Detection of the Traffic Light in Challenging Environmental Conditions

Syeda Shahista and Afreen Khatoon Khan

February 3, 2021

"Detection of the traffic light in challenging environmental conditions"

SYEDA SHAHISTA
Electronics and Communication Department
SECAB Institute of Engineering and Technology
Vijayapur, Karnataka, India
syedashahista555@gmail.com

AFREEN KHATOON KHAN
Electronics and Communication Department
SECAB Institute of Engineering and Technology
Vjayapur, Karnataka, India
Afreenkhan7263@gmail.com

Abstract—**This paper is to recognize the current traffic light phase to reliably detect the status of the traffic light. eg; Red, Green and, Off-status and to focus cases where cameras have difficulties eg; Traffic light high above the ground, Out of field-of-view, the traffic light is front-or backlit by the sun and to achieve an improved redundancy leading to batter confidence in driving assistance systems and for future automated driving solutions. In this paper, two methods are used; The first method presents a vision-based traffic light structure detection and convolutional neural network (CNN) based state recognition method, which is robust under different illumination and weather conditions. The second method presents a deep fusion network for robust fusion without a large corpus of labeled training data covering all asymmetric distortions.**

Keywords-traffic light detection; CNN; deep fusion networks;

I. INTRODUCTION

In this paper, we present two different methods to detect the traffic light in challenging environmental conditions; the first method presents a purely vision-based traffic light state detection method which is robust under different illumination conditions. The proposed method uses color segmentation and area-based rejection filters for traffic light candidate region extraction. Localization of traffic light structure is done using MSER (Maximally Stable Extremal Regions). To further validate the potential traffic light candidate regions, HOG features are extracted. The detected traffic light structures were then used to determine the state of the signal using the CNN-based model. The second method presents a multimodal fusion method for traffic light detection in adverse weather, including fog, snow, and harsh rain. This method presents an adaptive single-shot deep fusion architecture that exchanges features in intertwined feature extractor blocks.

II. OVERVIEW OF THE PROPOSED SYSTEM

1.METHOD 1:

The main objective of this paper is to design an efficient purely vision-based traffic light detection and state recognition system. The pipeline of the proposed system is shown in Figure 1 and the steps are as follows.

• Candidate region extraction
– Input images are pre-processed to limit the number of candidate regions using HSV based color segmentation method.
– Further, these candidates are filtered using aspect ratio and area-based analysis.
– MSER is applied to the candidate regions to localize the structure of the traffic light in diverse background scenarios.
• Detection of traffic light structure
– HOG features are extracted from each candidate region.
– Detection of traffic light structure was done based on extracted HOG features using SVM.
• Recognition of traffic light state
– The state of the traffic light was recognized using CNN based methods.

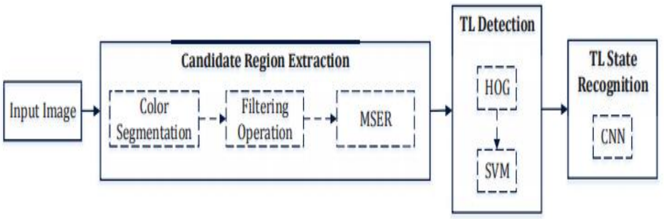


Fig. 1: Overview of the proposed system

A. Candidate Regions Extraction

Input images were pre-processed to limit the number of candidate regions using HSV based color segmentation method. Further these candidates were filtered using aspect ratio and area-based analysis. MSER was applied on the candidate regions to localize the structure of the traffic light in diverse background scenarios. Traffic lights are exposed to varied lighting conditions, therefore it is very important to make the system invariant to illumination.Hence we used HSV color space which separates luminance and chrominance components. Hue and saturation are the two features that we are more interested in HSV color space, thereby also reducing the feature space from 3-D in RGB color space to 2-D in HSV space. Color segmentation technique results in detection of objects which have chrominance similar to that of traffic light, thereby yielding many false positives (non-traffic attributes). Majority of these false positives are eliminated by applying traffic light specific rejection criteria (i.e. aspect ratio and area of the detected candidates). At the end of this process substantial amount of false positives were eliminated, however many false positives still remain and they will be removed using shape based features as explained in next section. The outcome of color segmentation and area based filtering process is shown in Figure 2.

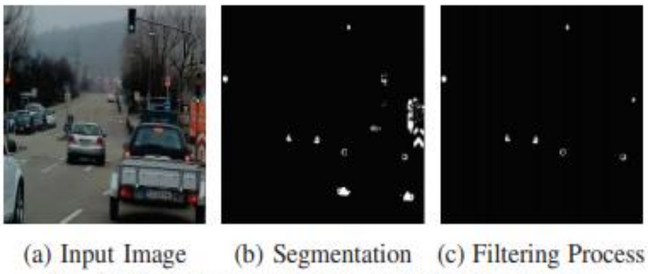

(a) Input Image (b) Segmentation (c) Filtering Process

Fig. 2: Color segmentation and filtering process

The contour detected using color segmentation contains traffic light and other similar colored objects. Size of the traffic light is pretty small and number of features resulting from traffic light is relatively insufficient to represent a complex structure of the traffic light. Therefore, we extract the entire traffic light structure for more relevant features, which results in more efficient detection and recognition phases. Figure 3(c) shows a contour containing a traffic light, whereas Figure 3(d) shows a contour containing an entire traffic light structure. Extraction of traffic light structure is done by initially fitting a bounding box over the traffic light. The size of the bounding box was fixed based on the maximum size of the traffic light structure that can be encountered by the system. The maximum size of the bounding box was determined from ground-truth. Figure 3(a) and 3(b) depict the area covered by a fixed size bounding box when the signal is far from camera and when the signal is close to camera.

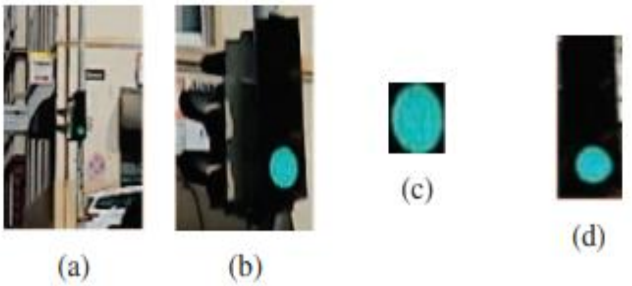
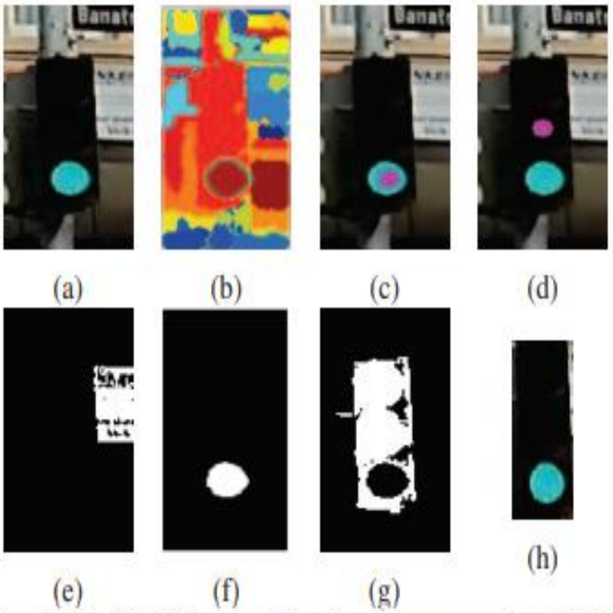

(a) (b) (c) (d)

Fig. 3: Example of fixed size bounding box

In absence of a localization method for traffic light structure, the bounding box for a signal which is far from camera will contain traffic signal along with diverse background which are non-traffic regions.This makes it difficult to effectively detect and recognize the state of a traffic signal. Thus it is very important to localize the traffic light structure effectively select two seed points. One seed point corresponding to the centre of the signal, as detected using color segmentation and another seed point corresponding to any point of the signal structure. Figure 4(a), shows the input image and Figures 4(c) and (d) indicate the seed points 1 and 2 respectively.The clusters that do not correspond to the seed points are eliminated (example, cluster in Figure 4(e)). Next step is to shortlist the clusters which best represents a traffic signal. Clusters which correspond to seed point 1 (Figure 4(f)) and seed point 2 Figure 4(g) are selected, then the two clusters are integrated and the resultant output is a contour corresponding to the localized traffic signal structure as shown in Figure 4(h). The outcome of this stage is processed in next stage for detection and recognition of traffic light.


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4: Traffic light localization process using MSER

B. Traffic Light Structure Detection

For each of the candidate regions resulting from the previous step, Histograms of Oriented Gradients (HOG) features were extracted to aid in the detection of the traffic light structures, thereby eliminating most of the false positives.HOG. The feature was developed with the aim of detecting humans in an image. Later on, the utility of the feature was extended to pedestrian detection, object detection and other computer vision problems. HOG features are relatively invariant to scale and rotation which is important for traffic lights. HOG features are computed by taking orientation histograms of edge intensity in a local region. As indicated previously, color is a major characteristic of a traffic signal. Many researcher suggested that performance of the conventional HOG can be improved by combining HOG features over multiple color channels Therefore, in this paper we have incorporated color information with HOG feature. A simple visualization of HOG feature extraction is illustrated in Figure 5. For extracting the HOG features, the detection window size was fixed at (*width* × *height* = 90 × 180) empirically. HOG feature descriptors are then fed to a non-linear SVM classifier for detecting traffic light structures. Nonlinear SVMs will create a space transformation, it will be a linear SVM in the feature space, but a non-linear separation border in the input space.Lower the number of input features, easier it is for the non-linear SVM to perform space transformation.Due to large amount of training data and relatively small number of HOG features, we chose nonlinear SVM classifier. SVM is a supervised learning model, it constructs a hyper-plane in a high dimensional space to separate the feature points into two or more classes. The feature points from which the separated hyperplane is located at the maximum margin are known as support vectors. For a test data consisting of a potential candidate regions, the HOG features are extracted and classified by calculating the distance between the extracted feature points of the test image with the support vectors found during training phase.

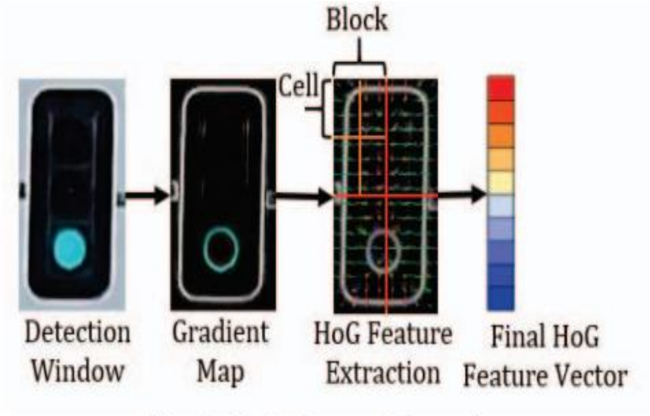


Fig. 5: HoG Feature Extraction

C. *Traffic light state recognition*

The state of the traffic light was recognized using CNN based methods. Thereby achieving high accuracy in classification. Hence we have opted for CNN based classifier for traffic light state recognition. However by incorporating application specific modifications in the network, we have been able to reduce the number of parameters to 76000 from the original 1400000 which amounts to reduction of complexity by about 95 percent. Block diagram of the network is given in Figure 6. Output of detection algorithm is resized to 48 × 96 pixel size before being given as an input to CNN. Initial convolution layer consists of 20 filters of size 5 × 5. Filter stride across the input image during each iteration is 2 pixels. After convolution layer we have a max pool layer with a pixel stride of 2. Max pool layer is followed by a second convolution layer with 50 filters of 5×5 with a stride length of 2 pixels. Output convolution layer is followed by second pooling layer. After the second pooling layer two consecutive fully connected layers map the network to three neurons which correspond to three classes of traffic lights. Parameter reduction is primarily achieved by reducing number of filters and increasing the stride length to two.

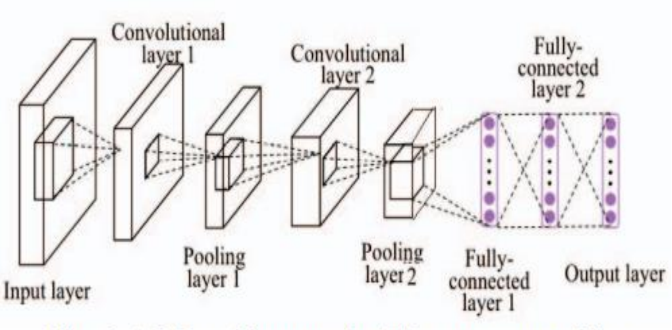


Fig. 6: CNN architecture for TLs state recognition

2. *METHOD 2:*

A. *Multimodal Adverse Weather Dataset*

To assess object detection in adverse weather,we have acquired a large-scale automotive dataset providing 2D and 3D detection bounding boxes for multimodal data with a fine classification of weather, illumination, and scene type in rare adverse weather situations. Table 1 compares our dataset to recent large-scale automotive datasets In Figure 2, we plot the weather distribution of the proposed dataset. The statistics were obtained by manually annotating all synchronized frames at a frame rate of 0.1 Hz. We guided human annotators to distinguish light from dense fog when the visibility fell below 1 km and 100 m, respectively. If fog occurred together with precipitation, the scenes were either labeled as snowy or rainy depending on the environment road conditions. For our experiments, we combined snow and rainy conditions. Note that the statistics validate the rarity of scenes in heavy adverse weather, which is in agreement to and demonstrates the difficulty and critical nature of obtaining such data in the assessment of truly self-driving vehicles, i.e. without the interaction of remote operators outside of geo-fenced areas. We found that extreme adverse weather conditions occur only locally and change very quickly.

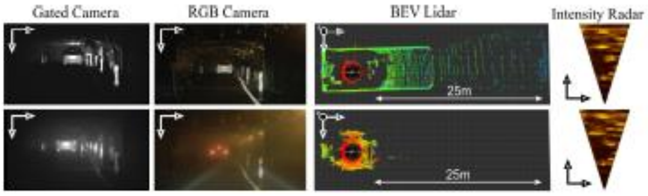


Figure 3: Multimodal sensor response of RGB camera, scanning lidar, gated camera, and radar in a fog chamber with dense fog. Reference recordings under clear conditions are shown in the first row, recordings in fog with visibility of 23 m are shown in the second row.

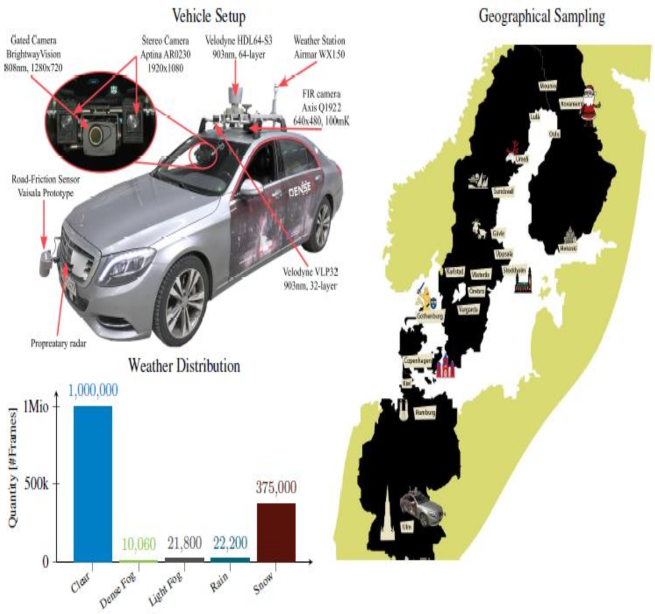


Figure 2: *Right:* Geographical coverage of the data collection campaign covering two months and 10,000 km in Germany, Sweden, Denmark, and Finland. *Top Left:* Test vehicle setup with top-mounted lidar, gated camera with flash illumination, RGB camera, proprietary radar, FIR camera, weather station, and road friction sensor. *Bottom Left:* Distribution of weather conditions throughout the data acquisition. The driving data is highly unbalanced with respect to weather conditions and only contains adverse conditions as rare samples.

B. Multimodal Sensor Setup

For acquisition we have equipped a test vehicle with sensors covering the visible, mm-wave, NIR, and FIR band, see Figure 2. We measure intensity, depth, and weather condition. Stereo Camera As visible-wavelength RGB cameras, we use a stereo pair of two front-facing high-dynamic rangeautomotive RCCB cameras, consisting of two On- Semi AR0230 imagers with a resolution of 1920 × 1024, a baseline of 20.3 cm and 12 bit quantization. The cameras run at 30 Hz and are synchronized for stereo imaging. Using Lensagon B5M8018C optics with a focal length of 8 mm, a field-of-view of 39.6 ° × 21.7 ° is obtained. Gated camera We capture gated images in the NIR band at 808 nm using a BrightwayVision BrightEye camera operating at 120 Hz with a resolution of 1280 × 720 and a bit depth of 10 bit. The camera provides a similar field-of-view as the stereo camera with 31.1 ° × 17.8 °. Gated imagers rely on time-synchronized camera and flood-lit flash laser sources . The laser pulse emits a variable narrow pulse, sand the camera captures the laser echo after an adjustable delay. This enables to significantly reduce backscatter from particles in adverse weather [3]. Furthermore, the high imager speed enables to capture multiple overlapping slices with different range-intensity profiles encoding extractable depth information in between multiple slices . Following , we capture 3 broad slices for depth estimation and additionally 3-4 narrow slices together with their passive correspondence at a system sampling rate of 10 Hz. Radar For radar sensing, we use a proprietary frequency modulated continuous wave (FMCW) radar at 77 GHz with 1 ° angular resolution and distances up to 200 m. The radar provides position-velocity detections at 15 Hz. Lidar On the roof of the car, we mount two laser scanners from Velodyne, namely HDL64 S3D and VLP32C. Both are operating at 903 nm and can provide dual returns (strongest and last) at 10 Hz. While the Velodyne HDL64 S3D provides equally distributed 64 scanning lines with an angular resolution of 0.4 °, the Velodyne VLP32C offers 32 nonlinear distributed scanning lines. HDL64 S3D and VLP32C scanners achieve a range of 100m and 120 m, respectively. FIR camera Thermal images are captured with an Axis Q1922 FIR camera at 30 Hz. The camera offers a resolution of 640 × 480 with a pixel pitch of 17 µm and a noise equivalent temperature difference (NETD) < 100 mK. Environmental Sensors We measured environmental information with an Airmar WX150 weather station that provides temperature, wind speed and humidity, and a proprietary road friction sensor. All sensors are timesynchronized and ego-motion corrected using a proprietary inertial measurement unit (IMU). The system provides a sampling rate of 10 Hz.

C. Adaptive Multimodal Single-Shot fusion Data Representation

The camera branch uses conventional three-plane RGB inputs, while for the lidar and radar branch, we depart from recent bird’s eye-view (BeV) projection schemes or raw point-cloud representations . BeV projection or point-cloud inputs do not allow for deep early fusion as the feature representatnios in the early layers are inherently different from the camera features. Hence, existing BeV fusion methods can only fuse features in a lifted space, after matching region proposals, but not earlier. Figure 4 visualizes the proposed input data encoding, which aids deep multimodal fusion. Instead of using a naive depth-only input encoding, we provide depth, height, and pulse intensity as input to the lidar network. For the radar network, we assume that the radar is scanning in a 2D-plane orthogonal to the image plane and parallel to the horizontal image dimension. Hence, we consider radar invariant along the vertical image axis and replicate the scan along vertical axis. Gated images are transformed into the image plane of the RGB camera using a homography mapping, see supplemental material. The proposed input encoding allows for a position and intensity-dependent fusion with pixel-wise correspondences between different streams. We encode missing measurement samples with zero value.

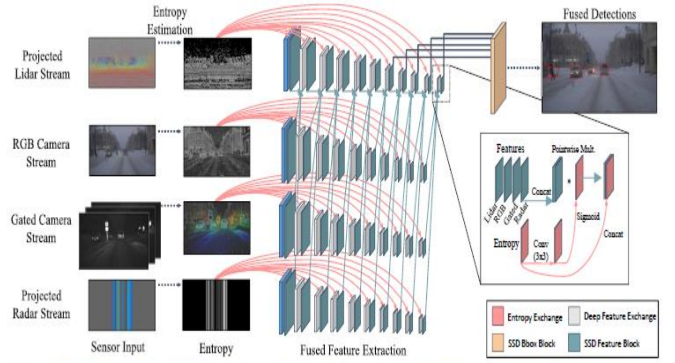


Figure 4: Overview of our architecture consisting of four single-shot detector branches with deep feature exchange and adaptive fusion of lidar, RGB camera, gated camera, and radar. All sensory data is projected into the camera coordinate system following Sec. 4.1. To steer fusion in-between sensors, the model relies on sensor entropy, which is provided to each feature exchange block *(red)*. The deep feature exchange blocks *(white)* interchange information *(blue)* with parallel feature extraction blocks. The fused feature maps are analyzed by SSD blocks *(orange)*.

D. *Feature extraction*

Feature Extraction As feature extraction stack in each stream, we use a modified VGG backbone. Similar we reduce the number of channels by half and cut the network at the conv4 layer. Inspired by , we use six feature layers from conv4-10 as input to SSD detectionlayers. The feature maps decrease in size1, implementing a feature pyramid for detections at different scales. As shown in Figure 4, the activations of different feature extraction stacks are exchanged. To steer fusion towards the most reliable information, we provide the sensor entropy to each feature exchange block. We first convolve the entropy, apply a sigmoid, multiply with the concatenated input features from all sensors, and finally concatenate the input entropy. The folding of entropy and application of the sigmoid generates a multiplication matrix in the interval [0,1]. This scales the concatenated features for each sensor individually based on the available information. Regions with low entropy can be attenuated, while entropy rich regions can be amplified in the feature extraction. Doing so allows us to *adaptively fuse features* in the feature extraction stack itself, which we motivate in depth in the next section.

III. EXPERIMENTAL RESULTS

The performance of different TLD algorithms were evaluated using the following steps:

- For the training and testing a common set of image sequence is selected.
- The detection output is quantified with the manuall annotated ground truth.
- Each of the candidates are placed under one of the 4 classes - True Positive(TP), False Positive(FP), True Negative (TN) and False Negative (FN), based on the comparison with ground truth.
- A set of statistics (Precision, Recall and F-measure) is calculated for each implemented algorithm. Each of the TL detection algorithm is evaluated based upon the following information retrieval measurements:
- Precision and Recall, as seen in equation 1
- F-measure, as seen in equation 2

Precision and Recall quantify how well the detected TLs matches with the ground-truth. High precision or high recall value means high performance. F-measure, indicates the overall accuracy of the system

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} \qquad (1)$$

Precision is the ratio of correct TL detections compared to the total number of detections. True positives indicate the number of traffic signals which were rightly classified as a traffic signal. False positives indicate the number of non traffic signals classified as a traffic signals. Recall is the ratio of correct TL detections compared to the actual number of TLs.

$$F = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (2)$$

F-measure is the geometric mean of precision and recall.

The proposed algorithm was tested under varying lights and different weather conditions. Figure 7 demonstrates the visual outcome of the proposed method for traffic light recognition in different environment conditions. The bounding box around the traffic light indicates the detection results and solid circle in the left corner of the image shows the state recognition results. The detection and recognition of the multiple traffic lights in urban scenario is shown in Figure 7 (a)-(d) while Figure 7(e)-(f) shows the detection and recognition of traffic lights in rainy conditions.

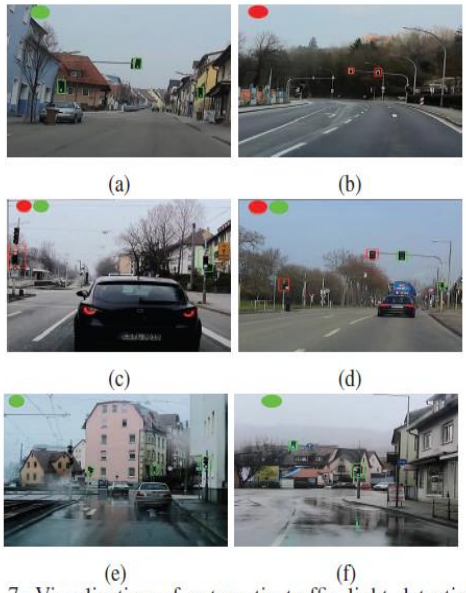

(a) (b)
(c) (d)
(e) (f)

Fig. 7: Visualization of automatic traffic light detection in various scenarios

III. CONCLUSION

This paper presents two methods to recognize the current traffic light phase to reliably detect the status of the traffic light. eg; Red, Green and, Off-status and to focus cases where cameras have difficulties eg; Traffic light high above the ground, Out of field-of-view, the traffic light is front-or backlit by the sun and to achieve an improved redundancy leading to batter confidence in driving assistance systems and for future automated driving solutions. The first method presents a vision-based traffic light structure detection and convolutional neural network (CNN) based state recognition method, which is robust under different illumination and weather conditions. And it is purely vision-based real-time traffic light detection and state recognition method which is robust under different illumination conditions. Results also shows that accuracy and consistency of the proposed method are better than conventional approaches The second method presents a deep fusion network for robust fusion without a large corpus of labeled training data covering all asymmetric distortions. This introduce a novel adverse weather dataset covering camera, lidar, radar, gated NIR, and FIR sensor data.

IV. REFERENCES

[1] M. Diaz, P. Cerri, G. Pirlo, M. A. Ferrer, and D. Impedovo, "A
survey on traffic light detection," in *International Conference on Image*
Analysis and Processing. Springer, 2015, pp. 201–208.
[2] 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," 2015.
[3] N. Fairfield and C. Urmson, "Traffic light mapping and detection," in
Robotics and Automation (ICRA), 2011 IEEE International Conference
on. IEEE, 2011, pp. 5421–5426.
[4] J. Levinson, J. Askeland, J. Dolson, and S. Thrun, "Traffic light
mapping, localization, and state detection for autonomous vehicles," in
Robotics and Automation (ICRA), 2011 IEEE International Conference
on. IEEE, 2011, pp. 5784–5791.
[5] Sanjay Saini1, Nikhil S1, Krishna Reddy Konda1, Harish S Bharadwaj1 and Ganesh" 2017 IEEE Intelligent Vehicles Symposium (IV)
June 11-14, 2017, Redondo Beach, CA, USA.
an N1"An Efficient Vision-Based Traffic Light Detection and State Recognition
for Autonomous Vehicles
[6] Mario Bijelic1,3 Tobias Gruber1,3 Fahim Mannan2 Florian Kraus1,3 Werner Ritter1 Klaus Dietmayer3 Felix Heide2,41Mercedes-Benz AG 2Algolux 3Ulm University 4Princeton University"Seeing Through Fog Without Seeing Fog:
Deep Multimodal Sensor Fusion in Unseen AdverseWeather".