21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Context-Driven Moving Vehicle Detection in Wide Area Motion Imagery

Xinchu Shi[1,2], Haibin Ling[2], Erik Blasch[3], Weiming Hu[1]
[1]*National Laboratory of Pattern Recognition, Institute of Automation, Beijing,China*
[2]*Department of Computer and Information Science, Temple university, Philadelphia, USA*
[3]*Air Force Research Lab, USA*
xcshi@nlpr.ia.ac.cn, hbling@temple.edu, erik.blasch@gmail.com, wmhu@nlpr.ia.ac.cn

Abstract

Detection of moving vehicles in wide area motion imagery (WAMI) is increasingly important, with promising applications in surveillance, traffic scene understanding and public service applications such as emergency evacuation and policy security. However, the large camera motion, along with low contrast between vehicles and backgrounds, makes detection a challenging task. In this paper, we propose a novel moving vehicle detection approach by embedding the scene context, which is a road network estimated online. A two-step framework is used in the work. First, with an initial vehicle detection, trajectories are achieved by vehicle tracking. Then, the road network is extracted and used to reduce false detections. Quantitative evaluation demonstrates that the proposed contextual model remarkably improves the detection performance.

1. Introduction

Detection of moving vehicles in airborne videos has a wide range of applications, from visual surveillance to security related tasks such as rescue and evacuation. There are many factors that make vehicle detection in WAMI a challenge task, such as large camera motion, low contrast between objects with backgrounds, illumination variation and so on. Large camera motion makes motion detection a tough task, since many incorrect motion blobs are caused by parallax errors in the image stabilization stage, producing distractions that reduce the detection, which can be seen in Figure 1. Appearance-based vehicle detection has little effect in discriminating objects with backgrounds, as they are similar on texture and intensity distribution. In this work, we propose a method using the scene context to aid vehicle detection.

In our approach, a two-step detection framework is proposed. After motion detection, the trained classifier is used to detect vehicles. In the second stage, trajectories are achieved by tracking detected objects, and a road network is extracted and fed back to the object detection in the second stage. By embedding the context information, most false alarms outside the road are filtered out.

Figure 1: Moving vehicle detection results. Top: without scene context. Bottom: using context

Our contribution is threefold. First, we introduce an effective road extraction method, which fits road network with the vehicle trajectories. This process is conducted online, thus can be used widely. Second, the scene context is used in the object detection and much better results are achieved. Third, two kinds of complementary features, shape and gradient distribution, are used to obtain a well-performed detector.

This paper is organized as follows. Section 2 gives related work on vehicle detection from aerial imagery. Section 3 describes details about the proposed approach. Experimental results are presented in Section 4 and Section 5 concludes the work.

2. Related work

Moving object detection in airborne video is especially difficult, since there exist parallax errors in frame stabilization. Yaclin et al [10] propose a Bayesian framework for detecting and segmenting moving objects from the background, based on statistical analysis of optic flow. Yu et al [11] discriminate the essential difference in motion patterns caused by parallax and moving objects, with a tensor voting technique. Appearance feature based classification is used widely in vehicle detection [7, 6], in [6], multiple kernel learning is used to fuse HOG and Haar descriptors for classifying vehicles and distracters.

Contextual information is useful in object detection. In [9], Geographical Information System (GIS) information of road network are used to filter out false alarms outside the road. However, the GIS source are not available universally. In [3], the image is segmented into streets, building, trees etc, and vehicle detections that are not present on the streets are discarded. A similar work [4] classifies the scene into vehicle, road and background, and the scene classifier is trained based on appearance and motion features. These approaches are promising. However, scene segmentation is a tough task itself, and the pre-learning needs a large set of images.

3. Our Approach

3.1. Overview

Figure 2 gives the framework of the proposed approach. It is a two-stage object detection: Initial vehicle detection and refined object detection with scene context. The whole framework can be roughly divided into three parts, which are motion detection, context estimation and two-stage vehicle detection. Moving vehicle detection is executed in a sliding temporal window. First, the road network is estimated when a batch of images are available. Second, the initial vehicle detections are refined by using the road context. The whole process is proceeded online and iteratively.

3.2. Motion Detection

Aerial images are achieved with a moving airborne platform, and large camera motion exists between consecutive frames, thus sequence stabilization is essential for motion detection. In this work, Feature point based registration is used for image alignment. In particular, SURF features [2] are exploited due to its efficiency, and we then fit affine models for image warping. With stabilized frames, background modeling is achieved through the median filter.

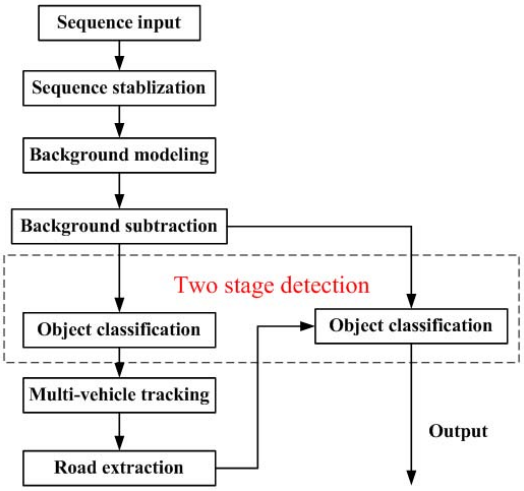

Figure 2: The framework of our approach.

Background subtraction performs well in the planar territories of the aerial scenes, while it degenerates sharply on the above-ground objects, such as buildings and trees. Due to the parallax errors introduced by image registration, buildings and trees produce many false motion detections along their edges. We solve this spurious detection problem with two procedures: a specific vehicle classifier and the scene context.

3.3. Vehicle detection

Motion detection produces the object candidates, including many false alarms. The task of vehicle detection is to discriminate real vehicles from backgrounds, therefore can be considered as a binary classification.

We construct a cascade of Support Vector Machine (SVM) classifiers for classifying object with background, and two kinds of features, shape (size) and histogram of orientated gradient (HOG), are used separately. Size feature is a four dimensional vector, which is represented as equation (1), where l and w denote the length and width of the object respectively. HOG feature is a 360 dimensional vector, which is extracted from the normalized 24×32 patch. The size based classifier is selected as the first stage of the classifier cascade, and followed by the HOG classification. In both stages, SVM is used as the basic classifier.

$$f = \{l, w, l/w, l * w\} \tag{1}$$

The reasons we selected these two kinds of features are as follows. First, HOG features represent the contour of the object in some extent, and psychological

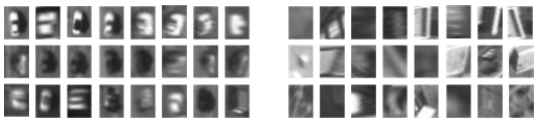

Figure 3: Positive (left) and negative (right) samples

tests [12] illustrate that the boundary and contour of the car body are the most prominent visual features for recognizing vehicles. Second, size features are powerful for discriminating vehicles and false alarms, since the former have consistent size representation such as the ratio of length with width, while the latter have no such traits. Generally, by integrating two kinds of complementary and effective cues, the learned classifier achieve the favorable performance.

In the training of the classifier cascade, all samples are extracted from the foreground images[1]. Some training samples are shown in Figure 3, from which we can see that some positive and negative samples are similar with each other. Therefore, though two kinds of effective features are utilized, it is difficult to distinguish objects with backgrounds by using individual representation only. In this work, scene context is extracted to aid moving vehicle detection.

3.4. Context extraction

Context is especially useful in aerial video analysis, because most of the vehicles move in a special area. Many methods [3, 4] estimate the road network using the scene classification, which needs a complicated training and many images are prepared in advance. In this work, the road context is extracted in a more flexible way. First, with the first stage object detections, trajectories of objects are achieved using multi-object tracking. Second, we estimate the road network from object trajectories.

In multi-object tracking, we follow the hierarchial multi-object association methods [5, 8]. Short but robust associations are achieved first, with appearance, spatial and size similarities. Then, short associations are joined into long trajectories further. We use Hungarian algorithm to fulfil the two layer association.

After multi-object tracking, many trajectories are obtained. We first discard the short-time trajectories, some of which are caused by false alarms, therefore are not reliable. We assume a long trajectory is $s^i_{1:t} = \{s^i_1, s^i_2, ..., s^i_t\}$, where $s^i_t = \{x^i_t, y^i_t\}$ is the spatial position of the i-th trajectory at instant t. For fragment $\{s^i_k, s^i_{k+1}\}(1 \le k \le t-1)$, the local road mask is estimated as equation (2).

[1]The scene is different from the test one

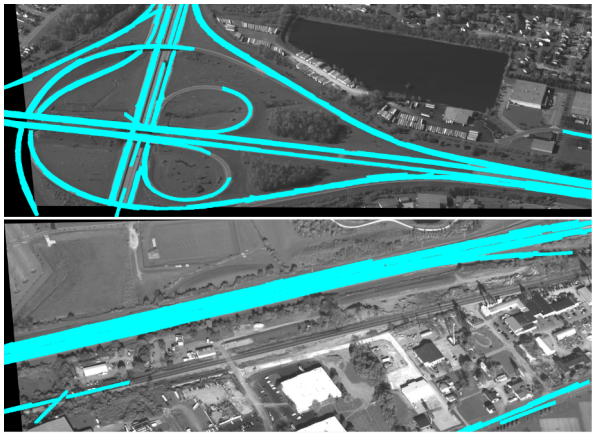

Figure 4: Road estimation. Top: Scene 1. Bottom: Scene 2

$$m^i_k(x,y) = \mathrm{I}(\varphi^i_k(x,y) < 0.5d^i_k)\mathrm{I}(\phi^i_k(x,y) < 1.5\sigma) \tag{2}$$

Where $\mathrm{I}(\cdot)$ is the indicator function, $\varphi^i_k(x,y)$ and $\phi^i_k(x,y)$ is the distance functions which are defined as equation (3); d^i_k is the length of $\{s^i_k, s^i_{k+1}\}$ and defined in equation (4); 3σ is the width of the fitting lane, and σ is the estimated width of a vehicle.

$$\begin{cases} \varphi^i_k(x,y) = \left|(x-p^i_k)\cos\theta^i_k + (y-q^i_k)\sin\theta^i_k\right| \\ \phi^i_k(x,y) = \left|(y-q^i_k)\cos\theta^i_k - (x-p^i_k)\sin\theta^i_k\right| \end{cases} \tag{3}$$

Where p^i_k and q^i_k is the center of $\{s^i_k, s^i_{k+1}\}$, θ^i_k is the orientation of $\{s^i_k, s^i_{k+1}\}$, their definitions are shown in equation (4).

$$\begin{cases} (p^i_k, q^i_k) = (\frac{x^i_k + x^i_{k+1}}{2}, \frac{y^i_k + y^i_{k+1}}{2}) \\ d^i_k = \left\|(x^i_k - x^i_{k+1}, y^i_k - y^i_{k+1})\right\|_2 \\ \theta^i_k = \tan^{-1}(x^i_k - x^i_{k+1}, y^i_k - y^i_{k+1}) \end{cases} \tag{4}$$

The single mask generated by trajectory $s^i_{1:t}$ is represented as equation (5). If there are N long trajectories, the final road mask RM is represented as equation (6).

$$M^i = m^i_1 \cup m^i_2 ... \cup m^i_{t-1} \tag{5}$$

$$RM = M^1 \cup M^2 ... \cup M^N \tag{6}$$

4. Experiments

We use the CLIF dataset [1] in all experiments. CLIF is a challenging WAMI dataset, with small object occupancy, low contrast between object and background etc. Two sequences are used, and there are 8888 and 8574 moving vehicles in each sequence respectively.

Results about road network extraction are given first. Figure 4 demonstrates the performances in two scenarios. It can be seen that the vehicle tracking based road

Table 1: Classification performance

	Size-SVM	HOG-SVM	Cascade
Positive rate	0.995	0.844	0.843
Negative rate	0.188	0.757	0.797

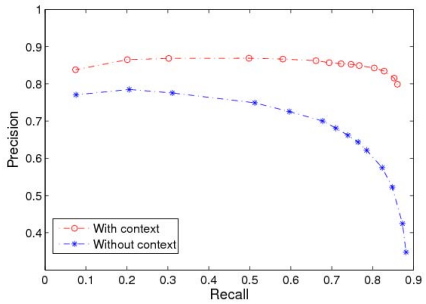

Figure 5: PRC on sequence 1

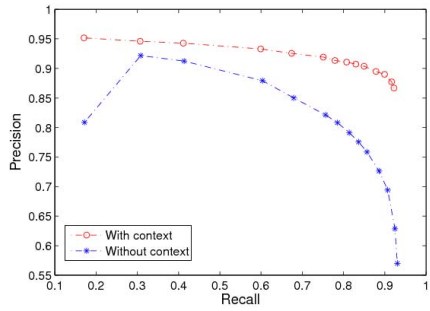

Figure 6: PRC on sequence 2

extraction achieves excellent results, especially in the first scene, which is complicated and consists of overpass and multiple highways. There are few small gaps left, mainly attributed to no vehicles pass through these areas. There are wrong road estimation in the second sequence, which can be seen in the left and bottom of Figure 4, which is actually a railway.

For evaluating the performance of the cascade classifier, quantitative results about vehicle classification at different stages in the test set are shown in Table 1. The size based classifier is constructed in a way to classify almost all positives correctly, meanwhile filtering out as much negative samples as possible. Generally, HOG classifier achieves the moderate performance alone. After embedding the shape features, the performance of the cascade has a notable advance.

With the extracted road context, most false alarms outside the road are filtered out directly. In this way, moving vehicle detection has a remarkable improvement. We compare the context aided detection with the counterpart without using the context information. "Precision-Recall Curve" is used in the evaluation measure, which are shown as Figure 5 and Figure 6. The proposed approach performs better than the method without using contextual information. The precision of context based approach is much higher since the contextual information is effective.

5. Conclusion

Detection of moving vehicles in WAMI images is a difficult task. In this work, we make use of two complementary and effective features, size and HOG, to detect vehicles. Further, the useful scene context, road network, is extracted to promote the motion detection. Our road estimation is very flexible and no pre-learning is needed. Experimental results illustrate that our approach achieves the remarkable performance.

Acknowledgement: This work is partly supported by NSFC (Grant No. 60825204, 60935002), and Beijing Natural Science Foundation (4121003). Ling is supported in part by NSF Grant IIS-1049032.

References

[1] Afrl: Columbus large image format (clif) 2006. www.sdms.afrl.af.mil/index.php?collection=clif2006.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *ECCV*, pages 404–417, 2006.

[3] H. Grabner, T. Nguyen, B. Gruber, and H. Bischof. On-line boosting-based car detection from aerial images. *ISPRS*, 63(3):382–396, 2008.

[4] C. Guilmart, S. Herbin, and P. Perez. Context-driven moving object detection in aerial scenes with user input. In *ICIP*, pages 1781–1784, 2011.

[5] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, pages 788–801, 2008.

[6] P. Liang, G. Teodoro, H. Ling, E. Blasch, G. Chen, and L. Bai. Multiple kernel learning for vehicle detection in wide area motion imagery. In *International Conference on Information Fusion (FUSION)*, 2012.

[7] P. Negri, X. Clady, S. Hanif, and L. Prevost. A cascade of boosted generative and discriminative classifiers for vehicle detection. *EURASIP Journal on Advances in Signal Processing*, 2008:136, 2008.

[8] V. Reilly, H. Idrees, and M. Shah. Detection and tracking of large number of targets in wide area surveillance. *ECCV*, pages 186–199, 2010.

[9] J. Xiao, H. Cheng, F. Han, and H. Sawhney. Geo-spatial aerial video processing for scene understanding and object tracking. In *CVPR*, pages 1–8, 2008.

[10] H. Yalcin, M. Hebert, R. Collins, and M. Black. A flow-based approach to vehicle detection and background mosaicking in airborne video. In *CVPR*, 2005.

[11] Q. Yu and G. Medioni. Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In *CVPR*, pages 2671–2678, 2009.

[12] T. Zhao and R. Nevatia. Car detection in low resolution aerial images. *Image and Vision Computing*, 21(8):693–703, 2003.