extract and encode texture motivated by the similarity between 2-D Gabor filters and the receptive fields in the visual cortex of the vision systems of humans [32]. Although cars do not possess distinguishable textured surfaces, Gabor texture descriptors play a great role in describing other textured areas, such as those of buildings' roofs and vegetation. Gabor filter kernels are formed by a 2-D Gaussian function modulated by a complex sinusoidal signal

$$\mathbf{F}_{\mathrm{Gabor}}(x, y; \lambda, \theta, \psi, \sigma_g, \gamma) = \exp\left(\frac{x'^2 + \gamma^2 y'^2}{2\sigma_g^2}\right) \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right) \quad \begin{cases} x' = x\cos\theta + y\sin\theta \\ y' = -x\sin\theta + y\cos\theta. \end{cases} \tag{8}$$

In this paper, nine Gabor kernels of size 64×64 are used, with a wavelength λ of 10 pixels, nine orientations $\theta_k \in \{10°, 30°, \ldots, 170°\}$, phase offset ψ of $-90°$, aspect ratio γ of 10, and the variance of the Gaussian envelop σ_g^2 of 13.671.

After filtering the detection windows using those nine Gabor kernels, windows are divided into one of the first set of the cell distributions, shown in Fig. 5, according to the estimated orientation of the detection window. The mean $\mathbf{t}_{\mathrm{cell,mean}}$ and energy $\mathbf{t}_{\mathrm{cell,energy}}$ of the response to the Gabor filters in each cell is computed at each orientation, resulting in 18 features per cell

$$\mathbf{t}_{\mathrm{cell,mean}}(k) = \frac{1}{\mathrm{size(cell)}} \sum_{x,y} (|\mathbf{F}_{\mathrm{Gabor}} * \mathbf{I}|) \mid \theta_k = (20k - 10)° \tag{9}$$

$$\mathbf{t}_{\mathrm{cell,energy}}(k) = \frac{1}{\mathrm{size(cell)}} \sum_{x,y} ((\mathbf{F}_{\mathrm{Gabor}} * \mathbf{I})(\mathbf{F}_{\mathrm{Gabor}} * \mathbf{I})^H) \mid \theta_k = (20k - 10)° \tag{10}$$

$$\mathbf{t}_{\mathrm{cell}} = [\mathbf{t}_{\mathrm{cell,mean}}, \mathbf{t}_{\mathrm{cell,energy}}]^T. \tag{11}$$

For rotation-invariance, the order of the nine orientations is changed such that rotation invariance is achieved by circularly shifting the response to different orientations θ_k. The final texture descriptor has 432 features and is formed from the concatenation of the $L2$-normalized eigenvalues of the covariance matrices $\mathrm{IR}_{\mathbf{tt},\mathrm{block}}$ of the cell responses of 24 groups of four adjacent cells.

C. Classification

An extended feature vector is formed from the aforementioned ensemble of descriptors by concatenation. The overall dimension of this vector is 3838-D (216 gradient features, 3190 color features, and 432 texture features). Since this problem of classification is characterized by having a relatively high dimensionality, it is recommended to adopt an SVM classifier that is known to work efficiently in high-dimensional scenarios [18]. Based on that and following Occam's razor, "Among competing hypotheses, the one with the fewest assumptions should be selected," a linear SVM classifier has been adopted in the proposed framework to discriminate the car subspace from the noncar subspace.

An SVM classifier assumes that a total of K positive and negative training data samples $s_k \in \{s_1, \ldots, s_K\}$ are given and that they are drawn from a hyperdimensional space with associated target values $t_k \in \{-1, 1\}$. It aims to solve the classification problem using a linear model in the form of

$$z(s) = \mathbf{w}^T \phi(s) + b \tag{12}$$

where $\phi(s)$ represents the transformation from the spatial domain to the hyper-dimensional space and b is a bias. In other words, the goal is to determine the hyperplane in the feature space $\phi(S) \in \Re^D$ that best separates the car and noncar classes. This can be achieved by maximizing the margin between the hyperplane and data samples that are close to it, which are known as support vectors. During the training stage, parameters $\mathbf{w}$ and b are optimized in order to maximize the margin and the solution is given by

$$\underset{\mathbf{w},b}{\mathrm{argmax}} \left\{ \frac{1}{\|\mathbf{w}\|} \min_k (t_k(\mathbf{w}^T \phi(s_k) + b)) \right\}. \tag{13}$$

Nevertheless, the above-mentioned assumptions are valid only for perfectly linearly separable classes. For overlapping classes, a slack variable $\xi_k \geq 0$, where $k = 1, \ldots, K$, is introduced to allow some training samples to be misclassified, converting the quadratic optimization problem in (13) to a constrained minimization problem of $\|\mathbf{w}\|^2$ as follows:

$$C \sum_{k=1}^{K} \xi_k + \frac{1}{2}\|\mathbf{w}\|^2 \tag{14}$$

subject to the following constraints:

$$\begin{cases} t_k(\mathbf{w}^T \phi(s_k) + b) \geq 1 - \xi_k \\ \xi_k \geq 0 \end{cases} \tag{15}$$

where C is a nonnegative regularization constant and it controls the tradeoff between the slack variable ξ_k and the margin. In this paper, $\phi(s_k)$ transforms the training samples to a hyperspace of 3838-D using a regularization parameter C of 15. More details about the theory and application of SVM classifiers can be found in [18].

IEEE Proof

D. Postprocessing

The aim of the postprocessing stage is to eliminate multiple detections per single true detection that are resulted from the use of an overlapping-sliding-window approach. Detection windows with positive score that overlap by more than 50% are eliminated except the one with the highest output confidence score.

IV. Data Sets

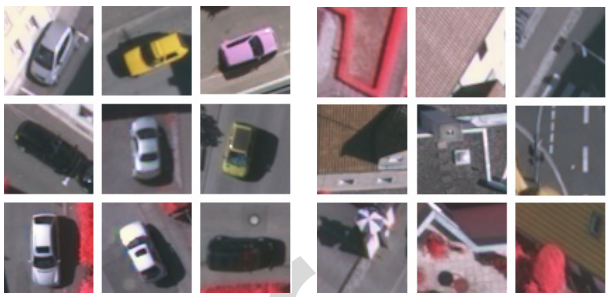
The Vaihingen data set [33] was used to train the SVM classifier and to evaluate the performance of the proposed framework. In order to evaluate the ability of the proposed framework to generalize on unseen data sets and compare its performance with the state of the art, the publicly available OIRDS data set [34] has been used and it was not included in the training data set. Description of both data sets is given below.

A. Training Data Set

The training process was carried out on a part of the Vaihingen data set that comprises a set of high-resolution aerial images of a GSD that is equal to ≈8 cm and a radiometric resolution of 11 b. Images of the Vaihingen data set are pan-sharpened colored infrared images and they were provided in TIFF RGB files. The data set was captured over the city of Vaihingen—Germany by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [33] using a Digital Mapping Camera with focal length of 120 mm and at a flying height above the ground of 900 m. This means that the requirement that the viewpoint must be at a perpendicular or a slightly oblique angle is met. Cars in the Vaihingen data set are almost of the same size and are contained inside square windows of 48×48 pixel size as the GSD is 8 cm.

Positive training data consisted of square patches of 64×64 pixels so that image descriptors can be computed efficiently over cars and a small part of the surrounding background. Patches that contain cars were chosen from areas other than the test areas. They were chosen from various regions such that a great diversity in the orientation of cars and backgrounds could be achieved. Data augmentation techniques were exploited by the inclusion of mirrored and horizontally flipped versions of these patches to the training data in order to richen the diversity in the training data set.

As for the negative training data, random patches were included that were selected from different backgrounds. In addition, bootstrapped hard negatives that were detected as false positives during the validation process were added. In total, 781 positive samples and 2473 negative samples were used in the training process. Sample positive and negative patches are shown in Fig. 6.



(a) (b)

Fig. 6. Samples of the patches that were extracted from the Vaihingen data set and used for training. (a) Positive training samples of car-class targets of different orientations and colors. (b) Negative training samples of non-car-class targets. Shown images were scaled for viewing.

B. Testing Data Set

1) Vaihingen Data Set: DGPF specified three testing areas for the evaluation of the developed algorithms that contained 128 nonoccluded cars. They contain typical urban backgrounds with different building styles, roads, and vegetation. Description of the three areas is as follows. Area 1 contains dense development consisting of historic buildings having complex shapes and it also has some trees, area 2 has a few high-rising residential buildings that are surrounded by trees, and area 3 is a pure residential area with small detached houses. Samples for the three different testing areas are shown in Fig. 7.

2) Overhead Imagery Research Data Set: The OIRDS data set [34] has more than 900 annotated images that contain approximately 1075 nonoccluded cars. The data set is a part of the freely available aerial imagery captured by the United States Geological Survey. Images were captured at a GSD of 15 cm under a variety of operating characteristics. Images were upsampled by a factor of 1.875 so that the final GSD is ≈8 cm. This means that the data set has a lower resolution compared with the training data set described in Section IV-A. Examples of the test images are given in Fig. 8. In this paper, no examples of the OIRDS data set were included in the training data set in order to test the ability of the proposed framework to generalize.

V. Experimental Results

The performance of the proposed framework was evaluated on the testing areas of the Vaihingen data set as well as on the OIRDS data set. No manual annotation of any testing data was carried out except to evaluate the algorithm (labeling ground truth in the testing data set only). The performance was assessed and evaluated using qualitative analysis and a number of quantitative measures, namely, the precision and recall rates, the average precision, and the interpolated average precision.

A. Qualitative Analysis

Visual results of the experiments of applying the proposed framework on sample regions of the test areas of the Vaihingen data set are shown in Fig. 7, whereas results of sample images of the OIRDS data sets are shown in Fig. 8. These two figures show that our framework is able to robustly detect cars in various environments.

In the center left image of Fig. 7, it can be seen that the proposed method is able to correctly classify detection windows that have poor contrast (fourth car from top), thanks to the local normalization of the gradient-descriptor responses of cells, discussed in Section III-B.1. This is a case that is even hard for the human eye. Nevertheless, in few situations, wherein the detection windows contain very dark cars that are covered by very strong building shadows and the contrast is too low, the proposed method fails to correctly classify

IEEE Proof

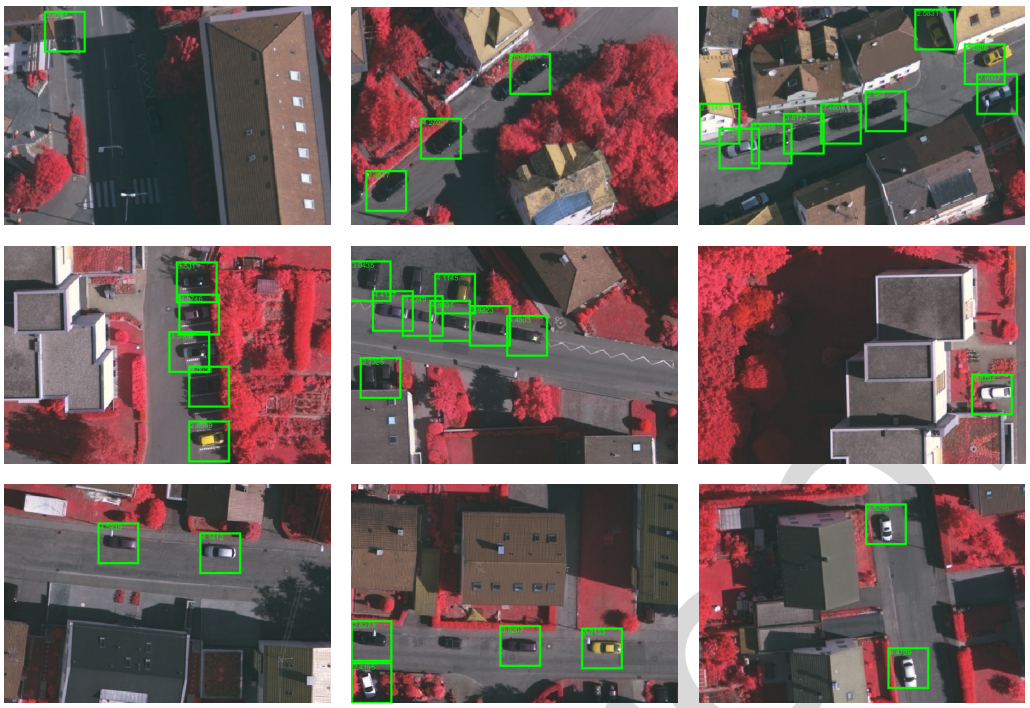


Fig. 7. Sample regions of the test area of the Vaihingen data set after applying the proposed framework. First row: sample regions from test area 1. Second row: sample regions from test area 2. Third row: sample regions from test area 3. Bounding boxes represent detections. Shown images were scaled for viewing.

them as true positives. For example, in the top-left image of Fig. 7, the black car parked next to the building was missed, because the strength of gradients in such detection windows is too low even after normalization and the color distribution is concentrated around the histogram bins that correspond to dark gray levels.

In the top-center and bottom-center images of Fig. 7, two cars were not detected as they have a very small size compared with other standard cars. However, this problem can be solved by extending the proposed method to multiscale analysis which is among our future goals.

From Fig. 8, it can be clearly deduced that the proposed framework has very good generalization as it achieves high true-positive rate while maintaining a low false-positive rate when it is applied to images with different acquisition conditions and a lower resolution than those of the data on which the SVM classifier was trained. More specifically, in challenging cases, where cars are partially overimposed by shadows, e.g., the first image of the first row in Fig. 8, our framework is able to detect cars despite the fact that the shadow alters the distribution of edges and colors. Moreover, in the case of complex backgrounds, for example, the third image of the first row in Fig. 8, the proposed algorithm does not produce a high rate of false positives.

B. Quantitative Analysis

Quantitative evaluation methods for target detection systems depend on the adopted approach [35]. As a window-based approach was adopted alongside a classifier that outputs a confidence score (distance from the separating hyperplane), metrics that depend on ranked outputs according to the confidence score in a descending order should be used. In this paper, to evaluate the performance and compare it with the state of the art that constitutes the baseline performance, the interpolated average precision $\mathrm{AP}_{\mathrm{inter}}$ was used.

In the adopted evaluation methodology, a detected bounding box B_D was considered true if it matched a corresponding ground-truth bounding box B_{GT} using (16) and there were no other bounding boxes matching the same ground truth [36]. In the case of having more than one bounding box matching the same ground truth, only one of them was considered a true detection and others were considered false positives. Ambiguous objects, i.e., objects, which are partially occluded, were not taken into account, when measuring the performance of the proposed framework following the works in [13]–[17] that also discarded occluded cars:

$$\frac{\mathrm{Area}(B_D \cap B_{\mathrm{GT}})}{\mathrm{Area}(B_D \cup B_{\mathrm{GT}})} \geq 0.5. \tag{16}$$

After defining the condition for considering true positives, precision PR, and recall RC can be computed at several levels according to the rank of the confidence score associated with a detection window using (17) and (18). The term "recall" is used here to indicate the proportion of the truly detected cars to the total number of ground-truth cars, whereas the term "precision" indicates the proportion of truly detected cars to the total number of detections above a given confidence score [36]

$$\mathrm{PR} = \frac{\#\mathbf{b}_{\mathrm{TP}}}{\#\mathbf{b}_{\mathrm{TP}} + \#\mathbf{b}_{\mathrm{FP}}} \tag{17}$$

$$\mathrm{RC} = \frac{\#\mathbf{b}_{\mathrm{TP}}}{\#\mathbf{b}_{\mathrm{GT}}} \tag{18}$$

where $\mathbf{b}_{\mathrm{TP}}$ represents the true positives, $\mathbf{b}_{\mathrm{FP}}$ represents the false positives, and $\mathbf{b}_{\mathrm{GT}}$ represents the ground truth.

IEEE Proof



Fig. 8. Sample regions of the test images of the OIRDS data set after applying the proposed framework. Bounding boxes represent detections. Shown images were scaled for viewing.

TABLE I

PERFORMANCE OF THE PROPOSED FRAMEWORK WITH DIFFERENT WINDOW EVALUATION SETTINGS ON THE VAIHINGEN DATA SET REPRESENTED BY THE AVERAGE PRECISION AP AND THE INTERPOLATED AVERAGE PRECISION AP_{inter}

Window Evaluation	AP	AP_{inter}
Des. + SVM	78.65%	64.88%
Histogram Difference + Des. + SVM	85.02%	63.78%
Texture + Des. + SVM	82.14%	68.01%
Maximum Eigenvalue + Des. + SVM	80.82%	66.79%
Proposed Wind. Eval. + Des. + SVM	**92.79%**	**68.03%**

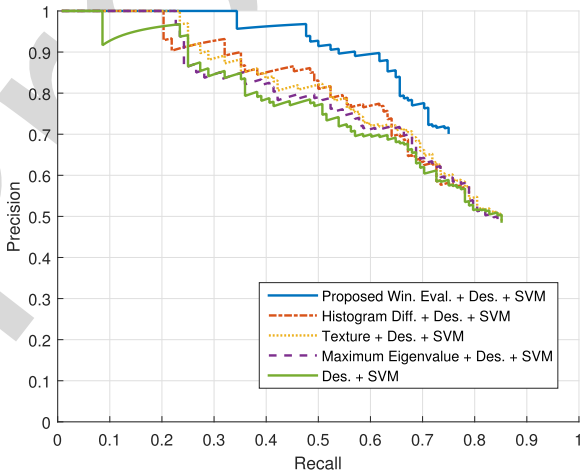


Fig. 9. Performance of the proposed framework on the Vaihingen data set represented by precision rates at different recall rates with different window evaluation settings and the proposed ensemble of descriptors alongside a linear SVM classifier.

The average precision AP and its interpolated version AP_{inter} can be computed using the following:

$$\text{AP} = \frac{1}{\#\mathbf{b}_{\text{TP}}} \sum_{k \in \{1,2,\ldots,\#\mathbf{b}_{\text{TP}}\}} \mathbf{PR}(\mathbf{b}_{\text{TP}}(k)) \tag{19}$$

$$\text{AP}_{\text{inter}} = \frac{1}{11} \sum_{k \in \{0,0.1,\ldots,1\}} \mathbf{PR}_{\text{interp}}(k) \mid \mathbf{PR}_{\text{interp}}(k) = \max_{\tilde{r}:\tilde{r} \geq r} [\mathbf{PR}(\tilde{r})] \tag{20}$$

where $\mathbf{b}_{\text{TP}}$ represents true positives at a given confidence score and **PR** represents the associated precision. It is worth to mention that the AP_{inter} penalizes methods that are not able to retrieve 100% of the ground-truth data [36], unlike the AP that does not take this into account. Accordingly, the AP_{inter} is more commonly used in quantitative analysis.

1) Performance of the Proposed Framework: The proposed framework uses the eigenvalues of the covariance matrix of the local gradient descriptor to estimate the orientation of cars as previously discussed in Section III-A. The achieved accuracy of estimating the orientation of cars in the test data set is 98.34%.

Regarding the contribution of each step in the window-evaluation stage, the performance of the detection system was assessed with adding only one of the proposed window filtering steps. Fig. 9 and Table I demonstrate the obtained results under five different scenarios: 1) the ensemble of descriptors was only used in conjunction with the SVM classifier; 2) window evaluation based on the difference in gray histograms and the ensemble of descriptors were used in conjunction with the SVM classifier; 3) window evaluation based on the level of texture and the ensemble of descriptors were used in conjunction with the SVM classifier; 4) window evaluation based on the value of the maximum eigenvalue and the ensemble of descriptors were used in conjunction with the SVM classifier; and 5) the overall proposed framework including the three steps of window evaluation and the ensemble of descriptors. Furthermore, Table II shows the effect of changing the recommended value of one of the three thresholds used in the window-evaluation stage while keeping the other two in their recommended values.

To evaluate the effect of each proposed image descriptor, the performance of the detection system was assessed using different combinations of descriptors. Fig. 10 and Table III demonstrate the obtained results under seven different

IEEE Proof

TABLE II

PERFORMANCE OF THE PROPOSED FRAMEWORK WITH DIFFERENT VALUES FOR THE THRESHOLDS USED IN THE WINDOW-EVALUATION STAGE ON THE VAIHINGEN DATA SET REPRESENTED BY THE AVERAGE PRECISION AP AND THE INTERPOLATED AVERAGE PRECISION AP_{inter}

$\pm x\%$	Histogram Difference		Texture		Maximum Eigenvalue	
	AP	AP_{inter}	AP	AP_{inter}	AP	AP_{inter}
-8.0%	90.46%	67.26%	91.59%	67.24%	92.40%	67.67%
-4.0%	91.82%	67.66%	92.92%	67.33%	**92.86%**	**68.18%**
0.0%	92.79%	68.03%	**92.79%**	**68.03%**	92.79%	68.03%
+4.0%	**94.39%**	**68.37%**	91.78%	67.19%	92.19%	67.32%
+8.0%	93.86%	68.09%	91.19%	67.02%	92.32%	67.50%

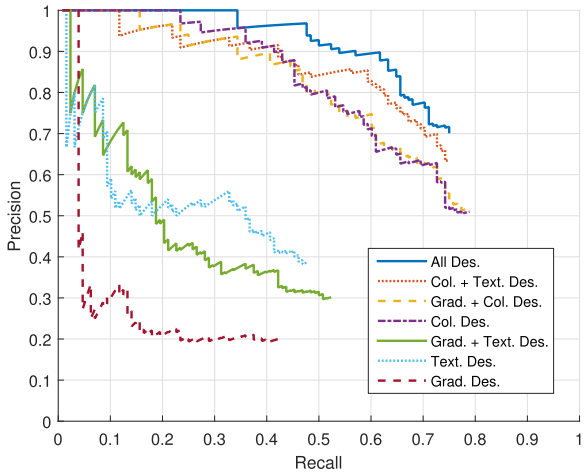


Fig. 10. Performance of the proposed framework on the Vaihingen data set represented by precision rates at different recall rates with the proposed three-step window-evaluation stage and different combinations of descriptors alongside a linear SVM classifier.

TABLE III

PERFORMANCE OF THE PROPOSED FRAMEWORK WITH THE WINDOW-EVALUATION STAGE AND DIFFERENT COMBINATIONS OF DESCRIPTORS ALONGSIDE THE SVM CLASSIFIER ON THE VAIHINGEN DATA SET REPRESENTED BY THE AVERAGE PRECISION AP AND THE INTERPOLATED AVERAGE PRECISION AP_{inter}

Descriptors	AP	AP_{inter}
Wind. Eval. + Grad. Des. + SVM	30.61%	17.41%
Wind. Eval. + Text. Des. + SVM	56.31%	28.39%
Wind. Eval. + Col. Des. + SVM	84.71%	63.65%
Wind. Eval. + Grad. & Text. Des. + SVM	50.45%	29.44%
Wind. Eval. + Grad. & Col. Des. + SVM	84.15%	63.32%
Wind. Eval. + Col. & Text. Des. + SVM	88.54%	65.35%
Proposed Framework	**92.79%**	**68.03%**

scenarios: 1) gradient descriptors were used alongside the window-evaluation stage and the SVM classifier; 2) texture descriptors were used alongside the window-evaluation stage and the SVM classifier; 3) color descriptors were used alongside the window-evaluation stage and the SVM classifier; 4) gradient and color descriptors were used alongside the window evaluation stage and the SVM classifier; 5) gradient and texture descriptors were used alongside the window evaluation stage and the SVM classifier; 6) color and texture descriptors were used alongside the window evaluation stage and the SVM classifier; and 7) the overall proposed car detection system.

The proposed framework is able to achieve an overall precision rate of 70.07% at a recall rate of 75%. It can be inferred that the three-step window filtering process has a significant effect in reducing the rate of false positives. More specifically, evaluating detection windows based on the variation of color distribution across the detection window results in a higher precision. Nevertheless, it has a slight effect on the recall rate as it can be observed from the decrease in the interpolated average precision AP_{inter}.

Changing the recommended values of the thresholds used in the window-evaluation stage by $\pm 4\%$ results in changing the average precision AP and the interpolated average precision AP_{inter} by less than $\pm 1\%$ indicating the robustness of the proposed window-evaluation stage to small variations in the acquisition conditions, e.g., changes in illumination. The recommended thresholds do not result in the highest average precision and interpolated average precision, as it can be inferred from Table II. Nonetheless, these thresholds were set using cross validation on the training data set to prevent overfitting as previously mentioned. It is also worth to note that it is important to have high values for both the average precision AP and the interpolated average precision AP_{inter} as this ensures having high precision and recall rates at the same time.

Combining the three window-evaluation steps with the proposed descriptors results in a high average precision AP and interpolated average precision AP_{inter}. Furthermore, it can be noticed from the performance comparison among different descriptors that the proposed color descriptor has the most significant contribution as compared with the gradient and texture descriptors.

2) Comparison With the State of the Art: In order to further validate the effectiveness of the proposed framework and its superiority to the state of the art, it was tested on the OIRDS data set. Its performance was compared with the state of the art using the well-known interpolated average precision AP_{inter} [36] as it takes into account the amount of ground-truth retrieval. The OIRDS data set has a wide range of image scenarios and it was used in several previous works. Table IV shows the achieved 11-point interpolated average precision by the proposed and baseline frameworks [15]–[17], [19]. A different number of test images of the OIRDS data set were chosen by each baseline framework. To overcome this problem, the evaluation of the proposed framework was carried out on the whole data set and on the subset that was used by Chen *et al.* [19]. We were not able to test the algorithm of [19] on the whole data set due to the unavailability of the source code. Clearly, the proposed algorithm outperforms the state of the art by achieving higher AP_{inter} on the OIRDS data set using only a linear classifier and one model as opposed to the nonlinear classifiers or the two models used in [15] and [17], despite the fact that this metric penalizes methods that retrieves only a subset of the ground truth [36]. It is worth to

IEEE Proof

TABLE IV

PERFORMANCE OF THE PROPOSED FRAMEWORK AND BASELINE FRAMEWORKS ON THE OIRDS DATA SET REPRESENTED BY THE INTERPOLATED AVERAGE PRECISION (AP_{inter})

Framework	Features	Classifier	Testing Criteria	AP_{inter}
Razakarivony and Jurie [15]	Gradients	Autoencoders and PCA manifolds	10 Folds ($\approx$ 900 images)	48.9%
ElMikaty and Stathaki [16]	Gradients, Spectrum, Colours	Linear SVM	One Fold ($\approx$ 900 images)	53.45%
Madhogaria *et al.* [17]	Gradients	Linear SVM and MRF	One Fold ($\approx$ 500 images)	54.45%
Chen *et al.* [19]	Gradients	Sparse Coding Based Dictionary	One Fold ($\approx$ 370 images)	62.15%
Proposed Framework	Gradients, Colours, Texture	Linear SVM	One Fold ($\approx$ 370 images)	**64.49**%
Proposed Framework	Gradients, Colours, Texture	Linear SVM	One Fold ($\approx$ 900 images)	57.34%

TABLE V

RUNNING-TIME ANALYSIS OF THE PROPOSED FRAMEWORK. THE REQUIRED COMPUTATIONAL TIME IS GIVEN BY A PERCENTAGE OF THE TOTAL TIME REQUIRED TO PROCESS A GIVEN IMAGE

Stage	Grad. Des.	Col. Des.	Text. Des.	Wind. Eval. & Other
Processing Time (%)	$\approx 5.09\%$	$\approx 11.35\%$	$\approx 69.99\%$	$\approx 13.57\%$

mention that the framework of [15] used a multiple-fold training/testing strategy, whereas the OIRDS data set was completely unseen to other baseline frameworks and the proposed framework. It is also worth to mention that the framework of [17] suffers from high-confidence score false positives. This makes it unreliable in some applications, such as in military fields.

C. Running-Time Analysis

The proposed framework has been implemented in a MATLAB environment and all of our experiments were run on an Intel-Core-i5 CPU with a speed of 2.5 GHz and a RAM of 6-GB size. Whilst the machine that was used has multiple processing cores, the current implementation uses only a single core. It is faster than the work of [16] by $\approx$5.5 times and is able to process $\approx$421 detection windows per second ($\approx$0.03 MB/s). This indicates that the current implementation has a near-real-time performance and is faster than the recent work of [19] that is able to process only $\approx$0.01 MB/s.

Table V shows the percentage of the processing-time consumed by each proposed descriptor. It can be inferred that the proposed color descriptor is very efficient in terms of its high discrimination power, as it was shown in Fig. 10 and Table III, and its relatively lower computational cost. The window-evaluation stage saves the required computational time by approximately 85.09% as it discards a huge number of candidate windows at an early stage.

VI. CONCLUSION

This paper presented a novel framework for the detection and localization of cars in high-resolution airborne imagery using an ensemble of image descriptors that depict the distribution of the gradients, colors, and texture of cars. The advantages of the proposed framework were investigated thoroughly and demonstrated clearly in its ability to achieve a higher precision compared with the state of the art. Several points can be learnt from the implementation and results of the proposed framework. First, eigenvalues of the covariance matrices of locally accumulated gradients can work effectively to identify the dominant orientation of a given detection window. Second, local normalization of the extracted feature vectors is essential for illumination invariance. Third, rotation-invariance properties can be achieved using the proposed adaptive cell distributions without altering the relative distribution of pixels. Fourth, the proposed color descriptor is a powerful single-cue descriptor. However, combining it with other descriptors results in better performance. Fifth, distinctive features of the foreground and background objects must be taken into account, when descriptors are chosen.

Furthermore, the proposed framework has the capacity to generalize as corroborated by testing on completely unseen data. The detection method used here can be further extended by adopting a multiscale analysis approach, e.g., the one adopted by Dalal and Triggs [11]. It has also been demonstrated that the use of an ensemble of image descriptors that efficiently show the visual characteristics of a particular object can lead to improved performance using low-dimensional feature vectors alongside a simple linear classifier.

ACKNOWLEDGMENT

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF).

REFERENCES

[1] V. Ramakrishnan, A. K. Prabhavathy, and J. Devishree, "A survey on vehicle detection techniques in aerial surveillance," *Int. J. Comput. Appl.*, vol. 55, no. 18, pp. 43–47, 2012.

[2] M. Liu, Y. Wu, P. Zhang, Q. Zhang, Y. Li, and M. Li, "SAR target configuration recognition using locality preserving property and Gaussian mixture distribution," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 268–272, Mar. 2013.

[3] X. Yuan, T. Tang, D. Xiang, Y. Li, and Y. Su, "Target recognition in sar imagery based on local gradient ratio pattern," *Int. J. Remote Sens.*, vol. 35, no. 3, pp. 857–870, 2014.

[4] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1. Jul. 2001, pp. 710–717.

[5] S. Hinz, "Detection and counting of cars in aerial images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Sep. 2003, pp. 997–1000.

[6] S. Hinz, C. Schlosser, and J. Reitberger, "Automatic car detection in high resolution urban scenes based on an adaptive 3D-model," in *Proc. 2nd GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion Over Urban Areas*, May 2003, pp. 167–171.

[7] J.-Y. Choi and Y.-K. Yang, "Vehicle detection from aerial images using local shape information," in *Advances in Image and Video Technology* (Lecture Notes in Computer Science), vol. 5414. Springer, 2009, pp. 227–236.

IEEE Proof

[8] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1197–1203.
[9] T. T. Nguyen, H. Grabner, H. Bischof, and B. Gruber, "On-line boosting for car detection from aerial images," in *Proc. IEEE Int. Conf. Res., Innov. Vis. Future*, Mar. 2007, pp. 87–95.
[10] P. Viola and M. J. Jones, "Robust real-time face detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. 2001, p. 747.
[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.
[12] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[13] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1250–1265, Jun. 2011.
[14] W. Shao, W. Yang, G. Liu, and J. Liu, "Car detection from high-resolution aerial imagery using multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 4379–4382.
[15] S. Razakarivony and F. Jurie, "Small target detection combining foreground and background manifolds," in *Proc. IAPR Int. Conf. Mach. Vis. Appl.*, 2013, p. 4.
[16] M. ElMikaty and T. Stathaki, "Car detection in high-resolution urban scenes using multiple image descriptors," in *Proc. IAPR Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4299–4304.
[17] S. Madhogaria, P. M. Baggenstoss, M. Schikora, W. Koch, and D. Cremers, "Car detection by fusion of HOG and causal MRF," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 1, pp. 575–590, Jan. 2015.
[18] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[19] Z. Chen *et al.*, "Vehicle detection in high-resolution aerial images via sparse representation and superpixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 103–116, Jan. 2016.
[20] H. Moon, R. Chellappa, and A. Rosenfeld, "Optimal edge-based shape detection," *IEEE Trans. Image Process.*, vol. 11, no. 11, pp. 1209–1227, Nov. 2002.
[21] H. Moon, R. Chellappa, and A. Rosenfeld, "Performance analysis of a simple vehicle detection algorithm," *Image Vis. Comput.*, vol. 20, no. 1, pp. 1–13, Jan. 2002.
[22] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*, 3rd ed. Thomson Engineering, 2007.
[23] H. Zheng, L. Pan, and L. Li, "A morphological neural network approach for vehicle detection from high resolution satellite imagery," in *Neural Information Processing* (Lecture Notes in Computer Science), vol. 4233. Berlin, Germany: Springer, 2006, pp. 99–106.
[24] J. Serra, *Image Analysis and Mathematical Morphology*. San Diego, CA, USA: Academic, 1983.
[25] L. Eikvil, L. Aurdal, and H. Koren, "Classification-based vehicle detection in high-resolution satellite images," *ISPRS J. Photogram. Remote Sens.*, vol. 64, no. 1, pp. 65–72, Jan. 2009.
[26] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, "Airborne vehicle detection in dense urban areas using hog features and disparity maps," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2327–2337, Dec. 2013.
[27] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.
[28] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.
[29] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[30] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[31] L. Shen and Z. Ji, "Gabor wavelet selection and SVM classification for object recognition," *Acta Autom. Sinica*, vol. 35, no. 4, pp. 350–355, Apr. 2009.
[32] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 36, no. 7, pp. 1169–1179, Jul. 1988.
[33] M. Cramer, "The DGPF-test on digital airborne camera evaluation—Overview and test design," *Photogramm.-Fernerkundung-Geoinf.*, vol. 2, no. 2, pp. 73–82, 2010.
[34] F. Tanner *et al.*, "Overhead imagery research data set—An annotated data library & tools to aid in the development of computer vision algorithms," in *Proc. IEEE Workshop Appl. Imag. Pattern Recognit.*, Oct. 2009, pp. 1–8.
[35] V. Y. Mariano *et al.*, "Performance evaluation of object detection algorithms," in *Proc. IAPR Int. Conf. Pattern Recognit.*, vol. 3. Aug. 2002, pp. 965–969.
[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.



Mohamed ElMikaty (M'17) received the B.Sc. degree (Hons.) in electronics and communications engineering from Mansoura University, Mansoura, Egypt, in 2009, and the M.Sc. degree (Hons.) in communications and signal processing engineering and the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2012 and 2016, respectively.

He is currently a Research Associate with the Department of Electrical and Electronic Engineering, Imperial College London. He has been involved in various industrial research projects in Egypt and U.K. His research interests include computer vision and machine learning, with an emphasis on image representation, analysis, and understanding techniques with applications to object detection, recognition, and tracking.



Tania Stathaki (M'17) was born in Athens, Greece. She received the master's degree in electronics and computer engineering from the National Technical University of Athens, Athens, Greece, and the Ph.D. degree in signal processing from Imperial College London, London, U.K.

She was a Lecturer with the Department of Information Systems and Computing, Brunel University London, London, and an Assistant Professor with the Department of Technology Education and Digital Systems, University of Piraeus, Piraeus, Greece. She is currently a Reader with the Department of Electrical and Electronic Engineering, Imperial College London. She has intensive research experience in image processing and computer vision and, more specifically, image fusion, image registration, change detection, object detection and recognition, and object tracking. She has been mainly involved in defense and security applications and has collaborated with various U.K. companies, such as Dstl, General Dynamics, Selex Galileo, and BAE Systems. She is actively involved in various European Union and U.K. research projects. She has authored or co-authored many papers on signal and image processing and computer vision.

IEEE Proof

Detection of Cars in High-Resolution Aerial Images of Complex Urban Environments

Mohamed ElMikaty, *Member, IEEE*, and Tania Stathaki, *Member, IEEE*

***Abstract*— Detection of small targets, more specifically cars, in aerial images of urban scenes, has various applications in several domains, such as surveillance, military, remote sensing, and others. This is a tremendously challenging problem, mainly because of the significant interclass similarity among objects in urban environments, e.g., cars and certain types of nontarget objects, such as buildings' roofs and windows. These nontarget objects often possess very similar visual appearance to that of cars making it hard to separate the car and the noncar classes. Accordingly, most past works experienced low precision rates at high recall rates. In this paper, a novel framework is introduced that achieves a higher precision rate at a given recall than the state of the art. The proposed framework adopts a sliding-window approach and it consists of four stages, namely, window evaluation, extraction and encoding of features, classification, and postprocessing. This paper introduces a new way to derive descriptors that encode the local distributions of gradients, colors, and texture. Image descriptors characterize the aforementioned cues using adaptive cell distributions, wherein the distribution of cells within a detection window is a function of its dominant orientation, and hence, neither the rotation of the patch under examination nor the computation of descriptors at different orientations is required. The performance of the proposed framework has been evaluated on the challenging Vaihingen and Overhead Imagery Research data sets. Results demonstrate the superiority of the proposed framework to the state of the art.**

***Index Terms*— Airborne imagery, automatic target recognition, car detection.**

NOTATIONS

X, x	Scalar.
$\mathbf{x}$	Column vector.
$\mathbf{X}$	Matrix.
$y(x)$	Function.
$(.)^T$	Transpose operator.
$(.)^H$	Hermitian transpose operator.
$*$	Convolution operator.
$\#\mathbf{x}$	Cardinality of $\mathbf{x}$.
IE{.}	Expectation operator.
$\mathrm{IR}_{\mathbf{xx}}$	Covariance matrix across $\mathbf{x}$.
d	Eigenvalue.
D	Dimensionality.
λ	Wavelength.
θ	Orientation of the normal to the strips of Gabor function.
ψ	Phase offset.
σ_g	Standard deviation of a Gaussian envelope.
γ	Spatial aspect ratio.

Manuscript received November 3, 2016; revised March 9, 2017 and May 2, 2017; accepted June 3, 2017. This work was supported by the Engineering and Physical Sciences Research Council Industrial Cooperative Awards in Science and Technology in collaboration with BAE Systems Military Air and Information, U.K., under Award 12220127. *(Corresponding author: Mohamed ElMikaty.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: m.elmikaty@imperial.ac.uk; t.stathaki@imperial.ac.uk).

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/TGRS.2017.2716984

I. INTRODUCTION

ROBUST detection and localization of small targets in airborne imagery have attracted many researchers in the fields of image analysis and computer vision over the past years. They have numerous useful applications in security, surveillance, defense, remote sensing, and other domains. Small targets, e.g., cars, shipping containers, aeroplanes, and military installations, are usually depicted within small areas that are represented by a small number of pixels in high-resolution aerial images of low ground sampling distance (GSD). Aerial images are commonly captured by sensors mounted on unmanned aerial vehicles, aeroplanes, and earth-observation satellites. In this paper, cars are considered the small targets of interest and the task is to automatically detect and locate cars relative to the rest of a given image of a typical urban scene. This is a tremendously challenging task. First, because of the huge intraclass variability that stems from the significant differences in the visual appearance among cars, as far as various features are concerned, such as colors, size, shape, contours, and boundaries. Furthermore, it is obvious that cars' appearance explicitly depends on the viewing angle. In order to overcome this problem of variability among cars, many researchers have focused on the task of the detection of cars in airborne imagery, wherein a collection of images are taken from a vertical or a slightly oblique viewpoint. This constrained viewpoint facilitates the detection process by eliminating that strong dependence on the viewing angle. In addition, most of the cars would possess almost similar sizes. Second, there is a significant interclass similarity between cars and other objects that arises as a result of the existence of various types of noncar objects in urban scenes, which have similar visual appearance. For instance, similar edge maps or color distributions to that of cars, such as those of buildings' windows, patios, and balconies, can be found.

In this paper, a novel framework is proposed that has the ability to carry out the robust detection and localization of cars in high-resolution airborne imagery of complex urban scenes. Original contributions of this paper include: 1) the design of a novel method that evaluates the likelihood of the detection windows to contain a target in order to reduce

IEEE Proof

0196-2892 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

search areas; 2) a new method to estimate accurately the dominant orientation of detection windows; 3) the introduction of adaptive distributions of cells over which image descriptors can be computed in a new fashion without rotating the original patch as it is a computationally expensive process and it alters the relative distribution of pixels; and 4) a novel encoding of the local distribution of complementary image cues for cars as well as the associated principal components into the feature vector. Points of strength of the proposed framework are: 1) robust performance in complex scenes of urban environments represented by the achieved higher precision and lower false-positive rate and 2) efficient encoding of image cues exploiting lower dimensionality than the state of the art.

This paper is organized as follows. In Section II, a brief discussion of previous works is presented. The proposed framework is introduced in Section III, followed by the description of the Vaihingen data set that was used for training and testing and the Overhead Imagery Research data set (OIRDS) that was only used for testing in Section IV. Results are demonstrated and discussed in Section III-B1. The conclusions are drawn in Section VI.

II. Related Work

Several methodologies have been proposed to solve the problem of the detection of cars in airborne imagery from a perpendicular or a slightly oblique viewpoint [1] and several types of images have been investigated, such as color images, synthetic aperture radar, and others [2], [3]. Nevertheless, our focus in this paper is on the detection of cars in color images and notable related past works will be discussed in this section.

Previous frameworks mainly adopted flat explicit or implicit models, i.e., models that do not use deep architectures. They can be broadly categorized into three main groups: the first group uses explicit models, the second group uses implicit models, and the third group exploits road maps alongside either an implicit or an explicit model.

A. Detection of Cars Using Explicit Models

Early works tried to solve the problem of the detection of cars using explicit models, wherein 2-D and 3-D templates that resemble image cues of cars are designed. Cars are claimed to be found, when the cost of matching between image descriptors of the patch under examination and those of the predesigned template is lower than a predefined threshold. For example, proposed works in [4]–[6] exploited a 3-D template that resembled the geometric structure of cars' contours and the distribution of the colors of cars' bodies, windscreens, and shadows. In a similar fashion, Choi and Yang [7] adopted a "bottom–up" approach, wherein blobs of pixels that possess similar geometric and radiometric properties to that of cars were formed using a mean-shift clustering algorithm [8].

Problems of such template-based models usually arise as a result of the high computational cost that is incurred by the matching process and their performance usually degrades, when they are applied in urban scenes with heavily cluttered backgrounds. This results in low precision rates because of the significant interclass similarity.

B. Detection of Cars Using Implicit Models

Unlike explicit models, implicit models do not exploit the predesigned templates. They use different image descriptors that are engineered to capture and encode different cues of cars. A classifier is trained on the extracted features and is then used to decide whether the patch under examination belongs to the car subspace or the noncar subspace. Several descriptors have been used. Nevertheless, there are no solid theoretical foundations about the choice of descriptors in the open literature, which can accommodate any possible image processing scenario. However, some guidelines can certainly be found. An example of the implicit models is the work that was conducted by Nguyen *et al.* [9], who used Haar wavelets [10], histogram of oriented gradients (HOG) [11], local binary patterns (LBPs) [12], and a mean-shift clustering algorithm [8] for postprocessing. Kembhavi *et al.* [13] followed a similar approach, i.e., they exploited color probability maps, pairs of pixels, and HOG to depict color and geometric structural properties. Likewise, Shao *et al.* [14] used HOG, LBP, and opponent histograms in their vehicle detector and used an intersection kernel support vector machines (SVMs) classifier. Following the basic idea of implicit models but in a different way, Razakarivony and Jurie [15] built two separate models: one for the HOG descriptors of vehicles and the other for those of the background using manifold learning. ElMikaty and Stathaki [16] addressed the problem using HOG, Fourier, and truncated pyramid color self-similarity jointly with a linear SVM classifier. Madhogaria *et al.* [17] used only HOG to detect cars in a two-stage classification process. In their first stage, HOG descriptors were computed over rotated images alongside a linear SVM classifier and a causal Markov random field (MRF) classifier [18]. Using superpixel segmentation, Chen *et al.* [19] extracted meaningful patches. Then, patches were rotated so that HOG descriptors could be computed. A sparse-coding-based dictionary was learnt and used to detect cars.

Although acceptable results of implicit models have been reported, their performance is significantly affected by the choice of features, the implementation, and the dimensionality of the descriptors used and classifiers. Moreover, it can be observed that in order to overcome the problems of intraclass variability and the interclass similarity, either a high-dimensional feature descriptor was used in conjunction with linear classification or a low-dimensional feature descriptor was used alongside nonlinear or multistage classification.

C. Detection of Cars Using Road Maps

In order to restrict search areas to regions that are more likely to contain cars, several previous works have exploited road maps to extract roads and parking areas as an initial step in order to decrease the false-positive rate. A representative example is the work conducted by Moon *et al.* [20], [21], wherein the authors made use of the locations of roads and buildings and utilized the derivative of the double-exponential filter [22] to extract gradients. Similarly, Zheng *et al.* [23] extracted roads and parking areas and used morphological structuring elements [24] and a classifier to distinguish actual car pixels. In a similar manner, Eikvil *et al.* [25] proposed a detection approach that consisted of a segmentation stage to

IEEE Proof

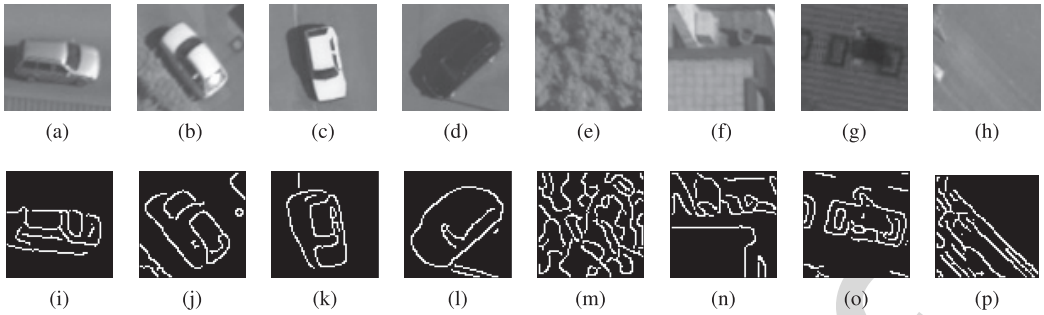


Fig. 1. Extraction of edges in target-class and non-target-class patches using Canny edge detectors [28] with an associated Gaussian filter of standard deviation of $\sqrt{2}$. Target-class patches are shown in (a)–(d), and their corresponding edges in (i)–(l), respectively. Non-target-class patches are shown in (e)–(h), and their corresponding edges in (m)–(p), respectively. Shown images were scaled for viewing.

separate regions that contain cars, followed by a two-stage object-classification framework exploiting road networks. Following the same idea, Tuermer *et al.* [26] proposed the use of road databases, disparity maps, and an HOG object detector.

Whilst exploiting road maps provides better performance in terms of true-positive and false-positive rates, it is based on the prior knowledge of accurate road maps, a precise map-projection method, which is computationally expensive, and the use of extra hardware, such as a global positioning system.

Unlike flat models, deep models for the detection of cars can hardly be found in the literature as they require vast training data sets, which are challenging to acquire. To the best of the our knowledge, only one work was proposed by Chen *et al.* [27] to solve the problem of car detection using deep models. In this paper, a hierarchal deep model was used in a hybrid fashion. Although high accuracy has been achieved, their testing data set compromised a set of images of vast parking areas containing a large number of cars and lacked the existence of complex shapes that characterize urban scenes.

III. Proposed Framework

Regarding a car in a given aerial image, its boundaries are expected to have almost rectilinear structures that correspond to the edges of the body and front and rear windscreens as shown in Fig. 1. Although these rectilinear structures form together a considerable number of crucial points of interest, e.g., high-curvature points and corners, they are incapable of discriminating cars from other objects. This is because in some cases, other noncar objects possess similar edge maps to those of cars, such as buildings' roofs and windows, and they can be easily misidentified as cars. Based on this observation, implicit models are seen to be superior to explicit models as they capture fine details and a simple linear classifier can learn from the training samples how to distinguish well-encoded combined image descriptors that correspond to the cues related to cars.

In this paper, three image descriptors are computed over a sliding detection window to characterize the distribution of gradients, colors, and texture. These image cues were chosen among other cues as cars possess: 1) almost rectilinear structures as aforementioned; 2) distinctive color distribution; and 3) mostly smooth surfaces with low amount of texture. The proposed framework adopts a gray-scale fixed-size window of 64×64 pixels with an associated stride length of four pixels as in a set of aerial images captured at a given GSD, cars have almost similar sizes. Square patches were chosen instead of rectangular patches as it was found after exhaustive experimental validations that square patches result in better performance. They incorporate some of the contextual information and they facilitate the computation of image descriptors at different orientations. The proposed framework comprises four stages, namely, window evaluation, extraction of features and encoding, classification, and postprocessing as shown in Fig. 2. Details of the implementation of each stage are discussed in Sections III-A–III-D.

Choice of Features and Parameters: All values of the parameters used in the implementation of the proposed framework were set empirically by testing excessively the properties of a large number of image samples from the training data set that will be described later in Section IV-A using cross validation.

A. Window Evaluation

In order to reduce search areas, it is necessary to investigate detection windows that are highly unlikely to contain cars at an early stage. This reduces the false-positive rate. In addition, it helps shorten the processing time by discarding a large number of windows, as it can be seen in Fig. 3. To achieve this, a novel three-step filtering process is proposed to measure the objectness [29] of the detection window and its likelihood to contain a target at an early stage based on several cues, including the local distribution of colors, texture, and gradients.

The first step is to investigate the variation in the distribution of gray levels across a detection window. This is examined by the computation of two 32-bin gray-level histograms $\mathbf{h}_w$ and $\mathbf{h}_c$, one for the whole detection window and the other for the 32×32 central square area. If the similarity between the two histograms is more than 50%, the window is discarded as this highly indicates that it belongs to a flat area with a constant or a slowly varying gray-level distribution, such as empty roads and vegetation areas. The Manhattan distance $r_{\text{Manhattan}}$ between these two color histograms is computed to measure their similarity with a threshold value of 2048 as the size of the detection window is 64×64 pixels

$$r_{\text{Manhattan}} = \sum_{k=1}^{32} |\mathbf{h}_w(k) - \mathbf{h}_c(k)|. \quad (1)$$

IEEE Proof

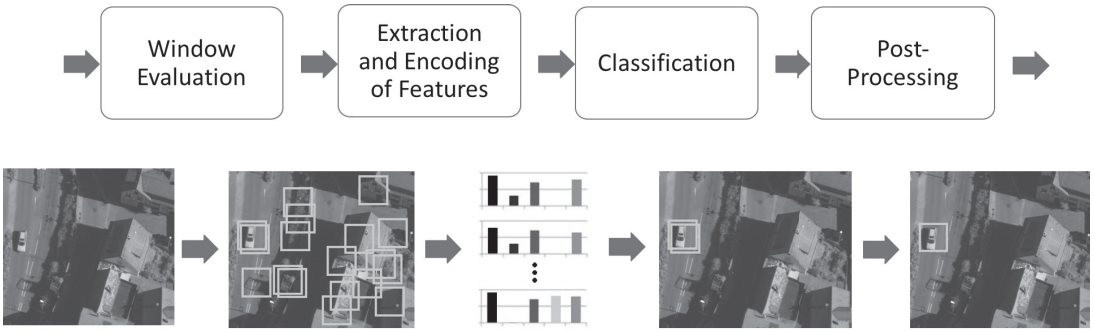
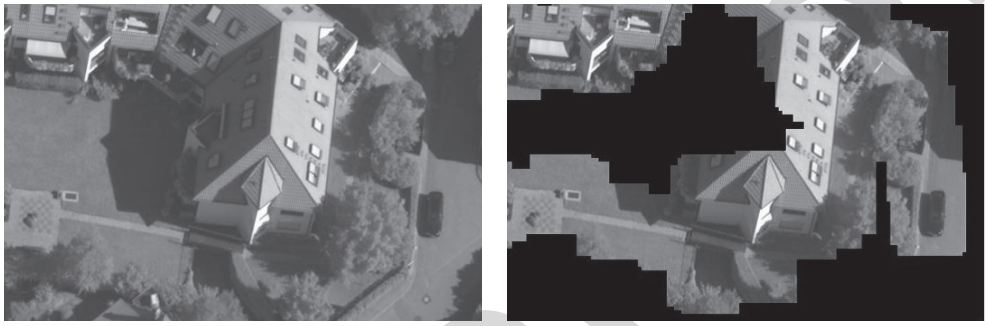


Fig. 2. Flow diagram of the adopted four-stage methodology. (Top) Block diagram shows the sequence of the four main stages of the adopted general methodology. (Bottom) Original image, processed image using the window-evaluation stage, computed image descriptors for each successful candidate window, output of the classification stage, and final detection after postprocessing (from Left to Right). Shown images were scaled for viewing.



(a) (b)

Fig. 3. Application of the window-evaluation stage on a sample region of the Vaihingen data set. (a) Original image. (b) Corresponding processed image. Areas that have been eliminated by the window-evaluation stage as they were highly unlikely to contain cars are marked with black. Shown images were scaled for viewing.

The second step is to measure the texture of the detection window. Pixels of each local cell of 4×4 pixels are divided into two groups according to whether the pixel value is larger than the average value of the cell or smaller. Then, the difference between the averages of the two groups is calculated. If the standard deviation σ_{texture} of the computed differences is larger than a predefined threshold value of 32, the detection window will not be considered for further processing. The intuition behind this is that a high standard deviation σ_{texture} of the mean values of the differences among average intensities (more than one eighth of the intensity range [0,255]) indicates a highly textured surface that cannot be a car as cars tend to have smooth surfaces.

The third step is to examine the local distribution of gradients. For a given gray image $\mathbf{I}$, magnitudes of intensity gradients $\mathbf{M}_g(x, y)$ at each pixel and their orientations $\mathbf{O}_g(x, y)$ are calculated over the detection window. This is done by the computation of the first-order discrete derivatives $(\partial\mathbf{I}/\partial x)$ and $(\partial\mathbf{I}/\partial y)$ of $\mathbf{I}$ using a nonsmoothing filter of $[-101]$ in both the vertical and horizontal directions

$$\mathbf{M}_g(x, y) = \sqrt{\left(\frac{\partial \mathbf{I}(x, y)}{\partial x}\right)^2 + \left(\frac{\partial \mathbf{I}(x, y)}{\partial y}\right)^2} \tag{2}$$

$$\mathbf{O}_g(x, y) = \arctan\left(\frac{\partial \mathbf{I}(x, y)}{\partial y} \Big/ \frac{\partial \mathbf{I}(x, y)}{\partial x}\right). \tag{3}$$

Then, the magnitudes of gradients of local cells of 4×4 pixels are accumulated independently in 9-bin histograms $\mathbf{g}_{\text{cell}}$ according to their unsigned orientation

$$\mathbf{g}_{\text{cell}}(k) = \sum_{x,y} \mathbf{M}_g(x, y) \mid \mathbf{O}_g(x, y) \in \mathbf{o}_k, \quad k \in \{1, 2, \ldots, 9\} \tag{4}$$

where $\mathbf{o}_k$ is the kth bin of orientation interval of width 20°, followed by computing the covariance matrix $\text{IR}_{\mathbf{gg}} = \text{IE}\{\mathbf{g}\mathbf{g}^H\}$ of the $L2$-normalized responses $\mathbf{g}_{\text{cell}} \leftarrow \mathbf{g}_{\text{cell}}/(\|\mathbf{g}_{\text{cell}}\|_2^2 + \epsilon^2)^{1/2}$ from local cells, where ϵ is a small constant and is equal to 0.01. For a typical detection window containing a car, there should be a dominant orientation of the gradients that corresponds to the direction, which is perpendicular to the orientation of the car in that particular window. Therefore, if the covariance matrix $\text{IR}_{\mathbf{gg}}$ is eigendecomposed, the maximum eigenvalue $d_{\max}$ must exceed a predefined threshold value of 0.09, otherwise the window is discarded. The intuition behind this is that the second-order statistics given by the covariance matrix $\text{IR}_{\mathbf{gg}}$ can be used to determine the number of the dominant orientations of gradients and their directions using the rank of $\text{IR}_{\mathbf{gg}}$ and the order of its significant eigenvalues d_k, respectively

$$[d_{1,\text{cells}}, \ldots, d_{9,\text{cells}}]^T = \text{eig}(\text{IR}_{\mathbf{gg},\text{cells}}). \tag{5}$$

Estimation of the Orientations of Cars: The orientation of cars O_{car} is discretized into nine orientations at 10°, 30°, 50°, 70°, 90°, 110°, 130°, 150°, and 170°. It is estimated to be in a perpendicular orientation to the bin orientation corresponding to the maximum eigenvalue $d_{\max}$ of the covariance matrix $\text{IR}_{\mathbf{gg}}$ of locally accumulated gradients computed over 4×4 cells.

IEEE Proof

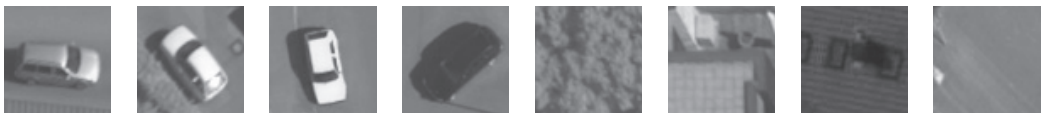


Fig. 4. Estimation of the dominant orientation in target-class and non-target-class patches. Target-class patches are shown in (a)–(d). Non-target-class patches are shown in (e)–(h). Estimated dominant orientations were (a) $O_{\text{dominant}} = 170°$, (b) $O_{\text{dominant}} = 130°$, (c) $O_{\text{dominant}} = 110°$, (d) $O_{\text{dominant}} = 50°$, (e) $O_{\text{dominant}} = 110°$, (f) $O_{\text{dominant}} = 170°$, (g) $O_{\text{dominant}} = 170°$, and (h) $O_{\text{dominant}} = 130°$. Shown images were scaled for viewing.

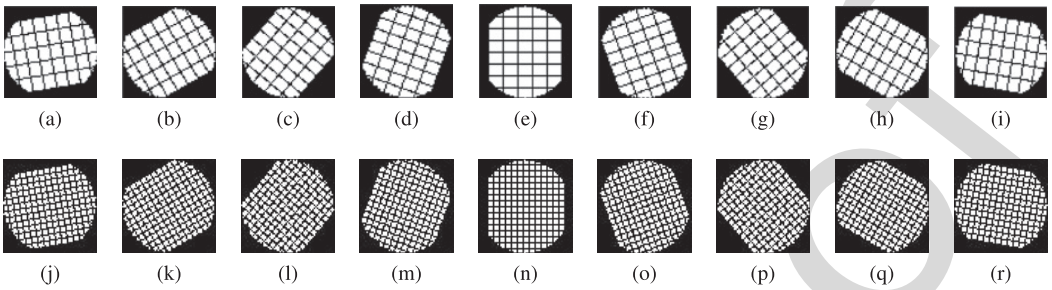


Fig. 5. Cells of original size of 8×10 pixels are shown in (a)–(i): (a) 10°. (b) 30°. (c) 50°. (d) 70°. (e) 90°. (f) 110°. (g) 130°. (h) 150°. (i) 170°. Cells of original size of 4×4 pixels are shown in (j)–(r): (j) 10°. (k) 30°. (l) 50°. (m) 70°. (n) 90°. (o) 110°. (p) 130°. (q) 150°. (r) 170°. Shown images were scaled for viewing.

Examples of estimating the dominant orientation of detection windows are shown in Fig. 4

$$O_{d_{\max}} = (10 + 20(k-1))^\circ | d_k = \max_{k=1,\ldots,9} (\text{eig}(\text{IR}_{\mathbf{gg}})). \quad (6)$$

B. Extraction and Encoding of Features

Detection windows are traditionally divided into square cells of equal size over which image descriptors are computed locally. For rotation-variant features, detection windows are rotated in most cases so that they would have a specific orientation. However, this rotation process is time-consuming and it affects the relative distribution of the original pixels due to the nonlinear transformation between coordinates. To overcome this problem, adaptive cell distributions are proposed, wherein the shape of cells is no longer fixed-size squares. Two sets of adaptive cell distributions are used in the proposed framework and they are shown in Fig. 5. In the first set, distributions of cells are originated from the rotation of a mask oriented horizontally of size 64×50 pixels and divided into ordered cells of size 8×10 pixels to the previously mentioned discretized orientations, whereas in the second set, distributions of cells are originated from the rotation of a mask oriented horizontally of size 64×52 pixels and divided into ordered cells of size 4×4 pixels. Corners of each mask are rounded by eliminating pixels that are located at a distance more than 31 pixels from the center of the window.

Image descriptors are computed over each candidate detection window after being divided into one of cell distributions according to the estimated dominant orientation of the detection window as follows.

1) Gradient Descriptors: Cars have rectilinear structures and a dominant orientation as previously mentioned. To encode this feature, gradient descriptors are exploited. They are among the well-established image descriptors and they are known as being powerful single-cue descriptors [30]. In this paper, gradient descriptors are computed over the first set of the cell distributions, shown in Fig. 5, based on the estimated dominant orientation of the detection window (described in Section III-A). As already mentioned, magnitudes of gradients $\mathbf{M}_g$ of local cells are accumulated in 9-bin histograms $\mathbf{g}_{\text{cell}}$ according to their orientation $\mathbf{O}_g$ independently without the trilinear interpolation process recommended by Dalal and Triggs [11] as it is a computationally expensive process and it has a minimal effect on the overall performance. The response for each cell $\mathbf{g}_{\text{cell}}$ is normalized using $L2-$normalization so that illumination-invariance properties can be achieved. The final gradient descriptor is formed from the concatenation of the $L2$-normalized eigenvalues $d_{k,\text{block}}$ of the covariance matrices $\text{IR}_{\mathbf{gg},\text{block}}$ of the cell responses of groups of four adjacent cells. It has a dimensionality of 216 as there are 24 blocks of cells, each has nine eigenvalues. This differs from the original implementation of [11] in the way local cells are distributed and that the final descriptor consists of the eigenvalues of the local covariance matrices rather than the absolute cell responses to the histogram of gradients.

2) Color Descriptors: Cars exhibit a distinctive distribution of colors: the color of windscreens is always dark in aerial images, and the color of the car boot is usually the same as the color of the car bonnet. Therefore, if a local descriptor that encodes the local similarity and variation of colors across pixels is used, it will facilitate the classification process. A new modified version of the Color Self-Similarity descriptor proposed by ElMikaty and Stathaki [16] is suggested. It incorporates the principal components of the covariance matrices of the local cell responses. The new color descriptor $\mathbf{c}_{\text{cell}}$ calculates the similarity among adjacent cells by computing the intersection between corresponding local-color-histogram bins

$$\mathbf{c}_{\text{cell}} = [\mathbf{h}_c \cap \mathbf{h}_n] = \left[\left[\sum_k \min(\mathbf{h}_c(k), \mathbf{h}_n(k)) \right] \right] \quad (7)$$

IEEE Proof

where $\mathbf{h}_c$ and $\mathbf{h}_n$ are the two gray-level histograms of the central cell and a neighboring cell, respectively, each has 16 bins, $\mathbf{h}_c(k)$ and $\mathbf{h}_n(k)$ are the scores of the kth bins of the two histograms, and $\lceil . \rceil$ indicates that the result is truncated at the size of an unsigned integer of eight bits as recommended by ElMikaty and Stathaki [16]. Local descriptors are computed on two levels in a pyramid approach using the two sets of cell distributions shown in Fig. 5. In the first level, descriptors are computed for 24 cells on a neighborhood of four cells, whereas in the second level, descriptors are calculated for 130 cells with a neighborhood of four cells, two cells with a neighborhood of three cells, and four cells with a neighborhood of two cells as small cells around the edges of the masks were not included. The final color descriptor has 3190 features and is formed by concatenating the $L2$-normalized local responses from all cells into a single vector and the $L2$-normalized eigenvalues of the covariance matrices $\mathrm{IR}_{\mathbf{cc},\mathrm{block}}$ of the local responses for each group of cells in each level of the pyramid.

3) Texture Descriptors: Texture is a common image cue that has been used, among other cues, to solve the problem of object detection [31]. Many researchers have used Gabor filters to extract and encode texture motivated by the similarity between 2-D Gabor filters and the receptive fields in the visual cortex of the vision systems of humans [32]. Although cars do not possess distinguishable textured surfaces, Gabor texture descriptors play a great role in describing other textured areas, such as those of buildings' roofs and vegetation. Gabor filter kernels are formed by a 2-D Gaussian function modulated by a complex sinusoidal signal

$$\mathbf{F}_{\mathrm{Gabor}}(x, y; \lambda, \theta, \psi, \sigma_g, \gamma) = \exp\left(\frac{x'^2 + \gamma^2 y'^2}{2\sigma_g^2}\right) \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right) \quad \begin{cases} x' = x\cos\theta + y\sin\theta \\ y' = -x\sin\theta + y\cos\theta. \end{cases} \tag{8}$$

In this paper, nine Gabor kernels of size 64×64 are used, with a wavelength λ of 10 pixels, nine orientations $\theta_k \in \{10°, 30°, \ldots, 170°\}$, phase offset ψ of $-90°$, aspect ratio γ of 10, and the variance of the Gaussian envelop σ_g^2 of 13.671.

After filtering the detection windows using those nine Gabor kernels, windows are divided into one of the first set of the cell distributions, shown in Fig. 5, according to the estimated orientation of the detection window. The mean $\mathbf{t}_{\mathrm{cell,mean}}$ and energy $\mathbf{t}_{\mathrm{cell,energy}}$ of the response to the Gabor filters in each cell is computed at each orientation, resulting in 18 features per cell

$$\mathbf{t}_{\mathrm{cell,mean}}(k) = \frac{1}{\mathrm{size(cell)}} \sum_{x,y} (|\mathbf{F}_{\mathrm{Gabor}} * \mathbf{I}|) \mid \theta_k = (20k - 10)° \tag{9}$$

$$\mathbf{t}_{\mathrm{cell,energy}}(k) = \frac{1}{\mathrm{size(cell)}} \sum_{x,y} ((\mathbf{F}_{\mathrm{Gabor}} * \mathbf{I})(\mathbf{F}_{\mathrm{Gabor}} * \mathbf{I})^H) \mid \theta_k = (20k - 10)° \tag{10}$$

$$\mathbf{t}_{\mathrm{cell}} = [\mathbf{t}_{\mathrm{cell,mean}}, \mathbf{t}_{\mathrm{cell,energy}}]^T. \tag{11}$$

For rotation-invariance, the order of the nine orientations is changed such that rotation invariance is achieved by circularly shifting the response to different orientations θ_k. The final texture descriptor has 432 features and is formed from the concatenation of the $L2$-normalized eigenvalues of the covariance matrices $\mathrm{IR}_{\mathbf{tt},\mathrm{block}}$ of the cell responses of 24 groups of four adjacent cells.

C. Classification

An extended feature vector is formed from the aforementioned ensemble of descriptors by concatenation. The overall dimension of this vector is 3838-D (216 gradient features, 3190 color features, and 432 texture features). Since this problem of classification is characterized by having a relatively high dimensionality, it is recommended to adopt an SVM classifier that is known to work efficiently in high-dimensional scenarios [18]. Based on that and following Occam's razor, "Among competing hypotheses, the one with the fewest assumptions should be selected," a linear SVM classifier has been adopted in the proposed framework to discriminate the car subspace from the noncar subspace.

An SVM classifier assumes that a total of K positive and negative training data samples $s_k \in \{s_1, \ldots, s_K\}$ are given and that they are drawn from a hyperdimensional space with associated target values $t_k \in \{-1, 1\}$. It aims to solve the classification problem using a linear model in the form of

$$z(s) = \mathbf{w}^T \phi(s) + b \tag{12}$$

where $\phi(s)$ represents the transformation from the spatial domain to the hyper-dimensional space and b is a bias. In other words, the goal is to determine the hyperplane in the feature space $\phi(S) \in \Re^D$ that best separates the car and noncar classes. This can be achieved by maximizing the margin between the hyperplane and data samples that are close to it, which are known as support vectors. During the training stage, parameters $\mathbf{w}$ and b are optimized in order to maximize the margin and the solution is given by

$$\underset{\mathbf{w},b}{\mathrm{argmax}} \left\{ \frac{1}{\|\mathbf{w}\|} \min_k (t_k(\mathbf{w}^T \phi(s_k) + b)) \right\}. \tag{13}$$

Nevertheless, the above-mentioned assumptions are valid only for perfectly linearly separable classes. For overlapping classes, a slack variable $\xi_k \geq 0$, where $k = 1, \ldots, K$, is introduced to allow some training samples to be misclassified, converting the quadratic optimization problem in (13) to a constrained minimization problem of $\|\mathbf{w}\|^2$ as follows:

$$C \sum_{k=1}^{K} \xi_k + \frac{1}{2}\|\mathbf{w}\|^2 \tag{14}$$

subject to the following constraints:

$$\begin{cases} t_k(\mathbf{w}^T \phi(s_k) + b) \geq 1 - \xi_k \\ \xi_k \geq 0 \end{cases} \tag{15}$$

where C is a nonnegative regularization constant and it controls the tradeoff between the slack variable ξ_k and the margin. In this paper, $\phi(s_k)$ transforms the training samples to a hyperspace of 3838-D using a regularization parameter C of 15. More details about the theory and application of SVM classifiers can be found in [18].

IEEE Proof

D. Postprocessing

The aim of the postprocessing stage is to eliminate multiple detections per single true detection that are resulted from the use of an overlapping-sliding-window approach. Detection windows with positive score that overlap by more than 50% are eliminated except the one with the highest output confidence score.

IV. Data Sets

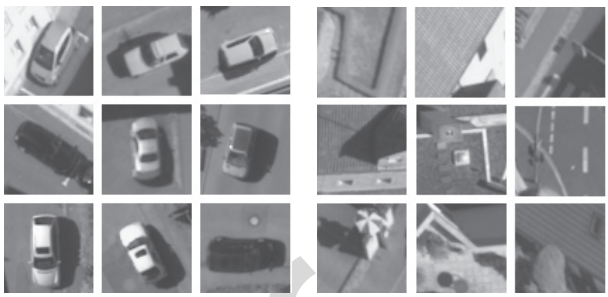
The Vaihingen data set [33] was used to train the SVM classifier and to evaluate the performance of the proposed framework. In order to evaluate the ability of the proposed framework to generalize on unseen data sets and compare its performance with the state of the art, the publicly available OIRDS data set [34] has been used and it was not included in the training data set. Description of both data sets is given below.

A. Training Data Set

The training process was carried out on a part of the Vaihingen data set that comprises a set of high-resolution aerial images of a GSD that is equal to ≈8 cm and a radiometric resolution of 11 b. Images of the Vaihingen data set are pan-sharpened colored infrared images and they were provided in TIFF RGB files. The data set was captured over the city of Vaihingen—Germany by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [33] using a Digital Mapping Camera with focal length of 120 mm and at a flying height above the ground of 900 m. This means that the requirement that the viewpoint must be at a perpendicular or a slightly oblique angle is met. Cars in the Vaihingen data set are almost of the same size and are contained inside square windows of 48×48 pixel size as the GSD is 8 cm.

Positive training data consisted of square patches of 64×64 pixels so that image descriptors can be computed efficiently over cars and a small part of the surrounding background. Patches that contain cars were chosen from areas other than the test areas. They were chosen from various regions such that a great diversity in the orientation of cars and backgrounds could be achieved. Data augmentation techniques were exploited by the inclusion of mirrored and horizontally flipped versions of these patches to the training data in order to richen the diversity in the training data set.

As for the negative training data, random patches were included that were selected from different backgrounds. In addition, bootstrapped hard negatives that were detected as false positives during the validation process were added. In total, 781 positive samples and 2473 negative samples were used in the training process. Sample positive and negative patches are shown in Fig. 6.



(a) (b)

Fig. 6. Samples of the patches that were extracted from the Vaihingen data set and used for training. (a) Positive training samples of car-class targets of different orientations and colors. (b) Negative training samples of non-car-class targets. Shown images were scaled for viewing.

B. Testing Data Set

1) Vaihingen Data Set: DGPF specified three testing areas for the evaluation of the developed algorithms that contained 128 nonoccluded cars. They contain typical urban backgrounds with different building styles, roads, and vegetation. Description of the three areas is as follows. Area 1 contains dense development consisting of historic buildings having complex shapes and it also has some trees, area 2 has a few high-rising residential buildings that are surrounded by trees, and area 3 is a pure residential area with small detached houses. Samples for the three different testing areas are shown in Fig. 7.

2) Overhead Imagery Research Data Set: The OIRDS data set [34] has more than 900 annotated images that contain approximately 1075 nonoccluded cars. The data set is a part of the freely available aerial imagery captured by the United States Geological Survey. Images were captured at a GSD of 15 cm under a variety of operating characteristics. Images were upsampled by a factor of 1.875 so that the final GSD is ≈8 cm. This means that the data set has a lower resolution compared with the training data set described in Section IV-A. Examples of the test images are given in Fig. 8. In this paper, no examples of the OIRDS data set were included in the training data set in order to test the ability of the proposed framework to generalize.

V. Experimental Results

The performance of the proposed framework was evaluated on the testing areas of the Vaihingen data set as well as on the OIRDS data set. No manual annotation of any testing data was carried out except to evaluate the algorithm (labeling ground truth in the testing data set only). The performance was assessed and evaluated using qualitative analysis and a number of quantitative measures, namely, the precision and recall rates, the average precision, and the interpolated average precision.

A. Qualitative Analysis

Visual results of the experiments of applying the proposed framework on sample regions of the test areas of the Vaihingen data set are shown in Fig. 7, whereas results of sample images of the OIRDS data sets are shown in Fig. 8. These two figures show that our framework is able to robustly detect cars in various environments.

In the center left image of Fig. 7, it can be seen that the proposed method is able to correctly classify detection windows that have poor contrast (fourth car from top), thanks to the local normalization of the gradient-descriptor responses of cells, discussed in Section III-B.1. This is a case that is even hard for the human eye. Nevertheless, in few situations, wherein the detection windows contain very dark cars that are covered by very strong building shadows and the contrast is too low, the proposed method fails to correctly classify

IEEE Proof

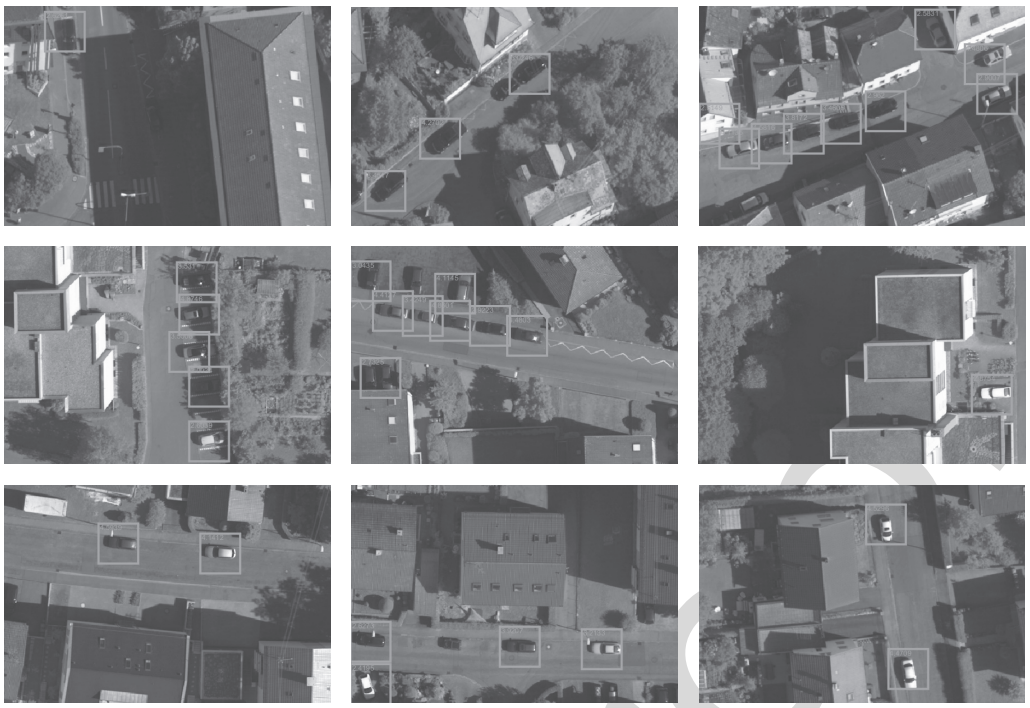


Fig. 7. Sample regions of the test area of the Vaihingen data set after applying the proposed framework. First row: sample regions from test area 1. Second row: sample regions from test area 2. Third row: sample regions from test area 3. Bounding boxes represent detections. Shown images were scaled for viewing.

them as true positives. For example, in the top-left image of Fig. 7, the black car parked next to the building was missed, because the strength of gradients in such detection windows is too low even after normalization and the color distribution is concentrated around the histogram bins that correspond to dark gray levels.

In the top-center and bottom-center images of Fig. 7, two cars were not detected as they have a very small size compared with other standard cars. However, this problem can be solved by extending the proposed method to multiscale analysis which is among our future goals.

From Fig. 8, it can be clearly deduced that the proposed framework has very good generalization as it achieves high true-positive rate while maintaining a low false-positive rate when it is applied to images with different acquisition conditions and a lower resolution than those of the data on which the SVM classifier was trained. More specifically, in challenging cases, where cars are partially overimposed by shadows, e.g., the first image of the first row in Fig. 8, our framework is able to detect cars despite the fact that the shadow alters the distribution of edges and colors. Moreover, in the case of complex backgrounds, for example, the third image of the first row in Fig. 8, the proposed algorithm does not produce a high rate of false positives.

B. Quantitative Analysis

Quantitative evaluation methods for target detection systems depend on the adopted approach [35]. As a window-based approach was adopted alongside a classifier that outputs a confidence score (distance from the separating hyperplane), metrics that depend on ranked outputs according to the confidence score in a descending order should be used. In this paper, to evaluate the performance and compare it with the state of the art that constitutes the baseline performance, the interpolated average precision $\mathrm{AP}_{\mathrm{inter}}$ was used.

In the adopted evaluation methodology, a detected bounding box B_D was considered true if it matched a corresponding ground-truth bounding box B_{GT} using (16) and there were no other bounding boxes matching the same ground truth [36]. In the case of having more than one bounding box matching the same ground truth, only one of them was considered a true detection and others were considered false positives. Ambiguous objects, i.e., objects, which are partially occluded, were not taken into account, when measuring the performance of the proposed framework following the works in [13]–[17] that also discarded occluded cars:

$$\frac{\mathrm{Area}(B_D \cap B_{\mathrm{GT}})}{\mathrm{Area}(B_D \cup B_{\mathrm{GT}})} \geq 0.5. \tag{16}$$

After defining the condition for considering true positives, precision PR, and recall RC can be computed at several levels according to the rank of the confidence score associated with a detection window using (17) and (18). The term "recall" is used here to indicate the proportion of the truly detected cars to the total number of ground-truth cars, whereas the term "precision" indicates the proportion of truly detected cars to the total number of detections above a given confidence score [36]

$$\mathrm{PR} = \frac{\#\mathbf{b}_{\mathrm{TP}}}{\#\mathbf{b}_{\mathrm{TP}} + \#\mathbf{b}_{\mathrm{FP}}} \tag{17}$$

$$\mathrm{RC} = \frac{\#\mathbf{b}_{\mathrm{TP}}}{\#\mathbf{b}_{\mathrm{GT}}} \tag{18}$$

where $\mathbf{b}_{\mathrm{TP}}$ represents the true positives, $\mathbf{b}_{\mathrm{FP}}$ represents the false positives, and $\mathbf{b}_{\mathrm{GT}}$ represents the ground truth.

IEEE Proof



Fig. 8. Sample regions of the test images of the OIRDS data set after applying the proposed framework. Bounding boxes represent detections. Shown images were scaled for viewing.

TABLE I

PERFORMANCE OF THE PROPOSED FRAMEWORK WITH DIFFERENT WINDOW EVALUATION SETTINGS ON THE VAIHINGEN DATA SET REPRESENTED BY THE AVERAGE PRECISION AP AND THE INTERPOLATED AVERAGE PRECISION AP_{inter}

Window Evaluation	AP	AP_{inter}
Des. + SVM	78.65%	64.88%
Histogram Difference + Des. + SVM	85.02%	63.78%
Texture + Des. + SVM	82.14%	68.01%
Maximum Eigenvalue + Des. + SVM	80.82%	66.79%
Proposed Wind. Eval. + Des. + SVM	**92.79%**	**68.03%**

The average precision AP and its interpolated version AP_{inter} can be computed using the following:

$$\text{AP} = \frac{1}{\#\mathbf{b}_{\text{TP}}} \sum_{k \in \{1,2,\ldots,\#\mathbf{b}_{\text{TP}}\}} \mathbf{PR}(\mathbf{b}_{\text{TP}}(k)) \tag{19}$$

$$\text{AP}_{\text{inter}} = \frac{1}{11} \sum_{k \in \{0,0.1,\ldots,1\}} \mathbf{PR}_{\text{interp}}(k) \mid \mathbf{PR}_{\text{interp}}(k) = \max_{\tilde{r}:\tilde{r} \geq r} [\mathbf{PR}(\tilde{r})] \tag{20}$$

where $\mathbf{b}_{\text{TP}}$ represents true positives at a given confidence score and **PR** represents the associated precision. It is worth to mention that the AP_{inter} penalizes methods that are not able to retrieve 100% of the ground-truth data [36], unlike the AP that does not take this into account. Accordingly, the AP_{inter} is more commonly used in quantitative analysis.

1) Performance of the Proposed Framework: The proposed framework uses the eigenvalues of the covariance matrix of the local gradient descriptor to estimate the orientation of cars as previously discussed in Section III-A. The achieved accuracy of estimating the orientation of cars in the test data set is 98.34%.

Regarding the contribution of each step in the window-evaluation stage, the performance of the detection system was assessed with adding only one of the proposed window filtering steps. Fig. 9 and Table I demonstrate the obtained

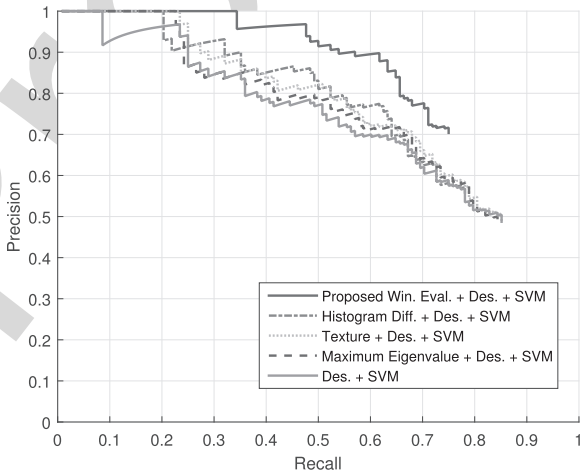


Fig. 9. Performance of the proposed framework on the Vaihingen data set represented by precision rates at different recall rates with different window evaluation settings and the proposed ensemble of descriptors alongside a linear SVM classifier.

results under five different scenarios: 1) the ensemble of descriptors was only used in conjunction with the SVM classifier; 2) window evaluation based on the difference in gray histograms and the ensemble of descriptors were used in conjunction with the SVM classifier; 3) window evaluation based on the level of texture and the ensemble of descriptors were used in conjunction with the SVM classifier; 4) window evaluation based on the value of the maximum eigenvalue and the ensemble of descriptors were used in conjunction with the SVM classifier; and 5) the overall proposed framework including the three steps of window evaluation and the ensemble of descriptors. Furthermore, Table II shows the effect of changing the recommended value of one of the three thresholds used in the window-evaluation stage while keeping the other two in their recommended values.

To evaluate the effect of each proposed image descriptor, the performance of the detection system was assessed using different combinations of descriptors. Fig. 10 and Table III demonstrate the obtained results under seven different

IEEE Proof

TABLE II

PERFORMANCE OF THE PROPOSED FRAMEWORK WITH DIFFERENT VALUES FOR THE THRESHOLDS USED IN THE WINDOW-EVALUATION STAGE ON THE VAIHINGEN DATA SET REPRESENTED BY THE AVERAGE PRECISION AP AND THE INTERPOLATED AVERAGE PRECISION AP_{inter}

$\pm x\%$	Histogram Difference		Texture		Maximum Eigenvalue	
	AP	AP_{inter}	AP	AP_{inter}	AP	AP_{inter}
-8.0%	90.46%	67.26%	91.59%	67.24%	92.40%	67.67%
-4.0%	91.82%	67.66%	92.92%	67.33%	**92.86%**	**68.18%**
0.0%	92.79%	68.03%	**92.79%**	**68.03%**	92.79%	68.03%
+4.0%	**94.39%**	**68.37%**	91.78%	67.19%	92.19%	67.32%
+8.0%	93.86%	68.09%	91.19%	67.02%	92.32%	67.50%

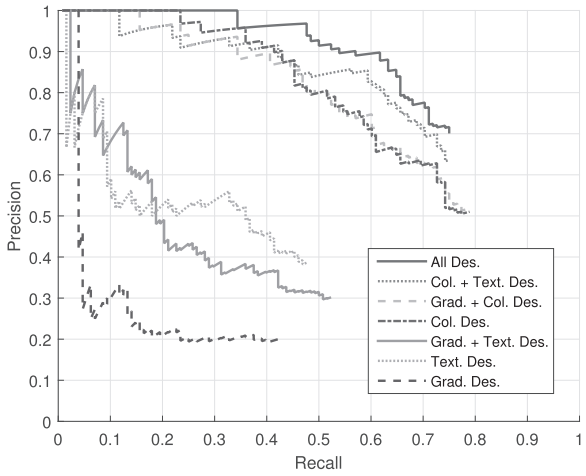


Fig. 10. Performance of the proposed framework on the Vaihingen data set represented by precision rates at different recall rates with the proposed three-step window-evaluation stage and different combinations of descriptors alongside a linear SVM classifier.

TABLE III

PERFORMANCE OF THE PROPOSED FRAMEWORK WITH THE WINDOW-EVALUATION STAGE AND DIFFERENT COMBINATIONS OF DESCRIPTORS ALONGSIDE THE SVM CLASSIFIER ON THE VAIHINGEN DATA SET REPRESENTED BY THE AVERAGE PRECISION AP AND THE INTERPOLATED AVERAGE PRECISION AP_{inter}

Descriptors	AP	AP_{inter}
Wind. Eval. + Grad. Des. + SVM	30.61%	17.41%
Wind. Eval. + Text. Des. + SVM	56.31%	28.39%
Wind. Eval. + Col. Des. + SVM	84.71%	63.65%
Wind. Eval. + Grad. & Text. Des. + SVM	50.45%	29.44%
Wind. Eval. + Grad. & Col. Des. + SVM	84.15%	63.32%
Wind. Eval. + Col. & Text. Des. + SVM	88.54%	65.35%
Proposed Framework	**92.79%**	**68.03%**

scenarios: 1) gradient descriptors were used alongside the window-evaluation stage and the SVM classifier; 2) texture descriptors were used alongside the window-evaluation stage and the SVM classifier; 3) color descriptors were used alongside the window-evaluation stage and the SVM classifier; 4) gradient and color descriptors were used alongside the window evaluation stage and the SVM classifier; 5) gradient and texture descriptors were used alongside the window evaluation stage and the SVM classifier; 6) color and texture descriptors were used alongside the window evaluation stage and the SVM classifier; and 7) the overall proposed car detection system.

The proposed framework is able to achieve an overall precision rate of 70.07% at a recall rate of 75%. It can be inferred that the three-step window filtering process has a significant effect in reducing the rate of false positives. More specifically, evaluating detection windows based on the variation of color distribution across the detection window results in a higher precision. Nevertheless, it has a slight effect on the recall rate as it can be observed from the decrease in the interpolated average precision AP_{inter}.

Changing the recommended values of the thresholds used in the window-evaluation stage by $\pm 4\%$ results in changing the average precision AP and the interpolated average precision AP_{inter} by less than $\pm 1\%$ indicating the robustness of the proposed window-evaluation stage to small variations in the acquisition conditions, e.g., changes in illumination. The recommended thresholds do not result in the highest average precision and interpolated average precision, as it can be inferred from Table II. Nonetheless, these thresholds were set using cross validation on the training data set to prevent overfitting as previously mentioned. It is also worth to note that it is important to have high values for both the average precision AP and the interpolated average precision AP_{inter} as this ensures having high precision and recall rates at the same time.

Combining the three window-evaluation steps with the proposed descriptors results in a high average precision AP and interpolated average precision AP_{inter}. Furthermore, it can be noticed from the performance comparison among different descriptors that the proposed color descriptor has the most significant contribution as compared with the gradient and texture descriptors.

2) Comparison With the State of the Art: In order to further validate the effectiveness of the proposed framework and its superiority to the state of the art, it was tested on the OIRDS data set. Its performance was compared with the state of the art using the well-known interpolated average precision AP_{inter} [36] as it takes into account the amount of ground-truth retrieval. The OIRDS data set has a wide range of image scenarios and it was used in several previous works. Table IV shows the achieved 11-point interpolated average precision by the proposed and baseline frameworks [15]–[17], [19]. A different number of test images of the OIRDS data set were chosen by each baseline framework. To overcome this problem, the evaluation of the proposed framework was carried out on the whole data set and on the subset that was used by Chen *et al.* [19]. We were not able to test the algorithm of [19] on the whole data set due to the unavailability of the source code. Clearly, the proposed algorithm outperforms the state of the art by achieving higher AP_{inter} on the OIRDS data set using only a linear classifier and one model as opposed to the nonlinear classifiers or the two models used in [15] and [17], despite the fact that this metric penalizes methods that retrieves only a subset of the ground truth [36]. It is worth to

IEEE Proof

TABLE IV

PERFORMANCE OF THE PROPOSED FRAMEWORK AND BASELINE FRAMEWORKS ON THE OIRDS DATA SET REPRESENTED BY THE INTERPOLATED AVERAGE PRECISION (AP_{inter})

Framework	Features	Classifier	Testing Criteria	AP_{inter}
Razakarivony and Jurie [15]	Gradients	Autoencoders and PCA manifolds	10 Folds ($\approx$ 900 images)	48.9%
ElMikaty and Stathaki [16]	Gradients, Spectrum, Colours	Linear SVM	One Fold ($\approx$ 900 images)	53.45%
Madhogaria *et al.* [17]	Gradients	Linear SVM and MRF	One Fold ($\approx$ 500 images)	54.45%
Chen *et al.* [19]	Gradients	Sparse Coding Based Dictionary	One Fold ($\approx$ 370 images)	62.15%
Proposed Framework	Gradients, Colours, Texture	Linear SVM	One Fold ($\approx$ 370 images)	**64.49%**
Proposed Framework	Gradients, Colours, Texture	Linear SVM	One Fold ($\approx$ 900 images)	57.34%

TABLE V

RUNNING-TIME ANALYSIS OF THE PROPOSED FRAMEWORK. THE REQUIRED COMPUTATIONAL TIME IS GIVEN BY A PERCENTAGE OF THE TOTAL TIME REQUIRED TO PROCESS A GIVEN IMAGE

Stage	Grad. Des.	Col. Des.	Text. Des.	Wind. Eval. & Other
Processing Time (%)	$\approx 5.09\%$	$\approx 11.35\%$	$\approx 69.99\%$	$\approx 13.57\%$

mention that the framework of [15] used a multiple-fold training/testing strategy, whereas the OIRDS data set was completely unseen to other baseline frameworks and the proposed framework. It is also worth to mention that the framework of [17] suffers from high-confidence score false positives. This makes it unreliable in some applications, such as in military fields.

C. Running-Time Analysis

The proposed framework has been implemented in a MATLAB environment and all of our experiments were run on an Intel-Core-i5 CPU with a speed of 2.5 GHz and a RAM of 6-GB size. Whilst the machine that was used has multiple processing cores, the current implementation uses only a single core. It is faster than the work of [16] by $\approx$5.5 times and is able to process $\approx$421 detection windows per second ($\approx$0.03 MB/s). This indicates that the current implementation has a near-real-time performance and is faster than the recent work of [19] that is able to process only $\approx$0.01 MB/s.

Table V shows the percentage of the processing-time consumed by each proposed descriptor. It can be inferred that the proposed color descriptor is very efficient in terms of its high discrimination power, as it was shown in Fig. 10 and Table III, and its relatively lower computational cost. The window-evaluation stage saves the required computational time by approximately 85.09% as it discards a huge number of candidate windows at an early stage.

VI. CONCLUSION

This paper presented a novel framework for the detection and localization of cars in high-resolution airborne imagery using an ensemble of image descriptors that depict the distribution of the gradients, colors, and texture of cars. The advantages of the proposed framework were investigated thoroughly and demonstrated clearly in its ability to achieve a higher precision compared with the state of the art. Several points can be learnt from the implementation and results of the proposed framework. First, eigenvalues of the covariance matrices of locally accumulated gradients can work effectively to identify the dominant orientation of a given detection window. Second, local normalization of the extracted feature vectors is essential for illumination invariance. Third, rotation-invariance properties can be achieved using the proposed adaptive cell distributions without altering the relative distribution of pixels. Fourth, the proposed color descriptor is a powerful single-cue descriptor. However, combining it with other descriptors results in better performance. Fifth, distinctive features of the foreground and background objects must be taken into account, when descriptors are chosen.

Furthermore, the proposed framework has the capacity to generalize as corroborated by testing on completely unseen data. The detection method used here can be further extended by adopting a multiscale analysis approach, e.g., the one adopted by Dalal and Triggs [11]. It has also been demonstrated that the use of an ensemble of image descriptors that efficiently show the visual characteristics of a particular object can lead to improved performance using low-dimensional feature vectors alongside a simple linear classifier.

ACKNOWLEDGMENT

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF).

REFERENCES

[1] V. Ramakrishnan, A. K. Prabhavathy, and J. Devishree, "A survey on vehicle detection techniques in aerial surveillance," *Int. J. Comput. Appl.*, vol. 55, no. 18, pp. 43–47, 2012.

[2] M. Liu, Y. Wu, P. Zhang, Q. Zhang, Y. Li, and M. Li, "SAR target configuration recognition using locality preserving property and Gaussian mixture distribution," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 268–272, Mar. 2013.

[3] X. Yuan, T. Tang, D. Xiang, Y. Li, and Y. Su, "Target recognition in sar imagery based on local gradient ratio pattern," *Int. J. Remote Sens.*, vol. 35, no. 3, pp. 857–870, 2014.

[4] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1. Jul. 2001, pp. 710–717.

[5] S. Hinz, "Detection and counting of cars in aerial images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Sep. 2003, pp. 997–1000.

[6] S. Hinz, C. Schlosser, and J. Reitberger, "Automatic car detection in high resolution urban scenes based on an adaptive 3D-model," in *Proc. 2nd GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion Over Urban Areas*, May 2003, pp. 167–171.

[7] J.-Y. Choi and Y.-K. Yang, "Vehicle detection from aerial images using local shape information," in *Advances in Image and Video Technology* (Lecture Notes in Computer Science), vol. 5414. Springer, 2009, pp. 227–236.

IEEE Proof

[8] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1197–1203.
[9] T. T. Nguyen, H. Grabner, H. Bischof, and B. Gruber, "On-line boosting for car detection from aerial images," in *Proc. IEEE Int. Conf. Res., Innov. Vis. Future*, Mar. 2007, pp. 87–95.
[10] P. Viola and M. J. Jones, "Robust real-time face detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. 2001, p. 747.
[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.
[12] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[13] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1250–1265, Jun. 2011.
[14] W. Shao, W. Yang, G. Liu, and J. Liu, "Car detection from high-resolution aerial imagery using multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2012, pp. 4379–4382.
[15] S. Razakarivony and F. Jurie, "Small target detection combining foreground and background manifolds," in *Proc. IAPR Int. Conf. Mach. Vis. Appl.*, 2013, p. 4.
[16] M. ElMikaty and T. Stathaki, "Car detection in high-resolution urban scenes using multiple image descriptors," in *Proc. IAPR Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 4299–4304.
[17] S. Madhogaria, P. M. Baggenstoss, M. Schikora, W. Koch, and D. Cremers, "Car detection by fusion of HOG and causal MRF," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 1, pp. 575–590, Jan. 2015.
[18] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[19] Z. Chen *et al.*, "Vehicle detection in high-resolution aerial images via sparse representation and superpixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 103–116, Jan. 2016.
[20] H. Moon, R. Chellappa, and A. Rosenfeld, "Optimal edge-based shape detection," *IEEE Trans. Image Process.*, vol. 11, no. 11, pp. 1209–1227, Nov. 2002.
[21] H. Moon, R. Chellappa, and A. Rosenfeld, "Performance analysis of a simple vehicle detection algorithm," *Image Vis. Comput.*, vol. 20, no. 1, pp. 1–13, Jan. 2002.
[22] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*, 3rd ed. Thomson Engineering, 2007.
[23] H. Zheng, L. Pan, and L. Li, "A morphological neural network approach for vehicle detection from high resolution satellite imagery," in *Neural Information Processing* (Lecture Notes in Computer Science), vol. 4233. Berlin, Germany: Springer, 2006, pp. 99–106.
[24] J. Serra, *Image Analysis and Mathematical Morphology*. San Diego, CA, USA: Academic, 1983.
[25] L. Eikvil, L. Aurdal, and H. Koren, "Classification-based vehicle detection in high-resolution satellite images," *ISPRS J. Photogram. Remote Sens.*, vol. 64, no. 1, pp. 65–72, Jan. 2009.
[26] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, "Airborne vehicle detection in dense urban areas using hog features and disparity maps," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2327–2337, Dec. 2013.
[27] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.
[28] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.
[29] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[30] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[31] L. Shen and Z. Ji, "Gabor wavelet selection and SVM classification for object recognition," *Acta Autom. Sinica*, vol. 35, no. 4, pp. 350–355, Apr. 2009.
[32] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 36, no. 7, pp. 1169–1179, Jul. 1988.
[33] M. Cramer, "The DGPF-test on digital airborne camera evaluation—Overview and test design," *Photogramm.-Fernerkundung-Geoinf.*, vol. 2, no. 2, pp. 73–82, 2010.
[34] F. Tanner *et al.*, "Overhead imagery research data set—An annotated data library & tools to aid in the development of computer vision algorithms," in *Proc. IEEE Workshop Appl. Imag. Pattern Recognit.*, Oct. 2009, pp. 1–8.
[35] V. Y. Mariano *et al.*, "Performance evaluation of object detection algorithms," in *Proc. IAPR Int. Conf. Pattern Recognit.*, vol. 3. Aug. 2002, pp. 965–969.
[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.



Mohamed ElMikaty (M'17) received the B.Sc. degree (Hons.) in electronics and communications engineering from Mansoura University, Mansoura, Egypt, in 2009, and the M.Sc. degree (Hons.) in communications and signal processing engineering and the Ph.D. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2012 and 2016, respectively.

He is currently a Research Associate with the Department of Electrical and Electronic Engineering, Imperial College London. He has been involved in various industrial research projects in Egypt and U.K. His research interests include computer vision and machine learning, with an emphasis on image representation, analysis, and understanding techniques with applications to object detection, recognition, and tracking.



Tania Stathaki (M'17) was born in Athens, Greece. She received the master's degree in electronics and computer engineering from the National Technical University of Athens, Athens, Greece, and the Ph.D. degree in signal processing from Imperial College London, London, U.K.

She was a Lecturer with the Department of Information Systems and Computing, Brunel University London, London, and an Assistant Professor with the Department of Technology Education and Digital Systems, University of Piraeus, Piraeus, Greece. She is currently a Reader with the Department of Electrical and Electronic Engineering, Imperial College London. She has intensive research experience in image processing and computer vision and, more specifically, image fusion, image registration, change detection, object detection and recognition, and object tracking. She has been mainly involved in defense and security applications and has collaborated with various U.K. companies, such as Dstl, General Dynamics, Selex Galileo, and BAE Systems. She is actively involved in various European Union and U.K. research projects. She has authored or co-authored many papers on signal and image processing and computer vision.

IEEE Proof