Vehicle Detection with Context

Yang Hu and Larry S. Davis
Institute for Advanced Computer Studies, University of Maryland, 20742 College Park, MD, U.S.A.

Keywords: Vehicle Detection, Context, Conditional Random Fields, Shadow, Ground, Orientation.

Abstract: Detecting vehicles in satellite images has a wide range of applications. Existing approaches usually identify vehicles from their appearance. They typically generate many false positives due to the existence of a large number of structures that resemble vehicles in the images. In this paper, we explore the use of context information to improve vehicle detection performance. In particular, we use shadows and the ground appearance around vehicles as context clues to validate putative detections. A data driven approach is applied to learn typical patterns of vehicle shadows and the surrounding "road-like" areas. By observing that vehicles often appear in parallel groups in urban areas, we also use the orientations of nearby detections as another context clue. A conditional random field (CRF) is employed to systematically model and integrate these different contextual knowledge. We present results on two sets of images from Google Earth. The proposed method significantly improves the performance of the base appearance based vehicle detector. It also outperforms another state-of-the-art context model.

1 INTRODUCTION

With the launch of new generation of earth observation satellites, more and more high-resolution satellite images with ground sampling distances of less than 1 meter have become publicly available. Small scale objects such as vehicles can be readily seen in these images. In this work, we consider the problem of detecting vehicles from such high-resolution aerial and satellite images. This problem has a number of applications in traffic monitoring and intelligent transportation systems, urban planning and design, as well as military and homeland surveillance. In spite of the increasing resolution of aerial and satellite images, vehicle detection still remains a difficult problem. In urban settings especially, the presence of a large number of rectangular structures brings significant challenges to the detectors.

Vehicle detection has been explored a lot in the literature. Most approaches only use the appearance of vehicles for detection. Due to the existence of the structures that resemble vehicles in the images, these methods typically generate many false positives. In this work, we investigate the use of context information to improve vehicle detection performance.

Context is a useful information source for visual recognition. Psychology experiments show that in the human visual system context plays an import role in recognition (Oliva and Torralba, 2007). In computer vision, using context has recently received significant attention. It has been used successfully in object detection and recognition (Rabinovich et al., 2007; Heitz and Koller, 2008; Divvala et al., 2009) as well as many other problems such as scene recognition (Murphy et al., 2003), action classification (Marszalek et al., 2009) and recognition of human-object interactions (Yao and Fei-Fei, 2012).

We explore useful context clues for the detection of vehicles. The first type of context information we use are shadows. Instead of using image meta-data to predict the expected location and shape of shadows, we apply a data driven approach to learn the typical patterns of vehicle shadows from examples. We also use the ground appearance around a vehicle as another contextual clue. Unlike previous work (Chellappa et al., 1994; Quint, 1997; Jin and Davis, 2007) that requires maps of road network registered to imagery, we use image appearance and a data driven approach to determine whether a not a putative vehicle detection is surrounded by "road-like" pixels. Finally, by observing that in urban areas vehicles often appear in parallel groups, we use the orientations of nearby detections to validate the initial detections. We employ a conditional random field (CRF) to systematically model and integrate these different contextual clues. The algorithms are evaluated on two sets of images from Google Earth. The results indicate that the proposed context model greatly improves vehi-

Hu Y. and S. Davis L..
Vehicle Detection with Context.
DOI: 10.5220/0004302907150722
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 715-722
ISBN: 978-989-8565-47-1

cle detection performance over a baseline appearance based detector. It also outperforms another recently proposed context model.

The rest of this paper is organized as follows. We first discuss related work in Section 2. Then, in Section 3, we briefly introduce the partial least squares baseline vehicle detector (Kembhavi et al., 2011) that we use to obtain the initial detections to build the CRF model. We present the CRF model, which is used as a general framework to integrate different context clues, in Section 4. We then discuss how we model the three kinds of contextual information, i.e. shadow, ground and orientations of nearby detections, in Section 5. Experiment results are discussed in Section 6. Finally, we conclude in Section 7.

2 RELATED WORK

Vehicle detection has previously been treated as a template matching problem, and algorithms that construct templates in 2D as well as 3D have been proposed. Monn et al. (Moon et al., 2002) proposed an approach to accurately detect 2D shapes and applied it to vehicle detection. They derived an optimal 1D step edge operator and extended it along the boundary contour of the shape to obtain a shape detector. Choi and Yang (Choi and Yang, 2009) first used a mean-shift algorithm to extract candidate blobs that exhibit symmetry properties of typical vehicles and then verified the blobs using a log-polar shape descriptor. Zhao and Nevatia (Zhao and Nevatia, 2003) posed vehicle detection as a 3D object recognition problem. They used human knowledge to model the geometry of typical vehicles. A Bayesian network was used to integrate the clues including the rectangular shape of the car, the boundary of the windshield and the outer boundary of the shadow.

The detection of vehicle has also been treated as a classification problem, and different machine learning algorithms have been exploited for it. Jin and Davis (Jin and Davis, 2007) used a morphological shared-weight neural network to learn an vehicle model. Grabner et al. (Grabner et al., 2008) proposed to use on-line boosting in an interactive training framework to efficiently train and improve a vehicle detector. Kembhavi et al. (Kembhavi et al., 2011) presented a vehicle detector that improves upon previous approaches by incorporating a large and rich set of image descriptors. They used partial least squares, a classical statistical regression analysis technique, to project the extremely high-dimensional feature onto a much lower dimensional subspace for classification.

Contextual knowledge has been exploited for vehicle detection in some previous systems. (Chellappa et al., 1994; Quint, 1997; Jin and Davis, 2007) integrate external information from site-models or digital maps to reduce the search for vehicles to certain image regions such as road networks and parking lots. Some use a vehicle's shadow projection as local context for vehicles (Hinz and Baumgartner, 2001; Zhao and Nevatia, 2003). In these works, meta-data for aerial images are used to compute the direction of sun rays and derive the shadow region projected on the road surface. Heitz and Koller (Heitz and Koller, 2008) present a ”things and stuff (TAS)” context model that uses texture regions (e.g., roads, trees and buildings) to add predictive power to the detection of objects and applied it to vehicle detection.

3 VEHICLE DETECTION USING PARTIAL LEAST SQUARES

Our context model is built on the detections from a sliding window vehicle detector. This detector slides a window over the image, scores each window according to its match to a pre-trained vehicle model, and returns the windows with locally highest matching scores. The vehicle model can be derived from most standard classifiers. In this work we use a partial least squares (PLS) based detector (Kembhavi et al., 2011) to generate the initial detections.

PLS is a method that uses latent variables to model the relations between sets of observed variables. The detector first uses PLS to project original features onto a more compact space of latent variables. Then quadratic discriminant analysis (QDA) is applied to classify the windows into vehicle and background. Although computationally simple, this detector has been shown to have good detection performance for both vehicles (Kembhavi et al., 2011) and human (Schwartz et al., 2009).

We use the Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) feature for the detector. HOG captures the distribution of edges or gradients that are typically observed in image patches that contain vehicles. Each detection window is divided into square cells and a 9-bin HOG feature is calculated for each cell. Grids of 2×2 cells are grouped into a block, resulting in a 36D feature vector per block. A multiscale approach that uses blocks at varying scales and varying aspect rations (1:1, 1:2, and 2:1) is employed (Zhu et al., 2006).

4 CONTEXT MODEL WITH CONDITIONAL RANDOM FIELDS

A detector that only relies on the appearance of the vehicles will trigger many false alarms at locations that show similar appearance patterns to vehicles. For example, in images captured by wide-area motion imagery (WAMI) sensors, the vehicle detector is always confused by electrical units and air conditioning units on the tops of buildings. We propose to use contextual information to reduce these false alarms.

One typical source of spatial contextual information is shadows. Shadows provide information to differentiate physical objects from texture regions with confusing appearance. The shape of the shadow area is closely related to the object casting it. These make shadows important context clue for vehicle detection. High level scene information is also very useful. Vehicles should always appear on the roads or parking lots instead of on trees or buildings. Therefore, investigating the type of the surrounding regions is also a useful way to validate a detection. Additionally, since nearby vehicles always move or park in the same orientation, they provide strong contextual support for each other.

To systematically employ all these sources of information, we use a conditional random field (CRF) to model and aggregate these contextual cues. After running the PLS based sliding window vehicle detector, we construct a graph with the top scoring (and locally maximal) detections from the detector as nodes and connect nearby detections (i.e. the distance between two detections is smaller than a threshold) by an edge. We then define a CRF on that graph, which expresses the log-likelihood of a particular label **y** (i.e. assignment of vehicle/non-vehicle to each window) given observed data **x** as a sum of unary and binary potentials:

$$-\log P(\mathbf{y}|\mathbf{x};\mu,\lambda) \sim \sum_i \sum_k \mu_k \phi_k(y_i,\mathbf{x}_i) + \sum_{(i,j)\in\varepsilon} \sum_l \lambda_l \psi_l(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j) \qquad (1)$$

where ε is the set of edges between detections, ϕ_k and ψ_l are the unary and pair-wise feature functions respectively, and μ_k and λ_l are weights controlling the relative importance of the terms.

Unary potentials measure the affinity of the pixels surrounding the detected locations to the presence of vehicles. The likelihood that a detection window contains a vehicle according to the PLS based vehicle detector can be encoded in a unary term:

$$\phi^p(y_i = 1, \mathbf{x}_i) = p_i \qquad (2)$$

where p_i is the confidence score for the ith window obtained from the detector. The likelihood that the detected object is accompanied by a vehicle shadow and the likelihood that the object is on the ground are also encoded in unary terms.

The binary potentials enforce the consistency of the labels assigned to neighboring detections according to their properties.

In the following sections, we describe the computation of these potentials in details.

5 CONTEXT CLUES

5.1 Shadow Clue

To use shadows as a context clue, we need to detect them. Since we are interested in the shadows near the detected objects, we only detect the shadows in the areas near the locations obtained from the sliding window vehicle detector. We use the appearance of local regions to detect shadows. When a region is in shadow, it becomes darker and less textured (Zhu et al., 2010). Therefore, the color and texture of a region can help predict whether it is in shadow. Taking a rectangular window centered at a detected location, following (Guo et al., 2011), we first segment the window into regions using the mean shift algorithm (Comaniciu and Meer, 2002). Then for each region, the color and texture are represented with a histogram in L*a*b space and a histogram of textons respectively. A SVM classifier with a χ^2 kernel, which is trained from manually labeled regions, is used to determine whether a region is in shadow. After classifying each region in the window, we obtain a corresponding binary image which indicates the shadow areas in it. We use these binary shadow images to compute the shadow potential in the CRF model.

The absence of shadows in a shadow image can help to filter out detections whose appearances are similar to vehicles but do not have casting shadows. For detections that have shadows, the position, shape and size of the shadow area further reveals the type of the object casting it. In some cases, some image meta-data may be available, which make it possible to calculate the shadows using the geometric relationship of the sun and the vehicles. Then we can verify a detection by comparing this theoretically computed shadow with the shadow image obtained by running the shadow detector. In general, however, we do not have the corresponding meta-data and therefore are not able to get the theoretical predictions for comparison. In such cases, we learn the characteristics of typical vehicle shadows from training images.

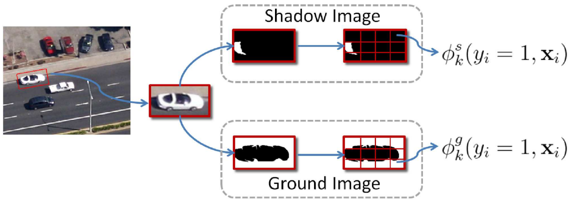

Shadow Image
$\phi^s_k(y_i = 1, \mathbf{x}_i)$
$\phi^g_k(y_i = 1, \mathbf{x}_i)$
Ground Image

Figure 1: Illustration of the computing of shadow and ground potentials.

Let I^s_i denote the binary shadow image of the ith detection; we assume that the likelihood that the shadow is from a vehicle is linear function of the pixels in I^s_i. A set of unary feature functions, each corresponding to a pixel in I^s_i is defined, i.e. $\phi^s_j(y_i = 1, \mathbf{x}_i) = I^s_{ij}$, where $I^s_{ij} \in \{0,1\}$ is the jth pixel in I^s_i. Then the coefficients μ^s_j learned by the CRF assign different weights to the pixels according to their positions in the window. This definition, while precisely differentiating each pixel, greatly increases the complexity of the CRF model. On the other hand, nearby pixels play similar roles for the prediction. To achieve a better performance-cost trade off, we can assume that they share the same weight. Among the many different potential patterns of sharing the weights, we simply divide the shadow image into a uniform grid of cells and have all the pixels in a cell weighted by the same coefficient. Then the unary potential of the shadow clue can be expressed as

$$\sum_i \sum_j \mu^s_{c(I^s_{ij})} I^s_{ij} \qquad (3)$$

where $c(I^s_{ij})$ indicates the cell pixel I^s_{ij} belongs to. This is equivalent to

$$\sum_i \sum_k \mu^s_k \phi^s_k(y_i = 1, \mathbf{x}_i) \qquad (4)$$

where $\phi^s_k(y_i = 1, \mathbf{x}_i) = \sum_{c(I^s_{ij})=k} I^s_{ij}$. Here we define a set of new feature functions, each of which computes the sum of the pixels in a cell.

Note that the above feature functions are computed over cells, making them robust to some position variability of the shadows. This is very important since the sliding window vehicle detector usually moves the windows with step size larger than 1 pixel. It is also not practicable for the detector to consider every orientation. Therefore, the detected vehicle may not lie in the center of the detection window, and its orientation estimate is subject to some sampling error. By only counting the number of pixels that are in shadow for each cell, we make the computation of the shadow potential tolerant to these sources of variance.

5.2 Ground Clue

Besides shadows, another important contextual clue for vehicles is they are typically on roads, driveways or parking lots. To utilize this information, we analyze the surrounding regions of the detected locations. Specifically, we consider a rectangle window centered at a candidate location, segment it into regions and characterize their appearance using color and texton histograms. Then, the regions are classified as ground or non-ground by a classifier. A binary image, which indicates pixels that are classified as belonging to ground, is obtained. We refer to this as the "ground image" for the candidate location.

The ground potential is calculated in a similar way as the shadow potential. Let I^g_i denote the ground image of the ith detection. After dividing it into a uniform grids of cells, the ground potential is expressed as

$$\sum_i \sum_k \mu^g_k \phi^g_k(y_i = 1, \mathbf{x}_i) \qquad (5)$$

where $\phi^g_k(y_i = 1, \mathbf{x}_i) = \sum_{c(I^g_{ij})=k} I^g_{ij}$, which corresponds to the number of pixels that are assigned to ground in the kth cell.

This method for computing ground potential is based on a local analysis of the ground. One may also first detect all the ground areas in the entire image and then check the spatial relationships between the candidate detections and the ground. The TAS model (Heitz and Koller, 2008) operates in this fashion although the ground areas are detected through an unsupervised procedure. Since it explicitly considers spatial relationships, it is effective at filtering out detections that are not near roads. Our method, on the other hand, not only expects a detection to be mostly surrounded by ground, but it also can penalize the situation in which ground appears in the center of the detection window. This is important for removing false positives that are on ground but do not contain vehicles. This crucial difference between two methods will be illustrated in the experiment results.

5.3 Orientation Clue

In addition to the unary potentials, the frequent co-occurrence of vehicles can be used to develop a binary potential.

Vehicles, while moving, typically move in the direction of road lanes; in parking, there are also regularities in the patterns of parking. Therefore, nearby vehicles are usually oriented in the same orientation. We can use this observation to validate nearby detections. Specifically, when two nearby detection windows have the same orientation, it is more probable that both of them contain vehicles. On the other hand, when two nearby detection windows have quite different orientations, the probability that they both are true vehicle windows should be low. Although the specific probabilities for different label combinations are hard to assign manually, they can be estimated from training data by maximizing the likelihood of the data.

Let $d(\mathbf{x}_i) \in (-180, 180]$ denotes the orientation of the ith detection window; we classify the orientation relations of two windows into three categories. In the first case, the two windows are in exactly the same orientation, i.e. $|d(\mathbf{x}_i) - d(\mathbf{x}_j)| \in \{0, 180\}$. In the second, their orientations are only slightly different, i.e. $|d(\mathbf{x}_i) - d(\mathbf{x}_j)| \in (0, d_0] \cup [360 - d_0, 360) \cup [180 - d_0, 180) \cup (180, 180 + d_0]$, where d_0 is a threshold which is set to 20 in experiments. Otherwise, they are in the third category.

Based on this classification, we define a set of binary feature functions:

$$\psi_{1,\cdots,8}(y_i = 0, y_j = 0, \mathbf{x}_i, \mathbf{x}_j) = [1, a_1, a_2, a_3, 0, 0, 0, 0] \quad (6)$$

$$\psi_{1,\cdots,8}(y_i = 1, y_j = 1, \mathbf{x}_i, \mathbf{x}_j) = [0, 0, 0, 0, 1, a_1, a_2, a_3] \quad (7)$$

$$\psi_{1,\cdots,8}(y_i = 1, y_j = 0, \mathbf{x}_i, \mathbf{x}_j) = \psi_{1,\cdots,8}(y_i = 0, y_j = 1, \quad (8)$$

$$\mathbf{x}_i, \mathbf{x}_j) = \mathbf{0}$$

where

$$a_1 = \begin{cases} 1 & \text{if } |d(\mathbf{x}_i) - d(\mathbf{x}_j)| \in \{0, 180\} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

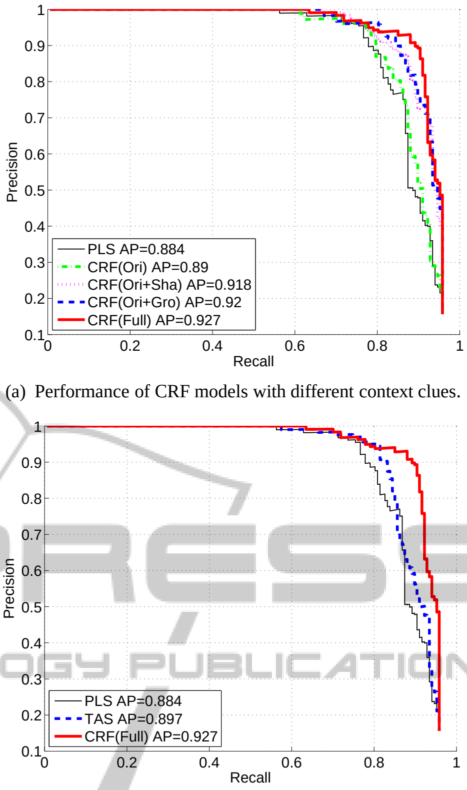
$$a_2 = \begin{cases} 1 & \text{if } |d(\mathbf{x}_i) - d(\mathbf{x}_j)| \in (0, d_0] \cup [360 - d_0, 360) \\ & \cup [180 - d_0, 180) \cup (180, 180 + d_0] \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

$$a_3 = \begin{cases} 1 \text{ if } |d(\mathbf{x}_i) - d(\mathbf{x}_j)| \in (d_0, 180 - d_0) \\ \quad \cup (180 + d_0, 360 - d_0) \\ 0 \text{ otherwise} \end{cases} \quad (11)$$

We can see that, based on their relative orientation, the probabilities that two windows both containing vehicles or not will be different. We also introduce a bias term, i.e. $\psi_i = 1$, to represent some baseline likelihood that is independent of the orientation clue.

6 EXPERIMENTS

To evaluate the CRF based context model, we perform experiments on two datasets. Although both of them are satellite images acquired from Google Earth, the appearance of the vehicles as well as the surrounding scenes are quite different in the images of these two sets.



(a) Performance of CRF models with different context clues.

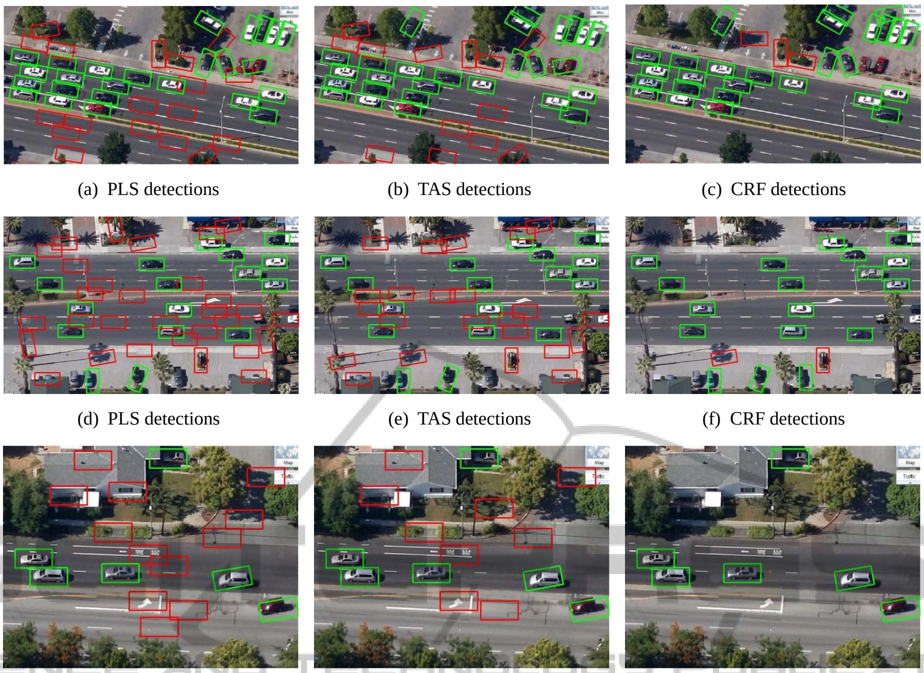
(b) Performance comparison with the TAS model.

Figure 2: Precision-recall (PR) curves for Google Earth Dataset I. AP stands for average precision.

6.1 Google Earth Dataset I

The first dataset contains 27 images of an area near Mountain View, California. There are 391 manually labeled cars in them. The vehicles are viewed obliquely with window size of 101 × 51 pixels. We use 14 images to train the CRF model and test the performance on the remaining 13 images.

We first compare the performance of the CRF models with different context clues. The PLS based detector (Kembhavi et al., 2011) was used to generate the initial detections and also serves as the baseline for comparison. Figure 2(a) shows the precision-recall curves of CRF with only orientation clue (CRF(Ori)), with both orientation and shadow clues (CRF(Ori+Sha)), with both orientation and ground clues (CRF(Ori+Gro)), and with all of the context clues (CRF(Full)). The scores from the PLS detector is included as a unary feature in all these models. We can see that although the orientation clue alone only slightly improved the performance, when combined with the shadow clue or the ground clue the detection performance is significantly improved. The effectiveness of shadow and ground clues are similar



(a) PLS detections (b) TAS detections (c) CRF detections

(d) PLS detections (e) TAS detections (f) CRF detections

(g) PLS detections (h) TAS detections (i) CRF detections

Figure 3: Example images of Google Earth Dataset I, with detections found by the PLS detector, the TAS model and our CRF(Full) model. The results at recall of 0.9 are shown. Green windows indicate true detections and red windows are false positives.

and also complementary to each other. When combined together (CRF(Full)), the detection accuracy is further improved.

We compare the performance of our CRF based context model with the things and stuff (TAS) context model (Heitz and Koller, 2008) in Figure 2(b). We provided the TAS model with the same initial detections as the CRF model. We can see that although the TAS model also improved the PLS result, the improvement is much smaller than our CRF based context model. This illustrates the advantage of the context clues we used.

We show in Figure 3 some example images, with detections found by the PLS detector, the TAS model and our CRF(Full) model respectively at a 90% recall rate. We can see that the PLS detector generates many false detections. The TAS model only filters out some of the false positives. With our CRF based context model, most of the false detections are removed.

6.2 Google Earth Dataset II

The second dataset is from TAS (Heitz and Koller, 2008). It contains satellite images of the city and suburbs of Brussels, Belgium. There are 30 images, of size 792 × 636 pixels. A total of 1319 cars are manually labeled in them. A car window is approximately 45 × 25 pixels. We use half of the images to train the context models and then test the performance on the other half of the dataset. The TAS model was trained with parameters suggested by (Heitz and Koller, 2008).

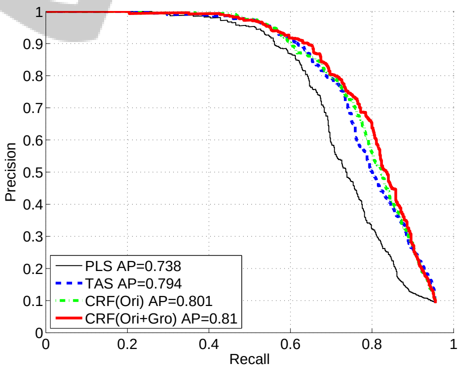
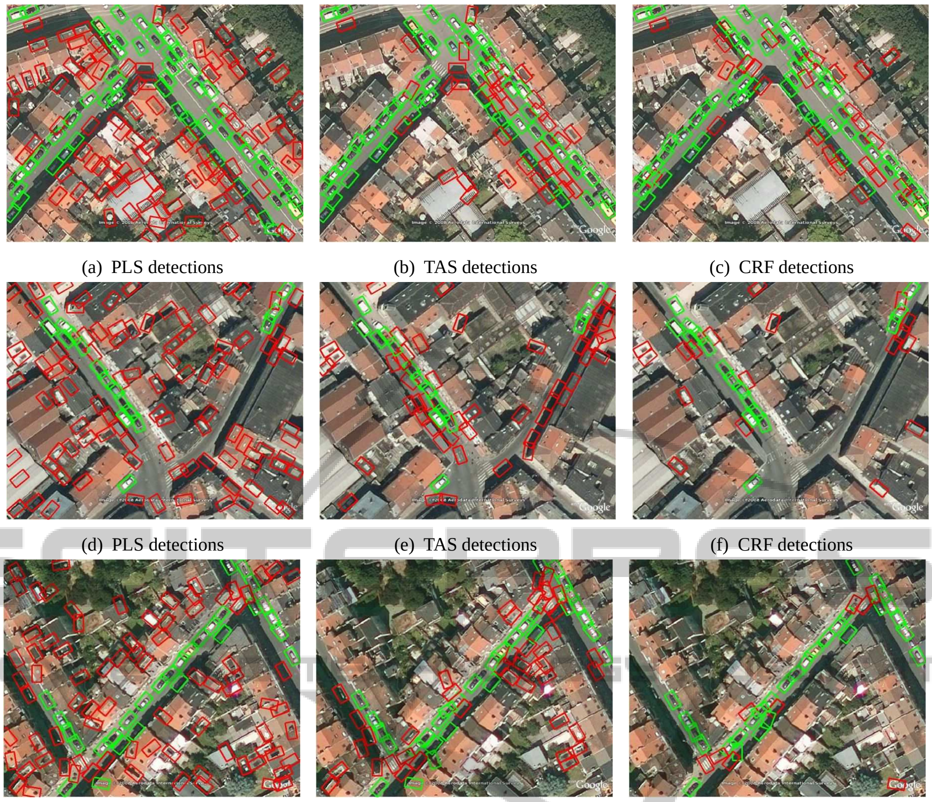


Figure 4: Precision-recall (PR) curves for Google Earth Dataset II. AP stands for average precision.

We show in Figure 4 the precision-recall curves of the PLS detector, the TAS model and our CRF based context models on this dataset. Compared with the previous dataset, a wider variety of surrounding environments other than the road occur in the images in this dataset. This enables the TAS model to better utilize the stuff, e.g. the roofs of houses, the trees and



(a) PLS detections (b) TAS detections (c) CRF detections

(d) PLS detections (e) TAS detections (f) CRF detections

(g) PLS detections (h) TAS detections (i) CRF detections

Figure 5: Example images of Google Earth Dataset II, with detections found by the PLS detector, the TAS model and our CRF(Ori+Gro) model. The results at recall of 0.8 are shown. Green windows indicate true detections and red windows are false positives.

water regions, to add predictive power to the detection of vehicles. Therefore, the TAS model achieved much larger performance improvement over the initial PLS results on this dataset than on the previous one. On the other hand, since the vehicles are more spatially proximate, the CRF model that only uses the orientation clue also achieved larger performance gain here than on the other dataset. After adding the ground clue, the performance was further improved. Since the sun was overhead, there are hardly any shadows around the vehicles. We therefore do not have result using the shadow clue for this dataset.

Figure 5 shows examples of the detections obtained by the three methods. Again we can see the PLS result includes many false alarms at the 80% recall point. The TAS model filtered out many of these false positives, especially those that are not near roads. The results of our CRF model are even better. In addition to the windows that are not on the road, those that are on the road but do not contain vehicles are also removed.

7 CONCLUSIONS

We explored the use of context information for vehicle detection in high-resolution aerial and satellite images. We presented an effective way to use both shadow and ground clues. The consistency of the orientations of nearby detections was also shown to be very useful context information. A CRF model was used to integrate the different types of contextual knowledge. Experiments on two very different sets of Google Earth images show that our method greatly improved the performance of the base vehicle detector.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Air Force Research Laboratory (AFRL) under Contract No. FA8750-11-C-0091. Any opinions, findings and conclusions or recommendations expressed

in this material are those of the authors and do not necessarily reflect the views of AFRL or the U.S. Government.

REFERENCES

Chellappa, R., Zheng, Q., Davis, L., Lin, C., Zhang, X., Rodriguez, C., Rosenfeld, A., and Moore, T. (1994). Site model based monitoring of aerial images. In *Image Understanding Workshop*.

Choi, J.-Y. and Yang, Y.-K. (2009). Vehicle detection from aerial images using local shape information. In *Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology*.

Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 18th IEEE Conference on Computer Vision and Pattern Recognition*.

Divvala, S., Hoiem, D., Hays, J., Efros, A., and Hebert, M. (2009). An empirical study of context in object detection. In *Proceedings of the 22th IEEE Conference on Computer Vision and Pattern Recognition*.

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

Guo, R., Dai, Q., and Hoiem, D. (2011). Single-image shadow detection and removal using paired regions. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*.

Heitz, G. and Koller, D. (2008). Learning spatial context: using stuff to find things. In *Proceedings of the 10th European Conference on Computer Vision*.

Hinz, S. and Baumgartner, A. (2001). Vehicle detection in aerial images using generic features, grouping, and context. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*.

Jin, X. and Davis, C. H. (2007). Vehicle detection from high-resolution satellite imagery using morphological shared-weight neural networks. *Image and Vision Computing*, 25(9):1422–1431.

Kembhavi, A., Harwood, D., and Davis, L. S. (2011). Vehicle detection using partial least squares. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1250–1265.

Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *Proceedings of the 22th IEEE Conference on Computer Vision and Pattern Recognition*.

Moon, H., Chellappa, R., and Rosenfeld, A. (2002). Optimal edge-based shape detection. *IEEE Transactions on In Image Processing*, 11(11):1209–1227.

Murphy, K., Torralba, A., and Freeman, W. (2003). Using the forest to see the trees: a graphical model relating features, objects, and scenes. In *Advances in Neural Information Processing Systems*.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527.

Quint, F. (1997). MOSES: a structural approach to aerial image understanding. *Automatic Extraction of Man-made Objects from Aerial and Space Images (II)*, pages 323–332.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2007). Objects in context. In *Proceedings of the International Conference on Computer Vision*.

Schwartz, W. R., Kembhavi, A., Harwood, D., and Davis, L. S. (2009). Human detection using partial least squares analysis. In *Proceedings of the 12th International Conference on Computer Vision*.

Yao, B. and Fei-Fei, L. (2012). Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Zhu, J., Samuel, K. G. G., Masood, S. Z., and Tappen, M. F. (2010). Learning to recognize shadows in monochromatic natural images. In *Proceedings of the 23th IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, Q., Avidan, S., Yeh, M.-C., and Cheng, K.-T. (2006). Fast human detection using a cascade of histograms of oriented gradients. In *Proceedings of the 19th IEEE Conference on Computer Vision and Pattern Recognition*.