Vehicle Detection in Wide Area Aerial Surveillance using Temporal Context

Pengpeng Liang[♮] **Haibin Ling**[♮] **Erik Blasch**[♯] **Guna Seetharaman**[♯] **Dan Shen**[‡] **Genshe Chen**[‡]
[♮]Computer & Information Science Department, Temple University, Philadelphia, PA, USA
[♯]Air Force Research Lab, USA
[‡]Intelligent Fusion Technology, Inc, Germantown, MD, USA
{*pliang,hbling*}*@temple.edu,* {*erik.blasch,Gunasekaran.Seetharaman*}*@rl.af.mil, gchen@intfusiontech.com*

***Abstract*—Moving vehicle detection from wide area aerial surveillance is an important and challenging task, which can be aided by context information. In this paper, we present a Temporal Context(TC) which can capture the road information. In contrast with previous methods to exploit road information, TC does not need to get the location of the road first or to use the Geographical Information System's (GIS) information. We first use background subtraction to generate the candidates, then build TC based on the candidates that have been classified as positive by Histograms of Oriented Gradient(HOG) with Multiple Kernel Learning(MKL). For each positive candidate, a region around the candidate is divided into several subregions based on the direction of the candidate, then each subregion is divided into 12 bins with a fixed length; and finally the TC, a histogram, is built according to the positions of the positive candidates in 8 consecutive frames. In order to benefit from both the appearance and context information, we use MKL to combine TC and HOG. To evaluate the effect of TC, we use the publicly available CLIF 2006 dataset, and label the vehicles in 102 frames which are 2672×1200 subregions that contain expressway of the original 2672×4008 images. The experiments demonstrate that the proposed TC is useful to remove the false positives that are away from the road, and the combination of TC and HOG with MKL outperforms the use of TC or HOG only.**

I. Introduction

Moving vehicle detection in Wide Area Motion Imagery (WAMI) is an important task, the result of which can be applied to monitoring traffic flow, identifying illegal behavior, etc. A common approach to detect moving vehicles in WAMI is to generate candidates using background subtraction [1, 2]; as WAMI frame rates increase the flux tensor model can be used as a more reliable motion detector [3]. Nevertheless, due to the large camera motion, 3D parallax and the low contrast between the vehicles and the background, there are many false positives among the candidates resulting from background subtraction.

Beyond using the appearance information, recent work has demonstrated that the contextual information is useful to boost the object detection task [4–6]. One reason for the effectiveness of context information is that objects in a scene always have a physical and reasonable layout, i.e., semantic context. Intuitively, for moving vehicle detection, the most useful context information is the road, if we have information about the road, we can eliminate the false positives which do not directly appear on the road.

One common way to explore the semantic context is to use a graphical method to model the relationships among objects or regions in the scene, such as [4, 5, 7–11]. A common requirement of these approaches is that in order to learn the relationships among objects or regions in a scene, we need to model objects or regions in the scene globally, i.e., objects or regions are used as context for each other. More specifically, for the vehicle detection task, if we want to use the road as context for vehicles, we need to first know where the road is.

In this paper, based on the motivation that along the direction of the road we can find a relatively large number of vehicles in several consecutive frames and these vehicles will cover a continuous region of the road, we propose a novel Temporal Context(TC) method, which can capture the road information without detecting the road. In order to build TC, we first use a background subtraction technique to generate the candidates, then, we build TC for those candidates that have been classified as positives by Histograms of Oriented Gradient(HOG) [12] with multiple kernel learning(MKL) [13–16]. For each such candidate, we divide a region around the candidate into several subregions based on the direction of that candidate, and each subregion is divided into 12 bins. Finally, the TC, a histogram, is built by calculating the number positive candidates in 8 consecutive frames lying in each bin. Figure 1 gives an example of the histograms of TC for positive and negative candidates classified by HOG with MKL, from which we can see that the TC for true positives have smooth consecutive bins that have relatively large values while the false positives do not. In order to benefit from both the appearance and context information, MKL [13] is used to combine TC and HOG. Our experiment is conducted on the Columbus Large Image Format (CLIF) 2006 dataset [17]. In order to get both qualitative and quantitative results, vehicles in 102 frames which are 2672×1200 subregions that contain the expressway road of the original 2672×4008 image are labeled. The experiment demonstrates that with the same recall, the combination of TC and HOG outperforms the use of TC or HOG only, and TC is useful to remove the false positives that are away from the road.

The rest of the paper is organized as follows. §II discusses the related work. §III presents the proposed Temporal Context(TC) method. §IV gives the details of the approach for vehicle detection. §V experimentally demonstrates the effectiveness of the proposed TC and §VI provides conclusions.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE
Vehicle Detection in Wide Area Aerial Surveillance using Temporal Context

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Temple University,Computer & Information Science Department,Philadelphia,PA,19104

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
11. SPONSOR/MONITOR'S REPORT NUMBER(S)

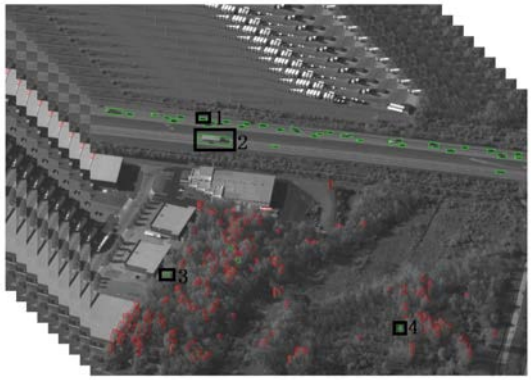
12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 16th International Conference on Information Fusion held in Istanbul, Turkey on 9-12 July 2013. Sponsored in part by Office of Naval Research Global.

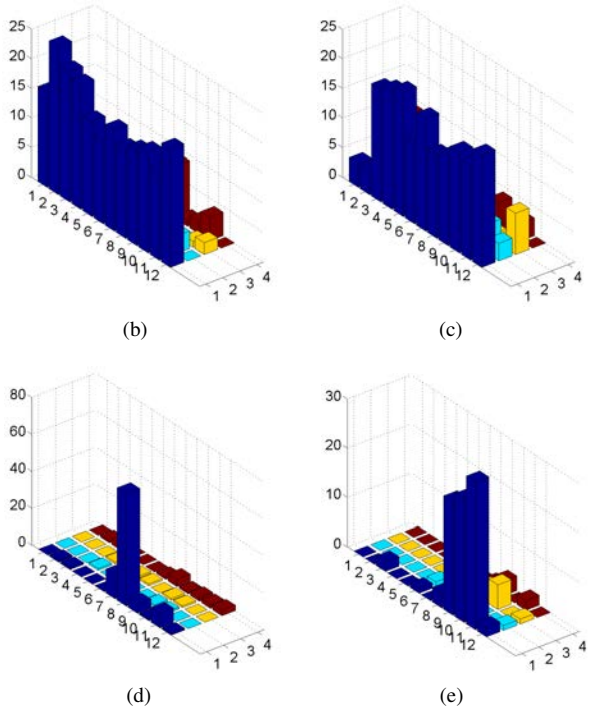
14. ABSTRACT
Moving vehicle detection from wide area aerial surveillance is an important and challenging task, which can be aided by context information. In this paper, we present a Temporal Context(TC) which can capture the road information. In contrast with previous methods to exploit road information TC does not need to get the location of the road first or to use the Geographical Information System?s (GIS) information. We first use background subtraction to generate the candidates, then build TC based on the candidates that have been classified as positive by Histograms of Oriented Gradient(HOG) with Multiple Kernel Learning(MKL). For each positive candidate, a region around the candidate is divided into several subregions based on the direction of the candidate, then each subregion is divided into 12 bins with a fixed length; and finally the TC, a histogram is built according to the positions of the positive candidates in 8 consecutive frames. In order to benefit from both the appearance and context information, we use MKL to combine TC and HOG. To evaluate the effect of TC, we use the publicly available CLIF 2006 dataset, and label the vehicles in 102 frames which are 2672 1200 subregions that contain expressway of the original 2672 4008 images. The experiments demonstrate that the proposed TC is useful to remove the false positives that are away from the road, and the combination of TC and HOG with MKL outperforms the use of TC or HOG only.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT	b ABSTRACT	c THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**8**	



(a)



(b) (c) (d) (e)

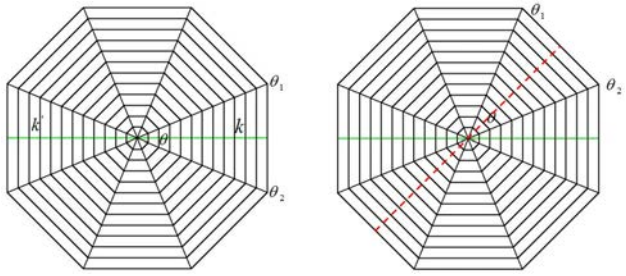
Fig. 1: An Example of TC. (a) shows the classification result of HOG with MKL, the green and red bounding box indicates the candidates classified as positive and negative respectively. (b),(c),(d),(e) are the histograms for candidates 1,2,3,4 in (a) respectively. Four different colors represent the consecutive bins for four different directions.

II. Related Work

The research about understanding of wide area motion imagery has become increasingly popular. Zhao et al. [18] used the boundary of the car, the boundary of the front windshield and the shadow as features and integrated these features in the structure of the Bayesian network. Also, several papers about vehicle detection and tracking have been published in recent years [1, 2, 19, 20]. These papers mainly focus on the tracking task as detection is just conducted as a prerequisite for tracking. Reilly et al. [1] used background subtraction to generate the candidates by modeling the background with 10 consecutive stabilized images using a median background model. Prokaj et al. [2] adopted the similar method as [1] to perform vehicle detection, while at the same time, they refined the detected vehicles using the tracking result. Xiao et al. [20] used a three-frame subtraction scheme to initialize the tracking, and the road information from an additional co-registered GIS database as a constraint. Shi et al. [19] first used the vehicle detection result to construct trajectories, then road information was estimated by these trajectories and used to refine the detection result. Aside from the above work, other studies on general moving object detection and analysis in WAMI using a variety of methods can be found in [21–26].

The effectiveness of context for object detection tasks has been well explored and studied in the community. Divvala et al. [6] gave an empirical study of context for the object detection task. Besides the spatial layout, the area surrounding the objects or the neighborhood of the objects can provide useful information [27, 28]. However, for moving vehicle detection in wide area motion imagery, the area surrounding the candidates of vehicles cannot provide enough information, due to the low contrast between the candidates and the background. Different approaches to explore contextual information have also been recently proposed, Song et al. [29] proposed the Context-SVM (Support Vector Machine) to boost the object classification and detection by using the outputs from one task as the context of the other one. Context-SVM is not suitable for the vehicle detection task in WAMI, since it is hard to categorize this kind of imagery into different categories. Felzenszwalb et al. [30] selected the highest score of detections from each of the k different models (for different object categories) to form a k-dimensional vector and rescoring a detection using this k-dimensional feature vector plus the original score, the position of the bounding box and the image context.

III. The Temporal Context



(a) (b)

Fig. 2: Diagram of the histogram bins of TC. (a)The difference between the estimated direction of a candidate and the direction of the road is not greater than a threshold, the horizontal green line is the estimated direction of the candidate; (b) The difference between the two directions is greater than the threshold, the horizontal green line is the estimated direction of the candidate and the red dashed line is the direction of the road.

Given T consecutive frames $I = \{I_1, I_2, I_3, ..., I_T\}$, and a candidate token p in frame i which is from the result of background subtraction and has been classified as positive by HOG [12] with MKL [13], we first divide the region around p into m subregions based on the estimated direction

characterizing a manifold centered at p. Each subregion is similar to a fan and all the subregions have the same angle at the center θ as shown in Figure 2. m is determined by θ, and $m = 360/\theta$. Then, each subregion is divided into 12 consecutive bins with a fixed length l. In our approach, every candidate obtained from background subtraction technique is a rectangle, and the direction d_p of a candidate is estimated using the direction of the longer edge of the rectangle. The value of each bin is the number of candidates lying in the region associated with each bin and satisfying a constraint. The constraint is that the difference between the estimated direction d_p of the candidate p and the estimated direction d_i of a candidate i that is not greater than a threshold ϵ, or d_i is between θ_1 and θ_2, the angles of the two edges of the subregion as shown in Figure 2(a). The second part of the constraint aims at solving the problem that the directions of a small number of candidates on the road might be quite different from most of the candidates on that road. In order to decrease the difference of TC between a candidate at the center of a frame and a candidate at the border of a frame, two bins which are symmetric, e.g. the bin k and bin k' in Figure 2(a), are combined into one bin k. The value of the kth bin of candidate p is

$$h_p(k) = \#\{i \in C \wedge i \neq p \wedge i \in R(k) \wedge (|d_p - d_i| \leqslant \epsilon \vee \theta_1 \leqslant d_i \leqslant \theta_2)\}, \quad (1)$$

$$h_p(k') = \#\{i \in C \wedge i \neq p \wedge i \in R(k') \wedge (|d_p - d_i| \leqslant \epsilon \vee \theta_1 \leqslant d_i \leqslant \theta_2)\}, \quad (2)$$

$$h_p(k) = h_p(k) + h_p(k'), \quad (3)$$

where C is the set of candidates of I, $\#$ is the cardinality function, and $R(k)$ is the region associated with the kth bin. So, given the number of subregions m, the dimension of TC is $\frac{m \times 12}{2}$.

Algorithm 1 Building TC for a Candidate

Input:
p: a candidate
$I = \{I_1, I_2, ...I_T\}$: T consecutive frames
C: the set of candidates of I
ϵ: a threshold
Output:
h: the TC, a histogram, for p

for each bin k of h **do**
  $count \leftarrow 0, count' \leftarrow 0$
  for each $i \in C$ **do**
    if $i \in C \wedge i \neq p \wedge i \in R(k) \wedge (|d_p - d_i| \leqslant \epsilon \vee \theta_1 \leqslant d_i \leqslant \theta_2)$ **then**
      $count \leftarrow count + 1$
    end if
    $k' \leftarrow$ the symmetric bin of k
    if $i \in C \wedge i \neq p \wedge i \in R(k') \wedge (|d_p - d_i| \leqslant \epsilon \vee \theta_1 \leqslant d_i \leqslant \theta_2)$ **then**
      $count' \leftarrow count' + 1$
    end if
  end for
  $h(k) \leftarrow count + count'$
end for
return h

Figure 2(b) is the situation that the direction of most candidates on a road cannot be estimated correctly because of the shadow. In this case, the road direction might be quite different from the estimated direction of candidates, but with some rotation of the estimated direction of the candidate; however, the road direction can still be found. The reason for using the candidates in T consecutive frames is that even when the traffic is not busy, the candidates in consecutive frames will be able to cover some consecutive bins which correspond to a consecutive region of the road due to the moving of the vehicles. Also, the subregion is similar to a fan, with the increase of distance from the center of a candidate, where the width of the region associated with a bin also increases, which allows a moderate change in the direction of road. From Figure 1, we can see that positive candidates have obvious patterns. The TC has only two parameters, the number of subregions m which can be determined by θ, the angle at the center, and the length l for dividing each subregion. Experiment in §V tests several θ and l, and there is a large range for choosing these two parameters.

IV. Implementation Details

A. Registration

Given a set I of T consecutive frames, we need to remove the camera motion first. As in [1], we use point-matching based algorithm. Given a reference frame t and another frame $t+i$, we first detect the keypoints in both t and $t+i$ using the Scale-Invariant Feature Transform (SIFT) [31] and extract a SIFT descriptor at each keypoint. Then, the keypoints in $t+i$ are matched with keypoints in t with FLANN (Fast Library for Approximate Nearest Neighbors) [32]. Finally, a robust homography H_t^{t+i} is estimated using RANSAC (RANdom SAmple Consensus) [33]. We use the first frame in I as the reference frame. Other piecewise simplex based approaches are also possible [34].

B. Generating Candidates

In order to detect the moving vehicles, we use background subtraction to first generate the candidate detections. As in [1], we use median image filtering to model the background. Due to the motion of camera, the more frames we use to build the background model, the smaller the active area is. To keep the active area as large as possible and also get a relatively satisfying background model, we use 8 consecutive frames to model the background B. Then, we can obtain the difference image $I_{dt} = |I_t - B|$. Since we use homography for registration, we make an assumption that the scene is planar which is not true for WAMI. So, pixels belonging to the areas that contain out of plane objects, e.g., trees, tall buildings, cannot be well aligned. There is a lot of noise along the edges of these out of plane objects due to parallax error. The work in [1] alleviated this problem by subtracting the gradient of the media background ∇B, i.e., $I_{dt}^r = I_{dt} - \nabla B$. Since we found that due to the misalignment, there is a offset between the pixels in ∇B that have obvious response and the pixels in I_{dt} where noise exists, we adopt a different approach. Given I_{dt} and ∇B, if the value of a pixel at position (i, j) of ∇B is greater than a threshold, we set the value of the corresponding

pixel at (i,j) of I^r_{dt} to 0,

$$I^r_{dt}(i,j) = \begin{cases} 0 & \text{if} \quad \nabla B(i,j) > \delta \\ I_{dt}(i,j) & \text{otherwise} \end{cases} \tag{4}$$

In our implementation, we found that our TC approach is better than the approach used in [1]. After getting I^r_{dt}, we filter out the blobs that are too large or too small, and the remaining blobs are used as the candidates.

C. Classification of Candidates

1) Generalized Multiple Kernel Learing: In order to benefit from both the appearance information and context information, multiple kernel learning(MKL) [13, 14] is used for classification. The main idea of MKL is to learn an optimal combination of a set of kernel matrices,

$$K_{opt} = \prod_k K_k(d_k) \tag{5}$$

The objective of the generalized MKL [13, 14] is to learn a function $f(\boldsymbol{x}) = \boldsymbol{w}^t \boldsymbol{\phi_d}(\boldsymbol{x}) + b$ with the kernel $k_{\boldsymbol{d}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}^t_{\boldsymbol{d}}(\boldsymbol{x}_i)\boldsymbol{\phi_d}(\boldsymbol{x}_j)$. The MKL not only estimates $\boldsymbol{w}$ and b which is the goal of SVM, but also estimates the kernel parameters $\boldsymbol{d}$ from the training data. The above problem can be formulated as the following optimization problem,

$$\begin{aligned} &\underset{\boldsymbol{d}}{\text{Min}} \quad T(\boldsymbol{d}) \quad \text{subject to} \quad \boldsymbol{d} \geq 0 \\ &\text{where} \quad T(\boldsymbol{d}) = \underset{\boldsymbol{w},b}{\text{Min}} \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + \sum_i l(y_i, f(\boldsymbol{x}_i)) + r(\boldsymbol{d}) \end{aligned}$$

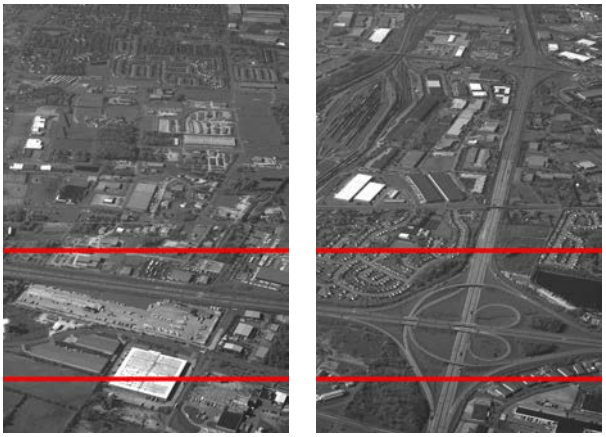
where l is the loss function and r is a regularizer for $\boldsymbol{d}$. The optimization includes two steps, in the outer loop, the kernel parameter $\boldsymbol{d}$ is estimated, and in the inner loop, the parameters of SVM are estimated with fixed kernel. Varma et al. [14] used projected gradient descent approach for the optimization. In order to deal with the inefficiency of the projected gradient descent optimizer, Spectral Projected Gradient(SPG) was proposed in [13] which can handle millions of kernels. In this paper, we choose to use SPG-GMKL (Generalized MKL), and each dimension of the feature vector is treated as a Radial Basis Function (RBF) kernel. Assuming the combination of HOG and TC is M dimensions, the optimal kernel is $k_{\boldsymbol{d}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \prod_{m=1}^{M} e^{-d_m(x_{im}-x_{jm})^2}$.

2) Classification: After background subtraction, we can get a set C of candidates for $I = \{I_1, I_2, ..., I_T\}$. To keep consistent with background subtraction, T is set to 8. For each candidate, we normalize it to 24(width) by 32(height) and use HOG with SPG-GMKL [13] to classify them first, then we only build TC for those candidates that have been classified as positive; and for those candidates that have been classified as negative, the value of each dimension of TC is 0. Then, we combine TC with HOG through SPG-MKL.

V. Experiment

A. Dataset

The dataset we use is Columbus Large Image Format (CLIF) 2006 [17]. The scene of this dataset is a flyover of the Ohio State University (OSU) from a large format monochromatic electro-optical platform which is comprised of a matrix of six cameras and the size of each image is 2672(width) by 4008(height) pixels. Since there is a large area in each image that does not contain an expressway road, a 2672×1200 subregion is used as shown in Figure 3 and Figure 4. The subregion is very challenging, including not only horizontal and vertical express ways, but also an overpass. For the test data, we labeled the vehicles in 102 frames of camera 3, and there are 9364 vehicles in total. For training, we labeled 1730 candidates obtained from background subtraction from 16 frames of camera 1.



(a) (b)

Fig. 3: (a) and (b) are the original images, where the region between the two red lines is the subregion.

B. Building Classifier

In our experiment, three kinds of classifiers are built, using Temporal Context(TC) only, using HOG only, and combing TC and HOG. For HOG, the block size is 12×12, the block stride is 4×4, the cell size is 6×6 and the number of bins is 6. The dimension of HOG is 576. To build the classifiers evaluated on the test data, all the 1730 candidates from camera 1 are used. Since the classification result of HOG is used to build TC, if we use 1730 candidates to train a classifier with HOG which is used to classify the same data to build TC, overfitting exists. So, we select 367 candidates from the 1730 candidates to build the classifier with HOG and MKL, which is used for building TC of the training data. After building TC, we linearly scale each attribute to the range $[0, 1]$ using the tool "svm-scale" provided by LIBSVM [35]. For the regularizer of the kernel weights d of the SPG-GMKL [13], we choose l_2 regularization. ϵ used in Equation 1 and Equation 2 is 10°.

C. Evaluation Metrics

We use the distance between a positive candidate and the groundtruth to judge whether it is a true positive. Given the groundtruth $G = \{g_1, g_2, ..., g_S\}$, for a true positive candidate c, there exists $g \in G$ and the distance between the center of g and the center of c is not greater than 10. For the evaluation of classifiers built with different kinds of features, we use a precision-recall curve and AUC(area under curve) to evaluate the performance. Since we classify the candidates obtained

from background subtraction, the recall can be calculated in two different ways. One way is to use the number of the actual groundtruth, S, i.e., the actual number of vehicles in each image; while another way is to use the number of candidates which are indeed vehicles as groundtruth, S'. In the following comparisons, both kinds of precision-recall curves are given. Without classification, the performance of background subtraction with the size and the gradient constraint is not satisfying, as the precision is only 0.398 at the recall 0.854.

D. Choosing θ and l for TC

TC has two parameters, the angle at the center of the subregion θ and the length l used to divide each subregion. To seek the best θ, we test $15°$, $30°$, $45°$ and $60°$ with fixed $l = 50$. Figure 5, Table I and Table II summarize the results. From the results we can see that when only TC is used, $45°$ performs best. One reasonable explanation is that small θ makes the pattern of TC too complicated, while large θ cannot capture enough information. So, a moderate θ performs best. When the combination of TC and HOG are used, $45°$ also performs best, but the advantage is not obvious.

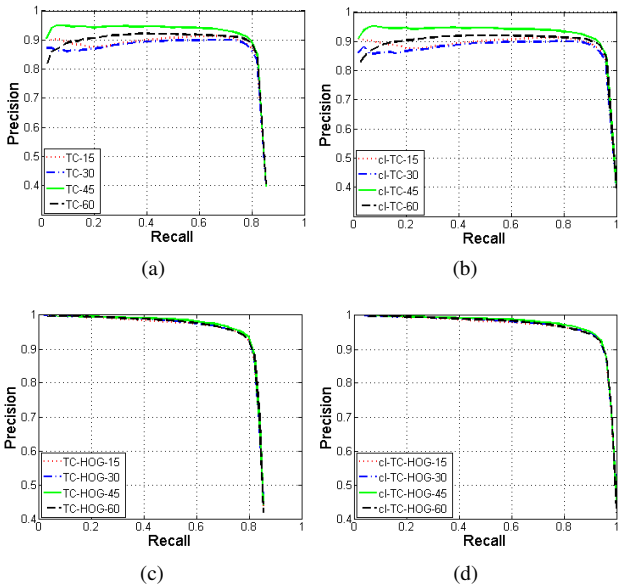


Fig. 5: The precision-recall curves for different θ. (a) and (b) use only TC and are plotted based on S and S' respectively; (c) and (d) use the combination of TC and HOG and are plotted based on S and S' respectively.

TABLE I: The AUC for different θ of TC

	$15°$	$30°$	$45°$	$60°$
S	0.738	0.729	**0.773**	0.748
S'	0.867	0.856	**0.908**	0.878

TABLE II: The AUC for different θ of TC+HOG

	$15°$	$30°$	$45°$	$60°$
S	0.807	0.808	**0.811**	0.808
S'	0.948	0.950	**0.953**	0.949

To seek the best l, we test 30, 40, 50, 60, 70 with fixed $\theta = 45°$. From Table III and Table IV, 30 performs best when only TC is used, and 40 performs best when TC and HOG are combined. However, from Figure 6(c), Table V and Table VI, we can see that when the recall is high, the precision of 50 is a little better. Since a better performance at low recall does not have practical use in the vehicle detection application, 50 is chosen for l. Though 50 is chosen for l, TC is not very sensitive to l, when the recall is high, the difference among 30, 40 and 50 is small.

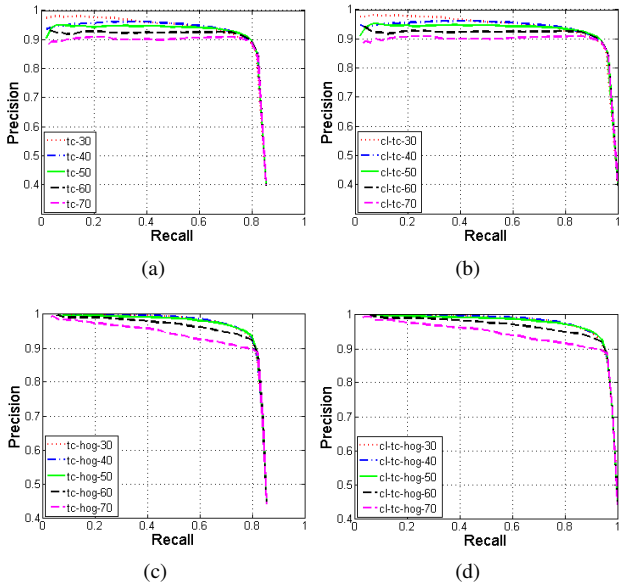


Fig. 6: The precision-recall curves for different l. (a) and (b) use only TC and are plotted based on S and S' respectively; (c)and (d) use the combination of TC and HOG and are plotted based on S and S' respectively.

TABLE III: The AUC for different l of TC.

	30	40	50	60	70
S	**0.783**	0.779	0.773	0.760	0.743
S'	**0.920**	0.915	0.908	0.892	0.873

TABLE IV: The AUC for different l of TC+HOG.

	30	40	50	60	70
S	0.812	**0.813**	0.811	0.802	0.783
S'	0.954	**0.955**	0.953	0.941	0.920

TABLE V: The precision for 3 different recall of TC with different l using S.

	30	40	50	60	70
0.78	0.899	0.905	**0.912**	0.906	0.898
0.80	0.886	0.891	**0.894**	0.893	0.888
0.82	0.856	0.854	0.854	**0.859**	0.854

E. Quantitative and Qualitative Comparison

To demonstrate that TC is useful to remove false positive candidates away from the road, the performance of TC, HOG, and the combination of TC and HOG are evaluated using SPG-GMKL [13]. We normalize the candidates from background subtraction to 24×32, and compute the HOG for each

(a) (b)

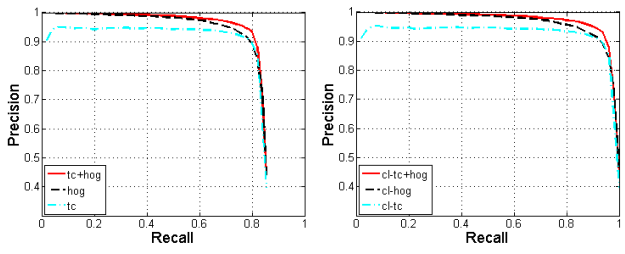
Fig. 4: (a) and (b) are subregions for 3(a) and 3(b) respectively.

TABLE VI: The precision for 3 different recall of TC+HOG with different l using S.

	30	40	50	60	70
0.78	0.942	0.945	**0.946**	0.931	0.900
0.80	0.923	0.928	**0.934**	0.923	0.896
0.82	0.863	0.873	0.881	**0.887**	0.886

candidate. For TC, 45° and 50 are used for θ and l respectively. The dimension of TC is 48. Figure 9, Table VII and Table VIII list the quantitative results. The combination of TC and HOG which can make use of both the appearance and context information outperforms the only use of TC or HOG. When comparing AUC, the advantage of TC+HOG is tiny. However, Figure 9 and Table VII shows that the advantage is obvious when the recall rate is high, which is useful in practice. Compared to the performance of background subtraction with the constraints on the size of candidates and the gradient of background model, the classification can obviously boost the detection performance. Using the combination of TC and HOG with SPG-GMKL [13], the precision can boost to 0.881 at the recall 0.82.

Figure 7 and Figure 8 are some qualitative results. There are very few false positives on the road for TC, HOG and the combination of TC and HOG. TC is useful to remove false positives that are away from the road with a very small number of misclassified candidates on the road; while HOG can almost capture all the vehicles on the road with some false positives away from the road. Figures 7(c), 7(f), 8(c) and 8(f) demonstrate the benefit of the combination of TC and HOG. By making use of both the appearance and context information, we can obviously reduce the number of false positives without sacrificing recall.

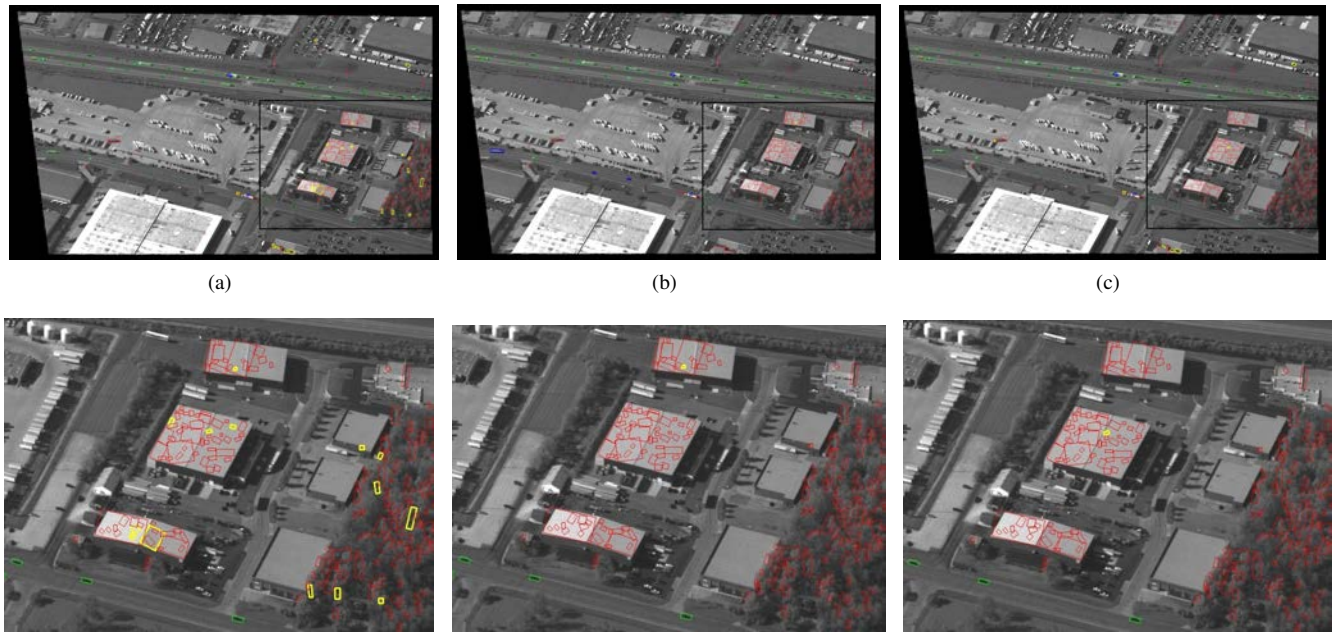


(a) (b)

Fig. 9: The precision-recall curve for TC, HOG and the combination of TC and HOG.(a) is plotted based on S, (b) is plotted based on S'.

TABLE VII: The AUC for TC, HOG and TC+HOG.

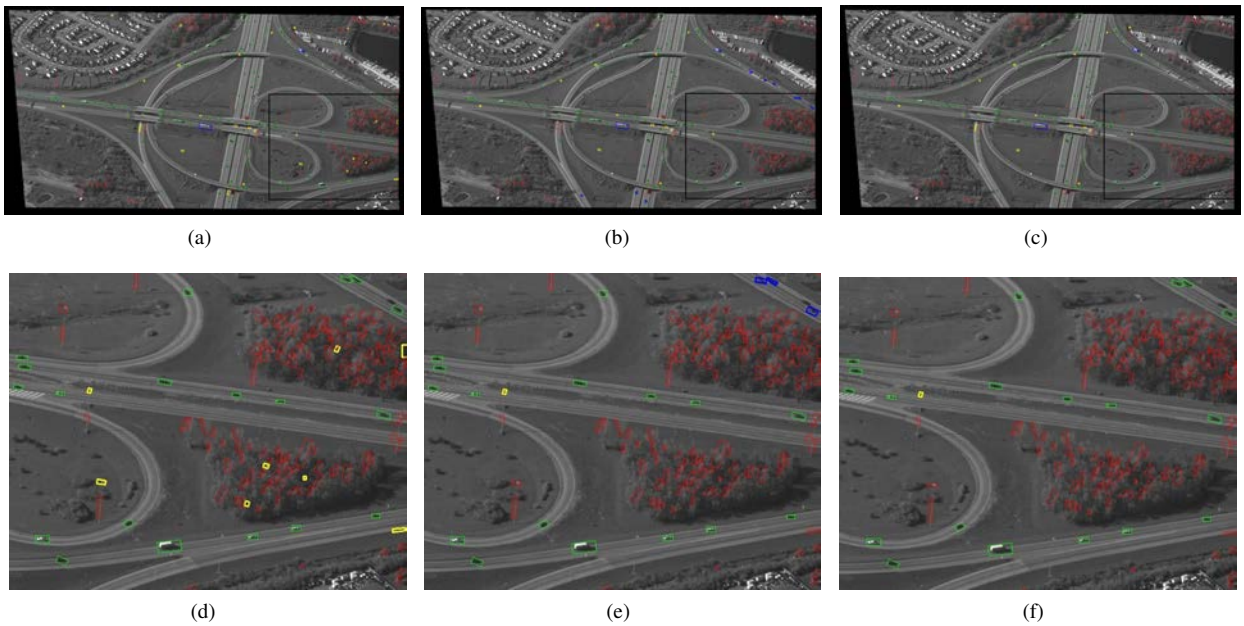
	TC	HOG	TC+HOG
S	0.773	0.805	**0.811**
S'	0.908	0.945	**0.953**



(a) (b) (c)

(d) (e) (f)

Fig. 7: The classification result. The green, yellow, red, and blue bounding boxes indicate true positive, false positive, true negative, false negative respectively. (a), (b), (c) are the results of TC, HOG, TC+HOG respectively. (d),(e),(f) are the enlargement for the part in the black bounding box in (a),(b),(c) respectively.



(a) (b) (c)

(d) (e) (f)

Fig. 8: The classification result. The green, yellow, red, and blue bounding boxes indicate true positive, false positive, true negative, false negative respectively. (a), (b), (c) are the results of TC, HOG, TC+HOG respectively. (d),(e),(f) are the enlargement for the part in the black bounding box in (a),(b),(c) respectively.

TABLE VIII: The precision of TC, HOG and TC+HOG at different recall using S.

	TC	HOG	TC+HOG
0.70	0.932	0.953	**0.970**
0.72	0.929	0.947	**0.966**
0.74	0.925	0.937	**0.960**
0.76	0.920	0.926	**0.955**
0.78	0.912	0.915	**0.946**
0.80	0.894	0.893	**0.934**
0.82	0.854	0.841	**0.881**
0.854	0.399	**0.442**	**0.442**

VI. Conclusion

We propose using the Temporal Context(TC) which can capture the road information in the moving vehicle detection task of wide area motion imagery. To make use of both the appearance and context information for classification, we use multiple kernel learning(MKL) to combine these two kinds of features. In order to demonstrate the effectiveness of the proposed TC, we label 9364 vehicles in 102 frames of the CLIF 2006 dataset, and the experimental results show that TC is very useful to remove false positives away from the road. In the future, we will investigate the idea for building TC further and extend TC to a more comprehensive robust TC method.

References

[1] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *ECCV (3)*, 2010.

[2] J. Prokaj, M. Duchaineau, and G. Medioni, "Inferring tracklets for multi-object tracking," in *Workshop of Aerial Video Processing joint with IEEE CVPR*, 2011.

[3] F. Bunyak, K. Palaniappan, S. K. Nath, and G. Seetharaman, "Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking," *J. Multimedia*, vol. 2, no. 4, pp. 20–33, August 2007.

[4] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *ECCV (1)*, 2008.

[5] A. Jain, A. Gupta, and L. S. Davis, "Learning what and how of contextual models for scene labeling," in *ECCV (4)*, 2010.

[6] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *CVPR*, 2009.

[7] H. Myeong, J. Y. Chang, and K. M. Lee, "Learning object relationships via graph-based context model," in *CVPR*, 2012.

[8] C. Galleguillos, B. McFee, S. J. Belongie, and G. R. G. Lanckriet, "Multi-class object localization by combining local contextual interactions," in *CVPR*, 2010.

[9] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *CVPR*, 2012.

[10] J. Porway, K. Wang, B. Yao, and S. C. Zhu, "A hierarchical and contextual model for aerial image understanding," in *CVPR*, 2008.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *ICCV*, 2007.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR (1)*, 2005.

[13] A. Jain, S. V. N. Vishwanathan, and M. Varma, "Spg-gmkl: generalized multiple kernel learning with a million kernels," in *KDD*, 2012.

[14] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *ICML*, 2009.

[15] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *ICCV*, 2007.

[16] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[17] "Clif 2006," https://www.sdms.afrl.af.mil/index.php?collection =clif2006.

[18] T. Zhao and R. Nevatia, "Car detection in low resolution aerial image," in *ICCV*, 2001.

[19] X. Shi, H. Ling, E. Blasch, and W. Hu, "Context-driven moving vehicle detection in wide area motion imagery," in *Int'l Conf. on Pattern Recognition (ICPR)*, 2012.

[20] J. Xiao, H. Cheng, H. S. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *CVPR*, 2010.

[21] 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 *Int'l Conf. on Information Fusion (FUSION)*, 2012.

[22] K. Palaniappan, F. Bunyak, P. Kumar, I. Ersoy, S. Jaeger, K. Ganguli, A. Haridas, J. Fraser, R. M. Rao, Seetharaman, and G. S., "Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video," in *Proc. of the Int's Conf. on Information Fusion (FUSION)*, 2010.

[23] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan, "Persistent target tracking using likelihood fusion in wide-area and full motion video sequences," in *Proc. of the Int's Conf. on Information Fusion (FUSION)*, 2012.

[24] H. Ling, Y. Wu, E. Blasch, G. Chen, and L. Bai, "Evaluation of visual tracking in extremely low frame rate wide area motion imagery," in *Proc. of the Int's Conf. on Information Fusion (FUSION)*, 2011.

[25] E. Blasch, G. Seetharaman, K. Palaniappan, H. Ling, and G. Chen, "Wide-area motion imagery (wami) exploitation tools for enhanced situation awareness," in *Proc. IEEE Applied Imagery Pattern Recognition (AIPR) Workshop: Computer Vision: Time for Change*, 2012.

[26] X. Shi, P. Li, W. Hu, E. Blasch, and H. Ling, "Using maximum consistency context for multiple target association in wide area traffic scenes," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[27] W.-S. Zheng, S. Gong, and T. Xiang, "Quantifying and transferring contextual information in object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 762–777, 2012.

[28] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *CVPR*, 2012.

[29] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *CVPR*, 2011.

[30] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Int's Conf. on Computer Vision Theory and Application*, 2009.

[33] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[34] E. C. Cho, S. S. Iyengar, G. Seetharaman, R. Holyer, and M. Lybanon, "Velocity vectors for features of sequential oceanographic images."

[35] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.