Anomaly Detection in Traffic Scenes via Spatial-Aware Motion Reconstruction

Yuan Yuan, *Senior Member, IEEE*, Dong Wang, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Anomaly detection from a driver's perspective when driving is important to autonomous vehicles. As a part of Advanced Driver Assistance Systems (ADAS), it can remind the driver about dangers in a timely manner. Compared with traditional studied scenes such as a university campus and market surveillance videos, it is difficult to detect an abnormal event from a driver's perspective due to camera waggle, abidingly moving background, drastic change of vehicle velocity, etc. To tackle these specific problems, this paper proposes a spatial localization constrained sparse coding approach for anomaly detection in traffic scenes, which first measures the abnormality of motion orientation and magnitude, respectively, and then fuses these two aspects to obtain a robust detection result. The main contributions are threefold, as follows. 1) This work describes the motion orientation and magnitude of the object, respectively, in a new way, which is demonstrated to be better than the traditional motion descriptors. 2) The spatial localization of an object is taken into account considering the sparse reconstruction framework, which utilizes the scene's structural information and outperforms the conventional sparse coding methods. 3) Results of motion orientation and magnitude are adaptively weighted and fused by a Bayesian model, which makes the proposed method more robust and able to handle more kinds of abnormal events. The efficiency and effectiveness of the proposed method are validated by testing on nine difficult video sequences that we captured ourselves. Observed from the experimental results, the proposed method is more effective and efficient than the popular competitors and yields a higher performance.**

***Index Terms*—Computer vision, video analysis, anomaly detection, motion analysis, sparse reconstruction, crowded scenes.**

I. Introduction

THERE are many potential dangers when driving, such as unsafe driver behavior, sudden pedestrian crossing, and vehicle overtaking. Fig. 1 shows some typical exemplars having potential dangers. Since the driver's attention can't focus in every second and notice all dangers, many traffic accidents occur every day. Therefore, it is necessary to auto-detecting potential dangers from a driver's perspective, and a surge of interests has been motivated in computer vision community. But it is almost impossible to design a system that can detect faultlessly all kinds of abnormal event, because the anomaly definition might be distinctive in different situations. Therefore, many researchers simplify the problem by focusing on specific objects and events, such as pedestrians, vehicles and crossing behaviors.

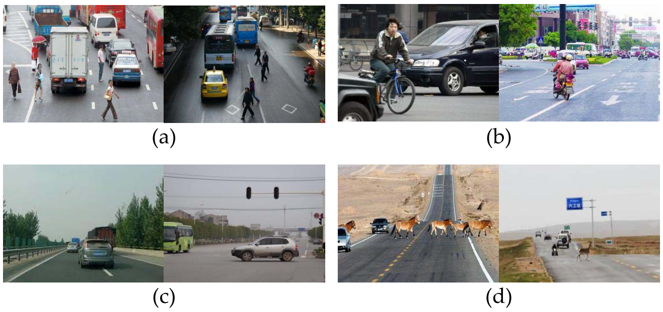


Fig. 1. Typical examples of anomaly in traffic scenes. (a) Pedestrian crossing the road. (b) Cyclists and motorcyclists on the road. (c) Vehicle overtaking. (d) Sudden appearance of animals. It is noticed that the abnormal objects have a different motion pattern compared to their neighboring object.

To tackle the above simplified problem, training object detectors is a straightforward method. To name only a few, Xu *et al.* [1] focus on detecting the sudden crossing pedestrians when driving, and learn a pedestrian detector to detect crossing pedestrians as early as possible. Sivaraman and Trivedi [2] propose a part-based vehicle detector to detect cars when driving. Moreover, to improving accuracy of the detector, Garcia *et al.* [3], [4] fuse vision-based pedestrian detection results and laser data to estimate the frontal pedestrian. Apart from these traditional methods, over recent years, the landscape of computer vision has been drastically altered and pushed forward through the adoption of deep learning, especially the Convolutional Neural Network (CNN) [5]. The CNN-based object detectors achieve state-of-the-art results in almost all object detection benchmarks. As an example, Region-based CNN [6] achieves excellent object detection accuracy by using deep ConvNet to classify object proposals. Based on the similar framework, there are quite a few works to speed up R-CNN such as Spatial Pyramid Pooling networks (SPPnets) [7] and Fast R-CNN [8]. Though the CNN-based object detection method is outstanding in static image, the trained models only capture appearance information and cannot be used to recognize specific actions immediately.

There is another clue to classification of different behaviors by contrast with static image, i.e., object motion information. A slice of papers investigate for action detections in this direction. Early work by Alonso *et al.* [9] detects the overtaking cars

Manuscript received November 29, 2015; revised April 20, 2016 and June 23, 2016; accepted August 10, 2016. Date of publication September 12, 2016; date of current version May 1, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61379094, by the Natural Science Foundation Research Project of Shaanxi Province under Grant 2015JM6264, and by the Fundamental Research Funds for the Central Universities under Grant 3102015BJ(II)JJZ01. The Associate Editor for this paper was J. Zhang. *(Corresponding author: Qi Wang.)*

The authors are with the School of Computer Science and with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2601655

in reference to the motion orientation of vehicles, which is obtained by calculating the optical flow of every frame. Along similar line, Kohler *et al.* [10] propose a Motion Contour image based on HOG-like descriptor (MCHOG) in combination with a SVM learning algorithm that decides within the initial step if a pedestrian at the curb will enter the traffic lane. Aside from these motion flow based methods, object trajectory is another technique for describing object motion information. As an example, Bonnin *et al.* [11] propose a generic model to predict pedestrians crossing behavior in inner-city, which predicts the pedestrain's motion orientation by tracking for a while. However, because object tracking is not credible all the time in fickle scenes, the object trajectory is misleading to object localization. This limitation makes it unfavorable in traffic scene. Besides, the tracking technique usually needs the target to be detected as an initial step, which makes the method also object-related.

A desirable property of a system which is able to identify threats when driving is to disentangle specific object classes. The detector-based and tracking-based methods invariably pour attention into quite a few object. Consequently, this work resorts to the motion flow based method. However, in order to make motion flow based method feasible, there are several difficulties should be considered carefully. First, since the camera is mounted on the moving vehicle, it is almost shaking all the time and the captured video is usually blurred. This makes the estimated motion information noisy and unstable. Second, in contrast to the static camera, the background of scene is all moving due to its relative movement to the camera, which makes the motion patterns of the scene very complex. Additionally, the ever-changing background makes the influence of background more serious. Third, there is some drastic variation of vehicle velocity, aggravating the difference of relative movements between objects. Due to dynamic uncertainty, the same behaviors such as sudden vehicle crossing, may show totally different motion patterns with different vehicle velocities.

In order to tackle the above problems, this work calculates two histograms to represent motion magnitude and orientation respectively, which makes a more comprehensive description of local motion pattern, and the separate descriptors have a clearer expression of motion patterns resulting in resistance of motion noise. Additionally, two anomaly maps are generated by spatial-aware reconstruction, which can alleviate the influence of dynamic background via spatial constraint. Finally, a Bayesian integration model is employed to fuse previously obtained anomaly maps to calculate the final anomaly map, which is robust to the drastic changes of vehicle velocity. Based on the obtained final anomaly map, the abnormal objects can be located.

The reminder of this paper is organized as follows: Section II reviews previous work on anomaly detection in computer vision. The main steps and contributions of the proposed method are clarified briefly in Section III. Section IV describes the strategy for motion region segmentation. Section V proposes the anomaly detection and localization using sparse reconstruction. The Bayesian-based integration method is elaborated in Section VI and experiments and discussions are given Section VII. The conclusion is finally summarized in Section VIII.

II. Related Work

The proposed framework in this paper bears some resemblance to region of interest (ROI) generation and selection methods, and measures the degree of anomaly via sparse reconstruction cost in conjunction with the integration of two motion clues that is inspired by multi-saliency evaluation. Hence the literature review for this work begins from these three aspects.

In the realm of the relative works for ROI generation and selection, there are several efforts [12]–[15] creating a relatively small set of candidate ROIs that cover the objects in the image. The "selective search" algorithm of van de Sande *et al.* [12] computes hierarchical segmentations of superpixel [16] and places bounding boxes around them. EdgeBoxes [13] outputs high-quality rectangular (box) proposals quickly, which are selected readily with a simple box objectness score computed from the contours wholly enclosed in a candidate bounding box. Additionally, BING [14] trains a two stages cascaded SVM [17] to measure generic objectness, and then produces a small set of candidate object windows. Finally, recent R-CNN [15] applies high-capacity convolutional networks to bottom-up region proposals in order to localize and segment objects, and gives more than a 50% relative improvement on PASCAL VOC. Our approach is inspired by the success of these ROI selection methods, and the difference is we filtrate ROIs according to measuring abnormality, rather than objectness.

There are quite a few alternatives to model the degree of anomaly, such as mixture of probabilistic principal component analysis (MPPCA) model [18], social force model [19], sparse basis [20]–[23], etc. However, based on the sparsity of unusual events, more and more sparsity based methods have emerged in this field recently. Cong *et al.* [20] calculate a multiscale histogram of optical flow to represent the local motion patterns for image sequences. Whether a testing sample is abnormal or not is determined by its sparse reconstruction cost, through a weighted linear reconstruction of the over-complete normal basis set. Zhao *et al.* [22] propose a fully unsupervised dynamic sparse coding approach for detecting unusual events in videos based on online sparse reconstructibility of query signals from an automatically learned event dictionary, which forms a sparse coding bases. Moreover, recent research has observed and validated that locality is more essential than sparsity [24]–[26]. The locality-constrained linear coding (LLC) [38] is a great advance in this aspect, which applies locality constraint to select similar basis of local image descriptors. Inspired by this work, we measure abnormality by spatial locality-constrained sparse reconstruction.

For obtaining robust and superior results, integration of multiple clues or factors is usually adopted in computer vision and machine learning community. Because of close similarity between anomaly map and saliency map, we review some work about multi-saliency fusion here. The straightforward and most intuitive scheme is linear fusion. Evangelopoulos *et al.* [27] apply this framework to fuse aural, visual and textual saliency. For more elaborate fusion, a Support Vector Machine is trained and used to predict the quality of each saliency map in [28], and then saliency maps are fused linearly using the quality measure of each map. Besides, Xie *et al.* [29] merge low and

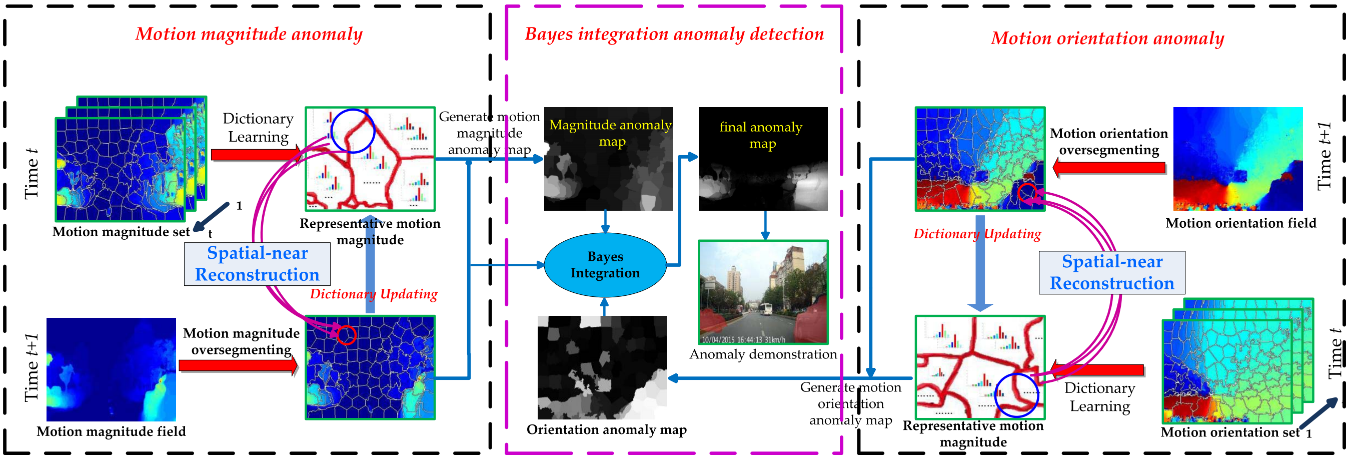


Fig. 2. Pipeline of the proposed method. First, with the obtained motion estimation, which is computed by a state-of-the-art dense flow method [31], the optical flow field is separated into two motion fields, i.e., motion orientation field and motion magnitude field. Then SLIC [32] superpixel segmentation is utilized to oversegment each motion field into superpixels. Second, with the superpixel motion features of both motion fields, this work learns two dictionaries, respectively, for the motion orientation and magnitude and updates the learned dictionaries to adapt to dynamic scenes. The newly observed superpixel motion feature is reconstructed by its top K nearest elements of the corresponding dictionary. The superpixel motion features with a large reconstruction error are not used to update the corresponding dictionary. Third, in order to give a more robust anomaly estimation, this work integrates the obtained two anomaly maps based on Bayes' model, which makes use of the complementarity between motion orientation and magnitude. In the end, the detected anomaly regions are superimposed on the original color image.

mid level visual saliency within the Bayesian framework, which generates more discriminative saliency map. Furthermore, the Bayesian integration method is also employed in [30] and performs better than the conventional integration strategy.

III. Overview

In this paper, an effective anomaly detection method for traffic scenes is designed, which is robust to the change of the camera movement. And the components and contributions of this method is illuminated schematically in this section.

A. Components of the Proposed Method

The main components are illustrated in Fig. 2, with a detailed description as follows.

1) Complementary Motion Description: Given a video sequence, this work calculates the optical flow field of each frame, which represents the motion characteristics of each pixel as a two-dimensional vector. With the obtained optical flow, the motion orientation and magnitude of each pixel is calculated and gathered together to form the motion orientation filed (MOF) and motion magnitude field (MMF) respectively. Since different parts of an object may have similar motion characteristics, the superpixel technique is employed to over-segment the obtained MOF and MMF, which can separate different objects well by preserving coherence of local motion patterns. With the segmented results, this work calculates a histogram for every superpixel to represent its motion orientation and magnitude. Because this technique takes these two aspects into consideration, the proposed method can detect motion orientation and magnitude anomaly simultaneously.

2) Abnormality Measurement Via Spatial-Aware Reconstruction: With the obtained motion orientation and magnitude histogram, this work detects the motion orientation and magnitude anomaly simultaneously via a dictionary-based method. To be specific, this work learns two normal dictionaries respectively for motion orientation and magnitude description by an incremental learning method, which finds the representative samples (histogram of motion orientation or magnitude) in the normal motion pattern set. And then we construct the dictionary via taking them as the bases of the learned dictionary. For the reason that the location of motion feature (i.e., the spatial location of the corresponding superpixel) is essential to anomaly detection in traffic scene, this work reconstructs the newly observed motion feature over the spatial-near subset of the learned dictionary, which is inspired by the locality-constrained linear coding (LLC) [24] method in image classification. Besides, in order to measure the difference of motion features more reasonably, the earth mover's distance (EMD) [33] is employed instead of traditional χ_2 distance. According to the reconstruction cost, two anomaly maps are generated and indicate abnormality of motion orientation and magnitude respectively.

3) Bayesian-Based Integration of Anomaly Detection: As mentioned above, this work measures the abnormality of motion orientation and motion magnitude simultaneously, and the behind idea is that some abnormal behaviors show a different motion orientation but some is motion magnitude, which is mainly caused by drastic changes of vehicle velocity. In order to tackle this problem, we integrate the two anomaly maps based on a Bayesian integration model via adaptive weights, which can make use of the complementarity between these two maps and obtain a robust detection result.

B. Contributions

In this work, we tackle the anomaly detection in traffic scenes via measuring the change of motion orientation and motion magnitude simultaneously and integrating these two complementarity aspects together to relieve the mobile camera problem. Additionally, the proposed method does not need any extra training video to pre-learn a off-line model. The main contributions of this paper are described as follows.

1) Explore different effects of motion orientation and magnitude on anomaly detection respectively and model them using a histogram-based method, which is suitable and reasonable to describe motion patterns in traffic video with mobile camera. Compared with the application scenes of traditional anomaly methods, which usually contain several simple motion patterns because of the static camera, the motion patterns in our scenery are more complex and noisy. The reason behind this is that the camera is shaking when driving and not in a constant velocity. Therefore, in order to increase the discriminability of the descriptor, this work calculates two histograms to represent motion orientation and magnitude respectively, which can eliminate the noise more easily.
2) Propose a spatial-aware spare reconstruction method to measure the abnormity of local motion patterns, which is achieved by reconstructing the newly observed motion pattern over its spatial-near dictionary elements. In previous literatures on anomaly detection, sparse reconstruction is utilized in some efforts, but they almost do not take the spatial information into consideration for the simplicity of application scenes. On the contrary, since the motion patterns in traffic video usually have a strong relationship with its spatial location, we reconstruct them with its spatial-near dictionary elements. It can eliminate the dynamic background influence and outperform the traditional sparse reconstruction method.
3) Introduce a Bayesian integration method to adaptively fuse the anomaly results from motion orientation and magnitude. Since the obtained two results usually have different efforts in different scenarios and are complementary to each other, this work integrates these preliminary results into a final detection result. Compared with the conventional integration strategy, such as addition and multiplication, which usually predetermine the integration weights, the employed Bayesian-based method takes the video content into consideration and allocate integration weights adaptively. Therefore, the Bayesian integration method can better reflect the video content and handle drastic changes of vehicle velocity.

IV. Complementary Motion Description

As we all know, traffic scenes are typically crowded. There is much occlusion when you driving on a road, which makes the trajectory-based approaches infeasible in this situation. As a main alternative, motion-based approaches show a promising result for anomaly detection. Therefore, our proposed approach makes use of motion information instead of tracking individuals in the scene. For describing motion patterns effectively, optical flow method [31] is employed.

A. Superpixel Motion Segmentation

Since motion orientation and magnitude of different parts that belong to one object are homologous, the superpixel technique, which has a powerful ability for preserving image local coherence, is employed to segment different motion regions. To be specific, the optical flow field is separated into motion orientation field and motion magnitude field and the superpixels are obtained from both fields respectively. In detail, as illustrated in Fig. 3, these two motion fields are converted into two gray-scale images, and then SLIC method [32] is employed to over-segment these two "images" because of its low computational cost and high performance.

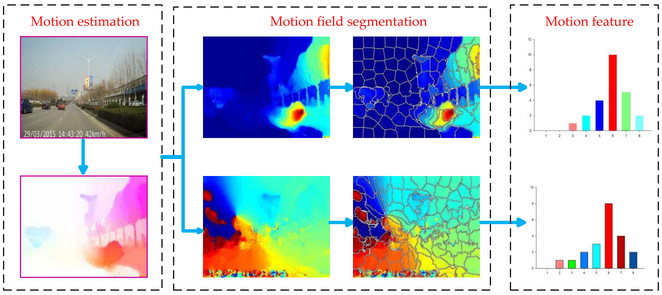


Fig. 3. Flowchart for motion feature extraction.

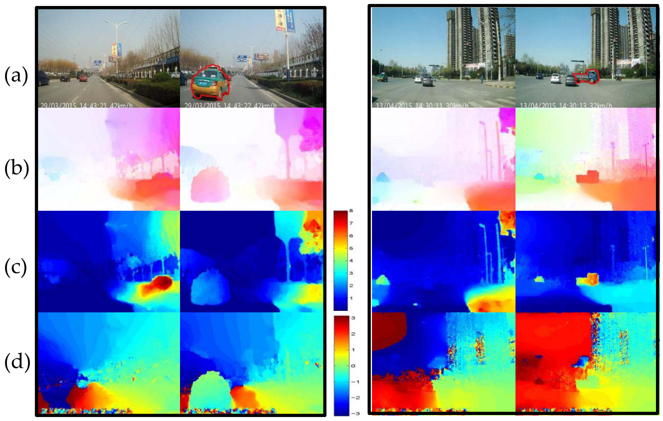


Fig. 4. Two abnormal events in traffic scenes, which show the complementarity of motion orientation and magnitude. (a) Original color image. The red circles denote the abnormal objects. (b) Optical flow field. It represents the motion information of every pixel. (c) Motion magnitude field. Different colors represent different motion magnitudes. (d) Motion orientation field. Different colors represent different motion orientations. It is obvious that motion orientation is more discriminative than motion magnitude in the first scenario, and motion magnitude is more important in the second scenario.

B. Complementary Motion Representation

With the obtained superpixels, a histogram-based descriptor is calculated to represent motion information. The traditional histogram of orientated optical flow (THOOF) [34] sums the magnitude of optical flow according to its orientation followed by a normalization operation, which loses the motion magnitude clue[35]. Considering that the anomaly definition in traffic scenes is usually different, as illustrated in Fig. 4, these two factors are measured simultaneously and integrated to detect anomaly efficiently.

Suppose the motion orientation field image is over-segmented into N superpixels. For i-*th* superpixel $sp_{oi}, i = 1, \ldots, N$, its motion feature is denoted as $y_{oi} \in R^{1\times d}$, where

d indicates the histogram dimension. In addition, the spatial location of i*-th* superpixel centroid is represented by a two-dimensional coordinate $z_{oi} \in R^2$. And the whole set of these superpixels are denoted as Y_o and Z_o. Similarly, the i-th superpixel of motion magnitude field is denoted as sp_{mi}, and its motion feature and spatial location are denoted as y_{mi} and z_{mi}, whose whole sets are denoted as Y_m and Z_m respectively.

The distance measurement between histograms is essential in the histogram-based method. Since the extracted optical flows are inevitably noisy and uncertainty, we adopt the earth mover's distance (EMD) as histogram distance function, which is a well-known robust metric in case of noisy histogram comparison. Specifically, the EMD between histogram P and Q is denoted as:

$$\begin{aligned} EMD(P,Q) &= \min_{f_{ij}>0} \sum_{i=1}^{d}\sum_{j=1}^{d} f_{ij} Dis_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{d} f_{ij} \leq P_j, \quad \sum_{j=1}^{d} f_{ij} \leq Q_i \end{aligned} \tag{1}$$

where f_{ij} denotes a flow from bin $P(i)$ to $Q(j)$, and Dis_{ij} is their ground distance. In general, the ground distance Dis_{ij} can be any distance measurement, such as l_1 and l_2. For simplification, l_1 distance is employed in this work, which is:

$$Dis_{ij} = |i - j|. \tag{2}$$

For reducing computation cost, we utilizes the EMD-l_1 instead of original EMD with l_1 ground distance. The equivalence of these two distances was verified in [33] and the EMD-l_1 has a lower time complexity.

V. Abnormality Measurement via Spatial-Aware Reconstruction

With the separated motion fields, the following task is to detect anomaly by measuring motion inconsistency. This paper formulates the problem of anomaly detection as the reconstruction of the newly observed local motion pattern by the historically collected normal motion patterns. Inspired by the Locality-constrained Linear Coding (LLC), more emphasis is laid on the spatial priors of the dictionary element. Moreover, the spatial prior is essential to alleviate the influence of the background motion patterns. Therefore, the reconstruction error of each superpixel's motion pattern is calculated by its spatially near elements in the dictionary, which is learned [36] via finding the representative normal motion patterns. In the following, the dictionary learning method is introduced firstly, and then the estimation approach of anomaly via spatial neighbor reconstruction is presented.

A. Dictionary Learning via Finding the Representative Motion Patterns

For the camera captured video in traffic scene, the motion pattern has a strong spatial dependency. Certain motion patterns usually arise at specific spatial locations and different regions are prone to show different motion prototypes. In order to describe them, we find a few representative motion patterns and retain its corresponding spatial localization.

To be specific, we measure the superpixel motion pattern's ability to reconstruct other normal motion patterns according to corresponding reconstruction coefficient, which is obtained by minimizing the reconstruction error of the all superpixel motion patterns. Similar to sparse reconstruction problem, the above optimization problem can be formalized as:

$$\min_{C} \frac{1}{2}\|Y - YC\|_F^2 \quad \text{s.t.}\ \|C\|_{1,2} < \varepsilon,\ \text{diag}(C) = \mathbf{0} \tag{3}$$

where $Y \in R^{c \times N}$ denotes the normal superpixels' motion patterns, c the dimensionality of motion feature and N the number of normal superpixels respectively. $\|C\|_{1,2}$ is defined as $\sum_{i=1}^{N} \|c^i\|_2$, which is the sum of l_2 norms of rows in coefficient matrix C. Moreover, the constraint $\text{diag}(C) = \mathbf{0}$ forces the diagonal elements of matrix C to be 0, which is to avoid self-reconstruction.

After solving the above optimization problem, the obtained coefficient matrix C is used to find the representative motion patterns. In detail, the ith row of matrix C denoted as c^i, indicates the reconstruction coefficient of the ith motion feature in matrix Y. Therefore, the motion feature in matrix Y whose corresponding reconstruction coefficient is nonzero has certain efforts to reconstruct other motion features and can be chosen as the representatives. Besides, the optimal coefficient matrix C also provides information about ranking, i.e., relative importance of the representatives to describe the other normal superpixels' motion patterns. More precisely, a representative is essential to reconstruct many superpixels' motion patterns. Thus, its corresponding row in the optimal coefficient matrix C contains many nonzero elements with large values. On the other hand, a representative only takes part in the reconstruction of few superpixels' motion pattern, hence, its corresponding row of C contains a few nonzero elements with smaller values. Therefore, we rank m representatives $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_m}$ according to the relative importance, i.e., $\mathbf{y}_{i_1}$ has the highest rank and $\mathbf{y}_{i_m}$ has the lowest rank. Whenever for the corresponding rows of C we have

$$\|c^{i_1}\|_2 \geq \|c^{i_2}\|_2 \geq \cdots \geq \|c^{i_m}\|_2. \tag{4}$$

According to the ranking result, we select the top M representatives to form the normal dictionary D, and the spatial localizations of the selected representatives denoted as L, are collected in the same order. Finally, the proposed optimization programs in Eq. (3) can be written as

$$\min_{C} \lambda_1 \|C\|_{1,2} + \frac{1}{2}\|Y - YC\|_F^2 \quad \text{s.t.} \quad \text{diag}(C) = \mathbf{0} \tag{5}$$

in practice.

B. Spatial-Aware Reconstruction for Abnormality Measurement

Denote the learned motion orientation dictionary as D_o^t at time t. For a newly observed superpixel motion orientation feature y^t_{oi}, we first calculate the spatial distance between this superpixel and every element in the dictionary, and then select the top K nearest elements to form a new spatial-near dictionary D^t_{ol}. To determine the motion orientation anomaly, the superpixel motion feature y^t_{oi} is reconstructed by D^t_{ol} and the reconstruction cost is viewed as anomaly degree of the examined superpixel. To be specific, the anomaly is defined as:

$$a^t_{oi} = EMD\left(y^t_{oi},\ D^t_{ol}\alpha^t_{oi}\right) \tag{6}$$

where a^t_{oi} is the anomaly degree of the ith superpixel in the flow orientation field and α^t_{oi} is optimal solution of the following sparse reconstruction problem:

$$\alpha^t_{oi} = \arg\min_{\alpha} \left\|y^t_{oi} - D^t_{ol}\alpha\right\|^2_F + \lambda_2\|\alpha\|_1. \tag{7}$$

With the calculated a^t_{oi} of each superpixel, we utilize the max-min normalizer to put a^t_{oi} into the range of [0, 1]. The anomaly degrees of all superpixels are gathered to construct a motion orientation anomaly map S^O_t for the tth frame in motion orientation level.

As for motion magnitude anomaly measurement, since the video is captured on a moving vehicle, their demonstrated motion is relative. This makes the abnormal motion magnitude might be very similar to the normal ones and utilizing the reconstruction strategy is unable to fulfill this task. In order to alleviate this problem, the abnormality of motion magnitude is measured by the difference between abnormal motion magnitude feature and elements of its spatial-near dictionary. Moreover, the highest weight is set to the nearest elements. In detail, suppose the y^t_{mi} denotes the superpixel's motion magnitude feature and D^t_{ml} denotes corresponding spatial-near dictionary, the anomaly degree of superpixel in motion magnitude field is calculated as follows:

$$a^t_{mi} = \frac{1}{K}\sum_{j=1}^{K} w_{ij} \times EMD\left(y^t_{mi}, D^t_{mlj}\right) \tag{8}$$

where D^t_{mlj} denotes the jth element of spatial-near dictionary and $w_{ij} = e^{-\|z^t_{oi}-l^t_{mlj}\|^2_2}$ gives the nearest element the highest weight. Similarly, after the normalization operation, we gather all the anomaly degrees of superpixels to construct a motion magnitude map S^M_t. Besides, for easier combination and visualization of the following Bayesian integration, we harness max-min normalizer to put S^O_t and S^M_t into range [0, 1]. The final anomaly map is generated by integrating these two maps and the integration strategy is described in Section VI.

To alleviate the influence of dynamic scene, the dictionaries need to be updated. We incrementally cumulate the new normal superpixels' motion features Y_{nor} and get the updated training set $Y_{\text{new}} = [D_e\ Y_{\text{nor}}]$, where D_e is the old dictionary. The obtained Y_{new} will subject to the dictionary learning procedure to obtain the updated dictionary every T frame, as discussed in Section V-A.

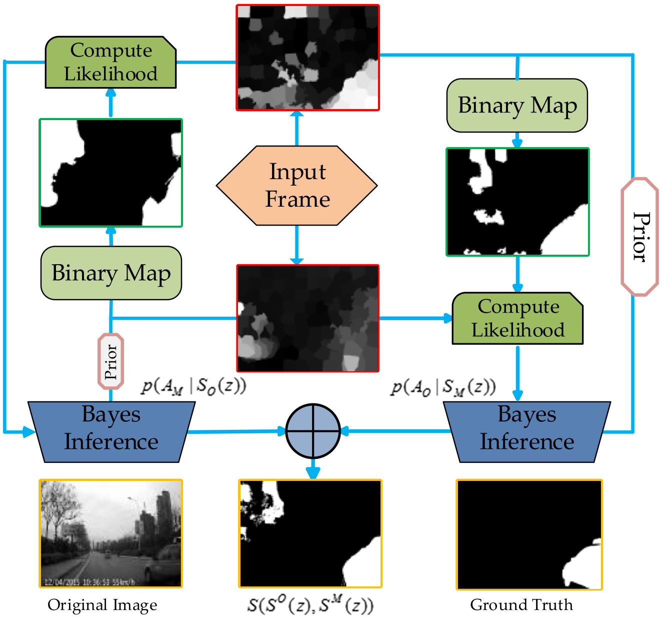


Fig. 5. Bayesian integration of anomaly maps. The two anomaly maps are measured via motion orientation and magnitude, respectively, denoted by S_O and S_M.

VI. Bayesian-Based Integration of Anomaly Detection

For anomaly detection in traffic scenes, the motion orientation and magnitude usually have different efforts in different cases, and are usually complementary to each other. Therefore, this work integrates the previously obtained two anomaly maps to generate the final anomaly map, which can address the change of vehicle velocity problem to some extent. To make full use of the complementarity between motion orientation and magnitude, this work employs an integration method based on Bayesian inference [30]. The posterior probability is formulated as:

$$p\left(A|S(z)\right) = \frac{p\left(S(z)|A\right)p(A)}{p(A)p\left(S(z)|A\right) + (1-p(A))\,p\left(S(z)|N\right)} \tag{9}$$

where the prior probability $p(F)$ is a anomaly map, $A(z)$ is the anomaly degree of pixel z, $p(S(z)|A)$ and $p(S(z)|N)$ represent the detected abnormal and normal likelihood of pixel z, respectively. It is noted that the prior probability and the likelihood probabilities are the key points for the result.

Given the motion orientation anomaly map S^O and the motion magnitude anomaly map S^M, we treat one of them as the prior $S^i (i \in \{M, O\})$ and use the other one $S^j (i \neq j, j \in \{M, O\})$ to compute the likelihood, as shown in Fig. 5. Specifically, first, S^i is thresholded by its mean anomaly value and a binary map B^i is obtained, the regions that having the value of 1 in binary map are denoted as A_i, which means abnormal regions. And the residual regions are normal regions, denoted as N_i. In each region, the likelihood probability at pixel z is calculated as:

$$\begin{aligned} p\left(S^j(z)|A_i\right) &= \frac{N_{A_i b(S^j(z))}}{N_{A_i}} \\ p\left(S^j(z)|N_i\right) &= \frac{N_{N_i b(S^j(z))}}{N_{N_i}} \end{aligned} \tag{10}$$

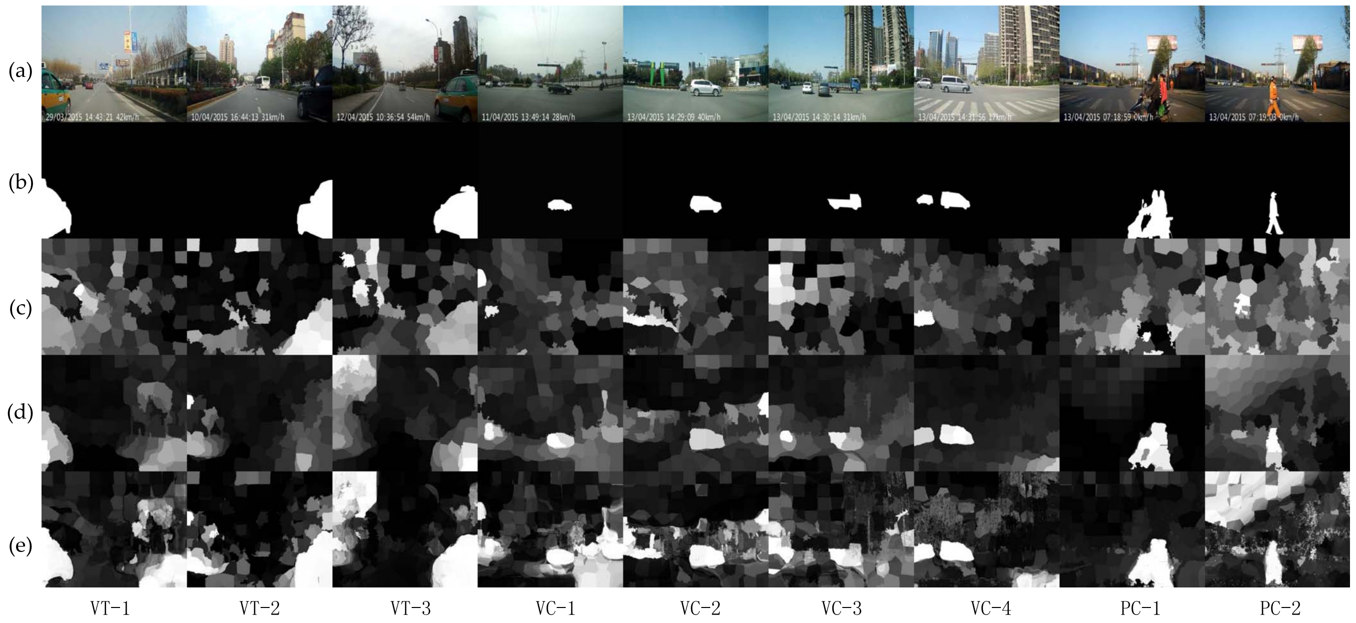


Fig. 6. Typical frameshots of the detected results by different competitors for each sequence. (a) Original color image. (b) Ground-truth anomaly. (c) Motion orientation anomaly map. (d) Motion magnitude anomaly map. (e) Integrated anomaly map.

where N_{A_i} and N_{N_i} are the number of the pixels in the detected abnormal region A_i and the normal region N_i in motion orientation map S^i. Moreover, the range [0, 1] divides into m intervals, and thus the i-th, $(i = 1, 2, \ldots, m)$ interval is $[(i-1)/m, i/m]$. $b(s^j(z))$ represents the interval where $s^j(z)$ falls into its range. $N_{A_i b(S^j(z))}$ denotes the number of detected abnormal region's pixels whose value falls into $b(s^j(z))$. Similarly, $N_{N_i b(S^j(z))}$ represents the number of normal region's pixels whose values fall into $b(s^j(z))$.

Consequently, the posterior probability is computed with S^i as the prior by

$$p\left(A_i|S^j(z)\right) = \frac{S^i(z)p\left(S^j(z)|A_i\right)}{S^i(z)p\left(S^j(z)|A_i\right) + (1 - S^i(z))\,p\left(S^j(z)|N_i\right)}. \tag{11}$$

Similarly, we can also get $p(A_j|S^i(z))$ by treating the two maps as the other. After obtaining the two posterior probabilities and specifying i, j with O, M, we compute an integrated anomaly map $S(S^O(z), S^M(z))$, based on Bayesian integration:

$$S\left(S^O(z), S^M(z)\right) = \left(p\left(A_O|S^M(z)\right) + p\left(A_M|S^O(z)\right)\right)/2. \tag{12}$$

The proposed Bayesian integration of anomaly maps is illustrated in Fig 5. It should be noted that Bayesian integration serve these two maps as the prior in turn and cooperate with each other in an effective manner, which uniformly highlights abnormal objects in a frame.

VII. Experiments and Discussion

In this section, we first introduce the datasets and implementation setups for the experiments. Then for demonstrating the effectiveness of the proposed method, we conduct experiments and compare the results with other competitors. Finally, analyses and discussions are made to explain the experimental results.

A. Datasets

Since the publicly available datasets are almost captured by a static camera, such as the car accident [37] dataset and QMUL Junction [38] dataset, this paper provides a dataset consisted of nine driving videos, which contains several kinds of abnormal events. The videos are captured by a vehicle mounted camera for daily driving, and its view of angle is consistent with the driver's. The anomaly that we considered here is a kind of threats, which have potential dangers, such as vehicle overtaking. To be more specific, based on the anomaly types, the captured video sequences can be divided into three categories: 1) "Three sequences have the vehicle overtaking (VT) behavior (We name them as *VT-1*,*VT-2*, and *VT-3*)", 2) "Four sequences consist of vehicle crossing (VC) behavior (They are named as *VC-1*, *VC-2*, *VC-3*, and *VC-4*)", 3) "Two sequences contain pedestrian crossing and motorcyclists crossing (PC) behaviors (They are denoted as *PC-1* and *PC-2*)". Due to the online application of our method, we do not split the overall dataset into training and test part. And the first 10 frames of sequences, which are always normal situation, are treated as training data for this sequence and the rest are utilized to test. There are 180 frames in each sequence averagely, and the frameshots of the video sequences are demonstrated in Fig. 6, some of which are very difficult for road anomaly detection because of complex background. In the captured dataset, the resolution of each frame is 480 × 640. The ground truth of each video sequence is manually labeled by ourselves.

B. Implementation Setup

1) Metrics: In order to prove the efficiency of the proposed method, the qualitative and quantitative evaluations are both considered. For qualitative evaluation, we demonstrate several typical snapshots of the detected anomaly in each video sequence. As for the quantitative evaluation, pixel-wise receiver

TABLE I
AUC (%) COMPARISON OF DIFFERENT DESCRIPTORS AND CLASSIFICATION METHODS FOR A CLEAR AND FAIRER COMPARISON. THE BOLD ONE IS THE BEST RESULT

Category	**Method**							
	SVM-THOOF	IF-THOOF	SRC-THOOF	SSRC-THOOF	SVM-TPMD	IF-TPMD	SRC-TPMD	SSRC-TPMD(our)
VT	65.94	69.96	72.75	81.89	70.88	78.91	68.47	**85.38**
VC	80.96	75.13	83.64	81.43	79.73	**89.23**	83.54	88.44
PC	78.16	87.54	82.75	88.65	74.47	**92.27**	91.16	90.61
Average	75.34	76.16	79.80	83.19	75.61	86.46	80.21	**87.90**

of characteristics (ROC) and area under ROC (AUC) are employed. Among them, ROC represents the detection ability of the proposed method, and its indexes are specified as:

$$TPR = TP/P, FPR = FP/N \tag{13}$$

where TP denotes the number of the pixels truly detected, FP is the number of pixels falsely detected, P and N represent the positive pixel number and negative pixel number, respectively.

2) Parameters: In our work, the SLIC superpixel [32] is employed, in which δ represents the compactness and N the number of the superpixels. The larger δ is, the more compact the superpixels are. In this paper, δ and N are set as 10 and 125 for all sequences respectively. The dimension of the motion feature c is specified as 30. Furthermore, the parameters λ_1 and λ_2 in Eq. (5) and Eq. (7) are set as 0.5 and 0.5 in all experiments, respectively. The number of basis M in the dictionary is set as 300 and the dictionary updating period T is set as 5, which makes the dictionary over-complete all the time. Additionally, the size of spatial-near dictionary K is set as 10.

3) Comparisons: Since the proposed method is fulfilled by the collaboration of the motion orientation and magnitude, the effectiveness of motion anomaly detection technique is firstly evaluated. In order to demonstrate the advantage of the proposed two-path motion description method, denoted as TPMD. we replace the proposed motion histograms with traditional histogram of oriented optical flow (THOOF) and measure the abnormality based on the proposed spatial-aware sparse reconstruction (SSRC-THOOF). Apart from spatial-near sparse reconstruction, we also make a comparison with other popular one class classification method. To be specific, we investigate one class SVM and Isolation Forest (IF) [39], which is a popular anomaly detection model based on random forest. These two variants are referred as SVM-THOOF and IF-THOOF respectively. Similarly, we retain TPMD and replace the proposed spatial-aware reconstruction method with traditional sparse reconstruction (SRC) [20], one class SVM and Isolation Forest (IF),which are denoted by SRC-TPMD, SVM-TPMD and IF-TPMD respectively. It should be noted that these three variants do not take the spatial information into consideration. Finally, we refer to our method as SSRC-TPMD and make a comparison between performances the proposed method and the above two variants and do some analysis according to the results.

As the second part of the proposed method, we integrate the two aspects to get the final result. To further validate the proposed integration method, we compare the detection results without integration and with different integration methods.

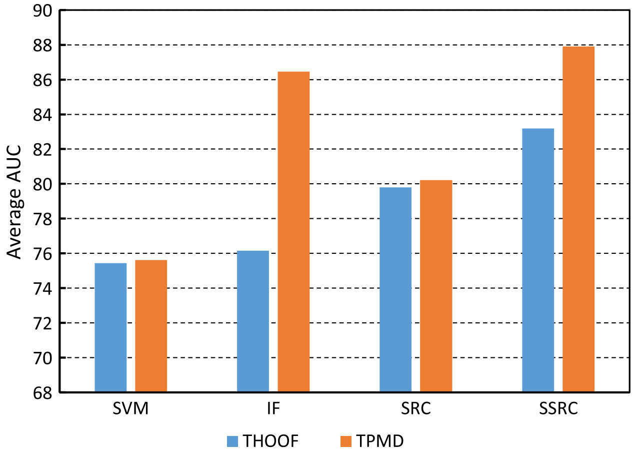


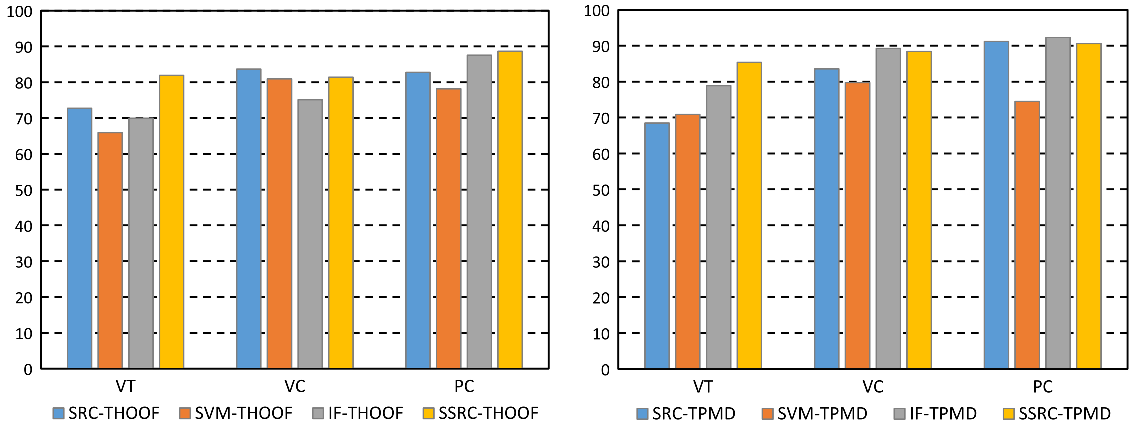
Fig. 7. AUC value comparison of SVM, IF, SRC [20], and our method for each sequence.

To be specific, the competitors are motion Magnitude (M) detection result, motion Orientation (O) detection result, integration result using inner-product of motion magnitude and motion orientation (MO), integration result using our Bayes' model (B-MO).

Last but not the least, for demonstrating the superiority of our method, it is in comparison with recent region proposal-based object detector Faster-RCNN [40], which outperforms significantly traditional object detection method. And it is noted that region proposal-based object detection technique can boost our system and achieve a higher performance.

C. Evaluation of Motion Anomaly Detection

1) Descriptor Comparison: The first experiment evaluates the benefits of the two-path motion descriptor (TPMD). The THOOF descriptor [34] is proposed to describe motion characteristic of sequences, and it pours attention into motion direction information [35]. Our motion utilization strategy is inspired by THOOF, and compute another histogram to describe motion energy information precisely. Therefore, in order to justify the superiority of the proposed TPMD, we combine these two descriptors with several popular classifiers, and average AUC values for each behavior category are listed in Table I. For a better visualized comparison, Fig. 7 illustrates the difference between THOOF and TPMD, and the performance of every method is represented by the average AUC value over the total nine sequences. It can be seen that every TPMD-based method



(a)

(b)

Fig. 8. Performance of all competitors. For a detailed comparison, the average AUC value of every category is plotted. (a) Performance of THOOF-based variants. (b) Performance of TPMD-based variants.

outperforms the corresponding THOOF-based variants, and there are 3.9% improvement medially. From this performance comparison, the superiority of our TPMD is apparently verified.

2) Classifier Comparison: We next investigated the advantage of the spatial-aware sparse reconstruction, with the Bayesian integration method. Apart from sparse reconstruction, there are a slice of wide-used classification methods, such as Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest (RF). Because ANN is usually utilized to classify two or more classes, and it is not suited for one class problem, in addition to traditional sparse reconstruction (SRC) [20], the SVM and RF are selected as competitors. In detail, traditional SRC, one-class SVM in [41] and Isolation Forest (IF) [39] replace the sparse-aware reconstruction, and other parts stay the same. IF explores the concept of isolation with random forest for anomaly detection and achieves pleasurable performances in many application. It should be noted that these three competitors do not take spatial information into consideration. The performance of overall dataset is summarized in Fig. 8. From the shown results, our method generates favorable accuracy for every behavior category regardless of the adopted descriptor. To be specific, in the right sub-figure in Fig. 8, our method performs best for VT behavior, and is comparable with best performer for other behaviors. A strong competitor is IF-TPMD and generates superior results in several sequences, but it is not robust to anomaly type. In particular, the IF classification method performs worse than our method in detecting vehicle overtaking, while our method is independent of specific events. Moreover, because our spatial-aware sparse reconstruction makes modification to traditional SRC, we conduct a comparison between SRC and SSRC. As shown in Fig. 8, SSRC significantly outperforms SRC in almost every behaviour (improvement of AUC by as much as 7 percent), and this suggests the spatial information is crucial to higher accuracy.

Similar conclusion comes with the left sub-figure in Fig. 8, where the THOOF is treated as motion descriptor. In addition, the SVM-based variants perform worse than others whatever the descriptor was used. This is not totally surprising, given the instability of optical flow. In other words, the noise of optical flow, which is caused by camera motion and dynamic background, makes SVM ineffective in this case. That is to say, our method can eliminate the influence of noise.

TABLE II
AUC (%) COMPARISON OF DIFFERENT CLUES AND INTEGRATION METHODS FOR A CLEAR AND FAIRER COMPARISON. THE BOLD ONE IS THE BEST RESULT

Sequence	Superpixels	O	M	MO	B-MO
VT-1	125	85.02	77.33	**85.16**	81.68
VT-2	125	76.07	81.79	82.17	**87.56**
VT-3	125	**89.03**	76.97	86.53	86.91
VC-1	125	50.97	81.67	59.89	**89.68**
VC-2	125	53.16	**90.03**	80.73	89.18
VC-3	125	39.20	80.64	63.59	**83.33**
VC-4	125	69.96	89.83	90.06	**91.54**
PC-1	125	57.11	83.96	73.10	**87.33**
PC-2	125	54.55	84.55	78.81	**93.89**
Average	-	63.89	82.97	77.78	**87.90**

D. Evaluation of Integration Method

To further explore the effectiveness of the Bayesian integrated model, the performance comparisons are presented in Table II. It can be seen that the Bayesian integration model is superior to the other integration techniques. In addition, we also make a comparison between motion magnitude (M) and motion orientation (O) anomaly detection result, and it is noticed that motion magnitude and orientation have different importance in different sequences. Specifically, for the sequences containing vehicle overtaking behavior, (i.e., VT-1, VT-2, and VT-3,) the motion orientation anomaly detection result is usually superior to the motion magnitude anomaly detection result, i.e., $O > M$. The reason is that the motion orientation of abnormal object is very different from the background or normal object. However, the motion magnitude may be very similar to background. But on the other hand, the motion magnitude anomaly detection result has a higher performance in other sequences. The reason is that motion magnitude of abnormal object is very different from background or normal object, but the motion orientation not. The above phenomenon is caused by the different relative

TABLE III
AUC (%) COMPARISON OF SSRC-TPMD AND FASTER-RCNN FOR A CLEAR AND FAIRER COMPARISON. THE BOLD ONE IS THE AVERAGE VALUE

	AVERAGE
Faster-RCNN	**86.15**
SSRC-TPMD	**87.90**
Faster-RCNN+SSRC-TPMD	**93.11**

speed between the abnormal object and the mobile camera. Generally, the overtaking vehicle usually has a faster speed than the camera. Therefore, the estimated optical flow can represent motion magnitude and motion orientation well. However, for vehicle crossing behavior, the crossing objects usually have a slower speed than the camera. Therefore, the estimated optical flow can only represent motion magnitude well, as illustrated in Fig. 4. Besides, it is noticed that the motion magnitude anomaly detection result has a high performance in all sequences, and it demonstrates the proposed motion magnitude descriptor is effective. In general, the method using only motion magnitude or motion orientation can not handle all kinds of abnormal events in traffic scenes because of the different relative speeds between abnormal object and the camera. In order to make use of these two aspects simultaneously, this work reasonably integrates both detection results.

For this purpose, this work integrates the motion magnitude and orientation detection results based on a Bayesian model. In order to demonstrate its effectiveness of the Bayesian model, we make a comparison between Bayesian integration (B-MO) and naive integration technique (MO), which is achieved by making inner-product using both aspects. It is manifest in Table II that the performance of MO sometimes is lower than M or O (for example, VC-1, VC-2, VC-3). This implies that the naive integration technique can not boost the performance, but weaken it. The reason is that a high performance using inner-product needs high performances in both aspects, but it is impossible for some sequences to get satisfying results in both aspects. In order to make use of their complementarity, this work integrates both detection results based on Bayesian model. From Table II, it can be seen that the performances based on Bayesian model are almost the highest in most sequences. Therefore, the integration technique can generate a high performance even though one single aspect has a very low performance. According to the above analysis, we can conclude that the Bayesian integration model is better than the naive method.

E. Performance Comparison

Recently, region proposal technique achieves a great success in detecting objects from a image and is adopted in different works such as Markus Enzweiler's pedestrian detector [42], Will Zou's work on regionlets [43], etc. For demonstrating the superiority of our method, region proposal-based object detector is regarded as competitor. Specifically, the Faster-RCNN is tested on dataset and its performances are listed in Table III. There are several reasons behind selecting Faster-RCNN as competitor, in the first place, the CNN-based Faster R-CNN achieves state-of-the-art performances on almost all public object detection datasets and outperforms Markus Enzweiler's pedestrian detector as well as Will Zou's work on regionlets. There is one more point, I should touch on, that traditional object detection methods are very dependent on specific dataset and is difficult to transfer to another dataset. Therefore, because of insufficient training data of our dataset, Faster-RCNN is our best choice. The last but not the least, albeit we do not fine-tune Faster-RCNN with our own data, the pre-trained model is robust to changing scenes and generates a promising results in our dataset. As shown in Table III, the performance of Faster-RCNN is inferior to our method with only 1 percent, and superiority of our algorithm is demonstrated.

Furthermore, because only appearance information is processed in Faster-RCNN, it is beneficial to incorporate it into our method, which just utilizes the motion information. From Table III, there is a significantly improvement after incorporation. As for incorporating strategy, we just add the object detection score on anomaly map and re-normalize it into range [0, 1].

F. Discussions

1) Range of Moving Objects' Speed: The motion estimation in our algorithm is highly dependent on the object's speed, and one basic assumption behind optical flow method is that object's movement is small between continuous two frames. Therefore, it is important to specify the range of moving objects' speed. Because we can not estimate the speeds of all objects in the scene accurately, we just record the speed of the camera. Moreover, there is two point that can explain the rationality of replacing objects' speed with camera's. First, the moving objects is quite fewer than static objects in the video frames, and their speeds in the video is just camera speed. Moreover, the moving objects's speeds in the videos are usually lower than static objects', and the reason behind this that objects almost are moving in the same direction as camera. Therefore, camera speed usually represents the highest speed in the frame and can be used to specify the range of moving objects' speed. Another important reason behind the collection of camera speed is that the absolute speeds of objects are useless. In detail, due to the mobile camera, the object moving speed in captured video is relative speed. For example, the static building is moving at 40 km/h in video when camera speed is 40 km/h. Therefore, instead of specifications of the range of moving objects' speed, camera speed, which is obtained according to the vehicle's speedometer, is recorded to explain the system's robustness to motion speed. Numerically, the camera speed varies from 0 km/h to 60 km/h in dataset video, which almost cover the highest speed limitation in urban road, and there is no problem with our system in this speed range. The effectiveness of our method with higher camera speed is not probed now, and a deeper investigation will be done in the future.

2) Runtime: In this paper, our method is achieved by a MATLAB-implementation on a machine with Intel i5-3470 3.2GHz CPU and 4 GB RAM. The main consumption is taken by SLIC superpixel segmentation whose average runtime at 125 superpixels is about 0.353 s. The spatial-aware reconstruction is very fast and only costs 0.083 s, and Bayesian integration of two anomaly maps takes away 0.169 s. Despite our work requires computing two anomaly maps, the superpixel segmentation and spatial-aware reconstruction are running in parallel, and will not double the time. Therefore, the total average runtime of this work is 0.605 s without code optimization. Albeit our method cannot achieve real-time speed, it is faster than many pedestrian detectors, such as ChnFtrs (0.845 s) [44], LatSvm-V2 (1.589 s) [45] and MultiFtr+CSS (37 s) [46]. For a real-time consideration, we will use some accelerating strategy to make the method perform in real-time.

VIII. Conclusion

This work addresses the problem of anomaly detection in traffic scenes from a driver's perspective, which is important to autonomous vehicles in intelligent transportation systems. In order to tackle three main difficulties caused by the mobile camera, this work describes motion magnitude and orientation respectively, and by measuring the abnormality of these two aspects simultaneously in conjunction with an adaptively weighted integration, the proposed method can alleviate the influence of the ever-changing scene and camera movement. Specifically, a new motion descriptor is presented to represent the motion magnitude and orientation by calculating a histogram respectively. It performs better than THOOF, which only describes the motion orientation information. With this motion descriptor, the motion anomaly is measured by the reconstruction cost of the spatial-near dictionary, and then these two clues are integrated by a Bayesian model to get a robust result. From the experimental results, the effectiveness and efficiency of the proposed method are proved. Some conclusions can be summarized through this work: 1) For describing the motion information more effectively, the calculated two motion histograms can describe motion magnitude and motion orientation respectively, and it is better than the THOOF. 2) Compared with the traditional anomaly detection, the spatial locations of motion patterns play an essential role in traffic scene anomaly detection. In order to utilize this spatial location information, this work measures the abnormality of the motion orientation and magnitude by reconstructing it over its spatial-near dictionary, and the experimental results demonstrates the rationality of the proposed method. Moreover, the influence of dynamic background is eliminated to some extent. 3) With the obtained two motion anomaly maps, this work fuses them based on a Bayesian-based integration method, which makes use of the complementarity of the two anomaly maps and the obtained result is robust to the change of vehicle velocity.

In the future, we would like to use more clues, for example, near-infrared information, depth information and so on, to improve the performance and robustness of the proposed method. Based on these new information, we would like to extend our method to handle more kinds of abnormal events. The key point is how to use these clues reasonably and integrate them efficiently.

References

[1] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossings for driving assistance systems," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 42, no. 3, pp. 729–739, Jun. 2012.

[2] S. Sivaraman and M. Trivedi, "Real-time vehicle detection using parts at intersections," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 1519–1524.

[3] F. García, A. de la Escalera, J. M. Armingol, J. G. Herrero, and J. Llinas, "Fusion based safety application for pedestrian detection with danger estimation," in *Proc. IEEE Int. Conf. Inf. Fus.*, Chicago, IL, USA, 2011, pp. 1–8.

[4] F. Garcia, B. Musleh, A. de la Escalera, and J. Armingol, "Fusion procedure for pedestrian detection based on laser scanner and computer vision," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2011, pp. 1325–1330.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 580–587.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[8] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[9] J. D. Alonso, E. R. Vidal, A. Rotter, and M. Mühlenberg, "Lane-change decision aid system based on motion-driven vehicle tracking," *IEEE Trans. Veh. Technol.*, vol. 57, no. 5, pp. 2736–2746, Sep. 2008.

[10] S. Kohler, M. Goldhammer, S. Bauer, K. Doll, U. Brunsmann, and K. Dietmayer, "Early detection of the pedestrian's intention to cross the street," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Sep. 2012, pp. 1759–1764.

[11] S. Bonnin, T. Weisswange, F. Kummert, and J. Schmuedderich, "Pedestrian crossing prediction using multiple context-based models," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2014, pp. 378–385.

[12] J. R. R. Uijlings, K. E. A. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[13] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[14] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 3286–3293.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[16] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[17] Z. Zhang, J. Warrell, and P. H. Torr, "Proposal generation for object detection using cascaded ranking SVMs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1497–1504.

[18] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[19] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 935–942.

[20] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, no. 7, pp. 1851–1864, 2013.

[21] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2720–2727.

[22] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 3313–3320.

[23] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 631–645, Apr. 2014.

[24] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.

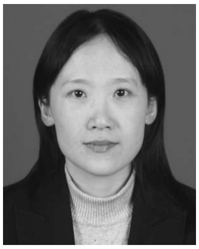
[25] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2223–2231.

[26] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2486–2493.

[27] G. Evangelopoulos *et al.*, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[28] J. Yang *et al.*, "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1070–1083, Jun. 2016.

[29] Y. Xie, H. Lu, and M. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.
[30] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2976–2983.
[31] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.
[32] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[33] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 5, pp. 840–853, May 2007.
[34] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1932–1939.
[35] Y. C. J. Y. J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 3449–3456.
[36] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 1600–1607.
[37] W. Sultani and J. Y. Choi, "Abnormal traffic detection using intelligent driver model," in *Proc. Int. Conf. Pattern Recog.*, 2010, pp. 324–327.
[38] C. C. Loy, T. Xiang, and S. Gong, "Modelling multi-object activity by Gaussian processes," in *Proc. British Mach. Vis. Conf.*, 2009, pp. 1–11.
[39] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.
[40] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[41] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[42] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.
[43] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 17–24.
[44] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.
[45] P. F. Felzenszwalb, R. B. Girshick, D. 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, Sep. 2010.
[46] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1030–1037.



Yuan Yuan (M'05–SM'09) is a Full Professor with the School of Computer Science and with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored more than 150 papers, including about 100 papers in reputable journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, as well as conference papers in Computer Vision and Pattern Recognition, British Machine Vision Conference, International Conference on Image Processing, and International Conference on Acoustics, Speech, and Signal Processing. Her research interests include visual information processing and image/video content analysis.

Dong Wang received the B.E. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2015. He is currently working toward the Ph.D. degree with the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University.

His research interests include computer vision and pattern recognition.

Qi Wang (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently an Associate Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.