Traffic Light Recognition for Complex Scene With Fusion Detections

Xi Li, Huimin Ma, *Member, IEEE*, Xiang Wang, *Student Member, IEEE*, and Xiaoqin Zhang

***Abstract*— Traffic light recognition is one of the important tasks in the studies of intelligent transport system. In this paper, a robust traffic light recognition model based on vision information is introduced for on-vehicle camera applications. Our contribution mainly includes three aspects. First, in order to reduce computational redundancy, the aspect ratio, area, location, and context of traffic lights are utilized as prior information, which establishes a task model for traffic light recognition. Second, in order to improve the accuracy, we propose a series of improved methods based on an aggregate channel feature method, including modifying the channel feature for each types of traffic light and establishing a structure of fusion detectors. Third, we introduce a method of inter-frame information analysis, utilizing detection information of previous frame to modify original proposal regions, which makes the accuracy further improved. In the comparison of other traffic light detection algorithms, our model achieves competitive results on the complex scene VIVA data set. Furthermore, an analysis of small target luminous object detection tasks is given.**

***Index Terms*— Traffic light recognition, fusion detectors, inter-frame analysis.**

I. Introduction

IN RECENT years, automatic driving technology has become the most talked about area of intelligent transport system [1]–[4]. Among many related studies, the task of traffic lights recognition and tracking is one of the most important work to study driving strategy for vehicle control. Taken the actual scene as an example, when a vehicle travels at a traffic junction, driver needs to keep an eye on the traffic lights, vehicles in front and surrounding environment, which makes it difficult for drivers to deal with complex scenes vehicle control. The task of TLR aims to provide a driving strategy, includes giving a start-up warning to drivers and a signal for vehicles control, which provides drivers a legal guidance of urban traffic, and a safer and wiser driving mode as well.

In existing research, to achieve vehicle control, some studies has been proposed to obtain vehicle position information by GPS [5], [6], while determine the vehicle's travel strategy according to the sensor installed in the traffic lights. However, this is difficult to implement, as it needs to rebuild the traffic light control system equipment of the entire driving area. At present, it is difficult to realize such large-scale engineering projects in vast majority of countries and regions. Furthermore, equipments such as radar is difficult to achieve mass production as a general vehicle devices, so the much inexpensive on-vehicle camera, to establish a completely visual-based system model has much universal significance and application value.

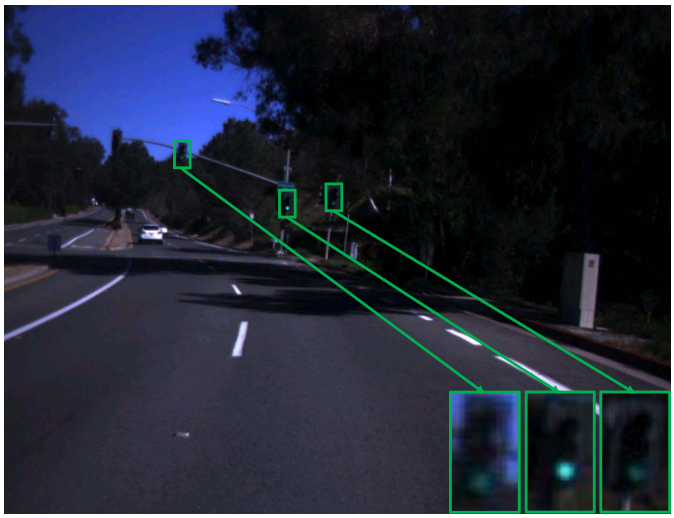


Fig. 1. A general traffic scene in VIVA dataset [7]. Traffic lights are elected and enlarged shown at the right corner in the image.

In this paper, we focus on the full use of visual information. The key problem of our research is to detect traffic lights accurately in video frames which are collected by on-vehicle camera, followed by judging the change of traffic lights and the movement situation of vehicles in front. Fig. 1 shows a general traffic scene include multiple traffic lights. As the target is small and ambiguous, establishing an accurate and robust TLR algorithm is a difficult task.

Meanwhile, a specialized TLR algorithm is established. We consider combining prior information of traffic lights in natural scene with a modified feature learning method, as well as introducing inter-frame information to improve detection performance. In this case, the analysis of prior information is used to improve the efficiency of the calculation, by limiting size, location and other parameters of traffic lights, a large number of meaningless areas can be removed. For the sake of robustness, we select feature learning method instead of a

Manuscript received December 8, 2016; revised March 7, 2017 and April 20, 2017; accepted April 27, 2017. This work was supported in part by the National Key Basic Research Program of China under Grant 2016YFB0100900 and in part by the National Natural Science Foundation of China under Grant 61171113. The Associate Editor for this paper was Q. Wang. *(Corresponding author: Huimin Ma.)*

The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: lixi16@mails.tsinghua.edu.cn; mhmpub@tsinghua.edu.cn; wangxiang14@mails.tsinghua.edu.cn; xiaoqin-15@mails.tsinghua.edu.cn).

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.2017.2749971

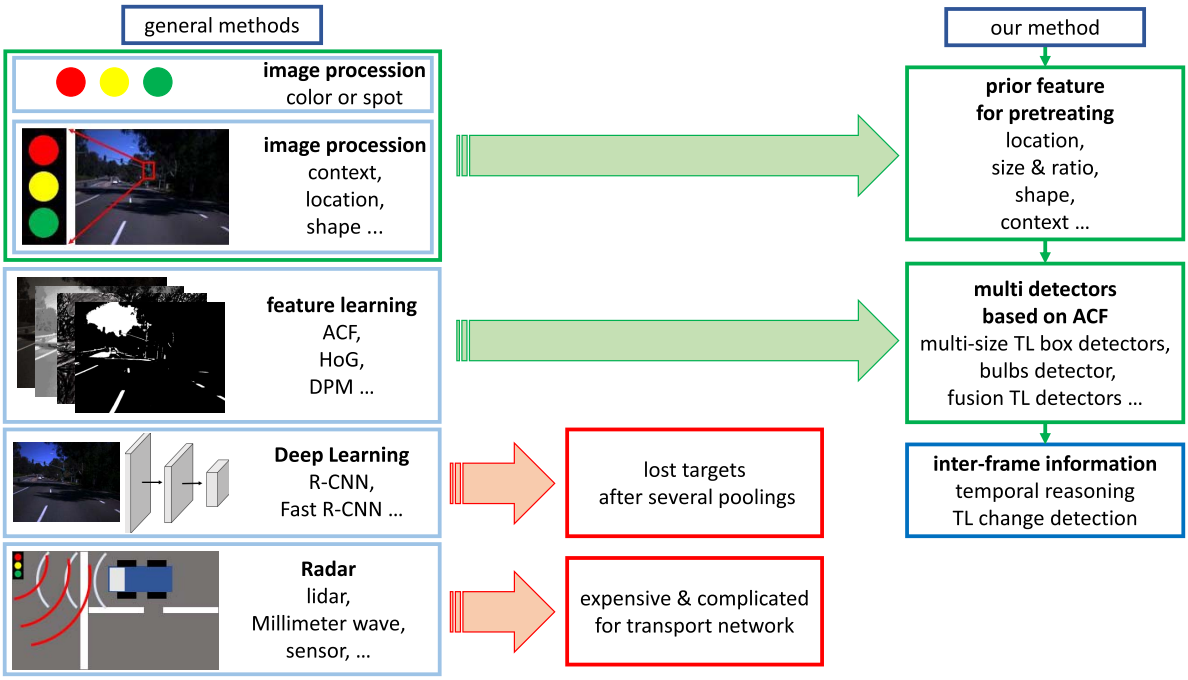


Fig. 2. A list of methods for TLR tasks. The left side enumerated 5 main ideas for TLR, while the right side shows the basic structure in our approach, which is a model combined prior feature with fusion detectors followed by inter frame analysis process.

simple image processing. Based on context and other features of different types of traffic lights, we improve ACF [8] method to make it specialized for TLR. Inter-frame information is used to correct the confident score for candidates, as we find that informations from adjacent frames can solve the problem of luminous fake flashing when imaging appropriately. For clearly demonstration, section III introduces the complete framework of our TLR algorithm and the three components, respectively.

II. Related Work

Aiming at the general object detection task, a lot of work have been proposed and they can be classified into two categories. The first kind of methods are based on bottom feature analysis, such as Deformable Parts Model (DPM) [9], Aggregate Channel Features (ACF) [8] and its improved version [10]–[12]. These methods are used to extract colour, gradient intensity, edge and other features, training a number of detectors to achieve target objects detection and recognition. These methods have high accuracy in traffic sign [3], [4], pedestrian [8], [10] and vehicle detection, and can achieve the detection speed at tens of frames per second on the CPU, which meets the real-time requirements. However, when they come to a small luminescent target, such as traffic light, it is difficult to achieve the accuracy required by the method, as their bottom features are weakly expressed and the detectors often do not possess sufficient specialization.

The second kinds of methods are based on deep learning method, such as R-CNN [13] and its accelerated version [14], [15]. Methods based on deep learning utilize the feature information of higher dimensions to realize the detection of target, and they can achieve higher accuracy than the method of learning bottom features on the general object detection task [1], [2], [16]–[21]. However, the number of pooling layers will be limited due to the tiny size of traffic lights in images taken by on-vehicle camera, which makes it difficult to design a sufficiently deep network [22]. Therefore, in TLR task, it is difficult for a deep learning network to take both the accuracy and parameters' storage into account.

On the other hand, in many studies of TLR, the method of image processing and morphological analysis are widely used [23]–[25]. As traffic lights have a distinctive feature in colour, shape and size, such methods have a certain detection accuracy rate in the case of high image clarity and simple scene. At the same time, the ideas of salient object detection also show the necessity of analysis of target object feature for region proposal [26], [27]. However, for an actual traffic scene, occlusion, weather conditions and other light sources will have a heavy impact for TLR task, the accuracy of these methods will be significantly reduced.

In the research of TLR, the study on VIVA dataset is of most interest. Methods in [7] and [28] proposed the Area-Under-Curve (AUC) results of the three methods on TLR, where AUC represents the area under the Precision-Recall curves of region proposal, which is an indicator both investigated the accuracy and recall rate of an algorithm. Under the requirement of overlap criteria of 50%, the method of colour information can only achieve an average AUC of 4% while using light spot can only reach 1%. Meanwhile, ACF method [8] can achieve about 40%, it can be seen that feature learning methods have more advantages than a simple image processing method, while they are still less than the actual demand.

Although it is difficult to achieve high accuracy by simply using image processing or morphological analysis, we find that the prior features and detection results from adjacent frame can still help to improve performance on the basis

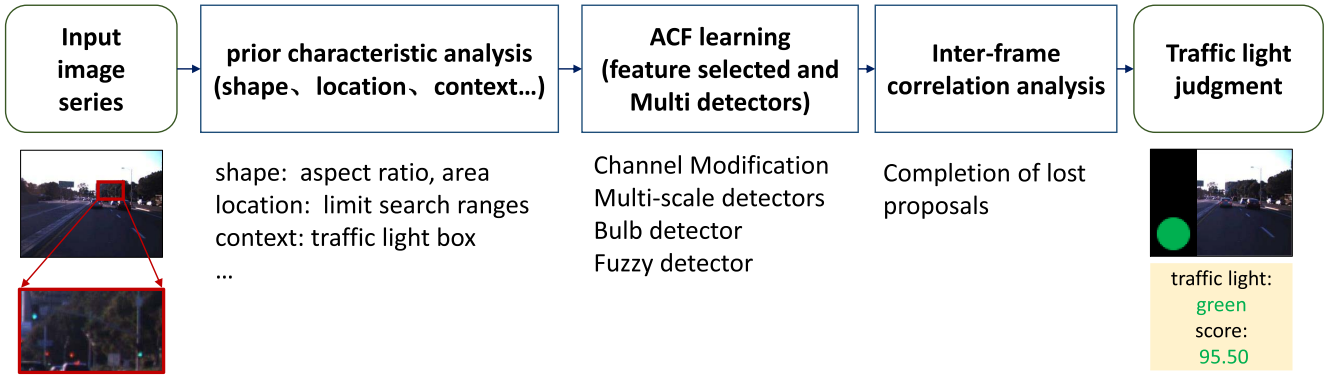


Fig. 3. The framework of our fusion detection model for TLR task. The flow chart shows basic steps: entering a natural traffic scene image, through a process by prior feature analysis, ACF learning and inter-frame correlation analysis, providing the positions and types of traffic lights in the image.

of feature learning algorithms, with a tiny time consumption increased [3], [4], [29]. As a result, we propose a more robust TLR method, which uses a modified feature learning method followed with prior information for specific task, to further improve the accuracy and recall of TLR. Inspired by [7] and [28], we also use AUC as an indicator to evaluate the accuracy and recall rate of our algorithm. Fig. 2 shows a list of methods and our approach for TLR tasks.

Beside, the actual needs of the system are focused on. For TLR task in the real driving scene, there is no need to detect all types of traffic lights and these which are not in front of driving detection, but the red and green lights in the nearest crossroad. Under the constraints of these conditions, a specific model is put forward for red and green light recognition, as well as a main traffic light judging method.

III. Approach

For a specialistic TLR task, we introduce a model which combines a prior feature analysis process with modified feature learning methods, as well as a main traffic light judging standard. Given an image taken from on-vehicle camera, through prior feature analysis, fusion detectors and the using of inter-frame information, finding out all traffic lights is the first step to implementing TLR. Furthermore, selecting the main traffic light which is the one need to observe is the second step for vehicle control. Fig. 3 shows the basic flow of our algorithm. In this section, we give a detailed description of the three contributions presented in our algorithm.

A. Prior Feature Analysis

It is found that traffic lights has a lot of prior information on location, size and scene context. Introducing a prior feature under the task model, followed by a preprocessing operation, will help reduce redundancy and improve algorithm efficiency. For TLR, we have introduced three pre-processing operations, including restricting the search range for sliding window, selecting template size with physical meaning and analysing the structure of traffic lights. Fig. 4 shows prior feature analysis process in TLR task.

In actual traffic scenarios, traffic lights tend to have a fixed frame structure to distinguish them from other sources of

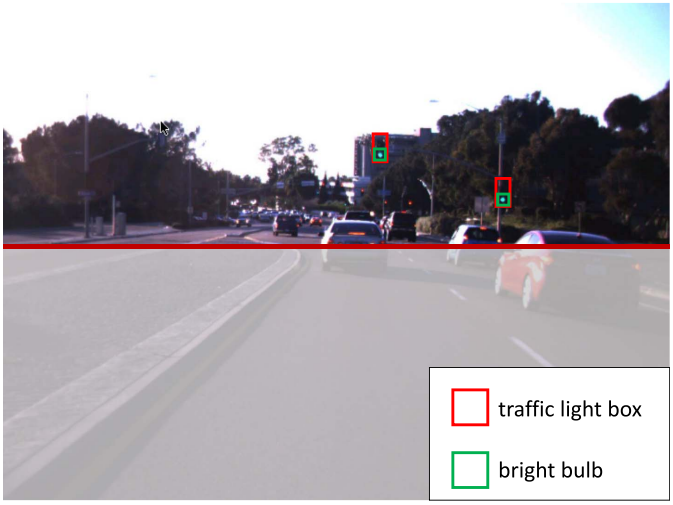


Fig. 4. A schematic of regions proposal based on prior feature. Several candidates with fixed sizes which have physical meaning in the image is selected in prior feature analysis process.

interference. In complex scene, since the bulbs of traffic lights is susceptible to interference from image quality and similar light sources, so it is inappropriate to be used as detection objects. As a result, when we perform feature learning, traffic light boxes with structural information are selected as target.

Second, due to the position of on-vehicle camera and traffic lights in natural scenes, traffic lights must appear in a certain area in acquired images. In most cases, they are present in top 40% of images in LARA dataset, and in top 50% of images in VIVA datasets. Images of the two datasets are continuous video frames from on-vehicle camera, and we will give details of these datasets in section IV. In practical application, the range of sliding window search is decided by the statistic result of locations of ground truth in training dataset, it can be expressed as a probability form for each pixel as $P(x, y)$, which represents for the probability that pixel (x, y) appears in the area that may contains traffic lights. For the simplest producing, we set the lowest position of ground truth as the $threshold$, $P(x, y) = 1$ if vertical coordinate y of the pixel is higher than $threshold$, while $P(x, y) = 0$ if it is the opposite.

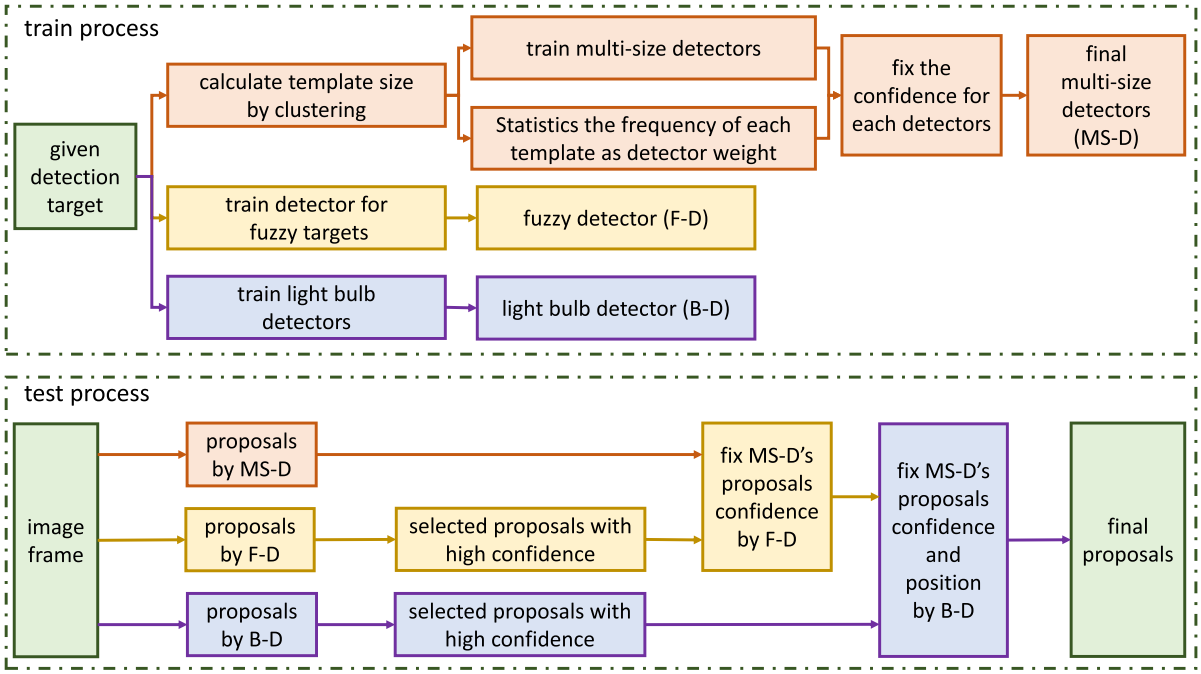


Fig. 5. The complete lifting algorithm for region proposals by the three types of detectors cascade. The upper dashed box shows the train process while the lower shows how to evaluate the proposal method's performance by multi detectors.

Third, as the size, aspect ratio and other information are relatively fixed for traffic lights, we argue that targets which could be identified in images can be included with very few types of template sizes. In order to get template sizes with physical meaning for sliding window operation, we cluster the sizes (length, wide of bound boxes) of ground truth in training dataset: first, recording the number of each different sizes, then sorting the number of clusters from large to small, and judging whether a size is close to others in front of it in this order. If the length and width ratio both are in the range of 0.8 to 1.25, then merging these two sizes to generate a new template size and using the number of clusters as weight, as formula (1), (2), (3):

$$N(l_{new}, w_{new}) = N(l_1, w_1) + N(l_2, w_2) \quad (1)$$

$$l_{new} = \frac{N(l_1, w_1)}{N(l_{new}, w_{new})} \times l_1 + \frac{N(l_2, w_2)}{N(l_{new}, w_{new})} \times l_2 \quad (2)$$

$$w_{new} = \frac{N(l_1, w_1)}{N(l_{new}, w_{new})} \times w_1 + \frac{N(l_2, w_2)}{N(l_{new}, w_{new})} \times w_2 \quad (3)$$

$N(l, w)$ represents the number of size in which length is l and width is w, then the new size (l_{new}, w_{new}) is obtained by weighting calculation, while its correspond number $N(l_{new}, w_{new})$ is obtained by summing.

B. Fusion Detectors Based on Prior

1) Channel Feature Modification: As traffic lights in the images obtained from on-vehicle camera are much smaller than the whole image size, and they do not have complex features, so the method of extracting basic features followed by a matching algorithm is more suitable for TLR tasks. Aggregation Channel Feature (ACF) model [8] is used as a baseline in our model, which belongs to a detection algorithm that achieves object recognition by template matching. The algorithm framework includes a clustering feature channels process, followed with a boosting tree for classification. And for the features of traffic lights, we redesign the channels base on general ACF model.

In the study, we find that the red light is more susceptible to similar light sources, the chroma is relatively stable, while the green light is more susceptible to image brightness. On the other hand, as the three-axis in CIE-Lab colour space represents brightness, red-green-axis and yellow-blue axis, which better expresses the colour-field features of traffic light, inspired by [30], we consider fusing CIE-Lab's channels into a single one, to replace one of LUV's chroma channel for generating red light's feature, while the luminance channel for green light. Formula (4) introduces the C_{lab} channel:

$$C_{lab}(x, y) = l(x, y) \cdot (a(x, y) + b(x, y)) \quad (4)$$

l, a and b represent the 3 channels defined in CIE-Lab, (x, y) is a pixel, its C_{lab} is calculated from the CIE-Lab channel feature of this point.

2) Multi-Size Detector: For complex scenes, a single ACF detector with channel feature modification doesn't achieve sufficiently performance. At the same time, much information for TLR can be obtained by ACF method. Here, we propose a method of fusing detection results, through ACF method, a number of different detectors for different target can be trained, according to the relationship between different types of lights, as well as bulbs and traffic light boxes, the candidate regions can be corrected by multiple sets of detectors. In particular, by introducing multi-size detector, fuzzy detector and bulb detector, we design a multiple detector model for TLR. Fig. 5 shows the process of three types of detectors cascade.

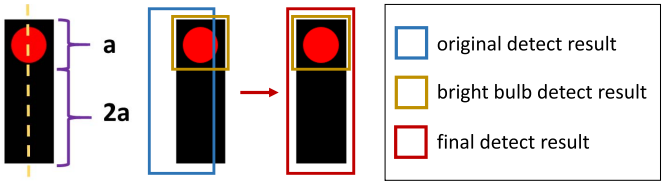


Fig. 6. Modify candidate regions by a weak prior constraint of the relative position of bright bulb and traffic light box.

For multi-size detectors, by the clustering method for template sizes mentioned in prior feature analysis, we can select multiple sizes for training ACF detectors. According to the proportion of samples that can be covered, the number of detectors of different template sizes can also be determined. For example, a single size can be used to satisfy the vast majority of samples for image scenes from LARA dataset, while template on VIVA dataset needs three. In addition, when merging candidate regions given by detectors of different template sizes, the confidence scores of candidates obtained by different detectors need to be weighted according to its corresponding number of clustering samples. This will give a meaningful sort for candidate regions. Formula (5), (6), (7) shows the process of merging proposals:

$$\frac{weight_1}{N(l_1, w_1)} = \frac{weight_2}{N(l_2, w_2)} = \frac{weight_3}{N(l_3, w_3)} = \cdots \quad (5)$$

$$c_x(pos, score) \in S_x \rightarrow c_x(pos, score \cdot weight_x) \in S_x \quad (6)$$

$$S_{final} = \{S_1, S_2, S_3, \cdots\} \quad (7)$$

$weight_x$ represents modification weights for different candidates, it is proportional to the number of clusters $N(l_x, w_x)$ of its detector's template sizes. By the parameter, we weight confidence scores for each element c_x belongs to the set of candidate regions S_x which obtained by detector x. Finally, we combine the updated candidates to get a final set S_{final}.

3) Bulb Detector: As mentioned above, when only using the bright bulb of traffic lights as detection target, the performance is always unsatisfactory because the tiny size and it's susceptible to interference. However, we find that there is a relatively fixed positional relationship between the bright bulb and traffic light box, which can also provide information for modifying candidate regions obtained by ACF detectors for traffic light boxes.In our research, candidates with high confidence score obtained from bright bulb ACF detector is used to compare with traffic light boxes candidate: if the upper 1/3 of red traffic light (lower 1/3 part of green traffic light) box has an IoU (the ratio of Intersection of Union) greater than 50% with a red (green) bulb candidate, then we add 20% of the bulb's confidence score to the traffic light box candidate, as well as modifying box's abscissa according to the bulb.

Fig. 6 shows the modify process by a weak prior constraint of the relative position between the bright bulb and traffic light boxes, while formula (8), (9), (10), (11) shows the process of modifying by bulb detector.

$$bulb_m(pos_m, score_m) \in score_{bulb} > threshold \quad (8)$$

$$TLbox_n(pos_n, score_n) \in S_{final} \quad (9)$$

$$score_n \rightarrow score_n + score_m \times 20\% \quad (10)$$

$$pos_n(x_n, y_n) \rightarrow pos(x_m \times 0.2 + x_n \times 0.8, y_n) \quad (11)$$

Here, $bulb_m$ belongs to the set of bulb candidates with confidence greater than $threshold$ (generally select 200 for greens while 150 for reds), $TLbox_n$ is a target from traffic light boxes candidates set S_{final}. If the IoU of $bulb_m$ and $TLbox_n$ is greater than 50%, (10), (11) are used to modify the postion $pos_n(x_n, y_n)$ and confident score $score_n$ of $TLbox_n$.

4) Fuzzy Detector: For low quality images, as a small target luminous body, some traffic lights are difficult to be classified, single-class ACF detector may lose such targets. However, in the task of continuous frame detection, even if it can not correctly determine its type, judging whether it is a traffic light is also meaningful. As a result, the fuzzy detectors which see multiple categories as ground truth are introduced (For VIVA dataset, we train a detector using red and red left lights as ground truth, as well as the greens). For a candidate with high confident score obtained by fuzzy detector: if it has an IoU greater than 50% with a candidate from single type detection, adding its confidence score with weighted to the candidate, which operation is the same as formula (10); otherwise adding it to corresponding single type traffic light sets with a reduced confidence score as formula (13) and (14).

$$fuzzy_m(pos_m, score_m) \in score_{fuzzy} > threshold \quad (12)$$

$$1c_x(pos, score) \in S_{fz} \rightarrow c_x(pos, score \cdot 0.5) \in S_{fz} \quad (13)$$

$$S_{final} \rightarrow \{S_{final}, S_{fz}\} \quad (14)$$

$fuzzy_m$ represents a fuzzy candidate with confidence greater than $threshold$ (generally select 200 for greens while 150 for reds), if it doesn't have an IoU greater than 50% with any candidates from single type detection, we reduce the confidence of these candidates to 20%, and record them as a set S_{fz}, then add this set to the original candidates set S_{final}.

5) Inter-Frame Correlation Analysis: In addition to using information from single image frame, the detection results provided by adjacent frames can also provide information. as it is impossible that the position of traffic lights changing abruptly during continuous video, comparing the candidate regions extracted from adjacent frames and their scores can help to modify the performance. In particular, false light flickering often occurs when luminous objects imaging, which makes the detection confidence of targets in adjacent images often obvious different. Therefore, the results from previous frame can be used to reduce the difference of false-flicker on detecting traffic lights in a single frame. Fig. 7 shows the differences of traffic light images in adjacent frames influenced by false flicker phenomenon. Referring to the detection result of adjacent frame, confidence scores of candidates extracted in current image can be recalculate. In our research, the modified method is designed as the same with fuzzy detector.

C. Main Traffic Light Definition

For actual driving control, after multiple traffic lights are detected, the task is to find out a "main traffic light" for vehicle control. Here, we put forward the definition of main traffic

Fig. 7. Different imaging quality of traffic lights in adjacent frames. The false flicker phenomenon makes objects missed when detecting in single frame.

Fig. 8. The definition of a "Main Traffic Light", which judging by the traffic lights' size, position and confidence score obtained from our algorithm.

light, which is the one really needs to be seen. According to the imaging rules under natural scenes, the main traffic light should be larger than majority of other lights which are observed, as well as a relatively high position in the image. In this case, we propose that the main traffic light should be present in candidate regions obtained by our algorithm and following requirements below: first, it must belong to targets in which score is greater than a threshold (In VIVA, we generally select 200 for greens while 150 for reds); second, its size needs to be greater than 80% of all regions that meet the first requirements; third, its position (measured by the ordinate of area's centre) needs to be the highest of all regions under the above conditions. Fig. 8 shows a result for selecting the main traffic light defined above.

IV. Experiments

A. LARA Dataset

LARA dataset is a video built in an urban traffic scene, the image quality is relatively high, and contains red, green lights' labels, which treat the whole traffic light boxes as ground truth. Based on the LARA dataset, we have studied the prior feature analysis and the feature channel modification

TABLE I
AUC Indicator of Red Light on LARA Validation Dataset by Channel Modification

Method	Colour	Amp	Hist	AUC indicator
ACF [8]	LUV	✓	✓	89.77%
Ours	LUV([1 2]) + C_{lab}	✓	✓	90.65%
Ours	LUV([1 3]) + C_{lab}	✓	✓	87.82%
Ours	**LUV([2 3]) + C_{lab}**	✓	✓	**91.97%**
Ours	LUV([2 3]) + C_{lab}	×	✓	91.06%
Ours	LUV([2 3]) + C_{lab}	✓	×	87.33%
Ours	LUV([2 3]) + C_{lab}	×	×	87.01%

TABLE II
AUC Indicator of Green Light on LARA Validation Dataset by Channel Modification

Method	Colour	Amp	Hist	AUC indicator
ACF [8]	LUV	✓	✓	84.03%
Ours	LUV([1 2]) + C_{lab}	✓	✓	86.57%
Ours	LUV([1 3]) + C_{lab}	✓	✓	87.98%
Ours	LUV([2 3]) + C_{lab}	✓	✓	84.04%
Ours	**LUV([1 3]) + C_{lab}**	×	✓	**89.32%**
Ours	LUV([1 3]) + C_{lab}	✓	×	86.18%
Ours	LUV([1 3]) + C_{lab}	×	×	80.24%

TABLE III
AUC Indicator Based on Multi-Size Detectors Method on VIVA Validation Dataset

Traffic Light	Method	AUC indicator	Improve
Red	general ACF detector [8]	63.29%	———
Red	Double-size detector + score fixed	66.83%	+3.54%
Red	**Treble-size detector + score fixed**	**68.02%**	**+4.73%**
Red Left	general ACF detector [8]	13.27%	———
Red Left	Double-size detector + score fixed	17.01%	+3.74%
Red Left	**Treble-size detector + score fixed**	**17.02%**	**+3.75%**
Green	general ACF detector [8]	40.26%	———
Green	Double-size detector + score fixed	48.33%	+8.07%
Green	**Treble-size detector + score fixed**	**48.98%**	**+8.72%**
Green Left	general ACF detector [8]	36.56%	———
Green Left	Double-size detector + score fixed	37.62%	+1.06%
Green Left	**Treble-size detector + score fixed**	**37.67%**	**+1.11%**

TABLE IV
AUC INDICATOR BASED ON MULTI DETECTORS ON VIVA VALIDATION DATASET

Traffic Light	Method	AUC indicator	Improve
Red	general ACF detector [8]	63.29%	———
Red	general ACF detector + Fuzzy detector	64.87%	+1.58%
Red	general ACF detector + Bulb detector	65.12%	+1.83%
Red	general ACF detector + Fuzzy detector + Bulb detector	65.47%	+2.18%
Red	Multi-size detectors	68.02%	+4.73%
Red	Multi-size detectors + Fuzzy detector	69.31%	+6.02%
Red	Multi-size detectors + Bulb detector	70.43%	+7.14%
Red	**Multi-size detectors + Fuzzy detector + Bulb detector**	**71.26%**	**+7.97%**
Red Left	general ACF detector [8]	13.27%	———
Red Left	general ACF detector + Fuzzy detector	17.82%	+4.55%
Red Left	general ACF detector + Bulb detector	15.49%	+2.22%
Red Left	general ACF detector + Fuzzy detector + Bulb detector	20.00%	+6.73%
Red Left	Multi-size detectors	17.02%	+3.75%
Red Left	Multi-size detectors + Fuzzy detector	22.09%	+8.82%
Red Left	Multi-size detectors + Bulb detector	19.23%	+5.96%
Red Left	**Multi-size detectors + Fuzzy detector + Bulb detector**	**23.14%**	**+9.87%**
Green	general ACF detector [8]	40.26%	———
Green	general ACF detector + Fuzzy detector	40.31%	+0.05%
Green	general ACF detector + Bulb detector	40.20%	-0.06%
Green	general ACF detector + Fuzzy detector + Bulb detector	41.07%	+0.81%
Green	Multi-sizes detector	48.98%	+8.72%
Green	Multi-sizes detector + Fuzzy detector	51.74%	+11.48%
Green	Multi-sizes detector + Bulb detector	48.95%	+8.69%
Green	**Multi-size detectors + Fuzzy detector + Bulb detector**	**51.83%**	**+11.57%**
Green Left	general ACF detector [8]	36.56%	———
Green Left	general ACF detector + Fuzzy detector	37.54%	+0.98%
Green Left	general ACF detector + Bulb detector	38.44%	+1.88%
Green Left	general ACF detector + Fuzzy detector + Bulb detector	39.39%	+2.83%
Green Left	Multi-sizes detector	37.67%	+1.11%
Green Left	Multi-sizes detector + Fuzzy detector	38.19%	+1.63%
Green Left	Multi-sizes detector + Bulb detector	39.80%	+3.24%
Green Left	**Multi-size detectors + Fuzzy detector + Bulb detector**	**40.20%**	**+3.64%**

TABLE V
AUC INDICATOR BASED ON INTER-FRAME INFORMATION ON VIVA VALIDATION DATASET

Traffic Light	Method	AUC indicator	Improve
Red	general ACF detector [8]	63.29%	———
Red	general ACF detector + Inter-frame information	66.05%	+2.76%
Red	Multi-detector	71.26%	+7.97%
Red	**Multi-detector + Inter-frame information**	**71.50%**	**+8.21%**
Red Left	general ACF detector [8]	13.27%	———
Red Left	general ACF detector + inter-frame information	18.40%	+5.13%
Red Left	Multi-detector	23.14%	+9.87%
Red Left	**Multi-detector + inter-frame information**	**27.67%**	**+14.40%**
Green	general ACF detector [8]	40.26%	———
Green	general ACF detector + inter-frame information	44.44%	+4.18%
Green	Multi-detector	51.83%	+11.57%
Green	**Multi-detector + inter-frame information**	**52.16%**	**+11.90%**
Green Left	general ACF detector [8]	36.56%	———
Green Left	general ACF detector + inter-frame information	36.89%	+0.33%
Green Left	Multi-detector	40.20%	+3.64%
Green Left	**Multi-detector + inter-frame information**	**40.47%**	**+3.91%**

method. For limiting the number of template, we select a single template as the clustering on LARA dataset is very concentrated. Furthermore, the influence of each channel on detection AUC of ACF method is analysis during testing on LARA dataset. We introduces the feature channel based on the colour domain of traffic light, and replaces the luminance and chroma channels of the original LUV colour channel model separately. The test accuracy is shown in Table I and Table II. LUV([m1, m2]) represent for $m1$-th and $m2$-th channels in LUV are chosen, while C_{lab} is the channel compute by our channel feature modification method.Here, AUC represent the area-under-curve on Precision-Recall curves, where precision and recall are calculate by formula (15), (16):

$$Presicion = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

Where TP represents the number of True Positive Samples (We define candidates with IoUs greater than 50% with ground truth as TP), while FP and FN correspond to false positive and negative value. When different confidence

Fig. 9. Vision results on VIVA validation dataset by our TLR model. The left side is the detection result on original image, with the main traffic light's guide (go, stop, and etc.) and confident score are indicated in the upper left corner. The right side shows the enlarging area where traffic lights is detected.

thresholds are taken, the PR curve can be drawn. A larger AUC value naturally corresponds to a better algorithm.

It can be seen that, as the object is small, the colour channel is still the most heavily dependent in traffic light detection, and the gradient feature expression is relatively vague. While when we introduce the channel which is closer to the traffic lights' colour feature expression, the accuracy is improved when the new channel replaces the luminance channel of the red light and the chroma channel of the green light. This also shows that, for different lights, the function of each feature channel and there sensitive interference are also different. As a result, we propose that higher performance could be achieved by detecting different traffic lights based on discriminate feature.

B. VIVA Dataset

VIVA dataset contains a number of different videos whose scenes are much complex, and the traffic lights' imaging quality is unstable. The VIVA dataset provides more types of traffic lights' labels(red, green, and their corresponding left turn lights), while it provides both the whole traffic light boxes

TABLE VI
THE COMPARISON OF OUR PROPOSAL MODEL WITH STATE-OF-THE-ART METHODS ON VIVA VALIDATION DATASET

Method	Red	Red Left	Green	Green Left	Platform	Time Cost(s)
Faster R-CNN [15]	13.61%	1.20%	19.30%	0.13%	GPU	0.020
SLD [24]	7.54%	———	10.01%	———	CPU	**0.014**
ACF [8]	63.29%	13.27%	40.26%	36.56%	CPU	0.043
Ours	**71.50%**	**27.67%**	**52.16%**	**40.47%**	CPU	0.081

and bright lights' labels. Based on the richer information on VIVA dataset, a total experiment using our complete model proposed could be done.

First, as the distribution of traffic lights' size dimension on VIVA dataset is scatter, so AUC indicator of the algorithm will be greatly reduced when only one size of template is used. Therefore, the use of multi-size template is considered. In the experiments, first three clustered templates with the largest data size are selected and the detectors are trained respectively. Here, due to the fact that different scales detectors are positively correlated with the amount of data clustered on the template scale, it is necessary to adjust the confidence score of candidate regions obtained from the different detection templates. Therefore, we record the number of different scales of templates obtained by clustering in training set as weights, then multiply them with the confidence scores of the candidate regions obtained by each template, followed with a re-ranking and integrating for the candidates. Table III shows the lifting AUC results for multi-size detectors model.

Second, we train red fuzzy detector which uses red and red left lights together as ground truth, and the same for green fuzzy detector. In addition, we also train four bulb detectors which mark the four bright lights as ground truth. The training methods of the above two detectors are the same as general ACF detector. Based on the method proposed in section III, the proposals and their corresponding confidence score obtained from single-size and multi-size detector model of the four kinds of traffic lights are modified by the two detectors. The improved AUC results are shown in Table IV.

Third, for each segment of continuous video stream (in addition to the first), we pick up candidate regions with high confidence which are detected from previous frame, and calculates the IoU of candidate regions detected from current frame. If a region has an IoU which greater than 50% with any candidate object selected in previous frame, we add 20% of corresponding target's score from previous frame to its candidate, while if a selected region in previous frame doesn't have an IoU greater than 50% with regions, we reduce its corresponding confidence score by 20% and add the region into candidates of the current frame, this is based on the assumption that traffic lights does not suddenly appear or disappear in a continuous video. Table V shows the lifting results by introducing inter-frame information, while several vision results are shown in Fig. 9.

Furthermore, we compare our algorithm with some famous object detection model, including the Faster R-CNN structure [15] with ZF-net [31], ACF model [8] and an image analysis algorithm for specific TLR SLD [24], the AUC indicator, platform we test on and time cost are shown in table VI. All the experimental content of this part is reproduced by us referring to [8], [15], and [24]. The Faster R-CNN model tests on TITAN X with GPU @ 2.5 GHz while others test on CPU @2.6 GHz. We verify that the deep neural network is very difficult to locate accurately because of the small size of traffic lights. Similarly, the method based on prior image analysis is difficult to obtain well performance in images with complex scenes and low quality. Our approach introduces a prior knowledge and a modified method on the basis of ACF, the AUC indicators of four traffic lights have been significantly improved. At the same time, each module can be used alone in our algorithm. In practical applications, the modules in our proposed model can be freely used according to the requirements for speed and performance.

V. CONCLUSION

In this paper, a TLR proposal method is proposed. Our detection model combination prior feature and inter-frame analysis with feature learning algorithms. Through statistics of shapes, location and context of traffic lights, only candidate regions with physical meaning are selected. Meanwhile, the fusion detection model introduce structural features of different types of traffic lights, our specific model for traffic light significantly improves the performance of standard ACF method. While the introduction of adjacent frame information solves the drawbacks when detecting the fake flashing luminous objects by single image frame.

Furthermore, to solve the task of detecting small luminous objects in complex scenes, our methods for TLR are also of reference. For a specific object recognition task, it is a good way to establish a model which combines prior feature with statistic learning. In particular, it is an important method to improve the recognition of small luminous objects with poor image quality, which is difficult to extract features. In this case, constructing a variety of detectors through introducing a prior knowledge, can be an essential way for luminous object detection and recognition.

REFERENCES

[1] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2147–2156.

[2] X. Chen *et al.*, "3D object proposals for accurate object class detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 424–432.

[3] Y. Yuan, Z. Xiong, and Q. Wang, "An incremental framework for video-based traffic sign detection, tracking, and recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1918–1929, Jul. 2016.

[4] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2016.

[5] D. Barnes, W. Maddern, and I. Posner, "Exploiting 3D semantic scene priors for online traffic light interpretation," in *Proc. IEEE Intell. Veh. Symp. (IV)*, Jun./Jul. 2015, pp. 573–578.

[6] A. E. Gomez, F. A. R. Alencar, P. V. Prado, F. S. Osorio, and D. F. Wolf, "Traffic lights detection and state estimation using hidden Markov models," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2014, pp. 750–755.

[7] M. P. Philipsen, M. B. Jensen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Traffic light detection: A learning algorithm and evaluations on challenging dataset," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 2341–2345.

[8] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[9] 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.

[10] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 82–90.

[11] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 424–432.

[12] Q. Li, Y. Yan, and H. Wang, "Discriminative weighted sparse partial least squares for human detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1062–1071, Apr. 2016.

[13] 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 Recognit.*, Jun. 2014, pp. 580–587.

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

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[16] S. Noh, D. Shim, and M. Jeon, "Adaptive sliding-window strategy for vehicle detection in highway environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 323–335, Feb. 2016.

[17] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 546–561.

[18] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3973–3981.

[19] M. Mathias, R. Timofte, R. Benenson, and L. Van Gool, "Traffic sign recognition—How far are we from the solution?" in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8.

[20] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1751–1760.

[21] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2587–2595.

[22] Q. Hu, S. Paisitkriangkrai, C. Shen, A. van den Hengel, and F. Porikli, "Fast detection of multiple objects in traffic scenes with a common detection framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1002–1014, Apr. 2016.

[23] Y. Shen, U. Ozguner, K. Redmill, and J. Liu, "A robust video based traffic light detection algorithm for intelligent vehicles," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2009, pp. 521–526.

[24] R. de Charette and F. Nashashibi, "Real time visual traffic lights recognition based on spot light detection and adaptive traffic lights templates," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2009, pp. 358–363.

[25] Z. Shi, Z. Zou, and C. Zhang, "Real-time traffic light detection with adaptive background suppression filter," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 690–700, Mar. 2016.

[26] X. Wang, H. Ma, and X. Chen, "Geodesic weighted Bayesian model for salient object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 397–401.

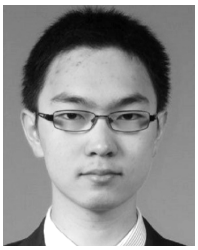
[27] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.

[28] M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for looking at traffic lights: Issues, survey, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1800–1815, Jul. 2016.

[29] R. K. Satzoda and M. M. Trivedi, "Multipart vehicle detection using symmetry-derived analysis and active learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 926–937, Apr. 2016.

[30] G. Siogkas, E. Skodras, and E. Dermatas, "Traffic lights detection in adverse conditions using color, symmetry and spatiotemporal information," in *Proc. VISAPP*, 2012, pp. 620–627.

[31] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.



Xi Li received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree.

His research interests are computer vision and machine learning, with particular interests in object detection and semantic segmentation.

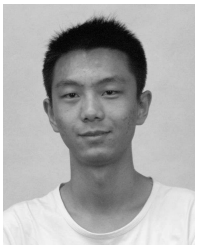
Huimin Ma (M'11) received the M.S. and Ph.D. degrees in mechanical electronic engineering from the Beijing Institute of Technology, Beijing, China, in 1998 and 2001, respectively.

She was a Visiting Scholar with the University of Pittsburgh in 2011. She is an Associate Professor with the Department of Electronic Engineering, Tsinghua University, and the Director of the 3-D Image Simulation Laboratory. She is also the Secretary-General of the China Society of Image and Graphics.

Her research and teaching interests include 3-D object recognition and tracking, system modeling and simulation, the psychological base of image cognition.

Xiang Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree.

His research interests are computer vision and machine learning, with particular interests in salient object detection and semantic segmentation.



Xiaoqin Zhang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree.

His research interests include multi-agent decision making, optimization, and 3-D simulation.