15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Detection of Cars in Complex Urban Areas

Mohamed ElMikaty Tania Stathaki
Department of Electrical and Electronic Engineering, Imperial College London
Exhibition Road, London SW7 2AZ, United Kingdom
{m.elmikaty, t.stathaki}@imperial.ac.uk

Abstract

Detection of cars in airborne images of typical urban areas has various applications in several domains, such as surveillance, military and remote sensing. It is a tremendously-challenging problem, mainly because of the significant inter-class similarity among various objects in urban environments. In this paper, a novel framework is introduced that adopts a sliding-window approach and it depicts, in a novel way, the local distribution of gradients, colours and texture. A linear support vector machine classifier is used to differentiate between descriptors that belong to cars and descriptors that belong to other objects in a hyperspace of 3838 dimensions. Descriptors are computed over a newly-proposed adaptive distribution of cells that enables the use of various rotation-variant image descriptors. The proposed framework has been evaluated on the Vaihingen dataset and results corroborate its superiority as it achieves a higher precision for a given recall than the state of the art.

1 Introduction

In airborne images of low ground sampling distance (GSD) of few centimetres, small targets, e.g., cars, shipping containers and aeroplanes, are usually depicted by a small number of pixels. Consequently, there is a significant inter-class similarity among different objects, posing a tremendous challenge to the task of detecting a particular target of interest. In this work, cars are considered the small targets of interest and the primary task is to detect and to locate them in aerial images of complex urban areas using a powerful image descriptor that is able to distinguish cars from other non-car objects that possess similar appearance, such as buildings' roofs and windows, whilst being robust to the huge intra-class variability among cars that stems from the significant differences in their appearance, as far as various features, such as colour intensities, size, shape and contours, are concerned.

Original contributions of this work include: (i) novel design of a method that evaluates the likelihood of windows to contain a target, (ii) new method to estimate accurately the dominant orientation of windows, (iii) introduction of adaptive cell distributions over which various rotation-variant 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 (iv) new encoding of the local distribution of complementary image cues as well as the associated eigenvalues of the covariance matrices of local descriptors in a robust single combined image descriptor.

2 Related Work

Several methodologies have been proposed to detect cars in airborne imagery. Early works [1], used explicit models, wherein templates that resemble cues of cars are designed. Windows are claimed to contain cars, when their cost of matching to the pre-designed templates is low. Problems of such models usually arise as a result of the associated high computational cost.

More recent works utilised implicit models, wherein image descriptors are engineered to capture different cues of cars. A pre-trained classifier is then used to discriminate descriptors that belong to a car. Different combinations of descriptors were proposed alongside several classifiers [2–4]. Although acceptable results of implicit models have been reported, their performance is significantly affected by the choice of features.

In order to restrict search areas to regions that are more likely to contain cars, road maps can be exploited [5]. Whilst this results in high precision rates, it is based on the prior knowledge of accurate road maps and on the use of a precise map-projection method.

Unlike flat models, deep models for the detection of cars can hardly be found in the literature as they require vast training datasets. Otherwise, models usually overfit the data. To the best of the authors' knowledge, only one work [6] was proposed, wherein a hierarchal deep model was used. Although high accuracy has been achieved, their testing dataset did not include typical complex urban areas.

3 Proposed Framework

Regarding a car in a given aerial image of low GSD, boundaries that correspond to the edges of its body and front and rear windscreens are expected to resemble rectilinear structures. These usually form together a considerable number of crucial points of interest. Nonetheless, they are incapable of discriminating cars in urban environments. This is because other non-car objects possess similar edge maps and they can be easily misidentified as cars. Based on this observation, implicit models could be superior to explicit models as they capture and encode fine details. Therefore, an implicit model and a linear classifier have been adopted. Three image descriptors are introduced to characterise the local distribution of gradients, colours and texture. These cues were chosen among others as cars possess: (i) almost rectilinear structures, (ii) distinctive colour distribution and (iii) mostly low amount of texture. The proposed framework comprises four stages, namely, window evaluation, extraction and encoding of features, classification and post-processing and it adopts a sliding grey-scale fixed-size window of 64×64 pixels with a stride length of four pixels.

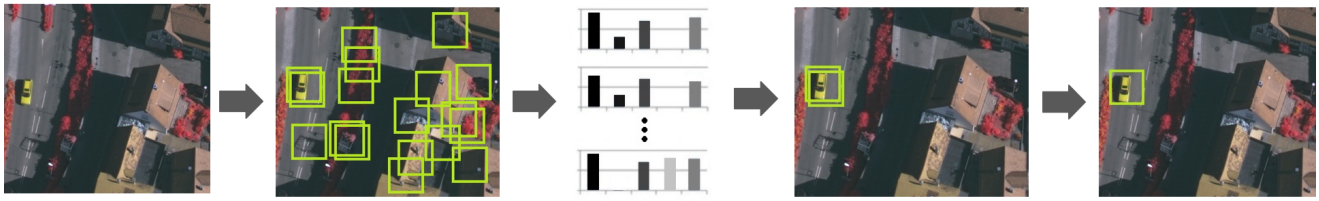
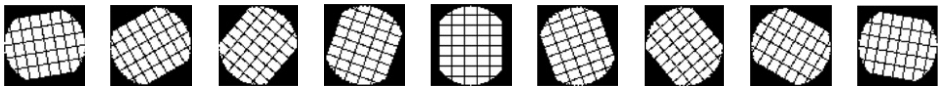


Figure 1: Processing a sample image using the proposed framework. From left to right: original image, processed image after evaluating windows, computed descriptors for each candidate window, output of the classification stage and final detection after post-processing.



(a) 10° (b) 30° (c) 50° (d) 70° (e) 90° (f) 110° (g) 130° (h) 150° (i) 170°

Figure 2: Adaptive distribution of cells over which image descriptors are computed at different orientations.

3.1 Window evaluation

In order to reduce search areas and the false-positive rate, the likelihood of windows to contain cars is investigated using a novel three-step evaluation process.

First, the variation in the distribution of grey levels across a window is examined by the computation of the Manhattan distance $r_{Manhattan} = \sum_{k=1}^{32} |\mathbf{h}_w(k) - \mathbf{h}_c(k)|$ between two 32-bin grey-level histograms $\mathbf{h}_w$ and $\mathbf{h}_c$, one for the whole window and the other for the 32×32 central square area. If distance is more than (2048), the window is discarded as this highly indicates that it belongs to an area with a constant or a slowly-varying colour distribution, such as empty roads or vegetation areas.

Second, the texture of the window is examined. Pixels of each cell of 4×4 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 of the computed differences is larger than a pre-defined threshold (32), the detection window will not be considered for further processing. The intuition behind this is that a high standard deviation indicates a highly-textured surface that cannot belong to a car.

Third, magnitudes of intensity gradients and their orientations are calculated. This is done by using a non-smoothing filter of [-1 0 1] vertically and horizontally. Then, magnitudes of gradients of cells of 4×4 pixels are accumulated independently in nine-bin histograms $\mathbf{g}_{cell}$ according to their unsigned orientation, followed by computing the covariance matrix $\mathbb{R}_{\mathbf{gg}} = \mathbb{E}\{\mathbf{g}\mathbf{g}^H\}$ of the $L2$-normalised responses $\mathbf{g}_{cell} \leftarrow \mathbf{g}_{cell}/\sqrt{\|\mathbf{g}_{cell}\|_2^2 + \epsilon^2}$ from local cells, where ϵ is equal to 0.01. For a typical detection window containing a car, there should be a dominant orientation of the gradients. Therefore, if the computed covariance matrix is eigen-decomposed, the maximum eigenvalue must exceed a pre-defined threshold (0.09), otherwise the window is discarded. The intuition behind this is that the second order statistics given by the covariance matrix can be used to determine the existence of a dominant orientation using the rank of the covariance matrix and the value of its maximum eigenvalue.

Estimation of the orientations of cars: The orientation of cars is discretised into equally-spaced nine orientations between 0° and 180°. It is estimated to be in a perpendicular orientation to the bin orientation corresponding to the maximum eigenvalue of the covariance matrix $\mathbb{R}_{\mathbf{gg}}$ of locally-accumulated gradients computed over 4×4 cells.

3.2 Extraction and encoding of features

In this work, not only do we introduce a new ensemble of image descriptors, but also a new method to derive and compute descriptors. Traditionally, detection windows are divided into square cells of equal size. Image descriptors are then computed over these cells. For rotation-variant features, windows are rotated in most cases so that they would have a specific orientation. However, this rotation process is time consuming and for small targets, it affects the relative distribution of the original pixels due to the non-linear transformation between co-ordinates. To overcome this problem, adaptive cell distributions are proposed, wherein the shape of cells is no longer fixed-size squares. According to the estimated dominant orientation of a window, a suitable cell distribution is chosen. Two sets of cell distributions are used in the proposed framework. 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 and they are shown in Figure 2. In the second set, distributions of cells are originated in the same way but using a mask of size 64×52 pixels 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 centre of the detection window.

Gradient descriptors: Gradient descriptors are exploited to encode the rectilinear structures of cars. They are computed over the first set of the cell distributions based on the estimated dominant orientation of the window. Magnitudes of gradients of local cells are accumulated in nine-bin histograms according to their orientation. The response of each cell is $L2$-normalised so that illumination invariance proper-

ties can be achieved. The final gradient descriptor is formed from the concatenation of the $L2$-normalised eigenvalues of the covariance matrix of the cell responses of each block of four adjacent cells. This differs from the original implementation of [7] 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 local histograms of gradients.

Colour descriptors: Cars exhibit a distinctive distribution of colours, the colour of windscreens is always dark in aerial images and the colour of the car boot is usually the same as the colour of the car bonnet. To capture this property, a new modified version of the colour descriptor proposed by [4] is introduced. This new colour descriptor $\mathbf{c}_{cell}$ depends on the calculation of the similarity among adjacent cells by computing the intersection between corresponding colour-histogram bins and it incorporates the principal components of the covariance matrices of the local cell responses.

$$\mathbf{c}_{cell} = [\mathbf{h}_c \cap \mathbf{h}_n] = \left\lceil \left[\sum_k \min[\mathbf{h}_c(k), \mathbf{h}_n(k)] \right] \right\rceil \quad (1)$$

where $\mathbf{h}_c$ and $\mathbf{h}_n$ are the two grey-level histograms of the central cell and a neighbouring cell, respectively, each has 16 bins, $\mathbf{h}_c(k)$ and $\mathbf{h}_n(k)$ are scores of the k^{th} 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 [4]. Local descriptors are computed on two levels in a pyramid approach using the two sets of adaptive cell distributions. The final colour descriptor is formed by concatenating the $L2$-normalised local responses and eigenvalues of the local covariance matrices of the colour histograms in each level of the pyramid.

Texture descriptors: 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 [8]. Although cars do not possess distinguishable texture, Gabor descriptors characterise textured areas, such as those of buildings' roofs and vegetation. Gabor filter kernels can be defined as:

$$\mathbf{F}_{Gabor}(x, y; \lambda, \theta, \psi, \sigma_g, \gamma) = exp\left(\frac{x'^2 + \gamma^2 y'^2}{2\sigma_g^2}\right) exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \quad (2)$$

$$\begin{cases} x' = x\cos\theta + y\sin\theta \\ y' = -x\sin\theta + y\cos\theta \end{cases}$$

In this work, nine Gabor kernels of size 64×64 are used, with a wavelength λ of 10 pixels, nine orientations $\theta_k = [10°, 30°, ..., 170°]$, 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 detection windows using those nine Gabor kernels, windows are divided into one of the first set of the cell distributions according to the estimated orientation of the window. The mean and energy of the response to the Gabor filters in each cell are computed at each orientation. For rotation-invariance, the order of the nine orientations is changed such that the rotation-invariance property is achieved by circularly shifting the response at different orientations θ_k. The final texture descriptor is formed from the concatenation of the $L2$-normalised eigenvalues of the covariance matrices of the cell responses of each block of four cells.

3.3 Classification

A linear support vector machine (SVM) classifier [9] has been exploited with a regularisation parameter of 15 to distinguish aforementioned ensemble of descriptors that belongs to cars. The overall dimension of this ensemble is $3838D$ (216 gradient features, 3190 colour features and 432 texture features).

3.4 Post-processing

Detection windows with positive score that overlap by more than 50% are eliminated except the one with the highest output confidence score in order to keep only a single true detection.

All values of the aforementioned parameters were set empirically by excessive testing on the training dataset using cross validation.

4 Datasets

Training dataset: From the training areas of the Vaihingen dataset [10] that has a GSD of 8cm, 781 positive samples that contain cars were chosen. Patches were chosen such that a great diversity in the orientation of cars and backgrounds could be achieved. Data augmentation techniques were exploited by including mirrored and horizontally-flipped versions. 2473 negative samples that included patches from different backgrounds and bootstrapped hard negatives were used.

Testing dataset: The proposed framework was run on the testing areas of the Vaihingen dataset that include various urban scenes with different structures.

5 Experimental Results

Qualitative analysis: Sample visual results of applying the proposed framework are shown in Figure 3. Our framework is able to robustly detect cars in various environments. Thanks to the local normalisation of descriptors, the proposed method is able to correctly classify detection windows that have poor contrast as shown in the bottom image of Figure 3 (fourth car from top), a case which is even hard for the human eye.

Quantitative analysis: Precision and recall curves are among the standard metrics to evaluate the performance of target-detection systems [11]. 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 (distance from the separating hyperplane) [11]. Figure 4 shows that the proposed framework achieves higher precision at a given recall rate using only linear classification and a lower dimensionality than the works in [3, 4], which outperformed traditional HOG-SVM frameworks. The accuracy of estimating the dominant orientation is 98.34%.

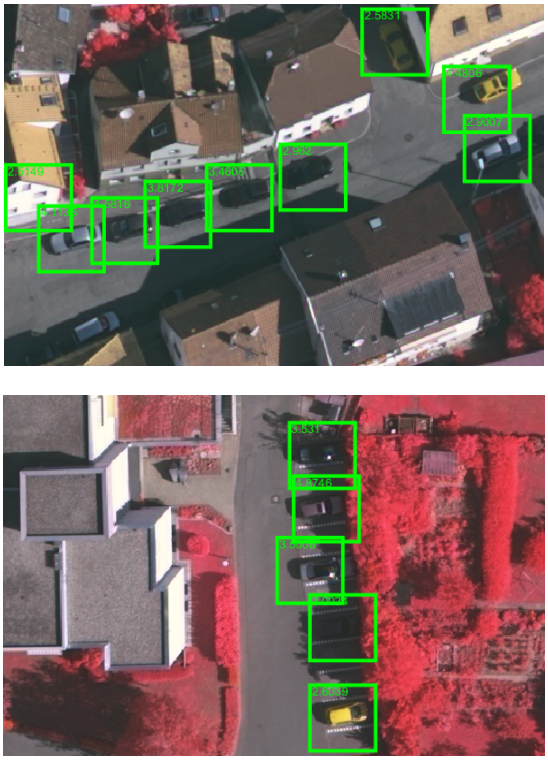


Figure 3: Sample regions of the test images of the Vaihingen dataset after applying the proposed framework.

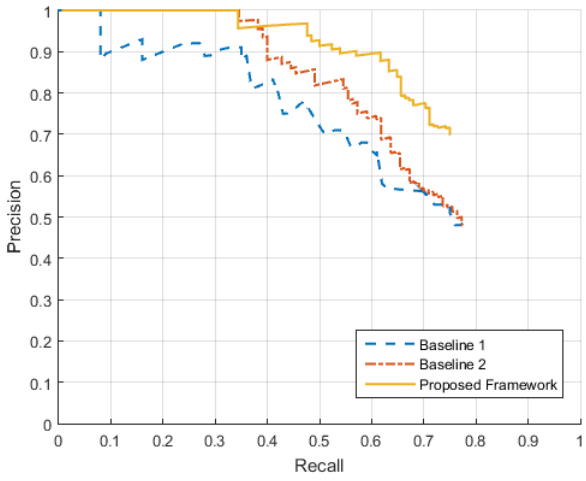


Figure 4: Performance of the proposed framework and baseline methods on the Vaihingen dataset. Baseline 1 [3] exploits $6760D$ and an IKSVM, baseline 2 [4] exploits $4996D$ and a linear SVM, whereas the proposed framework exploits $3838D$ and a linear SVM,

Running-time analysis: The proposed framework has been implemented in a MATLAB environment on a 2.5 GHz Intel-Core-i5 CPU with a RAM of 6.00 GB size. Using only a single core, 421 windows can be processed per second.

6 Conclusion

This paper presented a novel framework for the detection and localisation of cars in high-resolution airborne imagery using an ensemble of image descriptors that depicts the distribution of gradients, colours and texture of cars. The advantages of the proposed framework were investigated thoroughly and demonstrated clearly in its ability to achieve a higher precision than the state of the art. In addition, it can be learnt that eigenvalues of the local covariance matrices can work effectively to identify the dominant orientation of a given window. Rotation-invariance properties can be achieved using the proposed adaptive distribution of cells. Furthermore, it has been demonstrated that the use of image descriptors that efficiently depict the visual characteristics of a particular target of interest can lead to a significantly improved performance using a linear classifier.

Acknowledgements: This work has been funded by the Engineering and Physical Sciences Research Council (EPSRC) and BAE SYSTEMS Military Air and Information (UK). The Vaihingen dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF).

References

[1] T. Zhao and R. Nevatia, "Car detection in low resolution aerial image," in *IEEE International Conference on Computer Vision*, 2001, vol. 1, pp. 710–717.

[2] A. Kembhavi, *et al.*, "Vehicle detection using partial least squares," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1250–1265, 2011.

[3] W. Shao, *et al.*, "Car detection from high-resolution aerial imagery using multiple features," in *IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 4379–4382.

[4] M. ElMikaty and T. Stathaki, "Car detection in high-resolution urban scenes using multiple image descriptors," in *IAPR International Conference on Pattern Recognition*, 2014, pp. 4299–4304.

[5] S. Tuermer, *et al.*, "Airborne vehicle detection in dense urban areas using hog features and disparity maps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, , no. 99, pp. 1–11, 2013.

[6] X. Chen, *et al.*, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 10, pp. 1797–1801, 2014.

[7] N. Dalal and B.Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[8] J.G. Daugman, "Complete discrete 2-D gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[10] M. Cramer, "The DGPF-test on digital airborne camera evaluation-overview and test design," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2, no. 2, pp. 73–82, 2010.

[11] M. Everingham, *et al.*, "The PASCAL Visual Object Classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.