Neurocomputing 282 (2018) 202–217

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Incrementally perceiving hazards in driving

Yuan Yuan[a], Jianwu Fang[b,c], Qi Wang[d,*]

[a] Center for OPTical IMagery Analysis and Learning and School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[b] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
[c] School of Electronic & Control Engineering, Chang'an University, Xi'an 710064, China
[d] School of Computer Science, and Center for OPTical IMagery Analysis and Learning, and Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Article history:
Received 8 February 2017
Revised 14 September 2017
Accepted 7 December 2017
Available online 15 December 2017

Communicated by Marco Cristani

Keywords:
Computer vision
Hazards detection
Motion analysis
Saliency evaluation
Bayesian integration

ABSTRACT

Perceiving hazards on road is significantly important because hazards have large tendency to cause vehicle crash. For this purpose, the feedbacks of more than one hundred drivers with different experience for safe driving are gathered. The obtained feedbacks indicate that the irregular motion behaviour, such as crossing or overtaking of traffic participants, and low illumination condition are highly threatening to drivers. Motivated by that, this paper fulfills the hazards detection by involving motion, color, near-infrared, and depth clues of traffic scene. Specifically, an incremental motion consistency measurement model is firstly built to infer the irregular motion behaviours, which is achieved by incremental graph regularized least soft-threshold squares (GRLSS) incorporating the better Laplacian distribution of the noise estimation in optical flow into the motion modeling. Second, multi-source cues are adaptively weighted and fused by a saliency based Bayesian integrated model for arousing driver's attention when potential hazards appears, which can better reflect the video content and select the better band(s) for hazards prediction in different illumination conditions. Finally, the superiority of the proposed method relating to other competitors is verified by testing on twelve difficult video clips captured by ourselves, which contain color, near-infrared and recovered depth simultaneously and no registration or frame alignment is needed.

1. Introduction

The main goal of this work aims to implement the hazardous object detection in driving in an incremental manner, which only needs a small portion of training frames at the beginning of each video clip. By that we want to exploit the ability for detecting hazards in a certain video by exploiting itself. It is promising when there is no available well labeled training data for the valuable applications existing ambiguous definition [1], such as hazards. From the "Global status report on road safety 2015" launched by World Health Organization (WHO), about 1.25 million [2] people die each year owing to the numerous road traffic crashes, and half of these dying are "vulnerable road users" [2]: pedestrians, cyclists, motorcyclists and other moving objects. Fig. 1 demonstrates some typical examples having potential hazards. Analyzing the reason for this phenomenon [3], auto drivers are believed to be responsible for the fatal crash in about 92% traffic deaths. Major crash causing factors are *speeding, careless driving, driving in the wrong lane*, and *driving after drinking alcohol*.

Over the past decades, many researchers have dedicated significant efforts to the development of the Advanced Driver Assistance Systems (ADAS) and autonomous driving [4–6]. However, these techniques are still insufficient to achieve a fully non-human driving system [7]. The challenges are mainly caused by the dynamic, unpredictable traffic scenes being rich of uncertainty .

1.1. Motivations

Facing the hazards prediction, we start a questionnaire investigation for safe driving, by which we want to derive the most hazardous behavior that drivers want to avoid in driving. To be specific, we prepared four questions: (1) "*Which options of behaviour are dangerous when you are driving on highway?*" (2) "*Which options of behaviour are dangerous when you are driving on urban road?*" (3) "*Which options of the information are useful for you in driving?*" and (4) "*Which options of behaviour should be considered firstly in driver assistance systems?*". The answer options for these questions are selected by common consensus. In order to gather more feedbacks,

* Corresponding author.
E-mail address: crabwq@gmail.com (Q. Wang).

https://doi.org/10.1016/j.neucom.2017.12.017
0925-2312/

Fig. 1. Typical hazardous scenarios. (a) Pedestrian crossing; (b) Cyclist crossing; (c) Vehicle overtaking and (d) sudden appearance of animals.

we propagate the questions through instant-messaging apps. Fortunately, 145 participants took part in the investigation (99 men and 46 women, mean age of 33.3 years old ranging from 21 to 55). All drivers have a valid driving license, and averagely have 5 years driving experience, ranging from 0.5 to 20 years. The average driving mileage of participants is 114523.87 km.

The questionnaire investigation analysis for drivers is demonstrated in Fig. 2. From the results, we can discover that the irregular motion behaviour of frontal objects is highly threatening for drivers, such as overtaking from right side, pedestrian/vehicle crossing. In addition, the illumination and the spacing between vehicles are also important factors in safe driving. In order to confirm the observation, we investigate the latest reports of traffic fatalities[1,2] released by National Highway Traffic Safety Administration (NHTSA). These reports show that: (1) About 90.4% percent of pedestrian fatalities occurred when the pedestrian locates in front of the ego-vehicle and is with a crossing behaviour; (2) About 70% percent of pedestrian fatalities appeared in the nighttime condition; (3) Speeding, such as overtaking, changing lanes contributes the 27% percent of all fatal crashes, which is second only to drink driving, i.e., 29% percent. From these discoveries, this work will perceive the hazards that the object crosses or overtakes the ego-driving car, as well as a consideration for low illumination condition.

1.2. The way to success

Based on the above investigation, this work first contributes an incremental motion consistency measurement to distinguish the hazardous and normal situations. Consistently with the anomaly detection [8,9], motion consistency is the most efficient and straightforward cue for involving the irregular motion behaviour. In this procedure, optical flow is the most promising feature. Considering the infrequent occurrence of hazards, this work tackles the motion consistency measurement with a sparse representation paradigm. Different from the existing sparsity-based anomaly detection methods which usually model the motion noise with a Gaussian distribution, this work formulates the motion noise with a Gaussian–Laplacian distribution. It is inspired by that (1) Gaussian distribution is easy to be solved, such as the ordinary least squares (OLS) solution; (2) Laplacian is more adequate to model the noise of optical flow [10] while the Laplacian noise modeling is difficult to solve.[3] Although this strategy is coincident with the least soft-threshold squares (LSS) [11], our Gaussian–Laplacian distribution has more intrinsic physical meaning. Besides, LSS does not consider the spatial-correlations between different elements (i.e., superpixel[4] based image representation in this work). However, spatial-correlation exploitation is important to hazards prediction because a more reasonable hazards map should demonstrate consistent hazardous degree for different parts of the same object. It is in coincident with the relevance preservation of visual instances by graph-embedding ranking [12]. Motivated by that, we further introduce a graph-regularizer to infer the geometrical manifold of the motion feature in different image regions, which constitutes a new graph regularized least soft-threshold squares (GRLSS). Furthermore, different from LSS, we solve the objective function with a more efficient joint optimization.

To further boost the performance of hazards prediction, we not only consider the motion consistency principle, but also explore other visual cues. For one reason, motion sometimes cannot distinguish the hazardous target and background very clearly; for another, other cues such as appearance and location are also informative, as demonstrated in Fig. 2(c). Therefore, this work also provides the visual color, near-infrared spectral and visual depth bands to reflect the characteristics from aspects of color, physical material and position of front objects. To this end, we have to design a strategy for fusing them.

Traditional methods of tackling this fusion problem have two choices, feature level and decision level. In this work, we start from the feature level but with a more novel and efficient strategy: *saliency evaluation*. That is motivated by that (1) the occurrence of accident is always caused by the drivers' inattention and saliency can serve as an effective reminder; (2) saliency is the most successful simulation of human attention mechanism [13–15] and

[1] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811888.
[2] https://www.nhtsa.gov/risky-driving/speeding.

[3] "Due to illumination or occlusion effects, in optical flow estimation, matching the color or gray value is not always reliable, and the matching noise corresponds to a Laplace distribution which has longer tails than the Gaussian distribution [10]".
[4] A superpixel is a local image region grouping pixels with similar characteristics together, and its boundary is almost adhere to image boundaries.

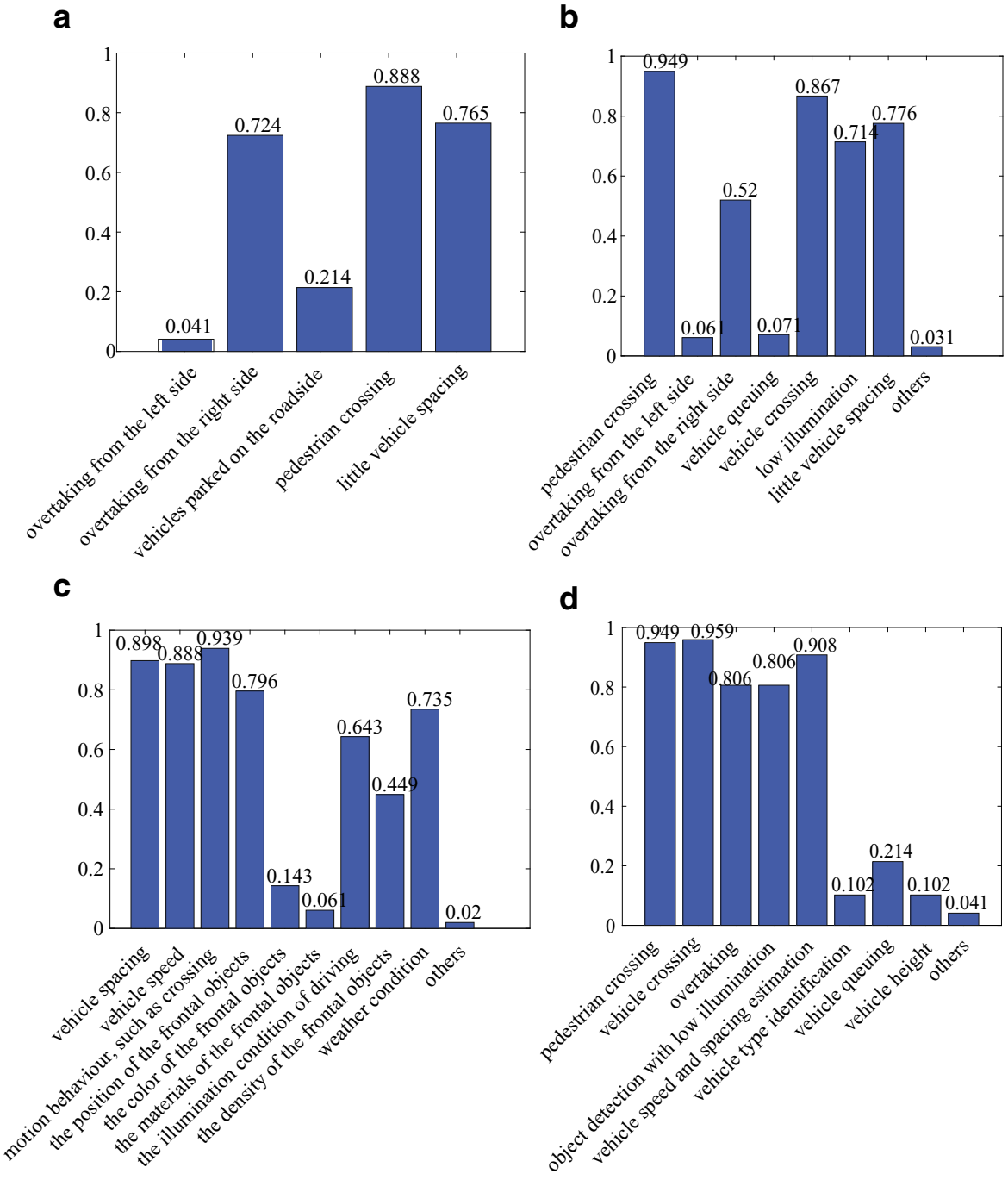

Fig. 2. The questionnaire investigation analysis on safe driving for drivers. The text in *x*-axis is the selected options for answering. (a) The statistical results on question1; (b) The statistical results on question2; (c) The statistical results on question3; (d) The statistical results on question4.

can significantly assist the danger detection on road [16]. By these means, each band is observed by a *saliency map* and the integration of different bands is converted into a *map fusion* problem. To fulfill this purpose, we consider using a novel Bayesian integration model. That is because we find that the content of video band can be effectively and robustly qualified by the prior probability inference in the Bayesian model.

Consequently, for incrementally **p**erceiving **h**azards in **d**riving (PHD), we propose an incremental motion consistency measurement in conjunction with an adaptive Bayesian integration of driver's attention to multiple visual bands.

1.3. Contributions

This work novelly involves motion, color, infrared, and depth clue together to roundly detect road hazards from different views. The contributions of this work are summarized as follows.

First, this work proposes an incremental graph regularized least soft-threshold squares (GRLSS) to infer the motion consistency. The novelty of GRLSS is to introduce a graph regularizer to make a more reasonable hazardous object detection by exploiting the geometrical relationship of different object parts, and model a more adequate Gaussian–Laplacian noise distribution of optical flow differing from other traditional methods.

Second, a novel Bayesian integration method for adaptively fusing multi-source cues into hazards prediction is introduced. This Bayesian model can better reflect the video content and assign more weight to more important cues in different scenes.

Finally, the superiority of the proposed method is verified by twelve video clips captured by ourselves containing RGB, near-infrared and recovered depth channels simultaneously. The provided video clips have the same view and resolution requiring no registration or frame alignment work.

The remainder of this paper is organized as follows. Section 2 thoroughly presents the literatures relating to this work. Section 3 illustrates the collection procedure of multi-source video data. Section 4 present the system overview and the datailed theory. Experimental validation is given in Section 5 followed by some discussions for this work in Section 6. The conclusion is summarized in Section 7.

2. Related works

Since this work addresses the road hazards prediction problem by utilizing motion and saliency, we review the relevant works in terms of the moving object detection in Advanced Driver Assistance Systems (ADAS), motion consistency measurement, and saliency based multiple source integration.

2.1. Moving object detection in ADAS

In ADAS, the most related methods to this work are for detecting the hazardous pedestrians or vehicles. As for this aspect, many researchers design robust detectors, such as pedestrian detectors [17,18] and vehicle detectors [16,19,20]. By automatically localizing the frontal pedestrians or vehicles, the detected response helps drivers to control their vehicles to avoid vehicle crashes. For example, Xu et al. [17] studied the problem of detecting sudden pedestrian crossing, and learned a novel pedestrian detector to localize the pedestrian as early as possible. Redmon et al. [18] proposed a YOLO detection module which predicted the coordinates of pedestrian directly using fully connected layers on top of the convolutional feature extractor. Some works for pedestrian detection used LADAR or Laser sensors [21,22] which were based on the hypothesis that front dynamic objects were all pedestrians. Sivaraman and Trivedi [23] proposed a part-based vehicle detector to localize cars. Sivaraman and Trivedi [23] built a vision-based system to detect and track vehicles. When a potentially hazardous situation arises, these systems trigger a warning. Though these methods can avoid danger to some extent, the detector-based methods always fail [7] because of the dynamic motion of frontal objects and camera. At the same time, these approaches still rely on external training data to generate target detector, and the detectors are all object-related. If some other moving objects move across the road suddenly, they will be missed and might cause vehicle crashes.

2.2. Motion consistency measurement

Among the approaches in the motion consistency measurement, histogram of optical flow [8,24,25], and histogram of blob change [26,27] are the main features utilized. For the motion pattern modeling, mixture of probabilistic principal component analysis (MP-PCA) model [28], social force model [24], sparse basis [9,29,30], etc. are the main alternatives. However, based on the sparsity of unusual events, more and more sparsity based anomaly detection methods have emerged in this field [9,29–31] recently. For example, Cong et al. [9] and Zhu et al. [31] decided the anomaly by the sparse reconstruction cost from an autonomously learned event dictionary. Zhao et al. [30] proposed an unsupervised dynamic sparse coding approach for detecting unusual events based on an online sparse reconstructor. Lu et al. [29] proposed an efficient structure preserving dictionary learning method to detect anomalies. It is worth noting that these sparsity based methods have been tested on the videos with static background. It is unclear whether they could deal with the dynamic camera motion and sudden object interactions. Besides, these sparsity based methods are all based on ordinary least squares (OLS) solution, but the least absolute deviations (LAD) solution [11] is proved to be more superior to OLS for outliers, which is introduced in this work.

2.3. Saliency based multi-source integration

It is worth noting that there are some related works attempting to integrate saliency of multi-source visual information into computer vision field. Different from the existing saliency methods focusing on the image only with RGB channels [32–34], the multi-source saliency integration exploits the collaborative mechanism between different visual sources. For example, Wang et al. [35] proposed a multi-spectral saliency detection method, which combined the near-infrared and RGB images together to obtain a more adequate saliency map. In addition, considering that the distance between camera and objects is often associated with the object's significance, and the relative position of objects in the scene is usually reflected by a visual depth map, Basha and Avidan [36] calculated a saliency map by incorporating geometrical consistency in the RGB and depth images. Shen et al. [15] proposed a seam carving method by a depth-aware saliency, whose central idea was that the important objects on the depth map had the large energy values. These multi-saliency methods utilized the advantages from either near-infrared or depth information, but the incorporation of more cues, such as near-infrared and depth cues, is probably better and has not been investigated. Besides, these methods fuse different source cues in linear regression [35] or joint feature learning [15], an adaptive integration which better considers the importance of cues may be superior to these methods.

3. Multi-source video data collection

To the best knowledge of the authors, there is no publicly available dataset simultaneously containing RGB, near-infrared and depth bands for hazards prediction on road. Therefore, we capture multi-source video data by a prism-based multi-spectral camera (JAI Inc., AD-080CL) mounted on a moving vehicle. The sensor can simultaneously capture RGB and near-infrared bands with the same resolution, where the wavelength of the near-infrared band is 790 nm.

Apart from the RGB and near-infrared bands, we also attempt to take the depth information of frontal object into account. Nevertheless, traditional depth map obtained from stereo cameras or radar lasers either needs to conduct tedious configuration work for two cameras or has limited perception range. Hence, this work extracts the depth information from the reconstruction of video clips captured by a moving camera. For the depth band in this work, it is recovered by the RGB channel because of the robust feature point detection in color band. Assume the video clip to be processed is $\mathcal{I} = \{I_t | t = 1, \ldots, n\}$, where $I_t(x)$ represents the color of pixel x at frame t. The recovery procedure is as following four steps:

The first step is to recover the camera parameters. The camera parameters containing the intrinsic matrix, the rotation matrix, and the translation vector are recovered by the shape from motion (SFM) technique. The estimated parameters are used for subsequent depth refinement.

The second step is to initialize depth map for each frame independently. By minimizing an energy function with a data term and a smoothness term with loop belief propagation, each pixel is

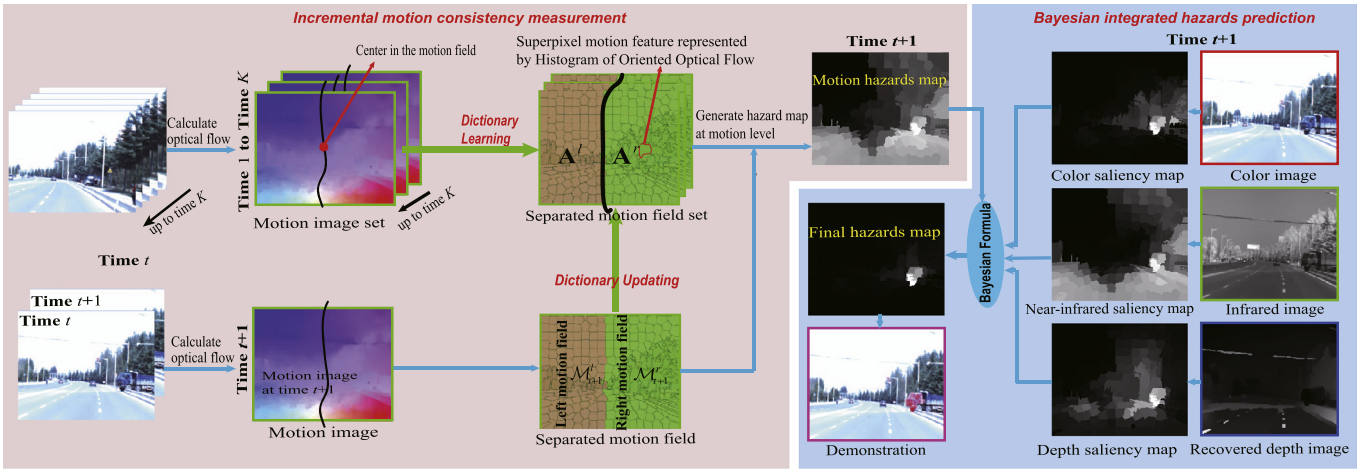

Fig. 3. The pipeline of the proposed method.

assigned a depth label. The energy function is:

$$E(D_t|\mathcal{I}) = \sum_x \left[1 - u(x) L_{init}(x, D_t(x)) + \sum_{y \in N(x)} \alpha(x,y) \cdot \beta(D_t(x), D_t(y)) \right], \quad (1)$$

where $u(x)$ is the adaptive normalization factor, $\alpha(\cdot, \cdot)$ is the adaptive smoothness weight, $N(x)$ is the neighborhood of x, $\beta(\cdot, \cdot)$ is the smoothness cost, and $D_t(x)$ is the estimated disparity. L_{init} is the disparity likelihood defined as $\frac{\pi_c}{\pi_c + \|I_t(x) - I'_t(x')\|}$, where π_c controls the shape of the differentiable robust function and x' is the corresponding pixel of x within frame I'.

The third step is the bundle optimization. The depth map obtained from the above step is a rough estimation. Here each frame is associated with others to refine the result. For a pixel x in frame t, its corresponding pixel x' in frame t' is computed by an epipolar geometry.

The final step is the space–time fusion. Though bundle optimization can improve the accuracy of depth maps greatly, there is still reconstruction noise. Hence, a space-time fusion algorithm is employed to reduce the disparity noises. The main idea is that spatial continuity, temporal coherence, and sparse feature correspondence are simultaneously considered to constrain the depth map. More details can be found in [37].

4. Overview of our system

4.1. System overview

The detailed pipeline of the proposed system is shown in Fig. 3, and the detailed procedure is described as follows. First, a multi-band video clip containing motion (obtained by calculating optical flow in color band), color, near-infrared and depth channels is acquired and each color frame is segmented into superpixels (explained in Section 4.2). The obtained superpixels' boundaries are superimposed on all the channels. Second, the motion consistency is incrementally inferred. Based on the observation that the motion flows of the left and right views are always different in driving, this work learns two dictionaries respectively for the left and right motion fields, and utilizes the learned dictionaries to represent the newly observed motion frame to generate a hazards map in motion band. The dictionaries initialized by the beginning K frames in each video, and are updated to adapt to the dynamic motion scene. Third, in order to give a more meaningful estimation, this work also extracts the human attentions to color, near-infrared and depth bands, which is achieved by saliency evaluation. Then, a novel *Bayesian integrated hazards prediction* is introduced by fusing the predicted result of motion band with these saliency maps. The adaptive integration helps the drivers pay attention to potential hazards more efficiently. Last, the predicted mask is superimposed on the original color image.

4.2. Incremental motion consistency measurement

As mentioned before, this work will formulate the motion consistency measurement with a sparse representation paradigm. The requirement for achieving this purpose is to incrementally learn the compact and consistent normal sparse basis. However, when driving on road, an interesting phenomenon is that the scene of left view moves at left-bottom direction and right-bottom direction for the scene of right view. Meanwhile, the degree of hazardous is rather different between left view and right view, as shown by the overtaking analysis in Fig. 2. Although this assumption is heuristic to some extent, the necessity is verified by our experiments. In order to learn a compact and consistent sparse basis, we treat the motion field $\mathcal{M}$ (computed by optical flow) as left and right part. After that, the following task is to sparsely represent the newly observed motion patterns with the learned sparse basis, in left and right views respectively. However, as claimed by Brox and Malik [10], the noise in the optical flow estimation is more adequately modeled by Laplacian than those Gaussian ones [38]. Nevertheless, Laplacian is difficult to solve and Gaussian is easier [11]. Hence, this work makes a trade-off that a Gaussian–Laplacian distribution is utilized to model the motion noise. Although this strategy is coincident with the least soft-threshold squares (LSS) newly proposed by Wang et al. [11], our Gaussian–Laplacian distribution has more apparent physical meaning. In addition, LSS does not consider the spatial-correlations between elements (i.e., superpixels in this work). Considering that spatial-adjacent superpixels in each frame might have similar motion patterns, and should receive a similar estimation of hazardous degree, this paper introduces a graph-regularizer into LSS and constructs a graph regularized least soft-threshold squares (GRLSS) to predict the hazards more effectively in motion band. Furthermore, different from Wang et al. [11], we solve the objective function with a more efficient joint optimization.

Before giving GRLSS, superpixel based motion representation is firstly given. Then a brief review of LSS is introduced, which forms the basis of our GRLSS. The estimation approach of hazards via GRLSS is presented finally.

4.2.1. Superpixel motion representation

For the motion band, we utilize the optical flow to obtain the motion cue. Considering the accuracy and efficiency of the existing optical flow approaches, Correlation Flow [39] is selected. To better capture the intrinsical structural information and lower the computational cost, simple linear interactive clustering (SLIC) superpixel [40] is employed to segment every input video frame. As for the motion representation, histogram of oriented optical flow (HOOF) [41] is utilized to represent the superpixel motion feature. Suppose the image is segmented into N superpixels. For each superpixel $sp_i, i = 1, \ldots, N$, its motion feature is denoted as $\mathbf{y}_i \in \mathbb{R}^{c\times 1}$, where c indicates the direction bins of HOOF. All the motion features at time t construct the image motion field $\mathcal{M}_t = \{\mathcal{M}_t^l, \mathcal{M}_t^r\}$, where $\mathcal{M}_t^l$ and $\mathcal{M}_t^r$ respectively specify the left and right motion field, and $\mathcal{M}^l$ is constructed by the superpixels whose centroid in x-axis are smaller than the one of the image center, and vice versa for $\mathcal{M}^r$.

4.2.2. LSS

As for data representation, sparse coding has been widely investigated in object tracking, action recognition, face recognition, etc. The objective function of sparse coding originates from the following [42]:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^{c\times 1}$ is a c-dimensional observation vector, $\mathbf{x} \in \mathbb{R}^{d\times 1}$ specifies the d-dimensional coefficient vector to be estimated, $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_c]^T \in \mathbb{R}^{c\times d}$ represents the dictionary containing base vectors, and $\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}$ denotes the error vector. Different distributions of $\mathbf{e}$ can generate different solutions of Eq. (2), such as ordinary least squares (OLS) solution $\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ if Gaussian distribution is considered, and least absolute deviations (LAD) solution $\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_1$ if Laplacian distribution is considered.

Based on the characteristics of OLS and LAD, we know that OLS is easy to solve but sensitive to outliers, while LAD is robust to outliers but difficult to solve. Inspired by Wang et al. [11], we aim to make a trade-off, which models the error as an additive combination of a Gaussian noise vector $\mathbf{g}$ and a Laplacian noise vector $\mathbf{u}$,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{g} + \mathbf{u}, \tag{3}$$

where Gaussian and Laplacian components handle small dense noises and outliers, respectively. In order to solve Eq. (3), it is rewritten as a standard least square and an ℓ_1 regularization term on $\mathbf{u}$,

$$[\hat{\mathbf{x}}, \hat{\mathbf{u}}] = \arg\min_{\mathbf{x},\mathbf{u}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{u}\|_2^2 + \lambda\|\mathbf{u}\|_1, \tag{4}$$

where $\mathbf{u}$ is a sparse vector. This function is easy to solve, but it does not consider the spatial-correlations between elements (i.e., superpixels in this work). However, correlation exploitation is important to hazards prediction because a more reasonable hazards map should demonstrate consistent hazardous degree for the different superpixels of the same object. Hence, this work embeds a graph regularizer into LSS and develops a graph regularized least soft-threshold squares (GRLSS) to conduct hazards estimation in motion band.

4.2.3. GRLSS

To exploit the intrinsic manifold structure of hazardous object efficiently, this work utilizes a graph construction method [43]. Given a motion field $\mathcal{M}_t$ at time t, rearrange $\mathcal{M}_t$ in label sequence of superpixels into a motion feature matrix $\mathbf{Y}_t = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N]^T \in \mathbb{R}^{N\times c}$, where N is the number of superpixels in the motion image. With the above definition, we construct the graph $\mathcal{G} =< V, E >$ on $\mathbf{Y}_t$, where V denotes the nodes representing $\mathbf{Y}_t$, and E specifies the edges weighted by an affinity matrix $\mathbf{W} = [w_{ij}]_{N\times N}$, where $w_{ij} = e^{-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|}{\sigma^2}}$. Denote $\mathbf{D} = diag(d_{11}, \ldots, d_{NN})$ as the degree matrix, where $d_{ii} = \sum_j w_{ij}$. Note that, for simplicity, the subscript t is omitted in the following description. In addition, the formulation in left and right views is the same. Therefore, for a more compact description of GRLSS, we will not make a distinction in the following.

Assume the motion dictionary $\mathbf{A}$ is generated by a PCA subspace with i.i.d Gaussian–Laplacian noise, and Laplacian noise on the observation $\mathbf{Y}$ is denoted as $\mathbf{U} = [\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^N]^T$. The objective function of GRLSS is written as:

$$[\hat{\mathbf{X}}, \hat{\mathbf{U}}] = \min_{\mathbf{X},\mathbf{U}} \mathcal{L}(\mathbf{X}, \mathbf{U}),$$
$$\mathcal{L}(\mathbf{X}, \mathbf{U}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{A}\mathbf{X} - \mathbf{U}\|_F^2 + \lambda_1\|\mathbf{U}\|_{1,1} + \frac{\lambda_2}{2}tr(\mathbf{X}\mathbf{L}\mathbf{X}^T), \tag{5}$$

where $\frac{\lambda_2}{2}tr(\mathbf{X}\mathbf{L}\mathbf{X}^T)$ denotes the term for regularizing the neighborhood of superpixels, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ denotes the Laplacian regularization matrix. To the best of our knowledge, there is no closed-form solution for Eq. (5). Therefore, the solution of Eq. (5) is presented as follows.

For solving $\mathbf{X}$, fix $\mathbf{U}$. The deviation of Eq. (5) is written as:

$$\mathbf{A}^T\mathbf{A}\mathbf{X}_{k+1} + \mathbf{X}_{k+1}\lambda_2\mathbf{L} = \mathbf{A}^T(\mathbf{U} - \mathbf{Y}). \tag{6}$$

Because $\mathbf{A}^T\mathbf{A} \neq \lambda_2\mathbf{L}$, the solving of Eq. (6) is fulfilled by Sylvester Equation [44].

For solving $\mathbf{U}$, fix $\mathbf{X}$. The optimal $\mathbf{U}_{k+1}$ can be obtained by a soft-threshold operation on each $\mathbf{u}_{k+1}^i = S_\lambda([\mathbf{y}^i - \mathbf{A}\mathbf{x}_{k+1}^i])$ in $\mathbf{U}_{k+1}$, where $S_\lambda(x) = \max(|x| - \lambda, 0)sign(x)$ and $sign(\cdot)$ is a sign function.

It is clear that Eq. (5) is convex, and the objective value reduces after every iteration.[5] Therefore, the objective function can obtain a global minimal solution.

After obtaining the solution $\hat{\mathbf{X}}$ and $\hat{\mathbf{U}}$ of Eq. (5), for each superpixel sp_i, its hazardous degree is determined by measuring its motion vector $\mathbf{y}^i$ with its related obtained optimal $\hat{\mathbf{x}}^i$ and $\hat{\mathbf{u}}^i$, and is represented as:

$$d(\mathbf{y}^i, \mathbf{A}) = \frac{1}{2}\|\mathbf{y}^i - \mathbf{A}\hat{\mathbf{x}}^i - \hat{\mathbf{u}}^i\|_2^2 + \lambda_1\|\hat{\mathbf{u}}^i\|_1. \tag{7}$$

With the defined distance measuring the newly observed motion patterns, the hazards map in this paper is calculated in the following subsection.

4.2.4. Hazards map calculation in motion band

As described before, the newly observed motion field is separated into left and right parts. Because of the difference between the left and right motion fields when driving, we learn two dictionaries via PCA subspace, respectively.

Denote the learned left and right dictionaries as $\mathbf{A}^l$ and $\mathbf{A}^r$. For a newly observed superpixel motion vector $\mathbf{y}_t^i$ at time t, we represent it both by $\mathbf{A}^l$ and $\mathbf{A}^r$ and generate $d(\mathbf{y}_t^i, \mathbf{A}^l)$ and $d(\mathbf{y}_t^i, \mathbf{A}^r)$. The behind idea is that for perceiving hazards in the left motion field, we treat the learned sparse basis in left motion field as *positive templates*, and *negative templates* for the ones in the right motion field, and vice versa. Another insight is that the normal region in the left motion field can be better represented by the linear combination of *positive templates* while the abnormal region in the left motion field can be better represented by the span of *negative templates*. Thus, we calculate the hazardous degree of superpixels in the left motion field as:

$$d_t^i = 1 - \exp(-(\varepsilon_l^i - \varepsilon_r^i)/\beta), \tag{8}$$

[5] $\mathcal{L}(\mathbf{X}_{k+1}, \mathbf{U}_k) = min_{\mathbf{X}}\mathcal{L}(\mathbf{X}, \mathbf{U}_k) \leq \mathcal{L}(\mathbf{X}_k, \mathbf{U}_k)$, and $\mathcal{L}(\mathbf{X}_{k+1}, \mathbf{U}_{k+1}) = min_{\mathbf{U}}\mathcal{L}(\mathbf{X}_{k+1}, \mathbf{U}) \leq \mathcal{L}(\mathbf{X}_{k+1}, \mathbf{U}_k)$.

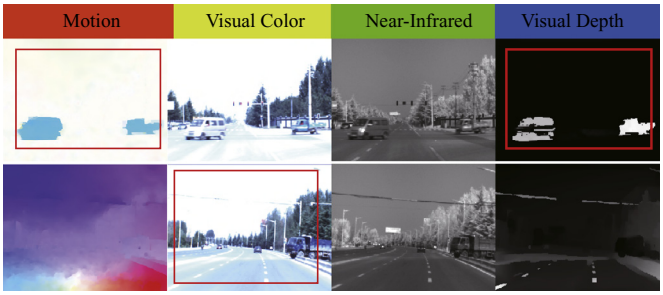

Fig. 4. Some typical frameshots with multi-band of different video clips. Obviously, the band marked by the red box is more informative to the other bands for a certain video clip. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the one of right motion field as:

$$d_t^i = 1 - \exp(-(\varepsilon_r^i - \varepsilon_l^i)/\beta), \tag{9}$$

where d_t^i is the hazardous degree of the ith superpixel in the tth frame of motion band, $\varepsilon_l^i = d(\mathbf{y}_t^i, \mathbf{A}^l)$ and $\varepsilon_r^i = d(\mathbf{y}_t^i, \mathbf{A}^r)$, and β (fixed as 0.4) is a small constant balancing the weight of left and right motion representation. Taking Eq. (8) as an example, it means that by measuring the distance difference $(\varepsilon_l^i - \varepsilon_r^i)$ between the observed motion pattern with left (right) and right (left) dictionaries, the hazardous degree of left (right) motion field can be obtained. If $\varepsilon_l^i - \varepsilon_r^i > 0$, it indicates that an abnormal object may appear in the left image region, and vice versa for $\varepsilon_l^i - \varepsilon_r^i < 0$. For easier combination with other cues in the following, we utilize max-min normalizer to put d_t^i into the range of [0, 1]. All the hazardous degrees of superpixels construct a motion hazards map $\mathcal{S}_t^M$ for the tth frame.

4.2.5. Dictionary updating

To adapt to the dynamic scene, the dictionaries need to be updated. Considering that the hazards always occurs occasionally, and the motion field calculated by optical flow has error, we respectively collect the left/right normal samples with $d_t^i < 0.5$. Then, the left/right-dictionary is updated by the related normal samples via an incremental principle component analysis (IPCA) [45].

4.3. Bayesian integrated hazards prediction

Although motion clue exploited above can predict hazards to some extent, there is still the situation that the motion cannot obviously infer because of the dynamic motion of camera. For example, in the second row of Fig. 4, the motion of the crossing truck has some confusion with the background, but the color band has a more obvious distinction. For a more general situation as illustrated in Fig. 4, the most informative band is different for various video clips. This is also proved by the fact that when driving, the drivers need to synthesize all the available cues to predict hazards, as claimed by the investigated results shown in Fig. 2(c). Therefore, apart from the motion consideration mentioned before, this work also provides color, near-infrared spectral and depth bands simultaneously. However, it generates an important problem that how to accurately evaluate and combine the better visual cues? To address this problem, this work builds an adaptive multi-source cue integration via a novel saliency based Bayesian inference.

We first exploit the human attention to color, near-infrared and depth bands by simultaneously evaluating their saliency maps. We take the method of Zhang et al. [46] as an attempt for saliency because of its efficiency. The salient object in [46] is detected by a graph based manifold ranking. With the hazards map at motion band and the saliency maps of color, near-infrared and depth bands, we then present a novel Bayesian integrated model to adaptively weight and fuse them.

In fact, many works integrating multi-source cues exist in the literature [15,19,47]. However, in many times, the fusions would result in a different, possibly conflicting or multi-face performance [48]. The reason behind this is that the data content is not reflected and considered properly. As for the multi-source based saliency map integration in this work, the most related is the high-level multi-spectral regression by Wang et al. [35] and joint RGB-D feature learning by Shen *et al.* [15]. Differently, we present an efficient and effective Bayesian integration model to fuse different maps, defined as:

$$\mathbf{Pr}(O|\mathcal{S}(z)) = \frac{\mathbf{Pr}(\mathcal{S}(z)|O)\mathbf{Pr}(O)}{\mathbf{Pr}(O)\mathbf{Pr}(\mathcal{S}(z)|O) + (1 - \mathbf{Pr}(O))\mathbf{Pr}(\mathcal{S}(z)|B)}, \tag{10}$$

where the prior probability $\mathbf{Pr}(O)$ is an object map, $\mathcal{S}(z)$ is the saliency value of pixel z, $\mathbf{Pr}(\mathcal{S}(z)|O)$ and $\mathbf{Pr}(\mathcal{S}(z)|B)$ represent the object and background likelihood probability of pixel z, respectively. Obviously, the prior probability and the likelihood estimation are the key. Next we will present the details of prior probability and likelihood estimation.

4.3.1. Prior probability

With respect to the prior estimation, because the predicted hazards is often reflected by objects, the object proposal [49] quantifying how likely it is for an image to contain objects may be an ideal strategy. Usually, objectness measurement is conducted by generating lots of object proposals represented as rectangle image regions which mostly contain objects. However, objectness measurement involves difficult task to determine the best candidate having potential hazards. Based on the assumption [50], the *elements* (i.e. superpixels) of an object always group as a particular region rather than *distributed* in the whole image, which means that the object is generally more compact than the background. Hence, this work introduces the unsupervised *element distribution map (EDM)* [50] to estimate the prior probability.

EDM: The element distribution of an examined superpixel is obtained by using the spatial variance of its color feature, i.e., computing the occurrence ratio of the color feature for the examined superpixel with respect to the elsewhere in the whole image. The smaller value of element distribution has higher probability belonging to an object.

Assume the spatial feature variance of the ith superpixel is v_i. Its color feature is denoted as $\mathbf{c}_i$. Then we compute:

$$v_i = \sum_{i=1}^{N} \|\mathbf{p}_i - \mu_i\|^2 \underbrace{w(\mathbf{c}_i, \mathbf{c}_j)}_{w_{ij}}, \tag{11}$$

where $\mathbf{p}_i$ denotes the position of the superpixel sp_i, w_{ij} specifies the similarity between color feature $\mathbf{c}_i$ and $\mathbf{c}_j$ of superpixel sp_i and sp_j, $\mu_i = \sum_{i=1}^{N} w_{ij}\mathbf{p}_i$ defines the weighted mean position of color feature $\mathbf{c}_i$, and N is the superpixel number. Because of the quadratic runtime complexity of Eq. (11), w_{ij} is set as a Gaussian modeling $\frac{1}{Z}\exp(-\frac{1}{2\sigma^2}(\|\mathbf{c}_i - \mathbf{c}_j\|^2))$, where Z is a constant for normalization, and σ is set as 20 empirically. The spatial feature variance $\{v_i\}_{i=1}^N$ of all the superpixels at time t constructs the EDM, denoted as $\mathcal{ED}_t$. Accordingly, we compute EDMs for motion, color, near-infrared and depth bands at time t and generate $\{\mathcal{ED}_t^M, \mathcal{ED}_t^C, \mathcal{ED}_t^I, \mathcal{ED}_t^D\}$, where the value of each element distribution map is normalized into [0, 1]. Fig. 5 demonstrates an example that for the same view represented by four kinds of visual bands, the calculated EDMs are apparently different, and $\mathcal{ED}_t^M$ is the best.

Prior probability computation: With the computed element distribution map of multiple visual bands at time t, we model the

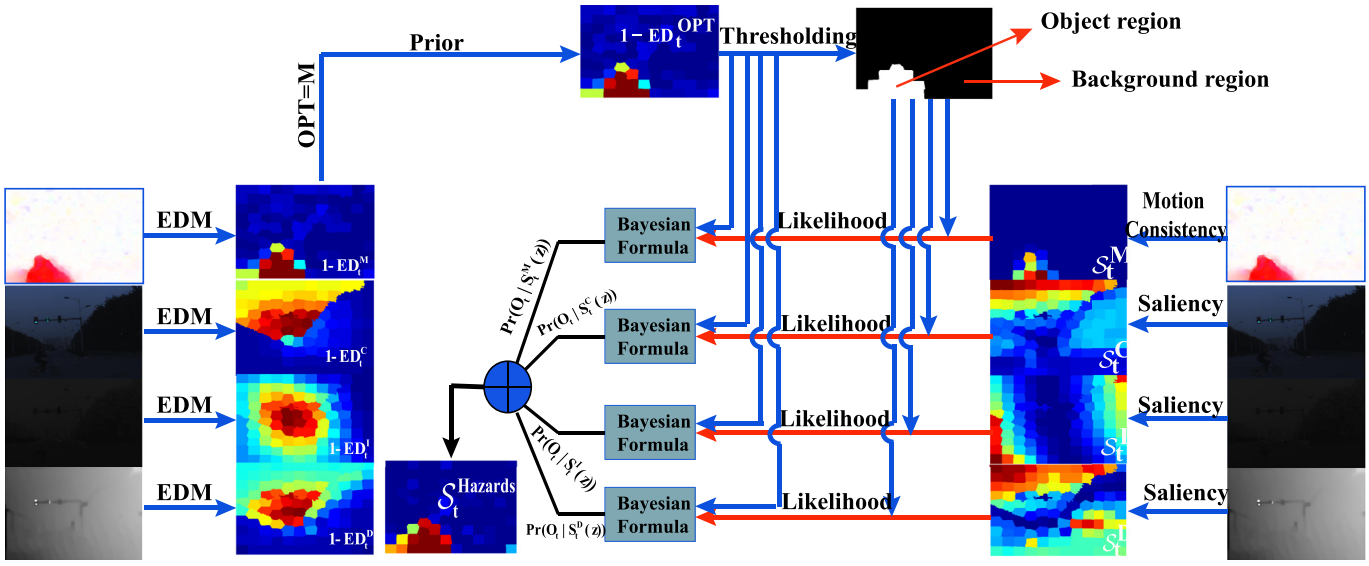

Fig. 5. The formulated Bayesian integrated hazards prediction model (best viewed in color mode).

prior probability $\mathbf{Pr}(O_t)$ by the optimal one $\mathcal{ED}_t^{\mathbf{OPT}}$, where the index **OPT** is selected by:

$$\mathbf{OPT} = \arg\min_i \overline{\mathcal{ED}_t^i}. \tag{12}$$

Here $\overline{\mathcal{ED}_t^i}$, $i \in \{M, C, I, D\}$ is computed by $\frac{\sum(\mathcal{ED}_t^i(z)<T)}{NUM(\mathcal{ED}_t^i(z)<T)}$, $\mathcal{ED}_t^i(z)$ is the element distribution value of pixel z, and $NUM(\cdot)$ calculates the superpixel number complying with the given condition. Here, we found $T = 0.5$ works well in all experiments. The purpose of the thresholding strategy aims to find the hazardous object as early as possible, where the object takes a small portion of the image.

Since the smaller value in $\mathcal{ED}_t^{\mathbf{OPT}}$ has a higher probability belonging to an object, $\mathbf{Pr}(O_t)$ is modeled by 1-$\mathcal{ED}_t^{\mathbf{OPT}}$, shown in Fig. 5.

4.3.2. Likelihood probability

In terms of likelihood probability estimation corresponding to the observed maps, we firstly denote the obtained map set at time t as the collection $\{\mathcal{S}_t^M, \mathcal{S}_t^C, \mathcal{S}_t^I, \mathcal{S}_t^D\}$ containing hazards map of **m**otion (M), saliency maps of **c**olor (C), near-**i**nfrared (I) and **d**epth (D), where the value of these maps is normalized into [0, 1]. Then, the following is to build the likelihood probability $\mathbf{Pr}(\mathcal{S}_t^i|O_t)$ and $\mathbf{Pr}(\mathcal{S}_t^i|B_t)$, $i \in \{M, C, I, D\}$.

First, $\mathcal{ED}_t^{\mathbf{OPT}}$ is thresholded by its mean value, and its object and background regions denoted as O_t and B_t are generated, as shown in Fig. 5. Second, calculate the object histogram $H_{O_t}^{bO_t}$ and background histogram $H_{B_t}^{bB_t}$ over the pixel value $\mathcal{S}_t^i(z)$ respectively in object region O_t and background region B_t, where bO_t and bB_t represent the bin number of the object histogram and the background histogram, respectively. Third, the likelihood probabilities at pixel z are calculated as:

$$\mathbf{Pr}(\mathcal{S}_t^i|O_t) = \frac{N_{bO_t(\mathcal{S}_t^i(z))}}{N_{B_t}}, \mathbf{Pr}(\mathcal{S}_t^i|B_t) = \frac{N_{bB_t(\mathcal{S}_t^i(z))}}{N_{B_t}}, \tag{13}$$

where N_{O_t} and N_{B_t} specify the number of the pixels in O_t and B_t, respectively, and $N_{bO_t(\mathcal{S}_t^i(z))}$, $N_{bB_t(\mathcal{S}_t^i(z))}$ denote the number of the pixels whose values fall into the object bin $bO_t(\mathcal{S}_t^i(z))$ and background bin $bB_t(\mathcal{S}_t^i(z))$. Here, $\mathbf{Pr}(\mathcal{S}_t^i|O_t)$ indicates the frequency of a given saliency value inside O_t, and $\mathbf{Pr}(\mathcal{S}_t^i|B_t)$ can be interpreted as the frequency of a given saliency value inside B_t.

Consequently, the posterior probability is computed by:

$$\begin{aligned}\mathbf{Pr}(\mathbf{O_t}|\mathcal{S}_\mathbf{t}^\mathbf{i}(\mathbf{z})) &= \frac{\mathbf{Pr}(O_t)p(\mathcal{S}_t^i(z)|O_t)}{\mathbf{Pr}(O_t)p(\mathcal{S}_t^i(z)|O_t) + (1-\mathbf{Pr}(O_t))\mathbf{Pr}(\mathcal{S}_t^i(z)|B_t)} \\ &= \frac{(1-\mathcal{ED}_t^{\mathbf{OPT}})\mathbf{Pr}(\mathcal{S}_t^i(z)|O_t)}{(1-\mathcal{ED}_t^{\mathbf{OPT}})\mathbf{Pr}(\mathcal{S}_t^i(z)|O_t) + \mathcal{ED}_t^{\mathbf{OPT}}\mathbf{Pr}(\mathcal{S}_t^i(z)|B_t)}.\end{aligned} \tag{14}$$

4.3.3. Integration

After obtaining the posterior probability, we compute an integrated hazards map $\mathcal{S}_t^{Hazards}$, denoted as:

$$\mathcal{S}_t^{Hazards} = \sum_i \mathbf{Pr}(O_t|\mathcal{S}_t^i(z)), \tag{15}$$

where $i \in \{M, C, I, D\}$. The integrated hazards map is normalized into [0, 1] by the max-min normalizer. The integrated hazards prediction model is illustrated in Fig. 5. We name this hazards prediction method as **PHD**.

5. Experimental validation

5.1. Dataset

In this paper, to evaluate the performance, we captured twelve video clips by the utilized imaging system in this work. It is worth noting that this work aims to address the hazardous prediction that there are some objects crossing or overtaking in front of the driving car, and infers the potential region that the hazardous object may appear. Based on the driving scenario, the captured video clips can be generally divided into four categories: (1) "three video clips with extremely low illumination (named as $LI-1$, $LI-2$, and $LI-3$)", having 208 frames, (2) "three video clips having overtaking behavior (specified as $OT-1$, $OT-2$, and $OT-3$)", owning 295 frames, (3) "two video clips containing the behavior of pedestrian crossing (denoted as $PC-1$ and $PC-2$)", having 191 testing frames, and (4) "four video clips consisting of vehicle crossing (named as $VC-1$, $VC-2$, $VC-3$ and $VC-4$)" with 319 frames. These video clips all begin with safe situation, and the number of safe frames in each video clip averagely take over about 10% percent of all video frames. The frameshots of the video clips are demonstrated in Fig. 6, some of which are very difficult for

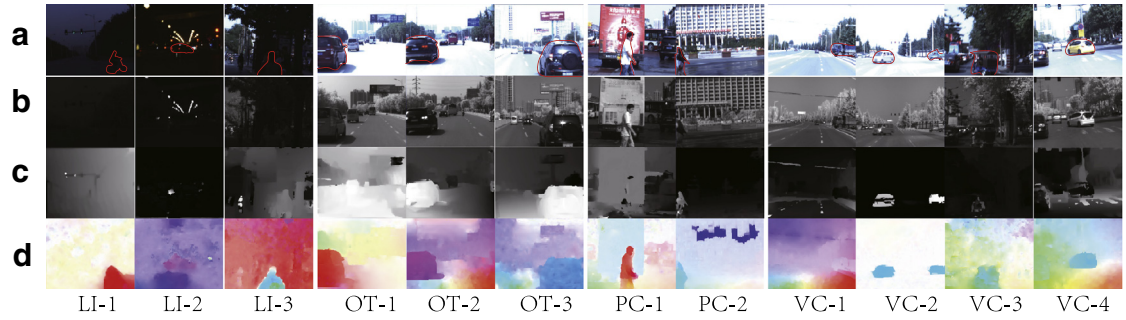

Fig. 6. Typical frameshots of the dataset. Among them, (a) represents the color band; (b) specifies the near-infrared band; (c) is the depth band and (d) denotes the motion band which is obtained by adopting Correlation Flow [39]. It can be seen that the importance of each band is different for each video clip.

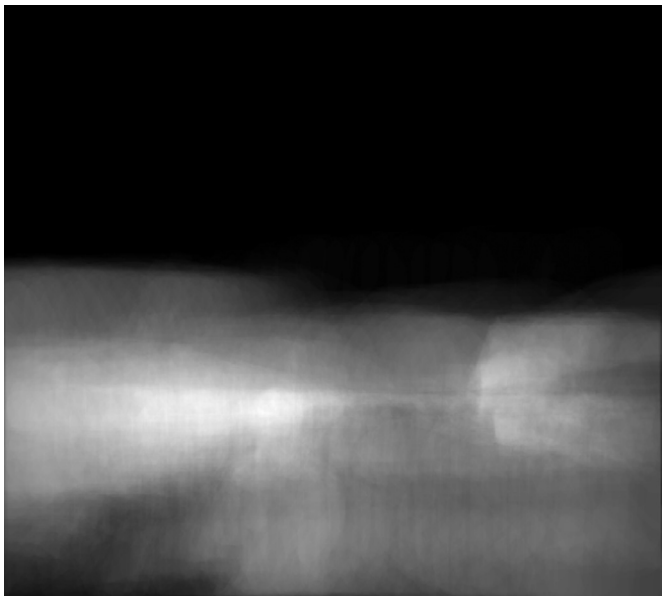

Fig. 7. Statistical distribution map of hazards. The statistical distribution map of hazardous region is obtained by averaging 700 ground truths in our dataset.

road hazard prediction. For example, it is hard to differ the hazardous object from the background in the color band of $LI-1$, $LI-2$, $VC-2$. In the captured dataset, the resolution of each frame is 443×553. The ground truth of each video clip is manually labeled by ourselves, which is obtained by computing the overlapping regions of ten drivers' labeling results. The key step in labeling is to accurately circle the frontal objects with hazardous behavior. Some examples of the ground truth marked by red line are shown on the color band in Fig. 6. It is worth noting that the resolution 443×553 is determined by the multi-spectral camera whose resolution is fixed. Therefore, the resolution in this work is pre-fixed and cannot be changed by ourselves.

In addition, we analyze the statistical distribution map of hazardous region of these video clips by averaging 700 ground truth images in the dataset. Interestingly, we find the hazardous objects always appear at the mid-bottom region of the image, as shown in Fig. 7. This statistical distribution map to some degree indicates context information for hazards prediction, which removes the influence of the disturbing scene, such as the crown of the tree and the sky. In this paper, the statistical distribution map is then utilized to refine the hazards prediction result by multiplying it with previously obtained integrated hazards map mentioned in Section 4.3.3.

5.2. Implementation setup

5.2.1. Metrics

In order to prove the efficiency of the proposed method, the qualitative and quantitative evaluations are both considered. For qualitative evaluations, we demonstrate several typical snapshots of the detected results in each video clip. As for the quantitative ones, pixel-wise receiver operating characteristic curve (ROC) and area under ROC (AUC) are employed. Among them, ROC represents the prediction ability of the proposed method, and its indexes are:

$$TPR = TP/P, FPR = FP/N, \tag{16}$$

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

5.2.2. Parameters

In our work, the SLIC superpixel [40] is employed, in which δ represents the compactness, and N is the number of the superpixels. The larger δ is, the more compact the superpixels are. In this paper, δ is set as 0.9 for all video clips. For determining N, this work runs each video clip for five trials with N at {125, 175, 225, 275, 325}. The average AUC curves according to different N as well as video categories are shown in Fig. 8(a). Interestingly, we find the superpixel number has little impact on the performance. Hence, the superpixel number is set as 125 for all the video clips because fewer superpixel needs less running time. Since this paper considers the motion consistency by specially exploiting the geometry relationship between different superpixels, we mainly examine the impact of λ_2 in Eq. (5) to the performance of motion consideration, and the examined results are shown in Fig. 8(b). Hence, λ_2 is set as 0.1. λ_1 is set as 0.1 in all experiments, which is similar to the LSS which is demonstrated in the work of [11]. The size of the basis in the left and right dictionaries in Eq. (5) is set as 20. The direction bin number of the HOOF [41] is specified as 30. As mentioned before, this work infers the motion consistency in an incremental way, and there is no training data available. For one video clip, the dictionaries are learned with the beginning 10% frames of the same video clip, and then used to infer the remainder frames of the same video clip, where the dictionaries are updated for every 10 frames. Note that the samples of the frames for constructing the dictionaries are far more than 30 the dimension of the HOOF. Therefore, the dictionaries in this work are all over-complete.

5.2.3. Comparisons

Since the proposed method is fulfilled by the collaboration of the motion consistency measurement and saliency based multi-source information integration, the motion consistency measurement is firstly evaluated. It is achieved by comparing the proposed

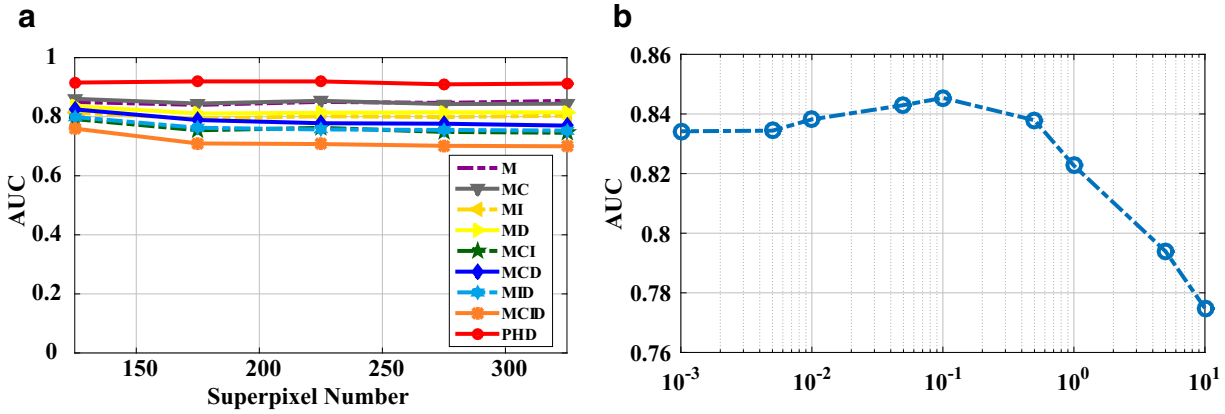

Fig. 8. Parameter selection. (a) Overall AUC comparisons w.r.t. N for all the video clips. (b)AUC comparisons w.r.t. λ_2 in Eq. (5).

GRLSS with the original LSS and the ℓ_1 in the SRC method [9] representing the state-of-the-art for anomaly detection. Note that the motion noise of ℓ_1 is modeled with *Gaussian* distribution, and our GRLSS and LSS utilize the *Gaussian–Laplacian* distribution. Besides, for the necessity validation of the motion field separation, we further present the experiments which treat the motion field as a whole, denoted as GRLSS-ws (GRLSS without separation).

Besides, another highlight of the proposed method is the multi-source information integration for hazards prediction. To validate the effectiveness for the proposed integration method, i.e., PHD model, we firstly exhaustively choose eight naive combinations into the evaluation. They are Motion (M), Motion-Color (MC), Motion-near-Infrared (MI), Motion-Depth (MD), Motion-Color-near-Infrared (MCI), Motion-Color-Depth (MCD), Motion-near-Infrared-Depth (MID), and Motion -Color-near-Infrared-Depth (MCID). These combinations treat the motion as a leader because that motion consistency to some extent provides a semantic knowledge for hazards understanding. Each kind of naive combination is achieved by cue inner-product which can verify which cues could boost or weaken the performance. To further validate PHD, we integrate different visual cues with PHD way. They are PHD-MC, PHD-MI, PHD-MD, PHD-MCI, PHD-MCD, PHD-MID, and PHD-ALL (fusing all the cues with PHD model).

5.3. Evaluation of motion consistency measurement

In this subsection, to prove the effectiveness of the GRLSS, we visually demonstrate typical snapshots of the predicted hazardous mask in motion band in Fig. 9. From the shown results, GRLSS generates a more meaningful result than LSS [11] and ℓ_1 [9]. Besides, because of the Gaussian modeling of motion noise, ℓ_1 can hardly differ the hazardous target and the background. As for our GRLSS and LSS with the Gaussian–Laplacian modeling, the performance is apparently boosted. In addition, in Fig. 10, we also demonstrate the AUC value comparison of ℓ_1, LSS and GRLSS for each video clip. From the demonstrated results, GRLSS is manifestly better than GRLSS-ws which treats the motion field as a whole, taking the snapshot of "$VC-2$" as an example. From this performance comparison, the superiority of our GRLSS is apparently verified. The reason is that except the appropriate modeling for motion noise, the geometrical manifold of elements (i.e., superpixel-based motion pattern) is novel for hazards prediction in motion. From the semantic consistency, GRLSS coincides with human perception better.

5.4. Evaluation of different clue combinations

To further explore the effectiveness of the multi-source integration in this work, the performance comparisons illustrated in Tables. 1 and 2, Figs. 11 and12 are presented according to different video categories in the dataset.

Low illumination: These video clips have extremely low illumination, in which the objects are very difficult to be found with color, near-infrared and depth bands, as shown in Fig. 6. From the analysis of information combination in Fig. 11 and Table. 1, the performance of M for these sequences is obviously superior to all the naive combinations. It concludes that our motion consistency measurement has strong robustness in the scenarios with low illumination. However, because of the tough color, near-infrared and depth cues, integrating them with M makes the performance weaker. One reason of this phenomenon is that the conventional saliency detection methods may not be feasible under low illumination. But the more important fact is that the naive combinations cannot reflect the importance of different cues, and are vulnerable to bad cues. In contrast, our PHD model can obviously boost the prediction performance. This is because that the robust prior probability estimation in PHD makes the integrations in these sequences assign larger weight to the motion cue.

Overtaking: In these sequences, there is a black car overtaking from the left or right view of our vehicle. Considering both the black and our vehicle move fast, this kind of hazards usually occurs on the mid-line of the road or highway, which has the largest impact on the human lives and properties reported by a German study [51]. As shown in Fig. 6, each band of these sequences has different discrimination, and color band is visually better than the others. It is verified by that MC (95.0%) > M (88.2%), MC(95.0%) > MCD (93.8%), MC(95.0%) > MCI (94.7%) in $OT-1$ and $OT-3$ with a fusion of color and other bands, as shown Table. 2. In addition, the comparisons, such as M(88.2%) < MC (95.0%), M(88.2%) < MD (92.2%), M(88.2%) < MI (94.4%) in $OT-2$, show that fusing more cues can boost the performance. However, different cues have distinct ability for boosting, such as M < MID < MCI < MC in $OT-1$ and $OT-3$. Therefore, color cue is better than near-infrared, depth and motion ones in $OT-1$ and $OT-3$ sequences. In addition, we observe that our PHD model can pay more weight to better cues, verified by PHD-MC > PHD-MI and PHD-MD in $OT-1$. Our PHD-ALL is the best than the naive combinations. Although PHD-ALL shows a little weaker than other integration of our PHD model, the performance between them has tiny difference.

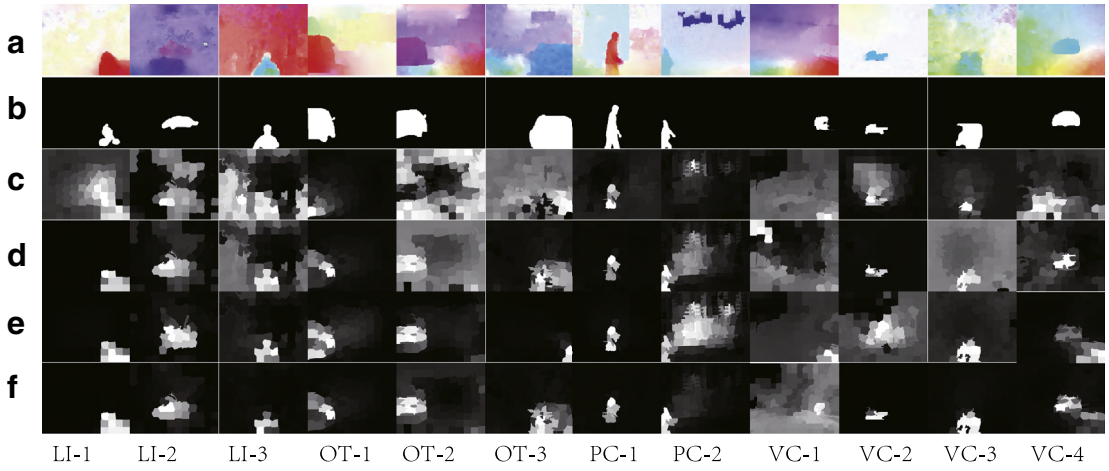

Fig. 9. Visual comparison of hazards prediction in motion band. (a) Optical flow image; (b) Ground truth of the hazards; (c) Detected results by ℓ_1 reconstruction error [9] (Gaussian modeling); (d) Detected results by original LSS [11]. (e) and (f) represent the GRLSS-ws and GRLSS (Gaussian–Laplacian modeling) without and with motion field separation, respectively. For a fairer comparison, the showed result here is not refined by the statistical hazards map, shown in Fig. 6.

Table 1
The AUC (%) comparison between the naive combinations and the adaptive Bayesian integration with all the available information (PHD-ALL). For a clearer and fairer comparison, the **bold** one is the best result. The **bold** one represents the second best and the *italic* one specifies the third best.

Seqs	M	MC	MI	MD	MCI	MCD	MID	MCID	PHD-ALL
LI − 1	**95.2**	92.0	91.3	*92.2*	86.0	89.9	88.9	82.6	**96.2**
LI − 2	**95.2**	80.6	81.6	*91.1*	65.0	77.6	77.8	65.7	**94.9**
LI − 3	**88.3**	79.7	77.3	*90.1*	53.3	78.6	75.4	54.6	**94.6**
OT − 1	88.2	**95.0**	*94.4*	92.2	**94.7**	93.8	91.7	90.4	**95.0**
OT − 2	**87.8**	*82.3*	75.2	76.0	81.3	81.3	74.9	76.6	**91.5**
OT − 3	88.6	**92.4**	88.5	77.8	**92.4**	84.3	81.2	84.1	*91.9*
PC − 1	*85.9*	83.1	**86.7**	82.5	77.8	77.0	80.3	73.4	**92.0**
PC − 2	*74.0*	**75.1**	69.2	70.7	61.8	65.9	61.9	55.4	**78.6**
VC − 1	72.0	**94.3**	61.7	79.8	**94.4**	92.3	82.1	*94.1*	**94.4**
VC − 2	**95.0**	89.2	92.7	*93.4*	89.1	86.1	87.9	84.2	**97.0**
VC − 3	*85.9*	**87.1**	80.8	81.5	76.7	80.3	74.0	69.6	**93.4**
VC − 4	*85.7*	84.6	**85.9**	**84.7**	85.0	**86.4**	**86.4**	83.9	85.3
Average	**86.8**	*86.3*	82.1	84.5	79.8	82.7	80.2	76.2	**92.0**

Table 2
The AUC (%) comparison of adaptive Bayesian integrations when different number of cues are provided. For a clearer and fairer comparison, the **bold** one is the best result. The **bold** one represents the second best and the *italic* one specifies the third best.

Seqs	PHD-MC	PHD-MI	PHD-MD	PHD-MCI	PHD-MCD	PHD-MID	PHD-ALL
LI − 1	95.2	**95.6**	94.6	*95.4*	95.3	**95.6**	**96.2**
LI − 2	94.9	95.1	94.9	*95.3*	**95.5**	**95.6**	94.9
LI − 3	93.5	*94.3*	93.8	94.2	93.9	**94.5**	**94.6**
OT − 1	**96.0**	*95.6*	**95.8**	**96.0**	94.8	94.4	95.0
OT − 2	88.9	**88.5**	88.7	89.4	**91.2**	*90.9*	**91.5**
OT − 3	**93.8**	*93.3*	91.2	**94.2**	92.0	91.4	91.9
PC − 1	**92.4**	**92.2**	91.6	91.8	*92.0*	91.9	**92.0**
PC − 2	76.1	76.5	74.9	*77.4*	*77.4*	**78.0**	**78.6**
VC − 1	82.4	*92.5*	87.5	**94.0**	88.3	89.8	**94.4**
VC − 2	94.7	**98.2**	**98.2**	96.1	**97.0**	*96.9*	**97.0**
VC − 3	92.2	*93.1*	92.5	**93.2**	92.7	93.0	**93.4**
VC − 4	*84.3*	83.6	*84.3*	*84.3*	**84.7**	84.3	**85.3**
Average	87.8	*91.5*	88.6	**91.7**	88.7	91.3	**92.0**

Pedestrian crossing: In these sequences, there is a person walking across the road. This type might appear in the urban road especially in the downtown area. Owing to the static camera, the motion of the pedestrian can be estimated efficiently. From the results of M in Figs. 11 and12, it indicates that motion consistency measurement is effective in these sequences. In *PC* − 1 sequence, the person and the red bus have an approximate distance away from our vehicle, which makes the depth information may be infeasible. Besides, because of the reflecting light of the bus, the color band might be influenced by the tough illumination. From the result of MI (86.7%) > MC (83.1%) > MD (82.5%) of *PC* − 1, these guesses are verified. In addition, because of the small size of the woman in *PC* − 2, it is much easily influenced by the complex background of color and near-infrared band. Therefore, it is nearly impossible to successfully predict the crossing woman with near-infrared cue in *PC* − 2, shown by MI (69.2%) in Table. 1 and Fig. 12. As for our PHD model, PHD-ALL is manifestly better than all the naive combinations. Besides, because of the efficient motion estimation, our

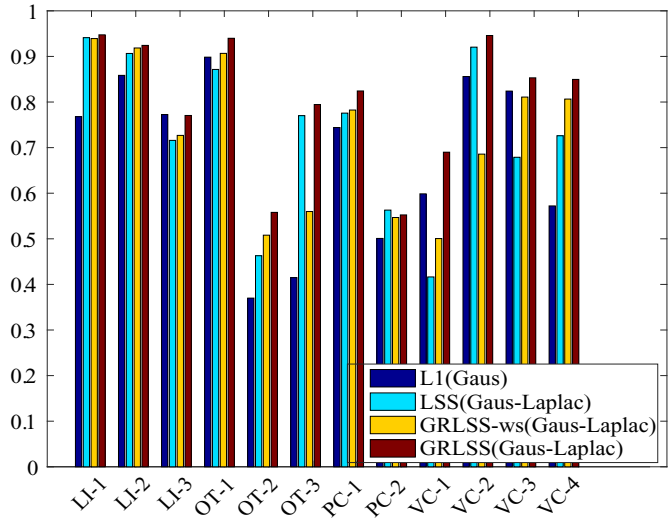

Fig. 10. The AUC value comparison of ℓ_1 [9] (Gaussian modeling), LSS [11], our GRLSS-ws (Gaussian–Laplacian modeling) without motion field separation and GRLSS (Gaussian–Laplacian modeling) with a separation.

PHD model with the integration of different cues can generate an equivalent detection.

Vehicle crossing: These sequences can be divided into two categories that: (1) our car drives fast and the hazardous vehicles cross slowly (i.e., $VC-1$ and $VC-4$), and (2) our car is almost static and the crossing vehicle is with a faster speed (i.e., $VC-2$ and $VC-3$). The former group usually occurs on the mid-line of the road or highway, and the latter kind always appears in the intersection or downtown region. In the first category, because of the dynamic motion of camera, M is weaker than other naive combinations. As for the second group, the motion consistency can be effectively conducted owing to the relatively static camera. Therefore, in $VC-2$ and $VC-3$ sequences, M (95.0%) is better than other naive fusions. From the results of MD (93.4%) > MC (89.2%), MD (93.4%) > MI (92.7%), as shown in Table. 1, the depth band is more discriminative than color and near-infrared bands in $VC-2$. As for $VC-1$ and $VC-3$, the results of MC > MI and MC > MD indicate that color band is superior to depth and near-infrared band. With respect to $VC-4$, MC = MI > MD. Specially, our PHD model is not the best in $VC-4$ sequence. We find the "yellow car" has large distance from our car, and puts little threat to our driving. Therefore, some missed detection is allowed in this sequence. As for our PHD-ALL, it almost can effectively predict the hazards in other sequences in which the hazardous objects have large threat degree.

From the experiments, we find that *different visual clues have distinct discrimination in varying scenarios*. Although fusing more clues can provide more information on prediction, the accurate weighting for them is very important. From above analysis, our PHD model can provide a promising fusion mechanism with an appropriate evaluation for quality of different visual clues, and can manifestly boost the performance than the one utilizing single cue.

Table 3
The running time and AUC (%) comparison of different methods. For a clearer and fairer comparison, the **bold** one is the best AUC result.

Methods	Time cost (ms/frame)	AUC (%)
Faster-RCNN [54]	235	81.55
RFCN [55]	225	80.12
PHD	261	**92.0**

5.5. Competing to previous fusions

To further evaluate the superiority of our PHD model, we here conduct more comparison experiments on the multi-cue fusion strategies. As claimed that the fusion of PHD is transferred to a *map fusion* problem. In the meanwhile, this work infers the motion consistency in an incremental way, and there is no training data available. To make a fairer comparison, this paper compares the PHD with other map fusion strategies. Here, we introduce the work of [52] as a competitor, which introduces the Dempster-Shafer theory into the saliency map fusion. In addition, Boujut et al. [53] claimed that multiplicative fusion performs the best in saliency map fusion. Besides, many state-of-the art saliency detection methods fuse multiple visual clues with an additive way. Therefore, in this work, we compare the performance of Dempster-Shafer fusion (DS-fusion), multiplicative fusion and additive fusion with our PHD model. We conduct the fusion experiments on all the sequences in this work, and the comparison results are shown in Fig. 13. From the results, we can observe that our PHD model is manifestly the best. The main reason is that the obtained hazards maps in different visual bands conflict each other sometimes. Therefore, the performance of multiplicative fusion and DS-fusion is weak. Although additive fusion may contain the true hazards as least, its fusion results introduce more background clutter. However, looking back to PHD, the prior knowledge of hazards makes a better fusion.

5.6. Competing to the state-of-the-art detectors

With an exhaustive investigation, for the road hazards detection, the most related works to this work are the detection based methods, such as pedestrian and vehicle detection. Therefore, this work selects the state-of-the-art pedestrian and vehicle detectors modeled by deep convolution neural networks to compare the performance. Specifically, Faster-RCNN [54] and RFCN [55] are chosen because of the efficiency and effectiveness. The performance comparison is demonstrated in Fig. 14 and Table. 3. From these results, we can observe that our PHD model fusing all the clues can

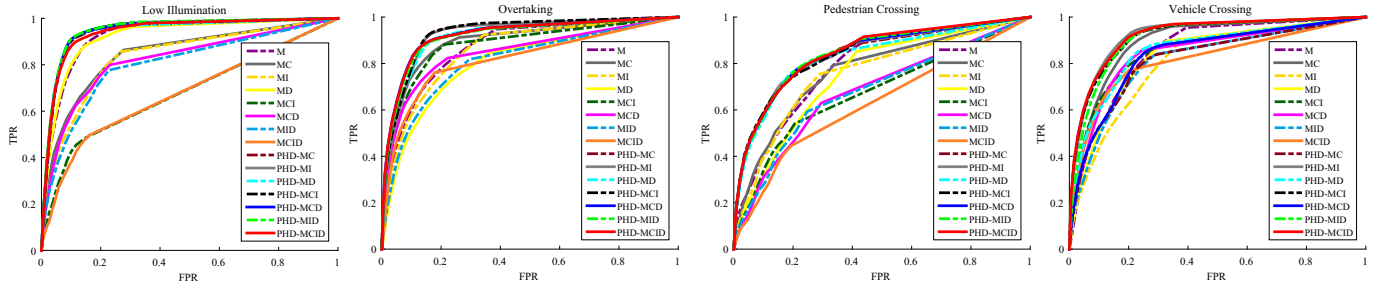

Fig. 11. The ROC curves of different naive information combinations and the proposed saliency based Bayesian fusion for each kind of video clips.

Fig. 12. Typical frameshots of the predicted results by different comparisons for each video clip. All the predicted results are refined by the statistical distribution map of hazards in Fig. 7. The last line marked by the red bounding box is the result by integrating all the cues by our PHD model. In addition, it is clear to find our PHD model can generate more compact, complete and semantical hazardous object region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

be manifestly superior to the RFCN and Faster-RCNN. Actually, the AUC values of RFCN and Faster-RCNN are also lower than PHD-MC, i.e., 87.8% shown in Table. 2 with the consideration of color and motion clues. From Fig. 14(b), we find that detector based methods can not handle the low illumination condition. In addition, the participants are indistinctively detected or lost by vision detectors, where there is no hazardous meaning when perceiving. Actually, these detection failures of vision detectors may be general because of no semantical reasoning for hazards detection.

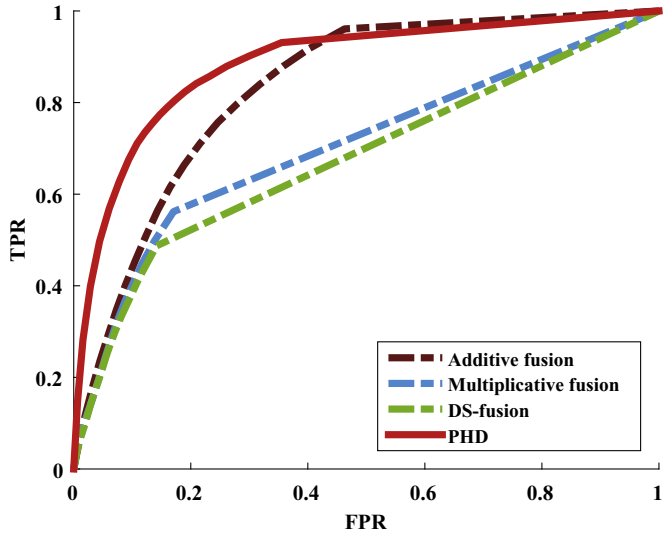

Fig. 13. Performance comparison of PHD model with other fusions for all the sequences. Note that all the clues are considered in this comparison.

In addition to detection performance, we also compare the computational complexity. Beside the SLIC superpixel segmentation having the linear computational cost $\mathcal{O}(n)$ [40], where n is the pixel number, the main cost is the computation of our GRLSS and the saliency detection. Actually, for GRLSS and saliency detection in this work, the largest time consumption is paid by the affinity matrix with a calculation of N (i.e., superpixel number) nodes. In our GRLSS, there is one affinity matrix which needs to be computed, and three for the multi-source saliency detection, i.e., the ones of color, near-infrared and depth. The largest computational complexity of affinity matrixes is $\mathcal{O}(N^2)$ without any refinement. Because $N \ll n$, the affinity matrix can be efficiently computed. That is proved by that with the runtime observation by a MATLAB compiler, the main time consumption is taken by SLIC superpixel segmentation whose average runtime with $N = 125$ superpixel number is about 0.080 s on a machine with Intel i7-6820 2.7 GHz CPU and 8 GB RAM. Despite our work requires computing four affinity matrixes, the segmented superpixel grid can be repeatedly used for the saliency computation and the motion measurement, and each of saliency maps only spends 0.043s. The solving of GRLSS only costs 0.026s. Therefore, the total average runtime of this work is 0.261s without code optimization. The detailed running time of our method and vision detectors is listed in Table. 3. We can see that our PHD model can generate a competitive efficiency.

6. Discussion

6.1. Reliability of the depth information

Depth information is a promising compensation in detection applications. However, from the experimental analysis for different information combinations, the recovered depth map plays different important role in hazards prediction. That is because the performance of Zhang et al. [37] still has some space for improvement, and makes a virtual depth information be obtained. Therefore, this work treats it as an assistant. For a more suitable and reasonable utilization of the recovered depth band, this work extracts the object in depth band in the same way with color and near-infrared bands (in saliency manner), instead of the physical distance measurement between driving vehicle and front objects. With the experimental evaluation, we find the complementary role of depth band for hazards prediction can be effectively demonstrated by the formulated saliency Bayesian integration model.

6.2. Feasibility of the near-infrared information

The wavelength of the near-infrared band in this work is 790 nm. It is a fixed parameter in the utilized multi-spectral camera. From the spectrum analysis, this band has strong reflection ability for organic materials, such as the cotton of clothes and green vegetation, which demonstrates a high intensity in the visualized near-infrared image. On the contrary, for the road and vehicle manufactured by the hybrid materials, because of the difference of the mixture ratio of hybrid materials, it is difficult to unify the reflection ability of near-infrared spectral to them. Therefore, in the visualized near-infrared image, the intensity of the vehicles is not the same. In this work, because the hazardous object extraction in near-infrared image is based on the intensity difference of the object and background, and the intensity difference may vary in distinct scenes, the fusion of near-infrared information is necessary, and a weighed fusion from Bayesian perspective is more reasonable and appropriate. It is verified by the experimental analysis above.

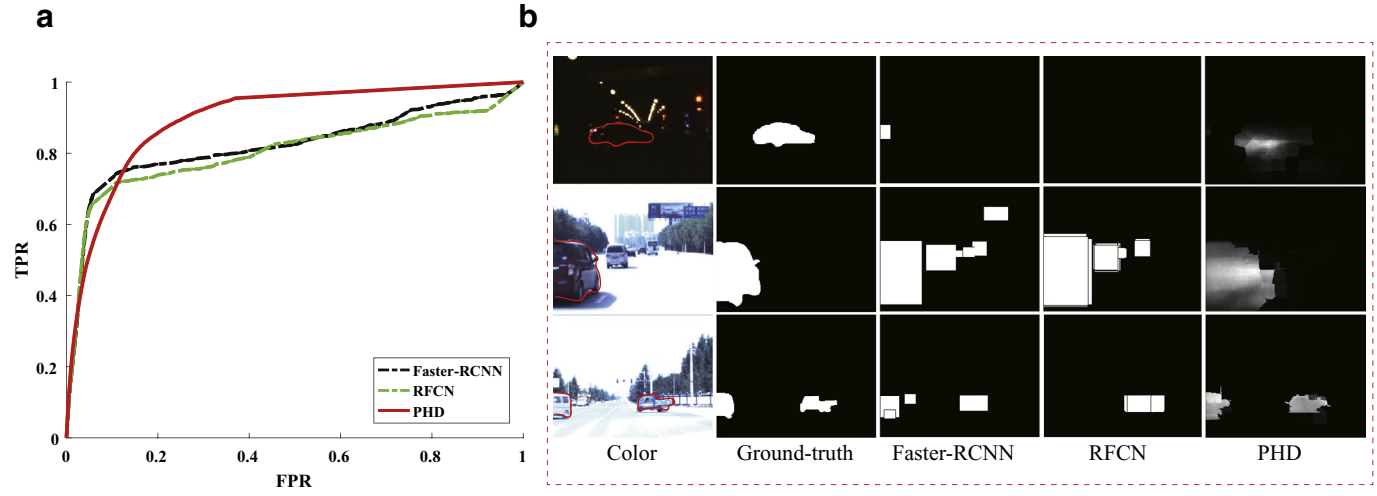

Fig. 14. Performance comparison of PHD model with other state-of-the-art vision detectors. (a) is the ROC curves of different methods on all of the sequences, and (b) is the typical snapshot for demonstrating the visual detection result by different methods.

6.3. Limitations and future works

The proposed system for perceiving hazards is currently validated with twelve video sequences, where the ground truth of each video sequence is obtained by computing the overlapping regions of ten drivers' labeling results. We recognize the importance of enlarging the dataset for verifying. Considering the different effect of visibility on the drivers with different experience [56], we will carry out the tests of the effectiveness of the method for informing drivers with more video sequences and more time to obtain a credible result by detailed designing and statistics. Therefore, the ultimate goal in this paper is to predict the hazards as early as possible. The technique in this system is implemented with MATLAB compiler. In the future, we will accelerate it with a C/C++ compiler. In addition, we find that the estimated optical flow is vulnerable to the motion blur in driving scenarios. Therefore, some works, such as [57] that can classify the blurred region of images, will more efficiently make the extracted frontal target clearer, which also can be regarded in our future works.

7. Conclusion

Based on investigation for safe driving launched in this work, the feedbacks of more than one hundred of drivers with different driving experience demonstrated that the irregular motion behavior and low illumination condition are highly threatening to drivers. In response to this observation, we have proposed and implemented a method for perceiving hazards in driving. It is implemented by an incremental motion consistency measurement and a novel Bayesian integration of multi-source clues. Specifically, a graph regularized least soft-threshold squares (GRLSS) is proposed to conduct the motion consistency measurement. It performs better than original LSS and other methods which utilize Gaussian error modeling. The method of multi-source clue integration is designed to adaptively fuse meaningful bands for the hazards prediction. The experimental results demonstrate the effectiveness and efficiency of the proposed method. Through the work, the advantages of our method can be summarized:

(1) Motion consistency measurement needs better motion noise modeling (Gaussian–Laplacian noise modeling showed better performance than the Gaussian-distribution only.), and has a meaningful reasoning for hazards detection.
(2) Multi-source video information is synthesized with a saliency based Bayesian model to predict hazards, and the content of multi-source video is accurately evaluated with a best band(s) selection in predicting. Besides, the proposed method demonstrated superior performance than vision based detectors.
(3) Twelve video sequences containing road hazards are captured by ourselves containing RGB, near-infrared and recovered depth channels simultaneously. The provided video sequences have the same view and resolution requiring no registration or frame alignment work.

In the future, we would like to fuse more spectral bands into hazards prediction and handle more kinds of hazards. The key point is how to select the appropriate spectral bands and integrate them efficiently. Besides, the powerful support of multi-source information could boost some traditional techniques, such as object tracking and video-based semantic segmentation.

Acknowledgments

This work is supported by the National Key R&D Program Project under Grant 2016YFB1001004, National Key R&D Program of China under Grant 2017YFB1002200, Natural Science Foundation of China under Grant Nos. 61603057, 61773316 and 61379094, and in part by the China Postdoctoral Science Foundation under Grant 2017M613152, Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2017JQ6041, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, Collaborative Research with MSRA, and the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

References

[1] A. Kuznetsova, S.J. Hwang, B. Rosenhahn, L. Sigal, Expanding object detector's horizon: incremental learning framework for object detection in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 28–36.
[2] Global status report on road safety 2015, http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/.
[3] I. Yankson, E. Browne, H. Tagbor, P. Donkor, R. Quansah, G. Asare, C. Mock, B. Ebel, Reporting on road traffic injury: content analysis of injuries and prevention opportunities in ghanaian newspapers, Inj. Prev. 16 (3) (2010) 194–197.
[4] W. Yao, Q. Zeng, Y. Lin, D. Xu, H. Zhao, F. Guillemard, S. Geronimi, F. Aioun, On-road vehicle trajectory collection and scene-based lane change analysis part ii, IEEE Trans. Intell. Transp. Syst. s18 (1) (2017) 206–220.
[5] R.K. Satzoda, M.M. Trivedi, Overtaking and receding vehicle detection for driver assistance and naturalistic driving studies, in: Proceedings of the IEEE Conference on Intelligent Transportation Systems, 2014, pp. 697–702.
[6] M.S. Kristoffersen, J.V. Dueholm, R.K. Satzoda, M.M. Trivedi, A. Mogelmose, T.B. Moeslund, Towards semantic understanding of surrounding vehicular maneuvers: a panoramic vision-based framework for real-world highway studies, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 1584–1591.
[7] J. Janai, F. Gney, A. Behl, A. Geiger, Computer vision for autonomous vehicles: problems, datasets and state-of-the-art 2017 arXiv: 1704.05519.
[8] Y. Yuan, J. Fang, Q. Wang, Online anomaly detection in crowd scenes via structure analysis, IEEE Trans. Cybern. 45 (3) (2015) 562–575.
[9] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Pattern Recognit. 46 (7) (2013) 1851–1864.
[10] T. Brox, J. Malik, Large displacement optical flow: descriptor matching in variational motion estimation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 500–513.
[11] D. Wang, H. Lu, M. Yang, Robust visual tracking via least soft-threshold squares, IEEE Trans. Circuits Syst. Video Technol. 26 (9) (2016) 1709–1721.
[12] X. Li, Y. Pang, Z. Ji, Relevance preserving projection and ranking based on one–class classification for web image search reranking, IEEE Trans. Image Process. 24 (11) (2015) 4137–4147.
[13] M. Jian, K. Lam, J. Dong, L. Shen, Visual-patch-attention-aware saliency detection, IEEE Trans. Cybern. 45 (8) (2015) 1575–1586.
[14] G. Zhu, Q. Wang, Y. Yuan, P. Yan, Learning saliency by MRF and differential threshold, IEEE Trans. Cybern. 43 (6) (2013) 2032–2043.
[15] J. Shen, D. Wang, X. Li, Depth-aware image seam carving, IEEE Trans. Cybern. 43 (5) (2013) 1453–1461.
[16] B. Lin, Y. Chan, L. Fu, P. Hsiao, L. Chuang, S. Huang, M. Lo, Integrating appearance and edge features for sedan vehicle detection in the blind-spot area, IEEE Trans. Intell. Transp. Syst. 13 (2) (2012) 737–747.
[17] Y. Xu, D. Xu, S. Lin, T. Han, X. Cao, X. Li, Detection of sudden pedestrian crossings for driving assistance systems, IEEE Trans. Syst. Man Cybern. Part B Cybern. 42 (3) (2012) 729–739.
[18] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[19] M. Rezaei, M. Terauchi, Vehicle detection based on multi-feature clues and Dempster–Shafer fusion theory, in: Proceedings of the Pacific-Rim Symposium on Image and Video Technology, 2013, pp. 60–72.
[20] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[21] L.E. Navarroserment, C. Mertz, M. Hebert, Pedestrian detection and tracking using three-dimensional Ladar data, Int. J. Robot. Res. 29 (12) (2010) 1516–1528.
[22] H. Wang, B. Wang, B. Liu, X. Meng, G. Yang, Pedestrian recognition and tracking using 3d Lidar for autonomous vehicle, Robot. Auton. Syst. 88 (2017) 71–78.
[23] S. Sivaraman, M. Trivedi, Real-time vehicle detection using parts at intersections, in: Proceedings of the IEEE Conference on Intelligent Transportation Systems, 2012, pp. 1519–1524.
[24] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935–942.
[25] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Pattern Recognit. 46 (1) (2013) 1851–1864.
[26] M. Thida, H. Eng, P. Remagnino, Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes, IEEE Trans. Cybern. 43 (6) (2013) 2147–2156.
[27] V. Saligrama, Z. Chen, Video anomaly detection based on local statistical aggregates, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2112–2119.

[28] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 18–32.
[29] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 fps in matlab, in: Proceedings of the IEEE Conference on Computer Vision, 2013, pp. 2720–2727.
[30] B. Zhao, L.F.-F. E. Xing, Online detection of unusual events in videos via dynamic sparse coding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3313–3320.
[31] X. Zhu, J. Liu, J. Wang, C. Li, H. Lu, Sparse representation for robust abnormality detection in crowded scenes, Pattern Recognit. 47 (5) (2014) 1791–1799.
[32] L. Wang, L. Wang, H. Lu, P. Zhang, R. Xiang, Saliency detection with recurrent fully convolutional networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 825–841.
[33] Q. Wang, Y. Yuan, P. Yan, Visual saliency by selective contrast, IEEE Trans. Circuits Syst. Video Technol. 23 (7) (2013) 1150–1155.
[34] L. Zhou, Y. Ju, J. Fang, Saliency detection via background invariance in scale space, J. of Electron. Imaging 26 (4) (2017) 1–14. 043021.
[35] Q. Wang, P. Yan, Y. Yuan, X. Li, Multi-spectral saliency detection, Pattern Recognit. Lett. 34 (1) (2013) 34–41.
[36] T. Basha, Y.S. Avidan, Stereo seam carving a geometrically consistent approach, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2513–2525.
[37] G. Zhang, J. Jia, T. Wong, H. Bao, Consistent depth maps recovery from a video sequence, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 974–988.
[38] G. Somasundaram, A. Cherian, V. Morellas, N. Papanikolopoulos, Action recognition using global spatio-temporal features derived from sparse representations, Comput. Vis. Image Underst. 123 (7) (2014) 1–13.
[39] D. Marius, N. Sergiu, Motion estimation using the correlation transform, IEEE Trans. Image Process. 22 (8) (2013) 3260–3270.
[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2281.
[41] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1932–1939.
[42] X. Mei, H. Ling, Robust visual tracking using ℓ_1 minimization, in: Proceedings of the IEEE Conference on Computer Vision, 2009, pp. 1436–1443.
[43] D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, in: Proceedings of the IEEE Conference on Computer Vision, 2007, pp. 1–8.
[44] A. Jameson, Solution of the equation AX+XB=C by inversion of an M*M or N*N matrix, SIAM J. Appl. Math. 16 (5) (1968) 1020–1023.
[45] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1–3) (2008) 125–141.
[46] L. Zhang, C. Yang, H. Lu, R. Xiang, M.H. Yang, Ranking saliency, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017) 1-1.
[47] B. Siciliano, O. Khatib, Multisensor data fusion, in: Computer Vision, A Reference Guide, Springer, 2014, pp. 516–536.
[48] K. Bahador, M. Alaa, K. Fakhreddine, N. Saiedeh, Multisensor data fusion: a review of the state-of-the-art, Inf. Fus. 14 (1) (2013) 28–44.
[49] G. Zhu, F. Porikli, H. Li, Beyond local search: tracking objects everywhere with instance-specific proposals, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 943–951.
[50] P. Federico, K. Philipp, P. Yael, H. Alexander, Saliency filters: contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.
[51] Auto Club Europa (ACE): Reviere der blinkmuffel, http://www.ace-online.de/fileadmin/user_uploads/Der_Club/Dokumente/10.07.2008_Grafik_Blinkmuffel_1.pdf).
[52] X. Wei, Z. Tao, C. Zhang, X. Cao, Structured saliency fusion based on Dempster–Shafer theory, IEEE Signal Process. Lett. 22 (9) (2015) 1345–1349.
[53] H. Boujut, J. Benois-Pineau, R. Megret, Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 436–445.
[54] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2015) 1137.
[55] J. Dai, Y. Li, K. He, J. Sun, R-FCN: object detection via region-based fully convolutional networks, in: Proceedings of The Conference on Neural Information Processing Systems, 2016, pp. 379–387.
[56] P. Konstantopoulos, P. Chapman, D. Crundall, Driver's visual attention as a function of driving experience and visibility. using a driving simulator to explore drivers eye movements in day, night and rain driving, Accid. Anal. Prev. 42 (3) (2010) 827–834.
[57] Y. Pang, H. Zhu, X. Li, X. Li, Classifying discriminative features for blur detection, IEEE Trans. Cybern. 46 (10) (2016) 2220–2227.

Yuan Yuan is currently a Full Professor with the Center for Optical Imagery Analysis and Learning, School of Computer Science, Northwestern Polytechnical University, Xian, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE Transactions and Pattern Recognition, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content.

Jianwu Fang received the Ph.D. degree in SIP (signal and information processing) from the Center for Optical Imagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China in 2015. He is currently a lecturer in the School of Electronic & Control Engineering, Chang'an University, Xi'an, China, and is also a postdoctoral researcher in the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision and pattern recognition.

Qi Wang received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, with the Unmanned System Research Institute, and with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.