Stereo cameras with applications to traffic scenarios, traffic lights detection

Oleksandra Baga
Master Computer Science, Freie Universität Berlin
Sommersemester 2021, Seminar KI / Autonomes Fahren
oleksandra.baga@gmail.com

***Abstract*— A traffic light recognition system is a very important building block in an advanced driving assistance system and an autonomous vehicle system. One of the major causes of traffic accidents worldwide is the disregard of traffic lights by drivers. When one vehicle collides with another in a car accident at the high speeds when running a red light, the effects can be catastrophic. A driving support systems must precisely detect and recognize traffic lights and give appropriate information to drivers and in the case of self driving car control the driving, stopping and navigation process respectively. The key sensor is a camera installed in a moving vehicle. A pair of cameras are required to implement stereo vision to enhance performance of detections algorithms. In this paper different modern systems and approaches for real-time detection and recognition of traffic signals are reviewed and analyzed.**

I. Introduction

Autonomous vehicles are able to perceive their surroundings (obstacles and track) and commute to destination with the help of a combination of sensors and radars such as RADAR, LIDAR, GPS and cameras for computer vision. Traffic light detection and distance measurement using stereo camera is a very important and challenging task in image processing domain of self driving vehicles. Last years various kinds of driving support systems have been developed and the driving environment has been improved by the recent research results using new techniques. The importance of the correctness of the detection can not be underestimated because according to the statistical reports the one ot the major causes of traffic incidents including humans deaths is a fact that driver has disregarded traffic lights [3]. Self-driving vehicles becomes part of humans' transport network starting from fully automated metro system and ending with private self driving cars. Thus integrated traffic light distance measurement system for self-driving car must be safe, robust and detect traffic lights without false positive and false negative outcomes because both of them could produce disastrous consequences.

A variety of algorithms and approaches have been used integrating different method of detection, colour recognition and distance calculation and they will be reviewed in this paper. Fairfield et al. [5] presented a traffic-light detection method using a prior map for indication when and where a vehicle should be able to see a traffic light this the predicted location can be used to improve detection rate of the traffic light state. Langner et al. [4] used during experiments with their autonomous car the known geometry of traffic lights and their officially defined exact positions on the street such a height above the ground to reach the detection rates above 95% in experiments. Langner et al. as well as Omar et al. [4] extracted traffic lights color from Hue component after converting a captured image from red green blue (RGB) to hue-saturation value (HSV) color model. Lindner et al. [6] implemented and tested system consisting from three parts for scanning each frame, grouping detected candidates over several frames and a verification of each single candidate, whether it is a traffic signal and what color it has. Finally Lindner et al. used and Cascade Classifier for increasing the suppression rate of false positive candidates

A. About a Stereo Camera

To be able to drive autonomously a vehicle has to perceive the world around and have eyes, not real eyes, but cameras to analyse these things to happen on the way. For a computer vision the stereo camera is used that is a type of camera with two or more image sensors. This allows a camera to simulate human binocular vision. By using stereo disparity the difference in image location of an object seen by the left and right cameras can be calculated using the cameras' horizontal separation. The principle underlying stereo vision is called triangulation and is used to estimate the position of an object by finding the intersection of the two lines passing through the centre of projection and the projection of the point in each image. Since cameras look at an image from different angles, the difference between the two point of view can be computed and a distance estimation established. That is what needed to estimate the distance to a traffic lights or other obstacles that occur during the process of driving. We have to distinguish the distance to the traffic lights since several traffic lights can be seen simultaneously at double intersections and only the near traffic light is important for determining the next maneuver.

B. Camera Setup

To achieve distance estimation using Stereo Vision the two cameras' intrinsic and extrinsic parameters must be first calibrated. Calibration means transformation a 3D point from the world coordinates [X,Y,Z] to a pixel coordinates [X,Y] going through camera coordinates. Extrinsic calibration can be done with parameters called R (rotation matrix) and

T (translation matrix) is using for conversion from World Coordinates to Camera Coordinates [1]. This transformation represents the cameras position and orientation relative to the cars coordinate frame and influences the sensitive accuracy of the mapping of captured picture [4]. Intrinsic calibration with matrix called K used for conversion from Camera Coordinates to Pixel Coordinates. The inner values for the camera such as focal length, optical centre and others which determine the lens distortion must be used for this conversion.

The timing delay between when the image is taken by the camera and when it is transmitted to be able processed can be estimated using precise hardware timestamps. This timing delay varies depending on the camera frame rate and Firewire bus scheduling allocation, but Fairfield et al. experimentally found it to be stable within a few hundredths of a second for a given configuration.

II. Traffic lights as a special perception problem

As traffic lights have to be detected up to passing the stop line at an intersection and on streets in Europe it is common that traffic lights are directly installed at stop lines [2] and, for example, in Germany it the exact positions of traffic-lights are defined by the RiLSA standard (Richtlinien für Lichtsignalanlagen) [4] what was mentioned by Langner et al. at their paper. Under this directive, the base of a hanging traffic-light frame lays at 4.5m. Signals mounted to a pole are positioned at 2.1m above the ground [4]. A human driver can easily turn their head and recognise the traffic lights above to the right. Opposite to this an autonomous vehicle needs to cover such cases with the sensor setup and a high field of view [2]. High field of views can decrease the resolution of objects in the camera image and must be considered during the camera setup. The detection of all traffic lights at a defined distance range must be also guaranteed. A stereo camera with a field of view of 60 degrees covers the distance range from approximately 20m up to infinity. The datasets recorded by Julian Müller et. al. [1] [2] on German streets mainly contain intersections, in which no or at most one lane must be monitored. Sometimes surprisingly at least two lanes must be monitored and this increases the minimum camera's angle need for a processing and a computational costs.

Traffic lights usually have startling colours so drivers can easily see them. Most of the existing methods for traffic lights detection also attempt to detect pixels of the typical colours of traffic lights, i.e.,red, yellow, and green combining the colour with other information, such as shape or position [3]. The most common failure conditions in a traffic light detection system are either visual obstructions or false detected objects. Brake red lights in front of the car can be recognized and thus are false positives but considered to be safe since the vehicle should already be braking for the obstructing object [5]. False positives for greens may arise from particular patterns of light on a tree, or from brightly lit billboards[5]. However, if the car fails to see the red lights in an intersection (false negatives) or falsely detect a green light (a false positive like sun shining through trees) it can take an incorrect action.

A traffic signal recognition system based on a color camera requires two inactive traffic lights to be visible for the verification of active traffic light [6]. To discern active traffic light from the background, the exposure time had to be long enough, which forced the active traffic lights into saturation hence producing white instead of the signal colors [6]. Additionally the preliminary experiments by Moizumi et al. revealed that the colours of traffic lights captured by a digital video camera are easily over-saturated in various conditions not only long exposure which renders traffic light detection using color information difficult [3]. For example the most of the pixels of the yellow traffic light in a case of over-saturation change to white. A method for detecting traffic lights even if the pixel colors of the traffic lights are over-saturated will be discussed below but many regions that are not traffic lights are also detected by this type of approach and all candidates other than traffic lights must be excluded efficiently [3].

The 3D position and orientation, or pose, of traffic lights can be estimated during the driving and used as a support system. Anyway position of the traffic lights acquired during the detection phase does not provide enough information for an autonomous vehicle to make a decision [7].

The rapid deployment of the new LED traffic lights has led to a new problem. In the case of LED traffic lights, since the LED lights blink at a high frequency, there are frames in which it appears that all the traffic lights are turned off. In that case, standard color-based detection also fails. Tracking results must be complemented with an additional frame to that in which the light appears to be turned off [3]. A blinking light can appear in a case of LED light even in successive frames.

III. CNN for object detection

Last years the title of the most widely used computational approach in the field of Machine Leaning gradually won Convolutional Neural Networks because they are able to achieve outstanding results in complex predictions combining with a speed as an important parameter for real-time object detection and decision making in fast changing conditions. CNN has features parameter sharing and dimensionality reduction. Because of parameter sharing in CNN, the number of parameters (dimensionality) is reduced thus the computational power needed also decreased. The main intuition is the learning from one part of the image is also useful in another part of the image.CNNs are able to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters. One of the benefits of CNN is the ability to learn massive amounts of data that is crucial for a traffic lights detection because network needed to be feed with a thousands of thousands images of real intersections in different weather conditions in different counties. Due to the reduction in the number

[1] https://medium.com/think-autonomous/pseudo-lidar-stereo-vision-for-self-driving-cars-41fa1ac42fc9

of parameters involved and reusability of weights after such learning the architecture performs a better fitting to the image dataset. Thus CNN is specifically designed to process input images and consist of a blocks. The typical architecture is shown on Figure 1. A first block is feature extractor, it performs template matching by applying convolution filtering operations. This process can be repeated several times, after filtering the new features maps obtained with new kernels, which gives again new features maps to normalize and resize and so on. The second block at the end of all the neural networks used for classification. There are four types of layers for a convolutional neural network: the convolutional layer, the pooling layer, the ReLU correction layer and the fully-connected layer. It is important to mention that the convolutional layer is the key component of convolutional neural networks, and is always at least their first layer.

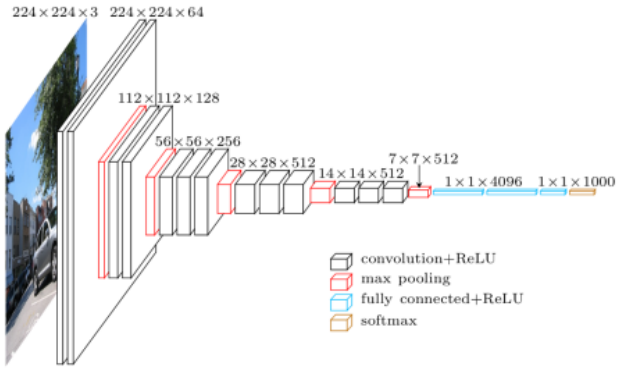

Fig. 1: Typical architecture of CNN Model (hier VGG16). Source: Smeda K. by www.towardsdatascience.com

A popular CNN object detector is Faster R-CNN, which consists of two CNNs: the first one proposes input image regions of interest, and the second one refines and classifies those regions [7]. In Zuo et al. [10], a plain Faster R-CNN was used to detect traffic signs in the research. The main idea of the algorithm is to design the Region Proposal Network ()RPN) network to extract the proposed regions and to generate the proposed regions with the convolution neural network [10]. Faster R-CNN claims to be able achieve the world's advanced object detection which can detect 20 types of objects including humans, animals, and vehicles. Classification model, is publicly trained on the ImageNet data set. Anyway, the experiments by Zuo et al. concluded that model struggled with detection because signs are commonly a small part of the image, making the detection task more difficult [7]. In the work done by Omar et al. the CNN YOLO was introduced. It improved some CNN Models that authors considered in their research, achieved 0,017 seconds per image 608x608 with a processing rate of 0,017 seconds, maintaining a high level of precision in the detection of traffic signs [7].

The futher detailed analysis of existing modern CNN is out of the scope of this work. However the author of this paper want to mention, that CNN nowadays is a computational standard for machine learning and predictions algorithm allowing to detect ans correct classify in real-time small objects and is a perfect choice for improvement of performance and safety of self driving vehicles regarding to the usual size of the traffic lights. There is still not exists a best CNN that performed ideally in urban environment with a huge amount od distracting false positive elements like billboards or illuminated windows in the background. For example, CNN by Omar et al. has two stages of detection and classification separetely. Although it has a great performance as a detection rate, the computational and processing result is slower if it was one stage. It is expected by author of the paper in the coming years to be new model announced that aims to detect, localize, and measure the distance between the camera and traffic light while performing real time traffic lights recognition even faster. The field of Deep Learning and CNN in particular is very lively and bringing new research results every year so that efficiently classify object proposals using deep convolutional networks can not be limited to some particular CNN implementation.

IV. Traffic Light Mapping

The color and lit/unlit state of standard traffic lights can only be perceived visually and active sensors such as sonar, radar, and lidar can not be used for this task. A traffic-light detection method able to recognize left and right turn traffic-lights is presented by Fairfield et al. [5]. They introduced in their paper a use of a prior map that can help to indicate when and where a vehicle should be able to see a traffic light. The prior map of traffic lights allow a vehicle to predict when it should see traffic lights and take conservative action, such as braking gradually to a stop while alerting the driver, when it is unable to observe any lights [5]. The process of estimating the 3D position and orientation, or pose, of traffic lights called in the paper a traffic light mapping. It is know that the precise position of traffic lights is not generally available and even the estimation of traffic light locations from aerial or satellite imagery provides the accuracy only within a few meters without information of the traffic light altitude. However, using cameras, GPS, IMU, lasers through intersections and collection of precisely timestamped camera images can estimate the traffic light positions by triangulating from multiple views. For this a large set of well-labeled images of the traffic lights is needed. To create this a huge log file that includes camera images and the car's precise pose is used as an input to the automatic mapping system [5]. Considering the fact that traffic lights will only occur at intersections, geospatial queries through the Google Maps API were used to discard images taken when no intersections are likely to be visible. It must be mentioned that Google Maps only includes information about intersections and not whatever it is controlled or uncontrolled intersection with any traffic lights to regulate traffic. While the car is approaching an intersection, a created by Fairfield et al. traffic light classifier runs over the entire captured image. Appropriate circle size and aspect ratios help classifier to find brightly-colored red, yellow, and green circles. The distortions of

captured objects due to changes in roll, pitch, and yaw of moving car can be corrected using the camera model thus the precise car pose is known for each image. Using a set of object classified labels in two or more camera images, the pose of the 3D object using linear triangulation is estimated [5].

When driving a car it is obligatory to know which lights are relevant to current lane and to desired trajectory and next maneuver. At double intersections several traffic lights can be seen simultaneously and semantics of new and complex intersections can be confused even for a human driver. A detailed prior map from the discussed approach can be used as an association between a traffic light and the different allowed routes through an intersection. Thus as explained above with the known car pose and an accurate prior map of the traffic light locations, it can be predicted when traffic lights should be visible, and where they should appear in the camera image. Car position is estimated from lidar localization with accurate elevation and then projected using the camera model into the image frame as an axis-aligned bounding box [5]. From the prior map, the type of the expected light elements is used to find appropriately-sized brightly colored red and green blobs. The type of the traffic signal is known thus its geometry is also known and can be used during detection such as geometric constraints on the low-level blobs to be green.

Fig. 2: Prior map with mapped traffic lights in San Francisco - San Jose area. Source: [5]

Thousands of human-labeled traffic lights are used as a base for a classifier training and authors declare to continue to add more labels as new situations were found, such as brightly colored billboards or inclement weather. The figure 2 shows mapped traffic lights from some intersections in San Jose area made in 2011 by Fairfield et al. In total are mapped about a thousand intersections and over 4000 lights. The described system has been deployed on multiple cars, and has provided reliable and timely information about the state of the traffic lights during thousands of drives through intersections [5]. According to a recent report from Droid-Life[2] in the August 2020, Google has started testing a new feature that will show traffic lights in Google Maps on Android. Traffic lights are shown just in some intersections of the several US cities. The traffic lights are visible both while using the traditional map view and while navigating, however, they do appear slightly bigger and more noticeable while navigating. For now, it does not seem like Google has a use for the traffic lights other than displaying them on a map[3]. It is expected that people with the same route to work every day will learn, over time, how to adjust the speed to reduce and even eliminate waiting times at traffic lights. It is expected that feature will be gradually available in more cities and counties when the prior maps will be created and extended to European cities, as local authorities and available infrastructure will allow. Anyway, in the June 2021, almost one year after feature was launched in some US cities, no further information about extension of this feature can be found.

V. Color pixels detection

Together with color of a detected traffic light the 3D position of each pixel have to be also considered during detection process [4]. The position of the red, yellow and red lights is in all Europe's countries is standardized by the RiLSA standard (Richtlinien für Lichtsignalanlagen) what was mentioned above. Thus the known geometry of traffic lights can be used to construct traffic-light entities consisting of one lit and two unlit circular light sources on a rectangular dark-colored frame [4]. With a known length ratios between the lights and the frame, positions of red, yellow and green traffic signals can be resolved, the area where the two other unlit light sources are can be found, and the correct lit traffic light can be detected. This method was used by Langner et al. with developed by them autonomous car and detection rates reached above 95% in experiments what considered already significant and useful for a warning system giving the driver a reaction time of 2.8s in case of a warning with in inner-city speed limit of 50km/h [4].

In their work Mu et al. [10] [7] proposed an image processing approach that converts the image color from red green blue (RGB) to hue-saturation value (HSV). Using the HSV color space for pixel classification can be considered because it can separate luminance and chroma [4]. Although the red and yellow colour's distributions overlap they have different modes. Potential areas can then identified by scanning the scene using transcendental colour threshold with prior knowledge of the image. Finally, the location of the traffic lights using the Oriented Gradients (HOG) and Support Vector (SVM) functionalities can be identified.

It has to be mentioned that most of color-matching pixels can be rejected very fast because of the height of their position on the street or lateral distance to the vehicle [4]. Thus brake red lights in front of the car can be distinguished from the traffic lights and depending from the distance control the braking to avoid the collision. Knowing the geometry and

[2]https://www.droid-life.com/2020/07/07/google-maps-is-starting-to-show-traffic-lights-on-android/

[3]https://arstechnica.com/gadgets/2020/08/google-maps-starts-widely-labeling-traffic-lights-in-the-us/

potential position of needed object it is possible to locate the traffic light accurately and cut out the area of interest around the location to reduce the calculation costs and use only this area in the image processing to achieve the final identification. The precise bounding box of the traffic light in the image must be tracked all the time and then can be smoothed and contrast can be increased for improving the detection rate [7]. After converting the cropped image from RGB to HSV colour space, the color data of the traffic lights results in H channel because the H (hue) component is separated from all data and thus it makes a traffic light determination possible.

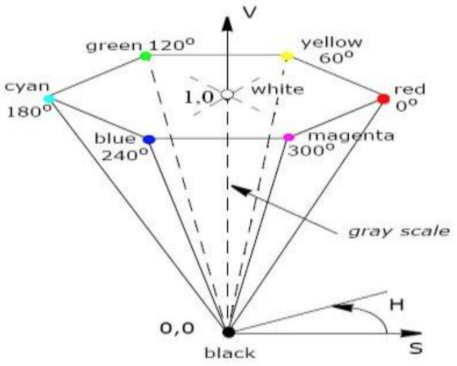

Fig. 3: HSV colour model. Source: [7]

HSV color model is shown on figure 3 where the Hue component describes the color itself in the form of an angle between [0,360] degrees. 0 degree mean red, 120 means green 240 means blue. 60 degrees is yellow, 300 degrees is magenta. The Saturation component signals how much the color is polluted with white color. The range of the S component is [0,1]. Experimentally, it was found by Omar et al. that saturation is high at most of the area as the traffic light body is very good shaped. Choosing the area with the high saturation and high hue values as the area to mask yielded a good result in detecting the traffic light color correctly [7]. However, using HSV yields a good accuracy but could not recognize and classify the colour of traffic lights from a far distance [7].

VI. Shape, texture and color detection

In the paper by Lindner et al. described a robust detection approach using either color, shape or/and texture as key features. The suggested system consists from three parts: detector scans each frame for traffic signal hypotheses, tracker groups these candidates over several frames to form an object/track hypothesis and, finally, a classifier verifies each single candidate, whether it is a traffic signal or not, and also classifies the status of the traffic signal (red, yellow or green) [6]. Authors claim that three stage system layout already proved its effectiveness in traffic sign recognition even with one camera and by using a stereo vision the performance and robustness of the complete system can be improved. Detector has to cope with a large amount of input data and could miss only very few traffic lights because on the back of it these gaps can be filled by the tracking module. As implementation platform for real time application authors have chosen FPGA for real-time image processing because FPGAs structure is able to exploit spatial and temporal parallelism to traverse the pixels of an image, and apply the filters to them.

The tracker part analyses the images captures by the detector, detected candidates and fill the gaps with missed by detector traffic light by taking into account the relative location, speed and acceleration between the moving camera and the traffic signal. It also suppresses spontaneous detection hypotheses of false candidates, which are not stable over several frames. The figure 4 shows that certain constellations in inhomogeneous regions such green trees with a sun shining through might look like a traffic light, but is usually not stable over several frames. The tracker also supports the analyse of at certain regions which had a positive detection in the last frame since the object of interest might temporarily be hidden by traffic participants or other obstacles in he following successive frames and suddenly important traffic signals may appear almost anywhere in the image, not only at the focus of expansion or at the image borders [6].

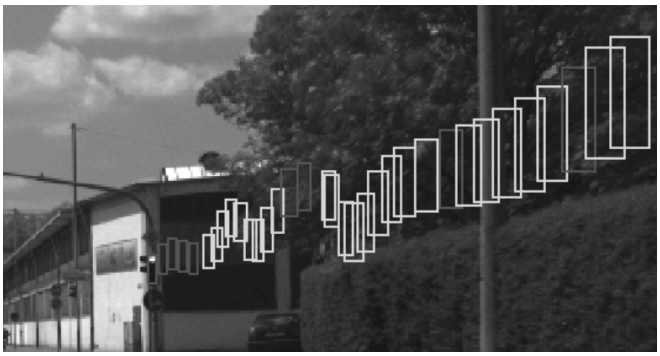

Fig. 4: Detection and tracking of a traffic signal. Source: [6]

The final stage of proposes approach is a classifier, which consecutively evaluates each element of a track by making a decision by soft majority voting from feed-forward neural networks with receptive fields. Rejecting the false candidates as robustly as possible is one of the important and complicated task for the classifier. A major impact on the improvement of the traffic signal recognition system comes from enhanced map information in combination with GPS and optical sensors. With high accuracy positioning systems and maps the detector can even be narrowed to image regions where to look for traffic signals. By using the enhanced map information the information about the lane, the vehicle is currently driving in, is available, and the different traffic signals to the according lane can be also assigned [6]. For the measurements of real world distance and size of the detected objects a second camera is installed that allow stereo algorithms to enhance the productivity of the method.

For the shape detection known geometry is used. The lights in traffic signals are round, except directional signals for left/right turning lanes (in Germany even the stop-state

is depicted by a black arrow on a red light, while the go-state is just a green arrow; in the USA colored arrows are used for all signal states) [6]. Generalized Hough transform using Sobel based gradient directions lay in the base of already implemented and tested by Lindner et al. circle finder. Traffic lights are detected if the diameter is larger than 6 pixels. The quality threshold for the circle was reduced that allowed experimentally to detect even slightly distorted lights. Reached recognition rates for the complete system is of over 80% that surely have to be improved in the next iterations of algorithm development. For the matching part of shape detector the same approach as was mentioned above in the method by Langner et al. is used. It is known that typical for any traffic light is the fact that a brighter spot surrounded by a darker box. Additionally using the Cascade Classifier increases the suppression rate of false positive candidates so that each classifier in the pipeline rejects as many candidates as possible without loosing any object searched for [6]. Only candidates not rejected can be passed into the next member of a cascade structure.

Thus a general system for real-time detection and recognition of traffic signals is reviewed. The key sensor is second camera for stereo vision installed in a moving vehicle that is used to enhance the performance and robustness of the system.

VII. Color Saturation Approach

VIII. High Frequency detection

IX. Conclusion

TADAM!
THE END

References

[1] Andreas Fregin, Julian Müller, Klaus Dietmayer. Three Ways of using Stereo Vision for Traffic Light Recognition. https://www.researchgate.net/publication/318810474 date accessed: 07.05.2021

[2] Andreas Fregin, Julian Müller, Klaus Dietmayer. Multi-camera system for traffic light detection: About camera setup and mapping of detections. https://www.researchgate.net/publication/323792381 date accessed: 12.05.2021

[3] Hiroki Moizumi, Yoshihiro Sugaya, Masako Omachi, Shinichiro Omachi. Traffic Light Detection Considering Color Saturation Using In-Vehicle Stereo Camera. https://doi.org/10.2197/ipsjjip.24.349 date accessed: 17.05.2021

[4] Tobias Langner, Daniel Seifert, Bennet Fischer, Daniel Goehring, Tinosch Ganjineh. Traffic Awareness Driver Assistance based on Stereovision,Eye-tracking, and Head-Up Display. https://www.drgoehring.de/bib/langner16icra-sst/langner16icra-sst.pdf date accessed: 19.05.2021

[5] Nathaniel Fairfield, Chris Urmson. Traffic Light Mapping and Detection. https://static.googleusercontent.com/media/research.google.com/de//pubs/archive/37259.pdf date accessed: 22.05.2021

[6] Frank Lindner, Ulrich Kressel, Stephan Kaelberer. Robust Recognition of Traffic Signals. https://www.researchgate.net/publication/4092560 date accessed: 30.05.2021

[7] Wael Omar, Impyeong Lee, Gyuseok Lee, Kang Min Park. Detection And Localization Of Traffic Lights Using Yolov3 And Stereo Vision. https://www.researchgate.net/publication/343746358 date accessed: 02.06.2021

[8] Kento Yabuuchi, Masahiro Hirano, Taku Senoo, Norimasa Kishi. Real-Time Traffic Light Detection with Frequency Patterns Using a High-Speed Camera. https://www.researchgate.net/publication/343091935 date accessed: 03.06.2021

[9] Guo Mu, Zhang Xinyu, Li Deyi. Traffic light detection and recognition for autonomous vehicles. https://www.researchgate.net/publication/274196452 date accessed: 06.06.2021

[10] Linxiu Wu, Houjie Li1, Jianjun He, Xuan Chen. Traffic sign detection method based on Faster R-CNN https://www.researchgate.net/publication/331901926 date accessed: 09.06.2021