REAL-TIME OBJECT DETECTION USING DEEP LEARNING TECHNIQUE

**A Project Work Synopsis**

*Submitted in the partial fulfilment for the award of the degree of*

**BACHELOR OF ENGINEERING**

**in Big Data Analytics**

**Submitted by:**

|  |  |  |
| --- | --- | --- |
| **Name** | **UID** | **Class/Section** |
| **Tatikonda Chandrashekharachari** | **20BCS4428** | **20BDA1-A** |

**Under the Supervision of:**

**Tanvi Mam**



**CHANDIGARH UNIVERSITY, GHARUAN, MOHALI – 140413, PUNJAB 2023.**

**Table of Contents**

1. [ABSTRACT 3](#_TOC_250013)
2. [INTRODUCTION 3](#_TOC_250012)
   1. [Problem Definition 3](#_TOC_250011)
   2. [Project Overview 4](#_TOC_250010)
   3. Hardware Requirements 4
   4. Software Requirements 4
3. [LITERATURE REVIEW 4](#_TOC_250009)
   1. [Existing System 4](#_TOC_250008)
   2. [Proposed System… 5](#_TOC_250007)
   3. [Literature Review Summary 5](#_TOC_250006)
4. [PROBLEM FORMULATION 8](#_TOC_250005)
5. [RESEARCH OBJECTIVES 8](#_TOC_250004)
6. [METHODOLOGIES 9](#_TOC_250003)
7. [CONCLUSION 13](#_TOC_250002)
8. [TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK 13](#_TOC_250001)
9. [REFERENCES 13](#_TOC_250000)

# ABSTRACT

In the current scenario, the fastest algorithm for detecting objects in images using a single-layer convolution network is the single-shot multi-box detector (SSD) algorithm. This article investigates object detection techniques to detect objects in real time on any device using the proposed model in any environment. In this article, we improve the object detection classification accuracy by improving the SSD algorithm while keeping the speed constant. These enhancements are achieved using deep separable convolutions in their convolution layers as well as spatially separable convolutions commonly referred to as multi-layer convolution neural networks. The proposed method uses these multi-layered convolution neural networks to develop a multi-layered system model to classify a given object into any defined class.

These schemes then use multiple images and detect objects from these images, labeling them with their respective class labels. To speed up computational performance, the proposed algorithm is applied with a multilayer convolution neural network using more default boxes, leading to more accurate detection. Check object detection accuracy with different parameters such as loss function, frames per second (FPS), average accuracy (mAP) and aspect ratio.

# INTRODUCTION

## Problem Definition:

While the human eye can instantly and accurately recognize a given image (its content, location, and nearby images) by interacting with it, robotic systems that support computer vision can sometimes be too slow and inaccurate. Any development in this area will lead to improvements in efficiency and performance, and could pave the way for smarter, human-like systems. Therefore, systems like advanced technology that allow humans to perform tasks with little or no awareness will certainly make our lives much easier.

Identifying each object in an image or scene using computer/software is called object detection. Object detection is one of the most important problems in the field of computer vision or wireless networks. It is the basis for complex vision tasks such as object tracking and scene understanding and is widely used in wireless networks. The task of object detection is to determine if there is an object belonging to a specified category in the image. If it exists, the next task is to identify its category and location information.

Traditional target detection algorithms are mainly aimed at detecting multiple targets, such as pedestrian detection and infrared target detection. Due to recent advances in deep learning techniques, especially after the emergence of deep convolution neural network (CNN) techniques, object detection algorithms have undergone a revolutionary development. Among these algorithms, the three main methods commonly used in this field are You Only Look Once (YOLO), Single Shot Multiple Box Detector (SSD) and Faster Region CNN (F-RCNN).

## Project Overview:

However, with the emergence of 5G, the data characteristics of wireless networks such as big data, changing business, data diversification, and uneven temporal and spatial distribution of data pose serious challenges for detection. Targets in real-time environments. Additionally, real-time object detection should be performed on any device and in any environment. To meet the challenge, the project's object detection technology detects objects in real time using a model that can run on any device in any environment.

More specifically, our proposed method applies a convolution neural network to develop a model composed of several layers to classify a given object into several defined classes. Based on recent advancements in deep learning for image processing, the proposed scheme then uses multiple images and detects objects from these images and labels them with their respective class labels.

These images can come from videos fed into our prepared model and the model will be trained until the error rate is reduced to an acceptable level. To accelerate the computational performance of object detection techniques, we use an improved single-shot multi-box detector (SSD) algorithm and a faster region convolution neural network.

* 1. **Hardware Specifications:**
     + CPU – Core i5 10Gen /Ryzen 5 or above
     + RAM – 8Gb or above
     + ROM - 500Mb or above
     + Web cam (either built-in or external)
  2. **Software Specifications:**
     + Python Compilers (Jupyter Notebook, Spyder, etc.)
     + Editors (Notepad, Visual Studio Code, Code Blocks, etc.)
     + Python packages for Image Recognition and Deep Learning Techniques

# LITERATURE REVIEW:

## Existing System:

Among the many uses of object detection, safety is one of the most important. Of course, there are also many object detection applications using machine learning in the security field. To our knowledge, the subjects we selected for our literature review have not been the subject of publications.

To our knowledge, no one has conducted an in-depth study of the literature on the subject we have chosen. The two studies we found closest to the survey were either specific or covered specific methods rather than analyzing other articles.

## Proposed System:

Use deep learning models to achieve real-time object detection and recognition in dynamic environments from images and video captured by webcams. The main objective is to detect and recognize objects in real time. All things considered, we need rich data.

We have to observe different types of objects that move relative to the camera. This will help us observe and identify different cooperating and interacting objects. In this project, we focused on precision. Use deep learning models to achieve real-time object detection and recognition in images and video captured by webcams in dynamic environments.

The main objective is to detect and recognize objects in real time. All things considered, we need rich data. We have to observe different types of objects that move relative to the camera. This will help us observe and identify different cooperating and interacting objects. In this project, we focused on precision.

## Literature Review Summary:

1. A study on several techniques for object detection and tracking in video surveillance footage was done in one work by Murugan et al. The paper emphasizes how video surveillance has been a technology since the 1950s and reinforces the point made in the Introduction section about how watching security cameras can be exhausting and have a negative impact on a person's mental health. It also demonstrates how, by introducing a sophisticated surveillance system, automating the procedure was the solution to that tedious chore. The primary goal of this paper is to describe the various techniques for object detection and tracking in videos. Backdrop subtraction is the first technique described in the study. According to the report, backdrop removal is one of the most popular techniques for detecting moving objects. Background subtraction involves identifying and removing the background in order to show only the pixels of the moving object. The issue with it is that the results of the procedure are impacted if the background is not static and

changes as a result of illumination or specific weather conditions. There are other algorithms for background subtraction as well.

1. A study by Flitton et al. compares different 3D interest point descriptors for CT images of baggage at airports. The main idea here is to find interesting things during baggage x-ray inspections. The paper compared five distinct methods: density, density histogram, density gradient histogram, rotation invariant feature transform, and scale invariant feature transform. Although the research was limited to a relatively narrow issue, the report does a good job of assessing each while providing key indicators. Both of the preceding studies do an excellent job of describing the various strategies, but they are too narrow in scope in comparison to our paper. Instead of focusing on machine learning and object detection separately, both publications provide a more in-depth explanation of the object detection techniques. The two articles cover fewer papers than our Systematic Literature Review does because of their narrower scopes.
2. Bezak, P. (September - 2016) suggests a deep learning approach for item recognition in historical architecture images in Trnava. For jobs requiring object recognition, it employs deep learning architectures based on CNN (Convolution Neural Network). Activation functions and a cascade of convolution layers are used to improve architecture. It is critical to determine the number of layers and the number of neurons in each layer. The TRNAVA LeNet 10 model was created and trained for this purpose. This model is based on 460 training images and 140 validation images, which is a 3:1 ratio. The images were color photographs that were 28x28 pixels in size and encoded in jpg format. In the image of the Trnava historical building, the model correctly identified the correct item. The prediction accuracy of the proposed model increased to 98.88%.
3. Jung, H., Lee, S. et al (Jan – 2015), suggested that instead of using manually produced characteristics, deep learning techniques should be used to recognize face expressions. Convolution neural networks (CNN) and deep neural networks (DNN) are two types of deep networks used to solve recognizing problems. Deep networks were quickly built using deep learning toolkits that support CUDA, such as I and CudaConvnet2. They also used the OpenCV library to develop the Haar-like face detection technique. The photographs were cropped and reduced in size to 64 \* 64. The 327 face photos were then divided into ten groups, one for training and the other nine for testing. The recognition rates for six emotions were high, but the disgust label had a low recognition rate. Because the FER 2013 database contained only 547 training photos for the disgust label. Over-fitting is a possibility with the DNN.
4. Tenguria, R., Parkhedkar et al (April – 2017), Convolution neural networks have been replaced by more precise yet sophisticated approaches that can recognize things in real-time, according to the study. This paper has the potential to make significant advances in object recognition and tagging. However, progress in this area has been relatively slow. It plans to merge the fields of computer vision and robotics, with a focus on the implementation of image description applications on an embedded system platform. Depending on the data set used to train the model, a fixed number of items are allowed in the image. According to Shaoqing Ren et al., the development of the Region Proposal Network (RPN) allows for the network to share the entire image’s convolution characteristics, resulting in nearly free region proposals. In this case, the region suggestion technique directs the algorithm to find objects in the image. Second, the application of this technique in our system makes it computationally efficient and tailored to function on low-powered platforms.
5. Etemad, E., & Gao, Q. (Sep. – 2017): In order to improve the effectiveness of current object recognition algorithms, the research presents an object localization method that makes use of image edge information as a cue to pinpoint the locations of the objects. The image’s Generic Edge Tokens (GETs) are extracted using the perceptual organization components of human vision. These edge tokens are parsed using the Best First Search method to precisely locate objects, with the detection score provided by the Deep Convolution Neural Network serving as the goal function. When the BFS is applied to the object localization and its search space, the search space is a collection of edge elements whose overlaps with the current candidate object are greater than zero. Real-time testing revealed that the model outperformed the RCNN, and there is still room for improvement by enhancing object localization by combining picture edge, color, and texture information, as well as the image’s learned properties.
6. Mazumdar, M., Sarasvathi, V. and Kumar, A. (Aug. – 2017), suggested a technique for creating an interactive application to identify things from films; upon user input, it is also capable of identifying the specific object now displayed on the screen. A sequential frame extraction technique for films, as well as a deep learning strategy utilizing Convolution and Fully Connected Neural Networks, is used for this challenge, which has a 77% accuracy rate. Even when the item is slightly warped, translated, rotated, or partially obscured from view and can still be easily spotted by humans, computer vision work remains difficult. Using the fact that videos are made up of frames synced with some playback audio, the video can be analyzed in much greater detail by looking at the objects present in the frame images themselves, running the classifier to obtain probabilities for various classes, and then classifying the genre as well as identifying any objects in the video. By adding more datasets and

optimizing the hardware setup, this model’s operational accuracy is increased, enabling faster and more accurate item categorization over a wider variety of classes.

1. Sujana, S. R., Abisheck, S. S., Ahmed, A. T., & Chandran, K. S. (2017), suggests a method for using convolution neural networks and the idea of deep learning to identify things. Using the input video, it generates an output with a collection of recognized objects. The convolution neural network computes a confidence score for each object. It makes use of the Single Shot Multi-box Detector, which has a high accuracy rate and uses convolution networks to identify multiple items at the same time. It employs Hard Negative Mining and Non-Maximum Suppression to increase the object’s confidence score and produce only one detection for each object, respectively. Thus, combining neural networks with deep residual networks improves the computing efficiency and accuracy of item identification.

Based on the results obtained from the literature review, the following conclusions are drawn.

* + From the performance results of various object detection models on the MS COCO dataset, it can be concluded that SSD and R-FCN models are faster than Faster R-CNN.
  + But if accuracy is more important than speed, Faster R-CNN outperforms SSD and R-FCN models.
  + Faster R-CNN is the most accurate model when using Inception ResNet, operates at 1 frame per second and meets the minimum requirements to perform real-time object detection and recognition.
  + Compared to other object detection models, SSD is faster but has difficulty detecting small objects.
  + The speed of Faster R-CNN increases as the number of proposals decreases, which also reduces the accuracy of the model.
  + According to Redmond et al, YOLOv3 detects 10 times faster than state-of- the-art methods. Therefore, YOLOv3 and its variant Tiny-YOLOv3 were chosen for the experiments.

# PROBLEM FORMULATION

There has been a rapid and successful expansion of computer vision research in recent years. Part of this success can be attributed to the adoption and adaptation of machine learning methods, while others can be attributed to the development of new representations and models for specific computer vision problems, or the development of efficient solutions. Object detection is one area that has made significant progress. There are many objects in this world that humans have identified. So, this one is to make the machines recognize them. The current work provides an overview of object detection.

Object detection, given a set of object classes, consists in determining the location and scale of all object instances, if any, present in an image. Thus, the goal of an object detector is to find all object instances of one or more given object classes regardless of scale, location, pose camera view, partial occlusions, or lighting conditions.

# RESEARCH OBJECTIVES

The goal of object detection is to recognize multiple objects in a single image and predict the corresponding bounding boxes as well as return the confidence of the class for each object.

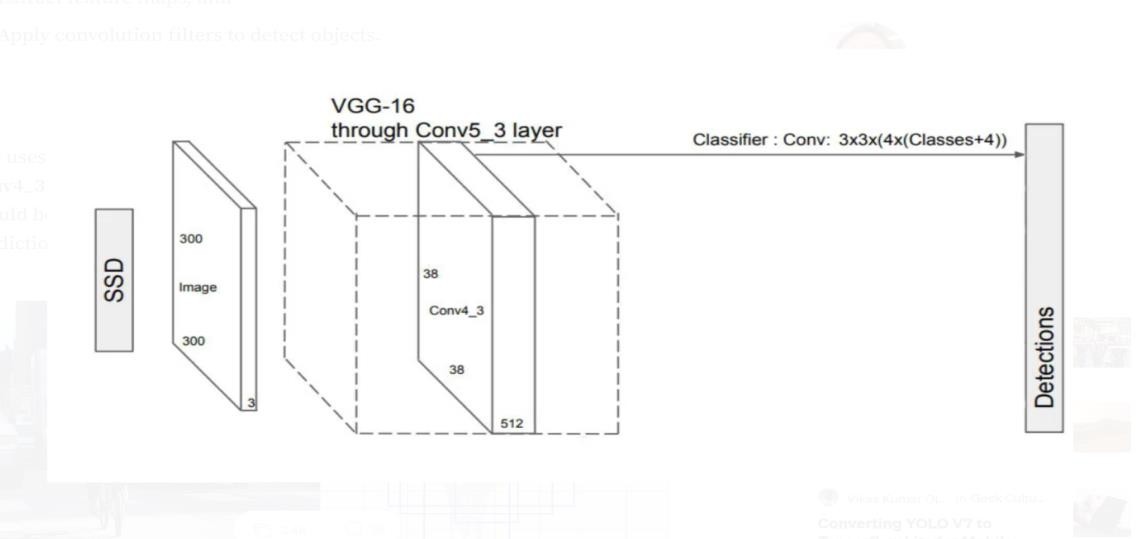
The following objectives have been identified:

* To find suitable and highly efficient deep learning models for real-time object recognition and construction vehicle tracking.
* Assess the classification performance of the deep learning models that have been chosen.
* Compare the classification performance of the chosen models and present the results.

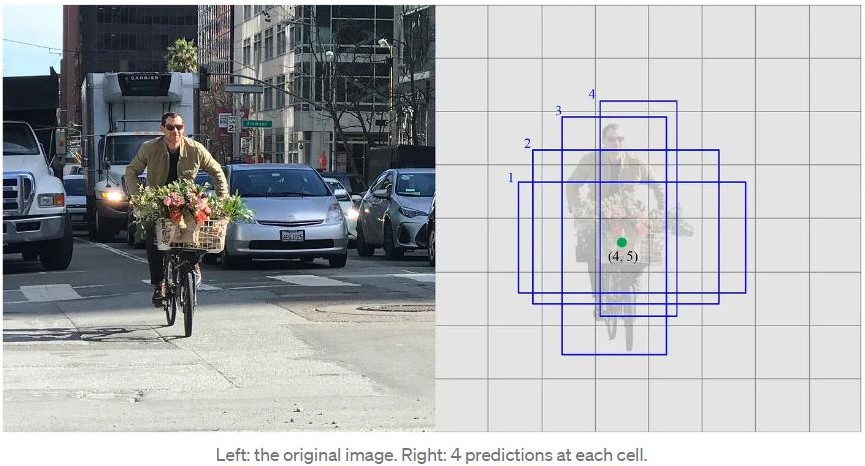
# METHODOLOGIES

Real-time object detection is a feature of SSD. Faster R-CNN uses boundary boxes that are generated by a region proposal network to classify things. The entire process moves at a speed of 7 frames per second, yet it is thought to be state-of-the-art in precision. Below the requirements for real-time processing. By doing away with the necessity for the regional proposal network, SSD accelerates the process. SSD implements a few improvements, such as multi-scale features and default boxes, to make up for the loss in accuracy. These enhancements enable SSD to match the Faster R-accuracy CNN’s while employing images of lesser quality, significantly increasing speed. The following comparison shows that it surpasses the accuracy of the Faster R-CNN and even achieves real-time processing performance.

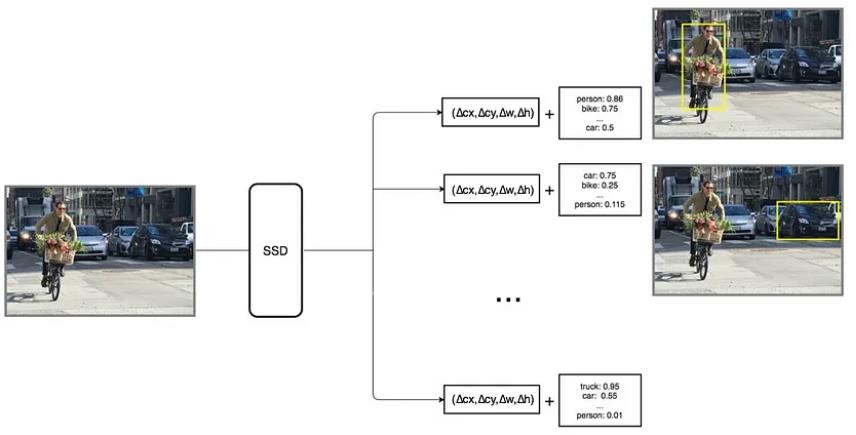
The SSD object detection composes of 2 parts:

* Extract feature maps.
* Apply convolution filters to detect objects.

SSD uses VGG16 to extract feature maps. Then it detects objects using the Conv4\_3 layer. For illustration, we draw the Conv4\_3 to be 8 × 8 spatially (it should be 38 × 38). For each cell (also called location), it makes 4 object predictions

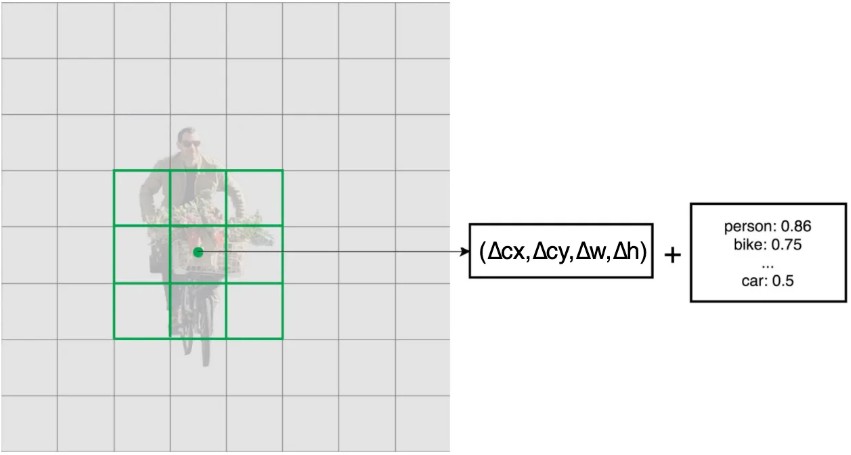


Each prediction composes of a boundary box and 21 scores for each class (one extra class for no object), and we pick the highest score as the class for the bounded object. Conv4\_3 makes a total of 38 × 38 × 4 predictions: four predictions per cell regardless of the depth of the feature maps. As expected, many predictions contain no object. SSD reserves a class “0” to indicate it has no objects.



**Convolution predictors for object detection:**

SSD does not use a delegated region proposal network. Instead, it resolves to a very simple method. It computes both the location and class scores using small convolution filters. After extracting the feature maps, SSD applies 3 × 3 convolution filters for each cell to make predictions. (These filters compute the results just like the regular CNN filters.) Each filter outputs 25 channels: 21 scores for each class plus one boundary box (detail on the boundary box later).

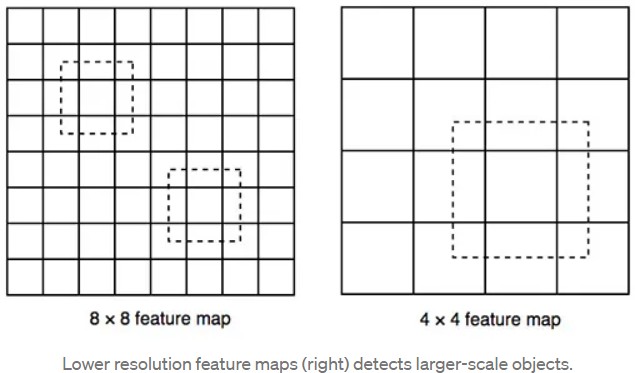


For example, in Conv4\_3, we apply four 3 × 3 filters to map 512 input channels to 25 output channels.



**Multi-scale feature maps for detection:**

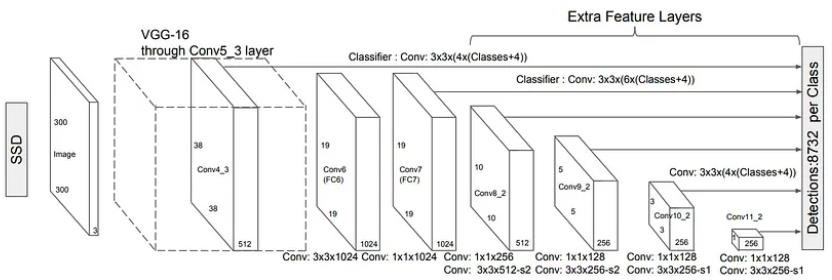
At first, we describe how SSD detects objects from a single layer. Actually, it uses multiple layers (multi-scale feature maps) to detect objects independently. As CNN reduces the spatial dimension gradually, the resolution of the feature maps also decrease. SSD uses lower resolution layers to detect larger scale objects. For example, the 4× 4 feature maps are used for larger scale object.



SSD adds 6 more auxiliary convolution layers after the VGG16. Five of them will be added for object detection. In three of those layers, we make 6 predictions instead of

4. In total, SSD makes 8732 predictions using 6 layers.

**Choosing default boundary boxes:**

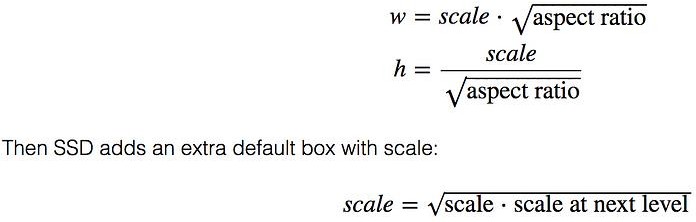


Default boundary boxes are chosen manually. SSD defines a scale value for each feature map layer. Starting from the left, Conv4\_3 detects objects at the smallest scale

0.2 (or 0.1 sometimes), and then increases linearly to the rightmost layer at a scale of

0.9. Combining the scale value with the target aspect ratios, we compute the width and the height of the default boxes. For layers making 6 predictions, SSD starts with

5 target aspect ratios: 1, 2, 3, ½, and 1/3. Then the width and the height of the default boxes are calculated as:



And aspect ratio = 1.

**Matching Strategy:**

SSD predictions are classified as positive matches or negative matches. SSD only uses positive matches in calculating the localization cost (the mismatch of the boundary box). If the corresponding default boundary box (not the predicted boundary box) has an IoU greater than 0.5 with the ground truth, the match is positive. Otherwise, it is negative. (IoU, the intersection over the union, is the ratio between the intersected areas over the joined area for two regions.)

# CONCLUSION

One of the newest and most fascinating areas of deep learning is object recognition. Face detection is a popular use of object detection, found in almost all smartphone cameras. On multiple deep learning models capable of performing real-time object detection and recognition. By studying the performance of these algorithms on standard datasets, YOLOv3, Tiny-YOLOv3 and Faster R-CNN have been identified as the most suitable and efficient deep learning models to perform detection and recognition of real-time objects on scale engineering vehicles. It is concluded that Faster R-CNN performs on par with SSD and FCN models in terms of speed, while exhibiting better accuracy compared to these models.

# TENTATIVE CHAPTER PLAN FOR THE PROPOSED WORK

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **December** | | | | | | **January** | | | | | |
| **16** | **17** | | **21** | **30** | | **01** | **02** | **03** | | **05** | |
| **Planning** |  |  | |  |  | |  |  |  | |  | |
| **Documentation** |  |  | | |  | |  |  |  | |  | |
| **Design** |  |  | |  |  | |  |  |  | |  | |
|  |  | | | |
| **Development** |  |  | |  |  |  | | |  | |  | |
| **Testing** |  |  | |  |  | |  |  |  | |  | |
|  | | |
| **Submission** |  |  | |  |  | |  |  |  |  | |  |

# REFERENCES

* S. Murugan, K. S. Devi (2018), A. Sivaranjani, and P. Srinivasan, “A study on various methods used for video summarization and moving object detection for video surveillance applications,” Multimer. Tools Appl., vol. 77, no. 18, pp. 23273– 23290.
* G. Flitton, T. P. Breckon, and N. Megherbi, “A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery,” Pattern Recognition., vol. 46, no. 9, pp. 2420–2436, 2013.
* Bezak, P. (2016, September). Building recognition system based on deep learning. In 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR) (pp. 1-5). IEEE.
* Jung, H., Lee, S., Park, S., Kim, B., Kim, J., Lee, I., & Ahn, C. (2015, January). Development of deep learning-based facial expression recognition system. In 2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV) (pp. 1-4). IEEE.
* Tenguria, R., Parkhedkar, S., Modak, N., Madan, R., & Tondwalkar, A. (2017, April). Design framework for general purpose object recognition on a robotic platform. In 2017 International Conference on Communication and Signal Processing (ICCSP) (pp. 2157-2160). IEEE.
* Etemad, E., & Gao, Q. (2017, September). Object localization by optimizing convolutional neural network detection score using generic edge features. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 675-679). IEEE.
* Mazumdar, M., Sarasvathi, V., & Kumar, A. (2017, August). Object recognition in videos by sequential frame extraction using convolutional neural networks and fully connected neural networks. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1485-1488). IEEE.

## LINKS:

[www.mindwise-groningen.nl/deep-learning-the-beautiful-mind/](http://www.mindwise-groningen.nl/deep-learning-the-beautiful-mind/)

https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-020-01826-x

https://[www.ijert.org/a-study-on-real-time-object-detection-using-deep-](http://www.ijert.org/a-study-on-real-time-object-detection-using-deep-) learning#:~:text=Object%20detection%20empowers%20recognizing%20instance,much%20like% 20human%20vision%20works.

http://www.diva-portal.org/smash/get/diva2:1414033/FULLTEXT02.pdf https://[www.ijraset.com/research-paper/moving-object-detection-using-ml](http://www.ijraset.com/research-paper/moving-object-detection-using-ml) https://youtu.be/ATw1Dy4p1GU

https://youtu.be/hMFx1TXjAJc https://youtu.be/1P6K9pVKgVo