Context-based Object-of-Interest Detection for a Generic Traffic Surveillance Analysis System

Xinfeng Bao[1], Solmaz Javanbakhti[1], Svitlana Zinger[1], Rob Wijnhoven[2] and Peter H. N. de With[1]

[1]Video Coding and Architectures Research Group (SPS-VCA), Electrical Engineering Faculty, Eindhoven University of Technology, 5600 MB Eindhoven, the Netherlands,

{xbao,s.javanbakhti,s.zinger,p.h.n.de.with}@tue.nl

[2]ViNotion B.V., Horsten 1, 5600 CH Eindhoven, the Netherlands

Abstract

We present a new traffic surveillance video analysis system, focusing on building a framework with robust and generic techniques, based on both scene understanding and moving object-of-interest detection. Since traffic surveillance is widely applied, we want to design a single system that can be reused for various traffic surveillance applications. Scene understanding provides contextual information, which improves object detection and can be further used for other applications in a traffic surveillance system. Our framework consists of two main stages: Semantic Hypothesis Generation (SHG) and Context-Based Hypothesis Verification (CBHV). In the SHG stage, a semantic region labeling engine and an appearance-based detector jointly generate the visual regions with specific features or of specific interests. The regions may also contain objects of interest, either moving or static. In the CBHV stage, a cascaded verification is performed to refine the results and smooth the detection by temporal filtering. We model the context by jointly considering spatial and scale constraints and motion saliency. Our proposed framework is validated on real-life road surveillance videos, in which objects-of-interest are moving vehicles. The results of the obtained vehicle detection outperform a recent object detection algorithm, in both precision (92.7%) and recall (92.0%). The framework is both conceptually and in the applied techniques of a generic nature and can be reused in various traffic surveillance applications, that operate, e.g. on a road crossing or in a harbor.

1. Introduction

Video-based traffic surveillance systems play a crucial role in traffic monitoring and management. Compared to other sensing techniques such as radar and lidar, video-based systems are easier to operate and significantly less expensive to install and maintain. Increasingly, cameras used in video-based systems offer high-resolution image capturing that enable automatic visual information analysis suited for semantic interpretation. This initiates advanced applications, such as traffic flow analysis, vehicle feature data retrieval, detection and understanding of abnormal events, etc.

In particular, the detection of moving objects-of-interest is an important research topic in computer vision for traffic surveillance. This importance is motivated by that the properties, the number and locations of moving objects are fundamental to semantic interpretation of traffic information. Such objects of interest can be defined depending on the scenario of surveillance, e.g. vehicles in road surveillance or vessels in port surveillance. In the past decades, significant research concentrated on object detection in the domain of traffic surveillance. Generally, the state-of-the-art approaches can be divided into two main categories: appearance-based and motion-based [10].

For appearance-based techniques, Haar-like and Histogram of Oriented Gradients (HOG) features are frequently applied. Hota *et al.* [6] have created a cascaded classifier combining these two features. A new set of image strip features is proposed by Zheng *et al.* [15] to model the structural characteristics of cars within a boosting framework. Robert [9] generates the hypothesis utilizing the pair of headlights as representative appearance for front/rear-view vehicles and trains so-called "Eigenvehicles" to verify the results. However, the performance of these methods decreases dramatically if the object to be detected is partially occluded, or when its appearance significantly differs from the model.

For motion-based detection techniques, a concept of

successive image differencing or background substraction is often used to find the moving blobs in consecutive frames [11, 13]. More complex techniques, such as background modeling [2, 4], are also explored in traffic surveillance. These techniques typically employ motion features and extract binary "blobs" at a very early stage, thus the merging and separation of blobs are difficult steps due to the lack of features. Another drawback of these methods is that they are sensitive to sudden changes in the background, e.g. lighting changes.

The algorithms discussed are mostly designed and implemented in a dedicated form for a particular traffic scenario: for example, they focus either on road surveillance or on port surveillance. Therefore, these dedicated algorithms lack generic characteristics, which hampers their re-use for object-of-interest detection independently of the domain. There are only a few research papers aiming at generic object detection [5, 14], which discuss typical objects (vehicles, pedestrians, etc.) in traffic surveillance. However, the proposed algorithms target a broad scope of objects and fail to emphasize the characteristics of objects that are specifically important for a traffic surveillance system. Particularly, they do not consider the motion features and contextual information in traffic surveillance, which is known to add robustness and semantic knowledge of the scene.

The use of context information is proven to be attractive in image and video analysis [8]. The context information is based on exploiting spatial and temporal *co-occurrence consistency* between different objects or between an object and its surroundings. This consistency can be used to reduce ambiguities of local appearances of different objects for improved detection and robustness performance. For real-time applications, it is important to avoid complex and therefore computationally expensive object detectors, especially when each object class requires a dedicated detector. Instead, we aim at employing context modeling in a generic framework to allow simpler detection algorithms.

In this paper, we propose a context-based generic framework combining scene understanding and detection of objects-of-interest for a traffic surveillance system. The framework is inspired by a ship detection algorithm [3] and conceptually further developed for completeness and generalization. We present 3 major contributions in our work. First, scene understanding is performed simultaneously with the object-of-interest detection, which is used for modeling context to improve the object detection performance. As a refinement, our context modeling can simultaneously be applied to further semantically analyze the image and add to semantic scene understanding in a surveillance system. Second, the designed framework is conceptually generic and can be applied to various surveillance scenarios in land transportation or waterway traffic. Third, although our framework is generalized, it can still handle challenging detection tasks, such as view-independent object detection and occlusion handling, because we are not fully dependent on appearance modeling.

This paper is organized as follows. In Section 2, we give a short overview of our framework. In Sections 3 and 4, we describe our algorithms in detail. In Section 5, we present the implementation of our framework on real-life road traffic surveillance and experimental results are compared with a recent object detection approach. Finally, in Section 6, conclusions are discussed.

2. Overview of Our Framework

Figure 1 illustrates the proposed framework. It consists of two cascaded main stages: (1) Semantic Hypothesis Generation (SHG); (2) Context-Based Hypothesis Verification (CBHV). The SHG stage aims at locating two groups of the candidate objects of interest as well as understanding the scene in the traffic surveillance video. Firstly, a semantic region labeling is performed to divide the video frame into labeled segments, such as road, vegetation, etc. Those segments are then grouped into regions according to spatial and motion similarities. Therefore, the scene is depicted as a set of semantic regions and the first group of candidate objects is located simultaneously. Meanwhile, a simple appearance-based detector is trained off-line and applied to the same frame to locate the second group of regions possibly containing objects of interest. The CBHV stage targets verifying the presence of such objects through a cascaded verification process. The context is first modeled and applied to remove the false detections. Then, we verify whether the motion of the obtained candidates is salient compared to their local surroundings. Finally, the detection results are merged and a temporal filter is applied to increase the robustness of the detection results. Each algorithmic step in the framework will be discussed into detail in the following sections. In Section 3, we present the details of Semantic Hypothesis Generation stage. In Section 4, we describe the Context-Based Hypothesis Verification stage.

3. Semantic Hypothesis Generation

3.1. Scene Understanding

The traffic scene is analyzed first, where a semantic region labeling is applied to divide the frame into meaningful segments followed by a region merging. The scene understanding sub-system provides two types of semantic information: (1) labeled regions in the surveillance video frame, characterized by distinguished color, texture and motion features; (2) locations of candidate objects.

3.1.1 Semantic Region Labeling

The region labeling stage explores the local features (color and texture) as well as global features [7]. In the frame-

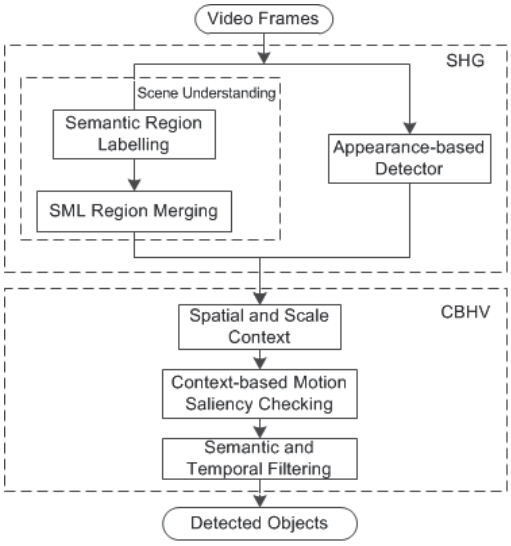

Figure 1. Framework of generic context-based traffic surveillance video analysis.

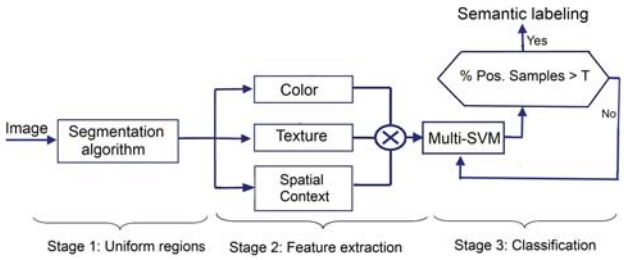

Figure 2. Semantic region labeling approach.

work, spatial location in the image is considered as a global feature, which reduces the ambiguities in contextual information from smooth regions, such as sky and water. The approach consists of three steps as depicted in Figure 2.

Step 1: Uniform regions. A graph-based segmentation is implemented to divide the frame into uniform patches based on color features.

Step 2: Feature extraction. Pixel-based features (HSV color space and a group of Gabor features) of each segment are extracted. Regarding global feature extraction, the normalized vertical position for each pixel is calculated, forming the spatial context.

Step 3: Classification. A set of specific Support Vector Machines (SVMs) is trained off-line for different classes. For each class, we measure the percentage of pixels classified as belonging to this class, in a given region. We then assign a specific label to a region when its percentage of positively classified pixels is above a threshold [7]. Considering the scenario of traffic surveillance, we train 5 classifiers: sky, vegetation, construction/buildings, road and water. All areas that do not belong to one of these classes are labeled as "unknown".

3.1.2 Region Merging

To interpret the scene in a traffic surveillance video, motion features are explored to merge the labeled segments into semantic regions. We assume that segments from the same object should have similar motion patterns, which distinguish them from segments belonging to other objects. Therefore, based on spatial adjacency, motion similarity and segment labels, we employ a reliable merging process to form consistent semantic regions [3]. Pixel-level motion is first calculated by optical flow, which is used in measuring motion similarity as well as in the later stages. Merging of two segments is performed if they: (1) are spatially adjacent; (2) have statistically similar region-level motion; and (3) have the same label.

After the region merging, the regions labeled as "unknown" in the frame are extracted as first group of candidate objects of interest.

3.2. Appearance-based Detector

An appearance-based detection is performed in parallel with the scene understanding (Figure 1). Haar-like features are used in our framework because they are more easily computed than HOG features. The boosted classifier is used for training the detector, considering this weak classifier is faster, compared to SVM. Since the designed framework should not emphasize appearance-based detection, we do not employ strong classifiers composed of HOG features and SVM. Instead, a cascade of weak, boosted classifiers based on Haar-like features, is applied to obtain real-time performance [12]. The stages of the cascade are limited to a small number for lower computational complexity. Three types of Haar-like features [12] are used in our system: a twofold-rectangle feature, a threefold-rectangle feature and a fourfold-rectangle feature. Figure 3 shows the visual properties of these features, the white and black regions are adjacent and of the same size and shape. The output value of each feature represents the difference between the accumulated sum of pixel intensities in the white region(s) and the accumulated sum of pixel intensities in the black region(s).

We employ the AdaBoost algorithm to train the detector

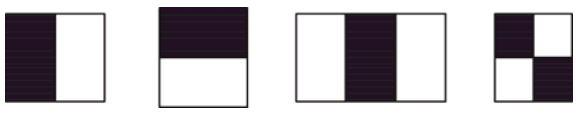

Figure 3. Haar-like features used in our appearance-based detector.

based on the above Haar-like features. The detector is a cascade of weak classifiers with increasing complexity in consecutive stages.

By applying the appearance-based detector, we generate detection windows possibly containing objects of interest, where the second group of candidates is located.

4. Context-based Hypothesis Verification

In this stage, the previously generated hypothesis is verified in order to remove the false positives. It is achieved through a cascaded context-based hypothesis verification by modeling the context, checking the motion saliency and applying semantic and temporal filters.

4.1. Spatial and Scale Context

Based on the outputs of scene understanding, we model the context of the scene by considering two aspects.
Aspect 1: Spatial Context Extraction. In a specific traffic surveillance scene, an object-of-interest is supposed to travel inside the traffic region, e.g. vehicles travel in/on a road region and ships travel inside the water region. Therefore, we assume that regions containing candidate objects should be surrounded by traffic regions or at least have common borders with them. The spatial context of the scene is modeled accordingly, which defines the region R containing a candidate R_{cand} if it satisfies the following condition: $R_{cand} \cap R_T \neq \oslash$, where R_T represents a traffic region.
Aspect 2: Scale Context Extraction. The size of each candidate object should fit into within a feasible interval which provides scale context. The number of pixels in R_{cand}, denoted by $|R_{cand}|$, should satisfy the following inequalities : $S_{min} < |R_{cand}| < w \times |R_T|$, where $|R_T|$ denotes the number of pixels in the traffic region. Parameter S_{min} indicates the smallest size of the candidate object. And weight w is always below unity, as it defines the allowed ration fraction between $|R_T|$ and $|R_{cand}|$. This implies that the upper bound of the candidate object is dependent on the size of the traffic scene, which is tunable according to the camera settings. Both S_{min} and w are set according to the application scenarios.

The modeled context is applied to filter the two groups of candidates from the previous stage.

4.2. Context-based Motion Saliency Checking

In traffic surveillance, we aim at detecting moving objects, which implies that they have salient motion compared to their surroundings. We examine the motion saliency of the candidates which remains after the filtering described in Section 4.1. We determine motion saliency at the region level to avoid expensive pixel-based saliency checking. We remove the false positives from our detections by context-based motion saliency checking, as described below.
Step 1: ROI (Region-Of-Interest) Extraction. We extract the ROI including the outer part R_{obj} of a candidate and its local background R_{bg} as defined in [3]. We consider only the outer part of an object, since the inner parts of objects are typically appear with the same color.
Step 2: Motion Calculation. Using the pixel-level motion obtained in the SHG stage, we calculate the region-level motion of R_{obj} as $\mathbf{v}_{obj}$ and of R_{bg} as $\mathbf{v}_{bg}$. We define global motion as $\mathbf{v}_T$ which is the average motion in $|R_T|$. For improved stability, the *relative* motion $\mathbf{rv}_{obj}$ of R_{obj} is defined by $\mathbf{rv}_{obj} = \mathbf{v}_{obj} - \mathbf{v}_T$. Similarly, $\mathbf{rv}_{bg}$ of R_{bg} is defined.
Step 3: Motion Saliency. We define two criteria to verify the candidate: $\frac{|\mathbf{rv}_{obj} - \mathbf{rv}_{bg}|}{|\mathbf{rv}_{obj}|} > T_1$ and $|\mathbf{rv}_{obj}| - |\mathbf{rv}_{bg}| > T_2$. In the first criterion, we remove/filter false positives whose motion contrast with the surroundings is not significant (e.g. traffic signs). In the second criterion, we remove false detections (e.g. swaying flags) with small distracting motion in a static background region. We set T_1 and T_2 according to the application.

4.3. Semantic and Temporal Filtering

Since we obtain two groups of candidates in the SHG stage, multiple detections are possibly generated for the same object. To remove the redundancy and refine the detections, a blob-map is created by imposing the spatial context to the first group of candidates. Multiple detections of the same object are merged by redefining the bounding box through averaging, if they are overlapping with the same blob.

The detections are further improved by applying a temporal filter based on the assumption that a moving object cannot disappear suddenly from the scene. For each detection in the previous frame, we search a pre-defined neighboring area for detections in the current frame. If no detection is found in the search area, the previous detection is propagated to the current frame. The recovered detection is re-verified through the verification stages in Sections 4.1 and 4.2. By simple reasoning and re-using the already explored context and motion, complex detection-by-tracking algorithms are avoided for smoothing the detections.

5. Experimental results

To validate our context-based traffic surveillance analysis framework, we consider a road traffic scene and on-road vehicles as an example. All videos have a resolution of 1280×960 pixels and are captured during daytime including sunny and cloudy weather. For scene understanding, we construct our own training set which consists of images from multiple data sets from the Internet and a number of scenes with road and vehicles, in total 173 images. The semantic region labeling aims at annotating 2 classes (vegetation, road) plus the class "unknown". For training the appearance-based detectors, we have chosen 100 images from the "cars 2001(Rear), Caltech" dataset. The training set is intentionally chosen to be limited to a single view with small number of training images, in order to evaluate the system performance when an imperfect appearance-based detector is applied. The base resolution of the detector is 40×40 pixels, consisting of a cascade with 10 stages.

Figure 4 visually demonstrates the processing flow of our system with outputs from major sub-stages. To evaluate the performance of vehicle detection in our framework, we compare our approach ("Context-Generic") with a state-of-the-art detection algorithm "SVM-HOG" [14]. For both the "Haar-AdaBoost" detector in our framework and the "SVM-HOG" detector we set the thresholds to actively detect vehicles, leading to many false alarms. We test the two approaches on two sequences, containing 150 frames and 200 frames. The test sequences contain various types of vehicles, including passenger cars, vans and trucks. The test video contains various scenarios involving multiple vehicles, occluded vehicles and/or turning vehicles occurring in the surveillance cases. To prove the advantages of including scene understanding and the CBHV stage within our framework, we provide in Table 1 the following detection results, based on various configurations: (1) our framework "Context-Generic"; (2) our "Context-Generic" with "SVM-HOG" instead of SHG (we call it "SVM-HOG-CBHV"); (3) only using "SVM-HOG"; (4) only using "Haar-AdaBoost".

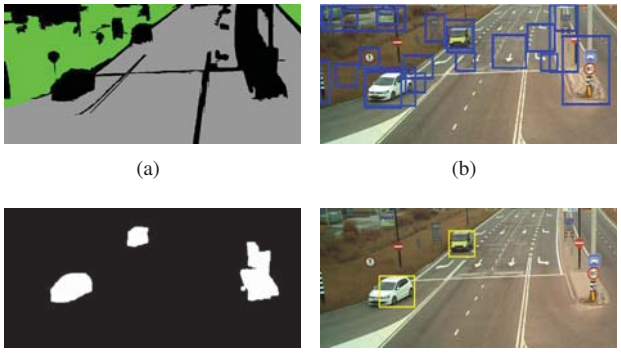

Figure 4. Flow of our system: (a) results of semantic region labeling; (b) candidate vehicles from the SHG stage indicated by blue bounding boxes; (c) blob-map described in Section 4.3; (d) final detection results (bounding boxes indicating false alarms are removed and overlapped bounding boxes are merged) indicated by yellow bounding boxes.

When CBHV is not applied, both of the two appearance-based detectors have poor performances, because the training set is deliberately chosen to be small with rear-view only. When CBHV is combined with 'SVM-HOG", the results show an increase in precision, but the recall drops. It means that the 'SVM-HOG-CBHV" has a high precision but misses nearly 30% of cars in the surveillance videos. On the other hand, our "Context-Generic" approach achieves a precision of 92.7% and a recall of 92.0%, which shows that the system can detect most of the vehicles with high precision. Although the integration of "SVM-HOG" can further improve the performance, we prefer lowering the computation requirements of the system, by employing the more efficient "Haar-AdaBoost" detector.

Table 1. Vehicle detection results. TP+FN = ground truth, TP+FP = detected vehicles, TP = correctly detected vehicles.

Methods	TP+FN	TP+FP	TP	Precision (%)	Recall (%)
Context-Generic	637	632	586	92.7	92.0
SVM-HOG-CBHV	637	478	461	96.4	72.4
SVM-HOG	637	901	502	55.8	78.8
Haar-Adaboost	637	2025	472	23.3	74.1

Table 2. Vehicle detection results on challenging scenarios.

Methods	TP+FN	TP+FP	TP	Precision (%)	Recall (%)
Context-Generic	244	253	221	87.4	90.6
SVM-HOG-HV	244	167	153	91.6	62.7

To further analyze our approach, we also apply "Context-Generic" system and "SVM-HOG" on a subset of the test sequences, where two challenging scenarios are addressed: (1) vehicles partially present in the video (occluded or partly out of the camera filed of view); (2) vehicles are turning (multiple views of vehicles present). The results are shown in Table 2. Our "Context-Generic" framework still keeps good performance in these difficult scenarios. Since there a number of vehicles clearly different from the training set, the performance of "SVM-HOG" dramatically decreases in recall value. Figure 5 provides a visual comparison of our "Context-Generic" system and "SVM-HOG". In this figure, the images at the left column show the results of our system, and the images at the right column illustrate the results of "SVM-HOG". In Figure 5(a), the "SVM-HOG" erroneously detects several traffic signs. In Figures 5(b) and 5(c), the turning cars and occluded cars are missed in the "SVM-HOG" results. In contrast, our "Context-Generic" shows successful detections in all three scenarios. The ex-

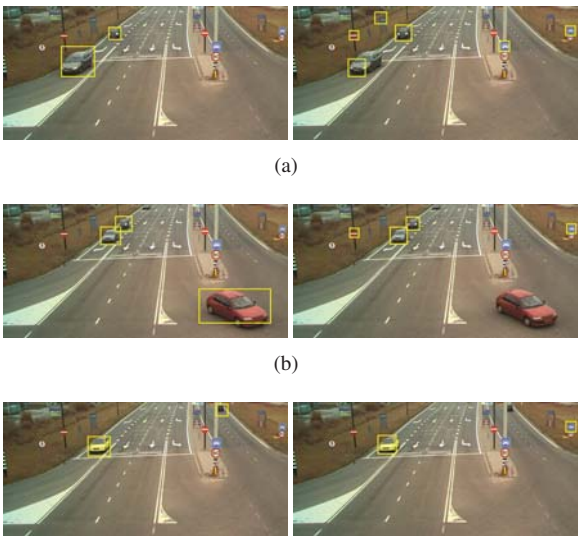

Figure 5. Detection results of two approaches in different scenarios, where left column shows results of our "Context-Generic" and right column shows "SVM-HOG": (a) multiple vehicles present; (b) a vehicle is turning (views are changing); (c) a vehicle is occluded by a lamp post.

periments show that the strong appearance-based detector "SVM-HOG" is not robust when the training set contains an insufficient number of training samples and limited views of the objects. The reason for this is that the pure appearance-based detector is intrinsically error-prone if the vehicles in the testing set differ from the training set. Since both the context and motion are explored, with our framework we obtain rich information for improving the vehicle detection in a surveillance system.

6. Conclusions

In this paper, we have presented a generic framework for traffic surveillance analysis, which combines scene understanding and detection of objects-of-interest. Semantic region labeling is performed to understand the scene of a surveillance video. The labeled regions are used to locate the candidate objects as well as extracting context information to improve the efficiency and reliability of the detection. Motion context is also well explored at pixel, segment and region levels. The motion analysis ensures not only a higher accuracy, but also improved robustness in analyzing the surveillance video. We have validated and applied our framework to various traffic scenarios. Compared to the well-known "SVM-HOG" detector, the total system provides an increase in detection performance from 55.8% to 92.7% in precision and from 78.8% to 92.0% in recall.

Besides the above performance growth, our framework offers further attractive key features for deployment. (1) *Re-use*: it is designed to be generic and can be applied to various surveillance scenarios. (2) *Context information to balance complexity*: semantic region labeling enables scene understanding, which explores the context for simplified and generalized appearance and motion modeling. This allows us to use an imperfect object detector that only detects a subset of the present objects; false detections are removed by the verification stage. (3) *Partial scene understanding*: the context information provides rich information for other semantic analysis if the framework is functionally extended. (4) *Robust detection*: it can handle challenging detection tasks like multi-view and occluded object detection.

The importance of the appearance-based detector is not emphasized in our framework, which requires less human efforts to train detectors. This makes the system easily configurable for other objects (vessels, pedestrians, etc.) in different traffic scenes. Future work will concentrate on introducing the motion context in early stage of the framework, so that the region labeling can also benefit from it. In the meanwhile, extensive tests will be performed, including tests on different traffic scenes and various lighting conditions. Comparisons with state-of-the-art algorithms will be also conducted using benchmark datasets [1].

References

[1] i-lids bag and vehicle detection challenge for AVSS 2007. http://www.eecs.qmul.ac.uk/ andrea/avss2007, 2007.

[2] N. Arshad, K. Moon, and J. Kim. Multiple ship detection and tracking using background registration and morphological operations. In *Signal Processing and Multimedia*, volume 123, pages 121–126. 2010.

[3] X. Bao, S. Javanbakhti, S. Zinger, R. Wijnhoven, and P. H. N. de With. Context modeling combined with motion analysis for moving ship detection in port surveillance. *Journal of Electronic Imaging*, 22:041114, 2013.

[4] A. Broggi, A. Cappalunga, S. Cattani, and P. Zani. Lateral vehicles detection using monocular high resolution cameras on terramax. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 1143–1148, 2008.

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8, 2008.

[6] R. N. Hota, K. Jonna, and P. R. Krishna. On-road vehicle detection by cascaded classifiers. In *Proceedings of the Third Annual ACM Bangalore Conference*, COMPUTE '10, pages 27:1–27:5, 2010.

[7] S. Javanbakhti, S. Zinger, and P. H. N. de With. Context-based region labeling for event detection in surveillance video. In *International Conference on Information Science, Electronics and Electrical Engineering (ISEEE 2014)*, 2014.

[8] O. Marques, E. Barenholtz, and V. Charvillat. Context modeling in computer vision: techniques, implications, and applications. *Multimedia Tools and Applications*, 51(1):303–339, 2011.

[9] K. Robert. Night-time traffic surveillance: A robust framework for multi-vehicle detection, classification and tracking. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)*, pages 1–6, 2009.

[10] S. Sivaraman and M. M. Trivedi. A review of recent developments in vision-based vehicle detection. In *IEEE Syposium on Intelligent Vehicles (IV 2013)*, pages 310–315, 2013.

[11] B. Tamersoy and J. K. Aggarwal. Robust vehicle detection for tracking in highway surveillance videos using unsupervised learning. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)*, pages 529–534, 2009.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001).*, volume 1, pages I–511–I–518, 2001.

[13] Y. K. Wang and S. Chen. A robust vehicle detection approach. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2005)*, pages 117–122, 2005.

[14] R. G. J. Wijnhoven and P. H. N. de With. Fast training of object detection using stochastic gradient descent. In *Pattern Recognition (ICPR), 20th International Conference on*, pages 424–427, 2010.

[15] W. Zheng and L. Liang. Fast car detection using image strip features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 2703–2710, 2009.