Object Detection And Recognition Techniques In Computer Vision

Kunal Mehta[1], Vineet Makharia[2], Rumil Dand[3], Siddhant Modi[4]
[1, 2, 3]Mukesh Patel School of Technology Management and Engineering
Mumbai, Maharashtra 400056

Abstract- *Making computers gain high level understanding from digital images and videos is a challenge that has needed constant refinement in solutions since its inception. There have historically been a vast number of approaches and methods to improve computer vision, some better than others, but there isn't currently a technology in the modern arena yet that provides a reasonable success rate so as to render all other methods obsolete. This is because the flexibility and freedom that image processing methods offer is matched by the large number of external factors that can affect the performance of computers when interpreting images of similar objects that differ in some characteristics. In this paper, we study computer vision in the areas of object detection and recognition by studying some of the most frequently used result-oriented algorithms and methods.*

Keywords- Computer Vision, detection, recognition, image processing

I. INTRODUCTION

Object detection is a process for identifying a specific object in a digital image or a video. Object Recognition includes identifying the instances of real-world objects such as faces, bicycles, and buildings in these images after detection and gaining useful information from them. Computer Vision as a field is challenging in the fact that it seeks to emulate the human visual sensory system, a system incredibly comlex and not yet completely understood by biologists and neroscientists as well. Problems faced during object recognition are:

1. Lightning: The lightning conditions may differ during the course of the day. Also the weather conditions may affect the lighting in an image. In-door and outdoor images for same object can have varying lightning condition. Shadows in the image can affect the image light.

2. Positioning: Position in the image of the object can be changed. If template matching is used, the system must handle such images uniformly.
3. Rotation: The image can be in rotated form. The system must be capable to handle such difficulty.
4. Mirroring: The mirrored image of any object must be recognized by the object recognition system.
5. Occlusion: The condition when object in an image is not completely visible is referred as occlusion. The image of car shown in a box in fig.3 is not completely visible. The system of object recognition must handle such type of condition and in the output result it must be recognized as a car.
6. Scale: Change in the size of the object must not affect the correctness of the object recognition system.

As a result, most of the approaches to detection and recognition of objects in images have resulted in a plethora of available methods and users tend to choose one based on their aims and results required and the performance of a particular method in the furtherance of that aim. These methods range from templating or direct database verification to using edge detection and morphological means to isolate objects in images. However, the abscense of one final solution, so to say, leaves room for exploration and innovation, and this remains of the biggest unsolved problems in the domain of computer science for now.

II. OBJECT DETECTION

Paul Viola and Michael Jones presented a fast and robust method for face detection in 2001 which was 15 times quicker than any technique at the time of release with 95% accuracy at around 17 fps. Viola Jones algorithm was initially used just for facial Detection, but it can be extended to almost all object detection systems. The basic idea is to slide a window across the image and evaluate a model at every location.

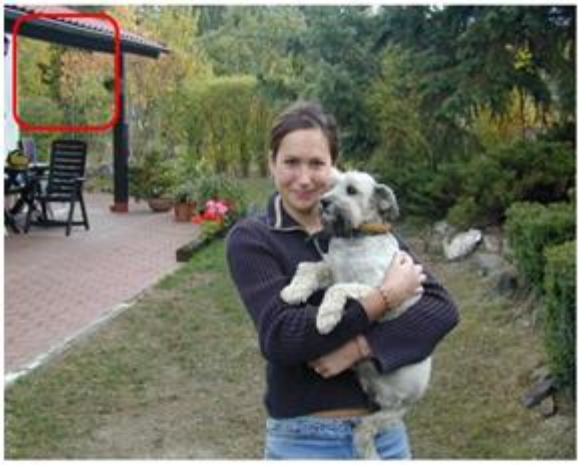


Fig. 1: Viola Jones Algorithm window

The algorithm has 4 main stages:

1) Haar Feature Selection
2) Creating an Integral Image
3) Adaboost Training
4) Cascading Classifiers

All human faces have some similar properties, for example the eyes are darker then the nose or the upper cheek. The location of the eyes, nose, cheeks are fixed in the face as well, and thus we can make use of Harr features. Each haar feature has a value and this can be calculated by taking the area of each rectangle and then adding the result.

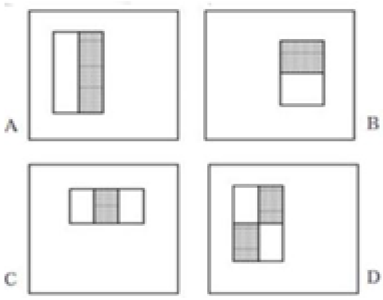


Fig. 2: Haar features

The integral image is defined as the summation of the pixel values of the original image. The value at any location (x, y) of the integral image is the sum of the image pixels above and to the left of location (x, y). The intensity sum of any rectangular-shaped area can be calculated by considering as few as four values of F. This allows for an extremely fast calculation of a convolution. The integral image F can be calculated in pre-processing stage prior to detection in a recursive manner in just one pass over the original image.

$$R(x; y) = R(x; y\ \ 1) + I(x; y) \qquad (1)$$

$$F\ (x; y) = F\ (x\ \ 1; y) + R(x; y) \qquad (2)$$

A Haar feature classifier uses the rectangle integral to calculate the value of a feature. The Haar feature classifier multiplies the weight of each rectangle by its area and the results are added together. We can eliminate false candidates quickly using stage cascading. The candidate is passed through a number of stages, and if any of them fails, we can conclude that the sub-window does not have a face. Every stage becomes more and more complex.

Another widely used approach in object detection and recognition systems, particularly in facial recognition and security systems is skin detection using fuzzy logic. Fuzzy logic, as opposed to boolean logic which suggests that there are just 2 truth values to a variable, true or false, suggests instead that there are an infinte number of truth values, and that the degree of the truth value (value between 0 and 1) decides whether it is suitable or not. There is a caveat to it though, as the algorithm assumes that human skin tones are distributed over a discriminate space in the RGB plane, when in practice, this may not always be true due to external factors affecting the image quality.

The logic works on intensity of the image, and the normal RGB image is first converted to an HSV image. This allows for clearer separation of skin hues from the background as compared to an RGB image.

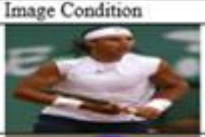
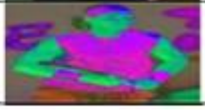
Description of the step	Image Condition
Sample input image (RGB Format)	
Sample input image (after converting it to HSV)	

Fig. 3: RGB to HSV conversion

After the image is converted, a threshold value is applied throughout its spatial resolution based on the intensity values that the skin colors are situated in. This results in isolation of the skin hued area from the general image.

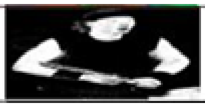
Output image after applying threshold to hue value	

Fig. 4: Threshold applied image

AFter thresholding, the image is reconstructed to HSV. The skin areas, unaffected by the thresholding return to their HSV values, but the surrounding non-skin hued area doesn’t.

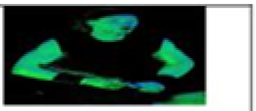
Output image after reconstruction of HSV image from the threshold image	

Fig. 5: Reconstruction after thresholding

Finally, the skin region is retrieved by converting to RBG. The final result of this process is that the skin region is isolated from the rest of the image by keeping it pristine while changing the intensity values of the background.

The final output image in which the background is eliminated and the skin region is retrieved	

Fig. 6: Skin Region Isolation

One of the most highlighted concepts of object detection is that it can make the machine understand the specifics of the object which is a very useful trait for an autonomous machine. A direct application based approach on the concept of Object Detection is that of a prosthetic robotic hand that can be used by an amputee.

III. OBJECT RECOGNITION

One of the methods of object classification is using deep learning, which is used extensilvely by systems aiming to operate fully autonomous vehicles. Convolutional Neural Net-works are some special multi-layer neural networks designed specifically for 2D data, like video and images. The CNNs are motivated by minimal data preprocessing requirements, and they largely receive raw input image and extract features on its own. During the real-time detection phase, the on-road objects are filtered such that the entire system is made to detect only the classes which correspond to on-road objects like animals, pedestrians and other vehicles. Prosthetic arms, while useful as an arm, wouldnt categorically be able to help in performing any functions of the hand, which is the gap the proposed system was designed to fill. The given system aims to be controlled by the EMG signals from our brain and depending upon the object type to held, optimizes the arrangement of the hand and fingers to provide the best possible grip to grab on to the object effectively.

The vision sensor is mounted on the prosthetic hand and moves with it. The image in the work environment is continu-ously captured while moving the hand. Therefore, the system can detect the image of the target object, which the prosthetic hand approaches to. The vision sensor attached takes up to 2000 to 3000 images of the object for accuracy verification. The EMG sensor detects the intention of the user of the model and the EMG signal is processed to get a better idea of the directive. The vision sensor tries to identify the object prematurely. Data from the EMG signals are compared with the data from the vision sensor and the object that the user wants to hold is identified. This control is sent to the motor controller in the system which maps the object in 3D space and accordingly makes adjustment to the arm so that object is grabbed by the arm as seamlessly and sub-consciously as a normal arm can.

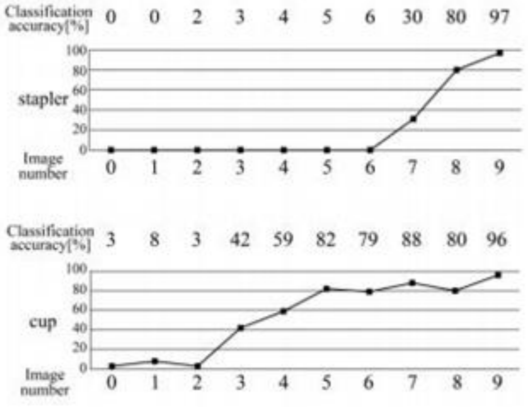


Fig. 7: Success rate

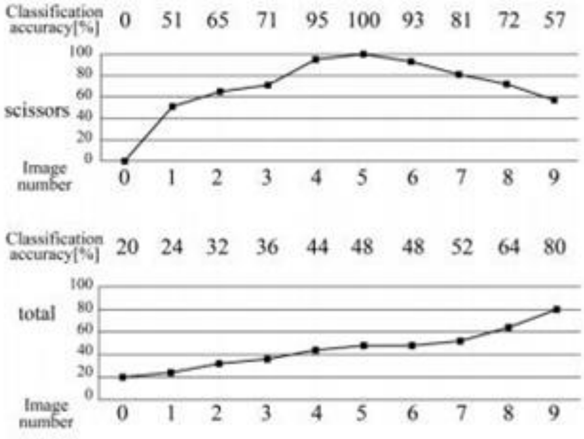


Fig. 8: Success rate

The method of object detection was verified in the approach cycle for the arm. The approach started at approximately 40 cm and up to 2000 to 3000 images were taken per object as a part of data preparation for identification. To quantify the ability of the mechanism to identify an object, images at 10 different instances are taken and the prediction success percentage is plotted as the number of images is increased. As the images are increased, the general trend is observed that the accuracy of getting the correct image is improved thus getting a point in favor of the system. In order to verify the validity of the proposed method, verification experiment we conducted. The images of 25-type objects were used and the recognition accuracy was confirmed for each object. The images were captured while the hand approaches to the object. From the given study it was observed that the general pattern was that of increasing accuracy with decrease in the distance between the camera and the object. The accuracy kept on increasing with moment generally however a few anomalies were observed for certain object types. Although global mean followed a steady growing pattern of increasing accuracy with decreasing distance.

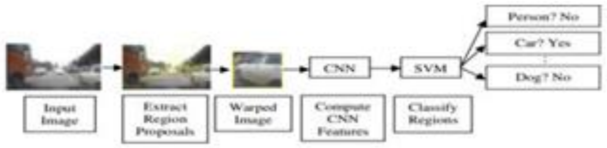


Fig. 9: Functioning of a R-CNN

Each image is fed to the system and the R-CNN trained module returns the processed image on which the on-road objects are recognized with bounding boxes tagged with the respective class name. The results are:

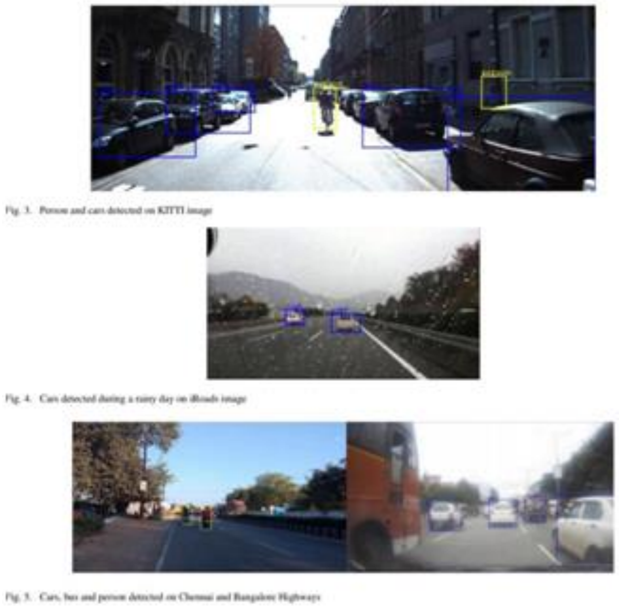


Fig. 3. Person and cars detected on KITTI image

Fig. 4. Cars detected during a rainy day on iRoads image

Fig. 5. Cars, bus and person detected on Chennai and Bangalore Highways

Fig. 10: Results of using R-CNNs for robotic vision

Another popular method of Object Recognitition is Tem-plate Matching. Template matching is a technique for finding small parts of an image which match a template image. The similarity within each image can be easily revealed with simple similarity measure, such as SSD (Sum of Square Dif-ferences), resulting in local self-similarity descriptors which can be matched across images. Object detection in a passive manner does not involve local image samples extracted during scanning. The window-sliding approach uses passive scanning to check if the object is present or not at all locations of an evenly spaced grid. This approach extracts a local sample at each grid point and classifies it either as an object or as a part of the background. However, in active scanning local samples are used to guide the scanning process. At the current scanning position a local image sample is extracted and mapped to a shifting vector indicating the next scanning position. The method takes successive samples towards the expected object location, while skipping regions unlikely to contain the object. The goal of active scanning is to save computational effort.

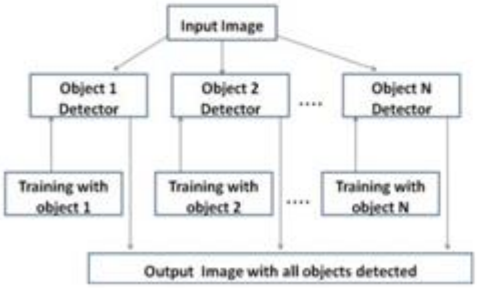


Fig. 11: Method for multi-object detection in an image

One of the methods of facial recognition is to use Eigen vectors. An eigenvector of a square matrix A is a non-zero vector v that, when the matrix is multiplied by v, yields a constant multiple of v, the multiplied being commonly denoted by (Av = v). Facial recognition using Eigenfaces involves:

1. Acquiring of an initial set of facial images.
2. Calculation of the mean of all images.
3. Eigenvector, eigenvalue and eigenface have to been calculated.
4. Signature for each and every image has to be calculated.
5. Subtraction of mean value from the random index image has to be performed.
6. By comparing the signature value with the subtracted value the face can be recognized.

Fig. 12: Successful case

Fig. 13: Unsuccessful case

License plate recognition is one of the most extensively researched subsections of object recognition. it’s applications mainly lie in the field of security and law and order. The proposed method was applied to new Abu Dhabi license plate. The vehicle images were obtained with varying conditions such as illumination, license angles, distance from camera to vehicle and different size and types of license plates. The first step in the implementation of this method is to convert the RGB image into gray scale. This is done in the pre-processing step along with contrast enhancement this helps with images with poor lighting and severe illumination conditions. The next step applied is edge detection, this is the one of the most crucial stage as it helps considerably reduce the candidate region for the license plate by a great margin. The gradient operator is chosen on its capability to detect vertical edges from the surrounding parts. Sobel operator was found to be ideal for this operation. After the edge image is collected, a closing operation is used with structuring element S(mxn) whose width is larger than the largest space between two consecutive characters on the license plate to guarantee that the license plate is not eliminated.

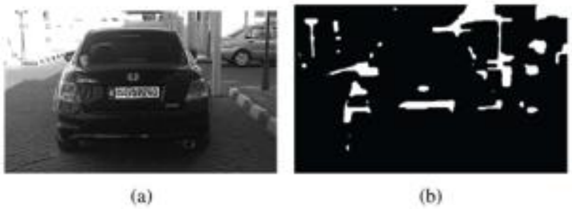


Fig. 14: (a) Enhanced Image (b) Output image of the morphological step

After this step for than one candidates can be present for LP region. To overcome this LP segmentation and verification. It uses 8 connected algorithms. Thus, the following rules must be followed:

1) The shape should be rectangular with aspect ratio w/h between 1.5 to 7.8.
2) The CCi should not be too small such that w ¡= 30 and h ¡= 11.
3) The CCi should not be connected to the image boundary.
4) The CCi orientation should be horizontal up to 30. This concludes the LP localization process.

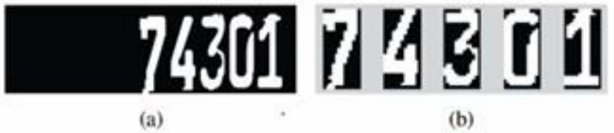


Fig. 15: (a) Image after applying character criteria (b) Segmented characters

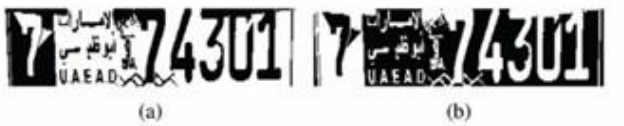


Fig. 16: (a) Binarized image using global thresholding (b) Complemented image

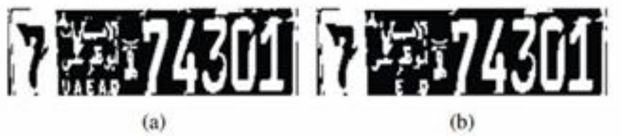


Fig. 17: (a) Binarized Image using adaptive thresholding (b) Image after applying morphological operations

From a testing cluster of 208 images, the results obtained for this method were 97.60% for license plate detection, 90.74% for License plate identification and 97.89% for Character recognition accuracy.

Automatic License Plate Recognition (ALPR), Red Light Camera are used in Russia, London, UK, Australia. But considering the Indian scenario its only restricted to CCTV cameras as of now. The problem with CCTV cameras are low resolution which results in poor footage. Lack of standardization of the number plates in terms of fonts and orientation is also problematic.

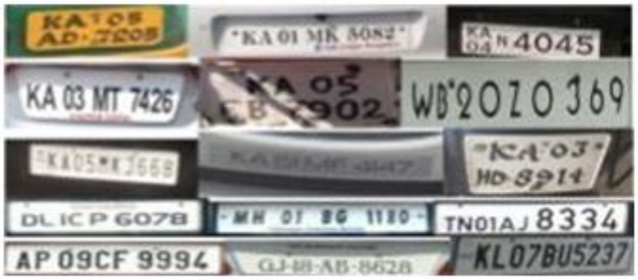


Fig. 18: Lack of standardized license plates

One way, although cumbersome and resource intensive, but effective is to use datasets to train the system, to supply it with a base with which to verify current inputs. The process used is as follows:

1) Image converted to grayscale and pre-processed.
2) Contours which could be characters are isolated.
3) Matching contours are listed, and redundancy removed.
4) Contours checked for longest list of matching characters.
5) Characters are verified with the files the system has been trained with before execution.
6) Possible plates with the most characters found are listed, and license plate found in the image is displayed with maximum characters recognized.

An abundance of information and details aren't always helpful, however. Content is so easily provided and thus the importance of cutting down on the content to enhance speed and reduce computational complexity is practical. Moreover, the aforementioned concept is needed in portable devices as they run on limited power and processing capabilities and in an environment where time is of the essence in taking decisions. While images contain a lot of data, processing that on the front end is not always feasible and hence it is useful to implement ways of server based image processing using an onboard communicator.

With the advent of faster processing capabilities, many algorithms are being developed with the idea of exploiting the exceptional processing powers of the high-end system to their fullest. However, once the algorithms are out of the Research and Development stage as there is a proof of working and a proof of concept, the optimization of the algorithm is necessary.

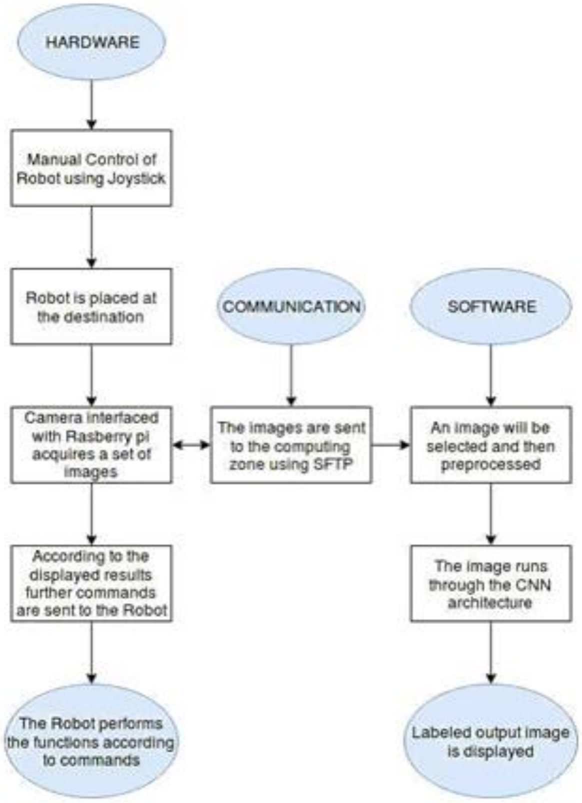

HARDWARE
Manual Control of Robot using Joystick
Robot is placed at the destination
COMMUNICATION
SOFTWARE
Camera interfaced with Rasberry pi acquires a set of images
The images are sent to the computing zone using SFTP
An image will be selected and then preprocessed
According to the displayed results further commands are sent to the Robot
The image runs through the CNN architecture
The Robot performs the functions according to commands
Labeled output image is displayed

Fig. 19: Overview of the integrated system

Current algorithms used for object recognition are on the path to being optimized for low powered systems that can be utilized at terminal ends and are portable in nature. Algorithms like R-CNN, CNN, pattern matching all require high compu-tational bias. Hence to tackle this a client server model was made for the system where in the job of the end computer is to capture the image in good quality for the server end to read and decode in terms of objects and send it back. This makes the model end-to-end and thus provides efficiency with respect to the computational power of the terminal computer by outsourcing the job and waiting for the output. It eliminates any complexities in the design by executing a single neural network by treating it as a regression problem. introduction of Region Proposal Network (RPN) allows sharing of whole image convolutional features with the network, thus, providing near cost-free region proposals. Wherein, region proposal tech-nique is used to guide the algorithm in order to locate objects residing in an image. Secondly, execution of this method in our system allows the system to be computationally efficient and customized to run on low-powered machines. The algorithm is confidence based and hence can have different recognizing confidence thresholds depending upon the application and required accuracy. The algorithm can be better improved to improve calculation on less computational power and also make the process less GPU dependent. An area map can be made if the processing is outsourced to a server using the latest data from the systems in the area and instead of parsing the image to identify objects, the new image can be cross verified with the previous image to register changes if any thus reducing the processing need. The computational model has the ability to learn and hence works much better than simple pattern matching. The obvious conclusion of having the ability to identify more classes already warrants the accuracy and versatility of the method also opening up new avenues for development that can recognize the object in a better way.

IV. CONCLUSION

Computer Vision is a field of research and innovation, and one that needs much more impetus and growth. The currently available methods do achieve a high degree of accuracy, but are much more specific instead of comprehensive. There is growing need for a technique or algorithm that provides high levels of accuracy and is not affected significantly by image quality factors, which in turn are affected by the image's environment at the time of image creation. Until this goal is achieved, however, computer vision applications will continue to use different detection and recognition technologies based on what application they are being used for, for instance, skin detection and eigenface facial recognition for security systems as opposed to R-CNN based systems for autonomous vehicle systems.

REFERENCES

[1] Rajanikant Tenguria, Saurabh Parkhedkar, Nilesh Modak, Rishikesh Madan and Ankita Tondwalkar, Design Framework for General Purpose Object Recognition on a Robotic Platform

[2] Yoshinori Bandou, Osamu Fukuda, Hiroshi Okumura, Kohei Arai, Nan Bu, Development of a prosthetic hand control system Based on general object recognition

[3] Gowdham Prabhakar, Bisnu Kailath, Sudha Natarajan, Rajesh Kumar, Obstacle Detection and Classification using Deep Learning for Tracking in High-Speed Autonomous Driving

[4] Mehul K Dabhi, Bhavna K Pancholi, Face Detection System Based on Viola - Jones Algorithm

[5] Ayman Rabee, Imad Barhumi, License Plate Detection and Recognition in Complex Scenes Using Mathematical Morphology and Support Vector Machines

[6] Khushboo Khurana, Reetu Awasthi, Techniques for Object Recognition in Images and Multi-Object Detection

[7] Arnav Chowdhury, Sanjaya Shankar Tripathy, Human Skin Detection and Face Recognition using Fuzzy Logic and Eigenface

[8] Sachin Prabhu B , Subramaniam Kalambur, Dinkar Sitaram, Recognition of Indian License Plate Number from Live Stream Videos

[9] Jianwei Gong, Shengyue Yuan, Jiang Yan, Xuemei Chen, Huijun Di, Intuitive Decision-making Modelling for Self-driving Vehicles

[10] Shahroz Tariq, Hyunsoo Choi, C.M. Wasiq, Heenim Park, Controlled Parking for Self-Driving Cars

[11] Pascale-L. Blyth, Milo N. Mladenovi, Bonnie A. Nardi, Norman M. Su, Hamid R. Ekbia, Driving the Self-Driving Vehicle

[12] Zhilu Chen, Xinming Huang, End-to-End Learning for Lane Keeping of Self-Driving Cars

[13] Jeremy Straub, Wafaa Amer, Christian Ames, Karanam Ravichandran Dayananda, Andrew Jones, Goutham Miryala, Dylan Shipman, An Inter-networked Self-Driving Car System-of-Systems

[14] Unghui Lee, Jiwon Jung, Seunghak Shin, Yongseop Jeong, Kibaek Park, David Hyunchul Shim, In-so Kweon, EureCar Turbo: a SelfDriving Car that can Handle Adverse Weather Conditions