Object Detection in a video based on Frame Differencing using Deep learning

Dissertation submitted in part fulfilment of the requirements for the degree of MSc in Data Analytics at Dublin Business School

Somasundaram Varadharajan
10394787

Declaration:

I, Somasundaram Varadharajan, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this work are fully compliant with the Dublin Business School's academic honesty policy.

SIGNATURE
Somasundaram Varadharajan
Date: 26/08/2019

ACKNOWLEDGEMENT

I would firstly want to thank my thesis guide Dr.Amir Sajad Esmaeily at Dublin Business School. The door to Amir's Office was constantly open at whatever points I kept running into an inconvenience spot or had an inquiry concerning my exploration or composing. He reliably enabled this paper to be my very own work, however controlled me in the privilege the bearing at whatever point he thought I required it. Without his enthusiastic interest and info, the research couldn't have been effectively finished.

I would likewise want to thank Mr. Shahram Azizi sazi at Dublin Business School as the second reader of this proposal, and I am thankfully obligated to his truly profitable remarks on this thesis. This achievement would not have been accomplished without them. Thank you.

Author – Somasundaram Varadharajan

ABSTRACT

Earlier, security guards were employed to safeguard the people and premises. Due to advancements in technology and cheap storage cost, surveillance cameras have been installed in many public and private properties to prevent crime and theft activities. Though, it requires human intervention for monitoring and if any suspicious activities are detected it should be manually reported to the relevant authority. This is a time-consuming process and may have a high possibility of human error. Hence, there is a requirement to automate this entire process and a lot of research has been made to build a highly accurate model to overcome this major problem. This project is implemented to detect object like vehicles, human beings, animals and many other classes required for security monitoring purposes with the help of Convolutional Neural Network by using the frame differencing technique in the Structural similarity Image Metric algorithm. The built model has been trained effectively to detect around 80 objects present in the COCO dataset.

Table of Contents

List of Figures:

CHAPTER 1: INTRODUCTION

Humans flash at a picture and at a glance they could able to detect objects present in the image, where they are and the way they interact. The human visual system and the neural system reacts quick and authentic, which enable us to execute accomplish various intricate tasks like driving. Development of quick and authentic algorithms would enable in developing a driver-less car without any sensor, enable assistive devices to pass the real time messages to the end users. Many computer vision and video analytics algorithms rely on background subtraction as the engine of choice for detecting areas of interest (change).

1.1 Objective

The main aim is developing a model for object detection and segmentation in a video, where the detection is performed with the help of the concept of frame differencing. Frame differencing is a technique where the algorithm computes the difference between two video frames. If the pixels have changed there apparently was found a motion difference between the frames (moving for example). Most techniques work with some blur and threshold, to differentiate real movement from noise.

1.2 Explanation of computer vision and object detection

The human brain processes visual information in semantic space mainly, that is, extracting the semantically meaningful features such as line-segments, boundaries, shape and so on. But by recent information processing techniques, these kinds of features cannot be detected by computers robustly so that in computer vision it's still difficult to process visual information as humans do. Computers must process visual information in data space formed by the robustly detectable but less meaningful features such as colours, textures etc. Therefore, the processing methodology in computers is quite different from that in human beings. In the talk, we will address the main principle of the image recognition (classification) approach in computer vision, its seedtime, main results and the difficulty faced recently.

Object detection mainly deals with identification of real-world objects such as people, animals, and objects of suspense or threatening objects. Object detection algorithms use a wide range of image processing applications for extracting the object's desired portion. It is commonly used in applications such as image retrieval, security, Medical field and defence.

Analysis and understanding of video sequences is an active research field in computer vision due to its applicability in diverse discipline. Important applications of moving object detection include automatic video surveillance system, optical motion capture, multimedia application and so on. Such kind of applications need to identify person, car or other moving objects in the scene. So, the basic operation needed is the separation of the moving objects called foreground from the static information called background. The process considered as unavoidable part for this kind of operation relies on background subtraction-based techniques. In the literature, many background subtraction methods can be found to handle the illumination changes and other background changes met in video sequences.

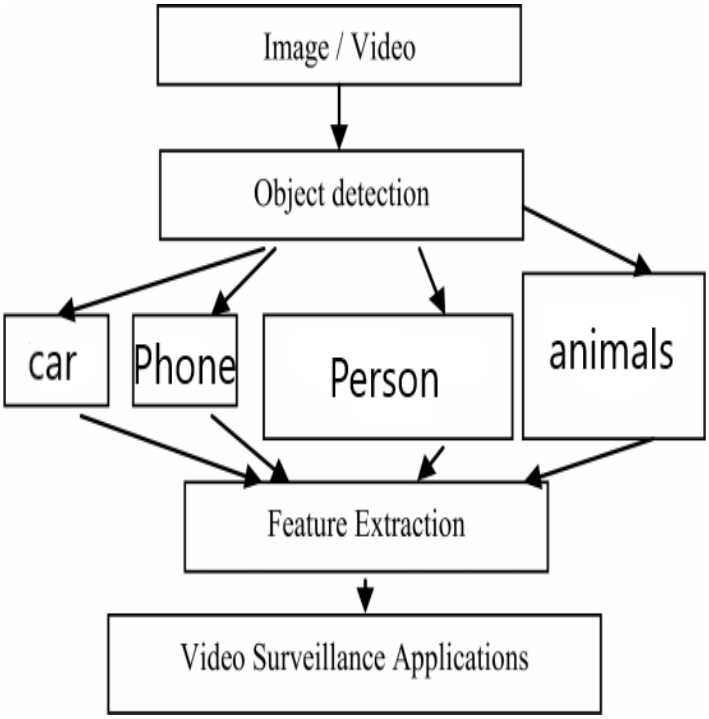


Fig. 1. Basic block diagram of object detection process

1.3 Fundamentals of Object Detection

Object detection is a technique of detecting a foreground object in a frame. The desired object could be person, animal or any other object or target of interest.

1.3.1 Foreground Object

A foreground object is distinct from the stationary background. It could be with respect to its appearance or local motion. It tends to change from frame to frame.

1.3.2 Background Object

Stationary objects in a frame which are part of the background are called background objects.

1.4 Motion Detection

Motion is an attribute associated with objects seen in the video scene. Therefore, tracking motions means to identify objects that are in motion. The task of tracking motion of moving objects, for example, in video surveillance takes the following functional steps to accomplish the goal:

(1) detection of moving segments of a video scene,

(2) segmenting such objects cut out of the whole image,

(3) detailed analysis of the segmented part of image identified as moving to find salient corners or feature points,

(4) triangularization of a set of obtained feature points to visualize the object and

(5) finding correspondence of the feature points from one frame to the next frame for tracking the motion of a moving object.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In recent years, analysis on video inclusive of image processing has become the state-of-the-art in the research domain. Development of object detection models which are built on handcrafted features are usually made with machine learning techniques, which are shallow learning architectures. The performance of the models lags because of the construction of composite ensembles which concatenates many low-lying image features and higher context from detectors and classifiers. Contemplating the fast-paced development in deep learning, more powerful tools, which can efficiently perform learning on semantic, high-level, deeper features and can address few other existing problems in existing conventional architectures. In this review, it was done on a focus with deep learning based on object detection frameworks using a predefined algorithm named Convolution neural network which belongs to a class of deep neural networks.

2.2 Deep Learning

The journal (Schmidhuber, 2015) explains the working of deep learning model and also distinguishes the difference between the deep and shallow learners. Deep learning algorithm allows computational models that are comprised with many layers for processing the requirement with multiple abstraction layers for filtering the data that are needed. The deep learning model has many state of the art applications(Lecun, Bengio and Hinton, 2015) object detection, voice processing, speech recognition and various other scrutinizing application. Deep learning recognizes the convolute or complex structure in the large datasets and shows the model to learn their own internal parameters which are observed from the previous layer which can be used in the current layer, this can be done using the backpropagation method. Convolution nets have brought advance development in processing audio, video, speech and image when compared to recurrent nets.

2.3 Convolution neural networks

Convolution neural networks are iterative neural network where convolution layers turn out in turn with subsampling layers(Cires *et al.*, 2003)(Schmidhuber, 2015), which is a resemblance of simple and complex cells present in the neuron. This depends on way the neural nets getting trained as well as identification of convolution and subsampling layers. Image processing layer

is an optional layer where all the pre-processing of the data is done and convolutional layer is parametrized by way of the dimensions and the wide variety of the maps, kernel sizes, skipping elements and connection table.

Convolution networks being a subset of deep learning model acts as a best and strong model for understanding the extraction of features and visual models. (Krizhevsky, Sutskever and Hinton) used convolution neural networks prosperously for grasp detection as a classifier in the detection pipeline using sliding window concept. The problem addressed here is coincidentally merges with the problem arises here, but the only difference is processing pipeline and different usage of network architecture, this increases the accuracy at greater speeds comparatively.

2.4 Grasp Detection

Contemporary work on grasp detection mainly spotlights the issues in detecting grasps solely from RGB-D data(Saxena, Driemeyer and Ng, 2008). These algorithms depend on the machine learning techniques to detect a good grasp from the data. Grasp visual models are the state of the work objects and it is well known for single object view, now not a complete bodily model. The core problem to be addressed in computer vision is object detection(Nowozin and Lampert, 2010). To start with detection of pipelines usually start with decoction of available robust features from input images (SIFT, HOG, Convolution features). Then, in the available feature space classifiers or localizers are passed to detect objects. After that the concept of sliding window is implemented either throughout the image or at a part of the image with the help of classifier or localizers.

2.5 Deformable parts models

Deformable parts models (DPM) is one of the algorithm which uses sliding window approach to identify objects (*IEEE Xplore: IEEE Transactions on Pattern Analysis and Machine Intelligence*). DPM uses a disjoint pipeline to extract static features, classify regions, predict bounding boxes for high scoring regions, etc. This can be replaced by a single convolution network, in which it may outperform the existing one. The layer performs feature extraction, bounding box prediction, nonmaximal suppression and contextual reasoning circumstantially.

In the place of detecting static features, the network can be implemented in a way that it finds the features in line and shape up themselves for the identification purpose.

2.6 R-CNN

The concept of sliding window is replaced by region proposal method in R-CNN and its variants. Selective search is a method used for generating possible bounding boxes, a convolutional network extracts features, an SVM scores the boxes, a linear model adjusts the bounding boxes, and non-max suppression eliminates duplicate detections. Each stage of this complex pipeline must be precisely tuned independently and the resulting system is very slow, taking more than 40 seconds per image at test time(Girshick, 2015).

Fast and Faster R-CNN focus on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions instead of Selective Search (Girshick, 2015). While they offer speed and accuracy improvements over R-CNN, both still fall short of real-time performance.

2.7 Deep Multi-Box

Instead of using selective search method, the method is trained to predict regions of interests with the help of convolution neural network (Saxena, Driemeyer and Ng, 2008; Wright *et al.*, 2010). Multi-Box detects objects in a single shot by replacing the prediction of confidence with single class prediction. Anyways, this cannot perform classification on general object detection and still it is just a smaller portion in larger detection pipeline, requiring further image classification. This predicts bounding boxes in an image with the help of convolution network.

2.8 Over-Feat:

Over-Feat algorithm is trained in a way to perform localization and make that localizer to learn itself to perform detection(Sermanet *et al.*, 2013). Even though it performs sliding window detection efficiently, it is considered to be a disjoint system. Localization is the only step optimized for Over-Feat, where detection performance is still not optimized. Similar to DPM, when making a prediction on the data the localizer only considers the local information. As this method lags in global context detection, to produce coherent detection it is mandatory to perform a significant post processing.

2.9 Multi-Grasp

This works on the principle of applying regression on grasps, so that it accurately detects the bounding boxes with the help of grid approach, which is very similar to the working on grasp detection(Redmon and Angelova, 2014). Object detection is comparatively a complex task than grasp detection. The advantage of Multi-Grasp model is that it predicts single graspable region for an image which has only one object. There's no need that it should estimate the size, location or boundaries of the object in an image, it should only identify a suitable region for grasping.

2.10 Shallow networks

To reduce the size of the network it is better to estimate a deep neural network with a shallower model. Before time concept proves that any decision boundary can be approximated by single hidden layer of sigmoid units formed by a network. But it is also proven that vision and speech recognition works better with deeper models compared to shallow models(*INTERSPEECH 2011 Abstract: Seide et al.*, 2011). (Dauphin and Bengio, 2013) has proven that it is very difficult to make a shallow network learn with large number of parameters, whereas the same builds a better output with smaller datasets.

2.11 Batch Normalization

This concept supports the elimination of other forms of regularization which also helps in remarkable improvements in convergence. Only possibility of improving the mean average precision is to add batch normalization on all the convolution layers(Ioffe and Christian Szegedy, 2015). This also regularize the model in a way that it does not overfit. With the help of this concept, the dropout from the model can be removed easily, which prevents the model from overfitting.

2.12 High Resolution Classifier

All currently developed model uses a classifier which is previously trained on ImageNet data. 256 x 256 tends to be the maximum resolution where most of the classifiers works on the resolution below the range. As the solution needs to be on detection, classifier network at 224 x 224 and for detection purpose, the resolution can be increased to 448 x 448. This makes the training model to switch between the classifier and detection and adjust to the input resolution

which is needed for the process. This grants the network enough time to switch and adjust the filters to adapt itself to higher resolution.

2.13 Convolution with anchor boxes

This is the process of a feature extractor where the bounding box coordinates are extracted using the fully connected layer on the top of the network. The bounding boxes using handpicked priors are directly predicted (Girshick, 2011), instead of predicting coordinates. The offsets and the confidence of the anchor boxes are predicted by the region proposal network using convolution layers. So nice convolution layer is kind of an iterative layer, as prediction layer is convolution offsets in the feature map are predicted at every location. To make the network learn easier, offset prediction the main thing to be simplified.

2.14 Alexnet

Alexnet is a convoluted neural network architecture which is a combination of convolution layers and fully connected layers. This was the first successful CNN architecture on image classification. This architecture is used in connection with batch normalization(Ioffe and Christian Szegedy, 2015) layers and compared with binary weight network and binary connect(Courbariauxécole and Bengio, 2015). The process of training the deep neural network based on the forward and backward propagations with binary weights. While updating the parameters, the real value weights are kept as it is.

2.15 Frame Differencing - Background Subtraction

To make the detection accurate, it is most important to perform effective background modelling and the values should be updated periodically. Though there is an improvisation performed on this topic every year, there is no unique way to categorize these methods(Bouwmans, 2011; Brutzer, Höferlin and Heidemann, 2011). Further, it is explained on the various mathematical model of background subtraction to detect the motion in objects.

2.15.1 Basic Model

The easiest and convenient way is to set the current frame at the front and the previous frame at the back and the subtract all the subsequent frames and then extract the details. The disadvantage of this model is that a shaky frame cannot be habituated in a single frame.

Furthermore, the method may falsely deject the first frame as background and the object in the first frame may be ignored. Instead of using the initial frame, the frame difference method uses the previous frame for subtraction purpose(Lai and Yung, 1998). This method easily adapts the lazy illumination variance, anyways the background will not be updated, when the moving object stops abruptly.

2.15.2 Statistical Models: Single Gaussian, Mixture of Gaussians

The temporal axis along with the initialization pixel sequence is designed using a univariate model distribution. The multivariate distribution which means the channels which use RGB colour and modelled as the product of those three RGB independent univariate gaussian distribution where each distribution is parametrized by the sample mean μ and standard deviation σ. The single gaussian model is unimodal and it does not have RGB factors, so it fails to accommodate the oscillating background. To overcome this, mixture of gaussian is incorporated to create a background with multi model. To manage with the illumination variance, the learning rate parameter has been introduced. The selection of the learning rate parameter plays a major role in subtraction, where high rate is more accurate and finds objects in the background, but with lower learning rate it will not adapt to sudden illumination variance(Stauffer and Grimson, 2000)(Stauffer and Grimson, 2016).

2.15.3 Non-Recursive Buffer Based Subtraction

(Lo and Velastin, 2001)store the recent pixel history in a finite buffer to represent the model location. The significant difference between the current pixel and buffer median decides if it were a foreground; else, the new background is enqueued inside the buffer. The first-in-first-out strategy is applied to tackle the situation when the overflow condition is encountered. Subsequently, (Cucchiara *et al.*, 2003) prefer the medoid rather than the median statistics to take the appropriate decision. In another work, the background is modelled using a linear predictive model through Wiener filtering; the covariance of pixel sequence estimates the filter coefficients. This work is further extended in a relevant subspace via PCA. Wang and Suter use the notion of consensus to model the background. Additionally, two algorithms are suggested to deal with rapid varying illumination and background relocation.

2.15.4 Fuzzy Model

During foreground extraction there may be some decision uncertainty, to overcome this fuzzy principle can be incorporated. The text and colour features along with the edge information to model the background, where the foreground pixels are applied to extract the choquet integral(El Baf, Bouwmans and Vachon, 2008). Fuzzy technique can also be used as a classification technique and used to model the shaky background.

2.15.5 Shadow Removal Model

Shadow darkens the scene illumination, and hence, the underlying region falsely appears as foreground. The literature includes two different ways to tackle this situation. The former group suggests various invariant colour model(Cucchiara *et al.*, 2003), to nullify the shadow effects, whereas another set of algorithms prefer the texture feature that remains indifferent in the presence of shadow(Calderara *et al.*, 2006). Wang and Suter, in their work, apply the normalized RGB colour space to model the background; however, it has been observed that the normalized RGB is very much noisy in case of low intensity. (Cucchiara *et al.*, 2003) apply the HSI model to suppress the shadow illumination and use a median filter to selectively update the established model; a second validation is further applied using both invariant colour and texture pattern of the underlying scene. (Huerta *et al.*, 2009).incorporate the colour and colour co-occurrence features to model the static and waving background respectively. Huerta et al. suggest a two-stage approach to counter the shadow illumination. The first stage combines the gradient details and colour information to detect the probable shadow pixels. A second validation is further applied based on the temporal and spatial analysis of chrominance measure, brightness content, texture distortion, and diffused bluish effect. Zhou et al. consider multiple cues such as motion details, object location, its shape, and colour feature to detect the objects in motion.

2.16 Image Quality Metrics

Image processing may have a lot of advancement daily, but still whatever manipulation is made to an image there will be an information loss or quality loss in the image. Evaluation of image quality is divided into objective and subjective methods. Subjective methods mainly depend on human judgement and works without any influence on explicit criteria (Horé and Ziou, 2010)(Smail Avcıbas¸uludag and Sayood, 2002). Objective methods are based on comparisons

using explicit numerical criteria and multiple references like previous knowledge about the statistical parameters and tests and on the ground truth.

The PSNR price processes infinity as the MSE procedures zero; this indicates that a higher PSNR fee gives a higher image quality. At the other end of the size, a small cost of the PSNR implies excessive numerical variations between images. The SSIM is a popular and widely used to measure the similarity between two images. It turned into advanced by Wang et al. (Wang et al., 2004) (Saxena, Driemeyer and Ng, 2008), and is taken into consideration to be correlated with the quality belief of the human visible gadget (HVS). Instead of the use of conventional blunders summation strategies, the SSIM is designed by using modelling any picture distortion as a mixture of 3 factors which might be lack of correlation, luminance distortion and evaluation distortion.

SSIM is a method which is used for evaluating the similarity of two images and use it to measure the quality of the image is proposed. A method based on SSIM is proposed to analyse the characteristics of SSIM value(Wang et al., 2004). Then analysed the possibility of moving target detection using SSIM and the area established to watch the SSIM in the cases of background and foreground with all frames, where the first frame is used as the template to obtain the SSIM between the template and object frames. Finally, SSIM is used for moving object detection directly and the SSIM is input into the Single Gaussian Model to finish the last detection.

2.17 Post-Processing Refinement

Foreground extraction may be erroneous owing to the cluttered background and the inherent sensor noise during image acquisition. It may so happen that a few portions of foreground pixels may be wrongly identified as the background and vice-versa. A post improvisation module should be incorporated to minimize such false alarms [58]. The median filtering is a suitable tool to reduce such false positives. Again, some methods apply connected component analysis to attach the disjointed regions. The size constraint as per the objects of interest can be incorporated to eliminate small foreground pixels. Many authors prefer morphological post-filtering for such improvisation. Morphological Opening is applied to reduce the scattered noise pixels. The closing operation connects the disjointed pixels. Moreover, the morphological filling can be applied to fill the camouflage gap.

The literature includes several articles on the use of background subtraction in identifying the moving entities. Parametric models are based on their underlying assumptions; the appropriate parameter selection can be cumbersome and moreover, it may vary with different scene structures. On the other hand, non-parametric models are more reliable, however, requires a long pixel history to estimate the underlying density function. Pixel-based methods usually apply the colour feature to compare the pixel intensities at the same location over the frame sequence, whereas block-based methods consider the inter-pixel neighbourhood characteristics, partition the image into several blocks, and apply both colour and texture cues to decide the pixel behaviour. The unimodal background outputs significant false positives in case of uninteresting waving motion that can be tackled by the multi-modal background at the price of higher space complexity. The recursive models update the model parameters in an iterative fashion, and thereby fast enough to deploy in real time applications. On the contrary, the non-recursive techniques store the recent pixel information inside a buffer to model the background. The latter one well adapts the gradual illumination variations at the cost of high memory overhead.

2.18 Tracking in Video

(Hati, Sa and Majhi, 2013) an efficient Background Subtraction Method for accurate Object Detection is proposed. Local Illumination based Background Subtraction (LIBS) method is used. Background modelling is done by defining an intensity range for each pixel, shadows are eliminated, which is an added advantage of this method. In monochromatic imagery for tracking people, W4 (Who? When? Where? What?) is used. For modelling the background each pixel is represented by three values, maximum, minimum intensity values and maximum intensity difference between consecutive frames is observed. The locations of these parts are verified and refined using dynamic template matching(Santosh and Mohan, 2015), three algorithms Gaussian Mixture Model (GMM), Extended Kalman Filter and Mean Shift Algorithm are compared in the context of multiple object tracking. The performance of GMM was observed to be good in the presence of occlusions. During Nonlinear transformation, random variables behaved in an abnormal manner, due to this Extended Kalman filter failed. Identification of Multiple objects becomes challenging when there are occlusions. For single object tracking Mean shift algorithm is best suited, which is very sensitive to window size. Jacinto Nascimento [4] In this paper the evaluation of object detection was performed taking five algorithms into consideration such as Lehigh Omni directional Tracking System (LOTS),

Basic Background Subtraction (BBS), Multiple Gaussian Model (MGM), W4 and Single Gaussian Model (SGM). Best results were achieved by LOTS and SGM algorithm in terms of the number of correct detections. False alarms, splits and merges were much less compared to other algorithms. The detection rate is improved without compromising on precision. This approach has been tested on a dataset of complex background scenes. The advantage of this method over other existing methods is that it improves the accuracy of foreground segmentation. This is evident from the results obtained by this method. Behaviour subtraction finds application in characterizing of dynamic events especially behaviour of the object(Jodoin, Saligrama and Konrad, 2009). Each event is composed of various moving objects which have been defined as spatial-temporal signatures. Modelling of these events has been done using stationary random processes.

CHAPTER 3: METHODOLOGY

3.1 Artificial Intelligence

The main objective of Artificial Intelligence is to fill the communication or understanding gap between the humans and machines. The growth of technologies in the Artificial Intelligence has been witnessed with the invention of Convolutional Neural Network and Recurrent neural network. Convolutional Neural Networks (CNNs) have been introduced to perform image classification which is used to solve many real time problems. With CNN different tasks like Voice recognition, classify objects within the given image can be performed. Identify the environmental features like day or night with respect to the light level, Classifying the objects in the video with the help of video frames, etc.

3.2 Convolution Neural Network

CNN comes under Deep Learning, and it is one of the concepts in the machine learning. Deep learning got inspired with the working methodology of the human brain neurons. With the older learning algorithms, it is not possible achieve the required performance level and the accuracy. And this situation got changed when Deep Learning technology comes in the picture post to the year 2013. Other Machine learning subsets performs the classification, prediction tasks on the organization data with the predefined algorithms. On the other hand, in the Deep Learning the system should be trained to learn more about recognizing the patters and structure of the real-time object by using many processing layers. With the knowledge of Deep learning, it can be improved performing different tasks like classifying, recognition, detecting the objects with patterns and describe the actions. Even advanced recommendation systems can be built with the help of Deep learning.

3.3 Layers of CNN

Let's discuss in detail on the working methodology and layers of Convolutional Neural Networks. The major building blocks of the CNN are Convolution and max-pooling Layer. The convolution layer acts as a filter that takes the image or video frames as an input and apply the required filters by preformation activations on the pixels in the input image. The process will be repeated for all the pixels in the input image to generate the feature map. The feature map will be given as an input to the Rectifier Unit (ReLU) or similar function to transform the non-linear values. The end output from the feature map is used to location the objects identified in the image and the accurate values of the feature map adds strengthen the object detection of the input image. The unimaginable ability of

convolutional neural networks is, it can automatically learn by itself with the help of large number of filters that can be formulated with the training dataset to perform the required prediction and classification. With the help of these advanced filters, the system will be able to identify objects anywhere in the given image or video frame.

3.4 General Working of CNN

The CNN classifies the object by assigning the weights and biases to the aspects of image so that the model will be able to classify and differentiate the same when the image is given as input to the model. In CNN, the model won't take much time for the pre-processing as it uses ConvNet which is comparatively better to other image classification algorithms. ConvNet is a layer which could perform the filtering and classifying the characteristic features of the object.

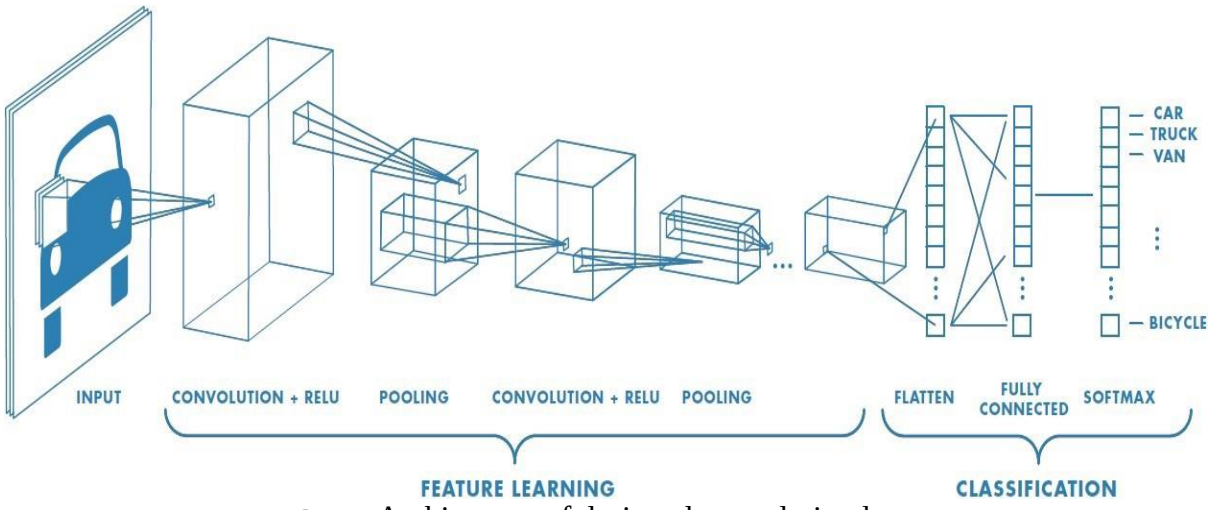


Fig. 2. Architecture of designed convolution layer

ConvNet, a component of CNN able to read the Spatial and Temporal dependencies of the given input with the help of Convolutional Filters. The alternate architecture of ConvNet is Feed-Forward Neural Nets, which can perform predictions on the basic images and that too with an average precision and accuracy on the prediction. Whereas in ConvNet, enhanced architecture can be used by reducing the number of parameters involved in performing the analysis.

In the CNN, RGB layers of the image can be taken as input, which separates the input image with the RGB - Red, Green and Blue colour planes. It also applies different colour spaces on the input image like RGB, Grayscale, HSV, CMYK, etc. The ConvNet is used to reduce the input image into simpler form which will ease the process. On the other hand, the reduction without the losing features in the image will be performed and end up good

prediction. ConvNet architecture is not only to perform classification with high accuracy and to scale to work with the large datasets.

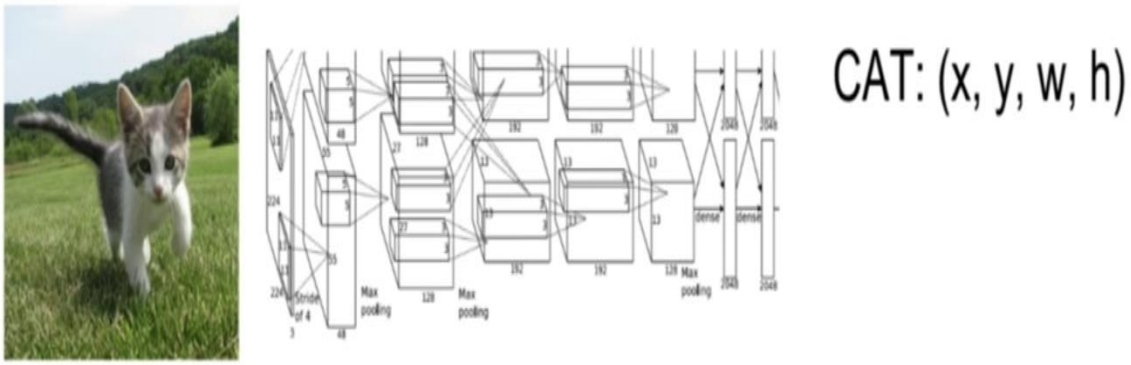


Fig. 3. Working and output of convolution layer with their co-ordinates

In CNN, there is a need to understand the terminology "Stride", which means the number of pixels that will be shifted when applying the filtering over the given input image which got converted into Matrix. When the stride value is set to 1, then the cursor can be moved to apply the convolutional filters to only one pixel at the given point and move to next pixels one-by-one.

You only look once model can be used as a base methodology for the research purpose. The chosen object detection algorithm is used to perform real-time object detection from the images and video frames. There are different models already available to perform the similar operations, and the reason behind the choice of you look only once as base model, is due to its accuracy and extreme speed to detect objects even with the live streaming video with 10k frames. Also, it has different version of chosen technique where it's getting enhanced with every version. The model can still be improved with the performance by limiting the dataset with the defined features that need to be trained.

The chosen object detection algorithm model uses different approach when compared to other older methodologies. It applies neural network concept on the input image and the neurons divides the images into regions and apply filtering with the layers. It first predicts and plot the bounding boxes. Then calculate the probabilities of each region and mapping with the trained model to know the weightage of objects with the calculated probabilities.

In the chosen base model, input image should be taken and divide into an S x S grid. And with the grid, the bounding boxes can be formed. The bounding box also results the probability and the offset values which is resulted from the model. These values are used

to locate the objects within the image. But there are some limitations in the model even though it is faster and good in accuracy. The limitation is that it might have difficulties in identifying the small objects like group of birds flying in the video frame and this problem may be fixed with the future enhancements in the model.

CHAPTER 4: IMPLEMENTATION

4.1 COCO dataset

Coco is a brand-new dataset which was uploaded with the aim of taking the object detection to the next dimension. It has lot of images that are related with our day to day activities which are bit complex but easily recognizable. The data set is properly labelled, and it contains more than 80 objects used by common people around the world for their daily activities and those objects can be easily identified and named even by a kinder garden child. Instance spotting segmentation and category detection were performed by a whole lot of people to label the objects accurately. The data set contains about 325000 images with 2.5 million labelled instances. The statistical analysis that are necessary for the labelling are also performed properly by comparing it with other common and publicly available datasets to enhance its reliability. This data set also enables the option for the performance analysis to analyse the performance of both bounding box segmentation and object detection using a proper model.

4.2 Model

Our model is completely built using multiple layers of CNN and SSIM algorithm is used to finally process the outputs obtained from the CNN model. Parameter sharing forms the base of this model as this is used at its distinctive feature. CNN is different from other common feedforward models in a unique way. CNN learns individual set if parameters at every location instead of learning separate parameters. Various members that are available in kernel are used as input at different locations. This plays an important role in capturing the image initially and the further processing is done using various available techniques which will be explained later. Our model consists of 27 CNN layers and each of the layer is used for various purposes. Initially the algorithms start working by segregating or separating the video into images and then converting those images into a standard size so that it can recognized properly. Despite using CNN, it also contains various other layers like residual, up sampling layer and multiple shortcut connections to effectively classify the objects present in the mirror into various classes. Our model makes use of the bounding box concept to recognize the image properly. CNN starts working by forming the coordinates and axis of the bounding boxes in which the image is present are framed initially and then using those coordinates of the bounding boxes, the probability of that bounding box is identified to detect the object in a frame. Then in the next stage the probability of an object belonging to a class is determined.SO we have 91 different classes and the probability for the object to be classified as a class is figured out. The class with

the highest probability if identified as the class to which the object in the image belongs. The detection of the object is done using three layers in our model thus making is highly efficient compared to other image detection models available in the market. The outputs that are received have almost the same length and width but there will be a slight change in the depth between the layers to differentiate it effectively.

The major advantage of this model is to identify multiple objects that are present in a frame. The various features that are available in the image are recognized and identified properly using these 27 layers of convolutional neural networks. In order to achieve more accuracy and to be more precise the model uses sliding window approach to accurately predict or identify the objects present in the image. The image that is needed to be recognized is broken down into multiple boxes where each grid can handle or accommodate 3 bounding boxes which are used to identify or classify the image properly. The probability of each bounding boxes is noted by the model and if the probability is less than 0.5, they are eliminated. This elimination is done using a specific algorithm called non max suppression algorithm. Using this specific algorithm, the box that has the maximum probability is taken as a trained box and it is compared with the other available box's probability. Then during this comparison, the intersection and the union probabilities are noted down based on the IOU rate the bounding boxes are rejected or approved.

To enable the detection of more than 2 objects within a frame a specific concept called anchor box is introduced which introduces one more dimension to the anchor label and thus this can be used to identify multiple objects within a frame. The figure below shows how multiple anchor boxes within a frame looks like.

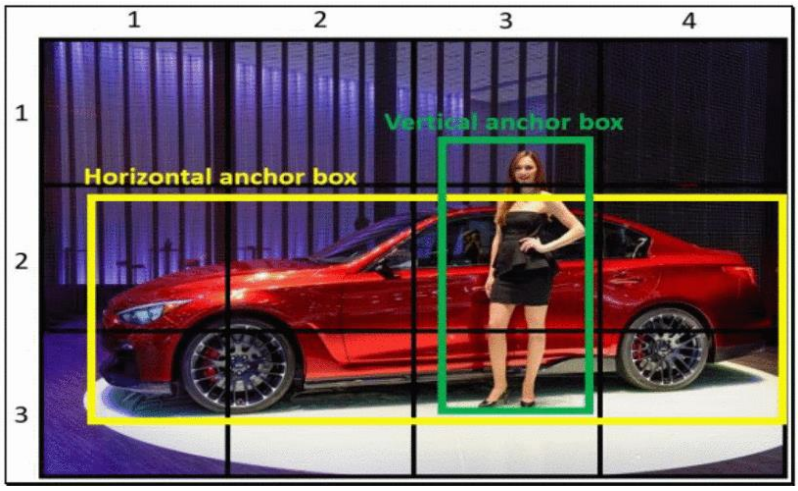


Fig. 4. Bounding box representation

Using this model, very detailed and important semantic information can be obtained from the sampled data and more granular information can be obtained from the feature map. Then along

with this certain CNN layers are also added to club this information together to detect the larger size images effectively.

4.3 SSIM

Structural similarity index is the widely used method to check and detect the quality of the image that is being detected using our model. In our model SSIM is used to identify the motion in the video. This is introduced to enhance the novelty of our model. SSIM has uses two kind situations where one is recorded during motion and the other is recorded when there is no motion in the frame.

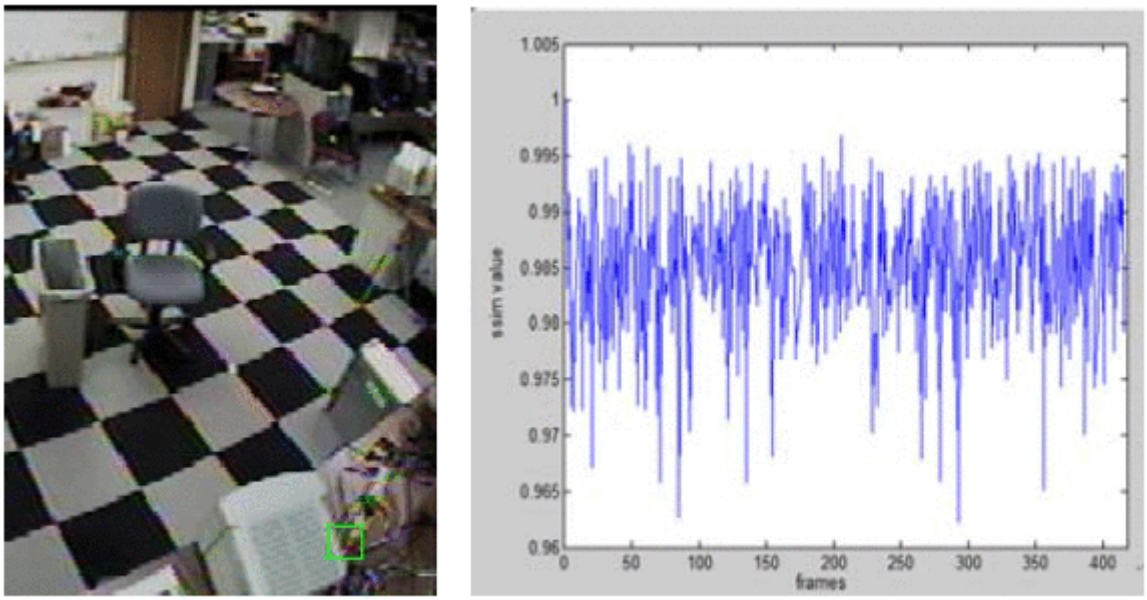
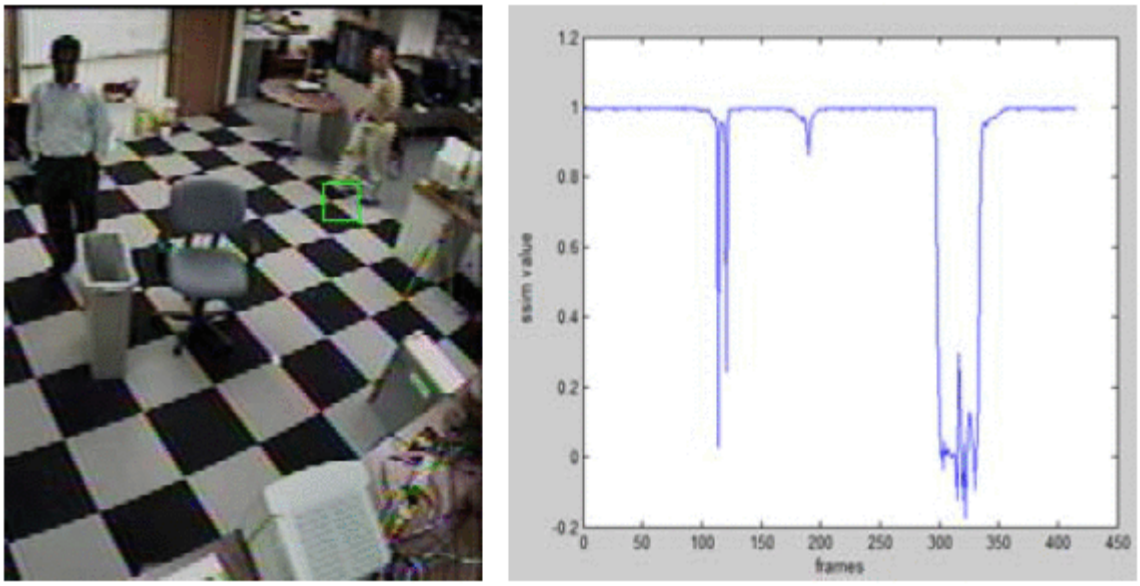


Fig. 5. SSIM Graph for Indoor Scene on selected Grid1

The SSIM value for all the frames will be calculated using the gain SSIM value as reference and based the deviation of this SSIM value, the motion in the frame is identified. If there is no movement in the frame the values of the SSIM will be uniform and will be approximately equal to one. In order to depict these two images taken inside a room and in the outdoor are taken and initially their SSIM value is noted. It was found that they are uniform and equal to one and later the SSIM value is recorder for those two images when there is motion.



(a)Original image in indoor scene (b) SSIM

Fig. 6. SSIM Graph for Indoor Scene on selected Grid2

When there is a motion the values were found to be abnormal. Thus, a concrete threshold value was set to differentiate between the object and the background in the feature. The value of the SSIM obtained is finally fed into gaussian model to detect the image clearly and to avoid confusion between various colour. This combination of SSIM with the single gaussian model is used to eliminate the noise present in our data and to enhance the quality of the detection.

4.4 Hardware and software specifications:

Detecting an object in a video with naked eyes are easy. But to make the machine detect an object in a video it takes time. In order to make it possible Python as a source code has been used. Python is a open source language that is used mainly for statistical and data analytics purpose. It has various packages and libraries inbuilt that can be used for this research. The ease of coding and understanding the bug and tracing the bugs are easier in python. That is major reason behind choosing python as base coding language.

The implementation was done in Visual studio code due to its ease in getting the packages installed. The terminal is inbuilt. It is platform independent and can be used in any Operating system however the installation steps may change.

Editing, building and debugging the code is a lot easier. It is easy to integrate with other tools to perform tasks that happens on daily basis. It is very robust and has extensible architecture so any type pf coding language can be written and executed.

Since we have more support and compatibility of software in Linux, we suggest using Ubuntu 18.04 LTS. The NVIDIA drivers as follows,

1. Display Drivers

2. CUDA 9

3. cuDNN 7.6.1 4. TensorRT 5. Video Codec SDK

Development software: 1. Anaconda Python 2.7 and 3.5 2. Visual Studio Code 3. PyCharm 4. Atom

When it comes to workstation hardware, we have multiple choices.

GPU, CPU, RAM and Hard Disk are the very important hardware required to perform any task.

Hardware specifications – Microsoft Azure – NC series+, CPU Xeon Platinum / Xeon W – 18/24 Core RAM 256GB HDD

As required GPU Basic Tier: 3X GeForce RTX 2080 Ti, Titan RTX Mid-Tier: Quadro RTX 8000

High-end: Quadro GV100”

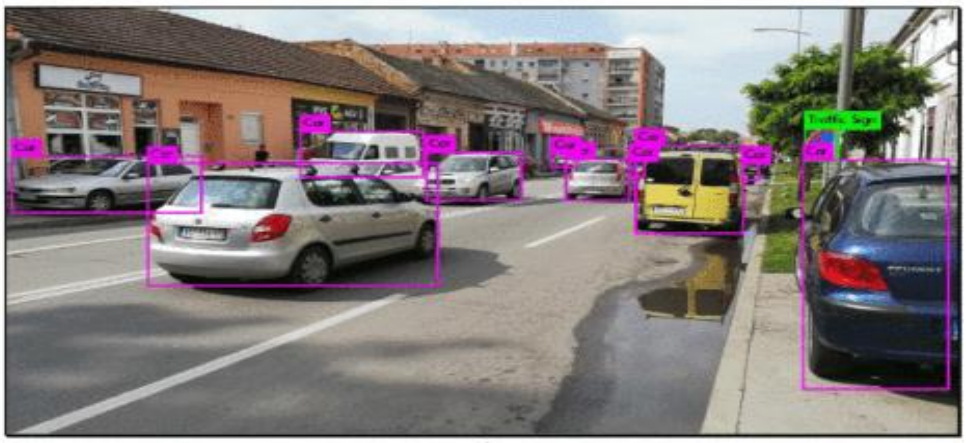
As these graphic cards are primarily used for gaming.

The Quadro graphics card can be scaled later, if required using NV-link.

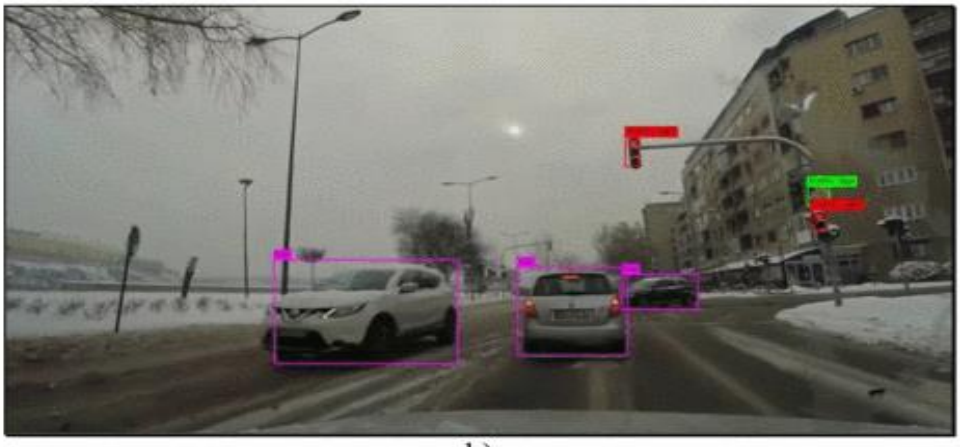
CHAPTER 5: RESULTS AND DISCUSSION

5.1 Results

Because of the too many small and occluded, but labelled objects in the dataset, precision and recall values were not reaching high values as expected. Common false detections found in the outputs of the neural network were further investigated on the smaller custom dataset and video. Most of the undetected objects were in heavy traffic situations where one cell in the detection grid would be responsible for detecting more than three objects. In the remaining cases, some of the objects were undetected because they were occluded by other detected objects. However, in both previous cases, all of the closest objects to the camera's position (vehicle) were successfully detected and classified as shown in *Fig 3*. Because of this accuracy, the safety of vehicles and passengers will not be put in danger at any moment if using this algorithm for the detection of traffic participants.



a)



b)

Fig. 7. Detection visualization examples in different weather

The model adopts the deep learning method to improve the network structure It has achieved good results in object detection. In this paper, object detection algorithm is combined with frame differencing algorithm where the current frame and the previous frame is compared among themselves and only if there is a difference in the frame or change in the pixel then there detects an object or else the frame will be skipped. This brings out a major advantageous thing where it could save a lot of time and cost in terms of deployment. Experiments also show that this algorithm improves the accuracy of object detection. The number of detection frames can reach 25 frames/s, basically meeting the requirements of real-time performance. The weights of the neural network were initialized using a pretrained model trained on the COCO dataset. The weights and configuration details of the model are stored as a trained model and used for further classification of the given input. Once the video is given as an input it is streamed out over the loops for detection and classifies the output. SSIM is a method which is used for evaluating the similarity of two images and use it to measure the quality of the image is proposed. A method based on SSIM is proposed to analyse the characteristics of SSIM value. Then analysed the possibility of moving target detection using SSIM and the area established to watch the SSIM in the cases of background and foreground with all frames, where the first frame is used as the template to obtain the SSIM between the template and object frames. Finally, SSIM is used for moving object detection directly and the SSIM is input into the Single Gaussian Model to finish the last detection.

Fig. 8. Frame taken for detection

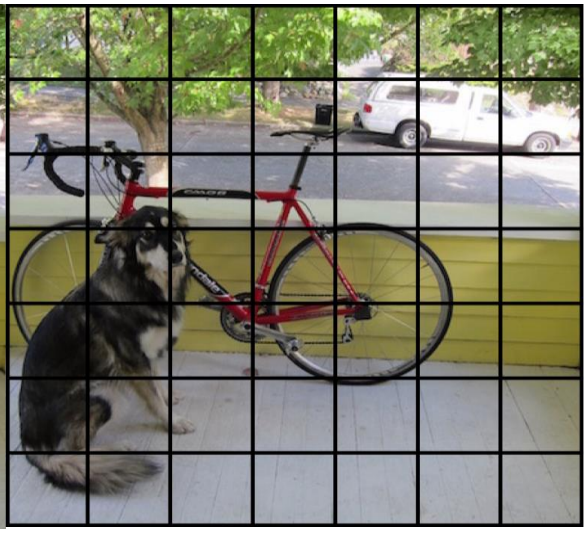


Fig. 9. Frame split into S x S grid

The model divides the input image into associate S×S grid. Every grid cell predicts just one object. For instance, the yellow grid cell below tries to predict the “person” object whose centre (the blue dot) falls within the grid cell. every grid cell predicts a set range of boundary boxes. during this example, the red grid cell makes a boundary box prediction to find wherever the

object. However, the one-object rule limits however shut detected objects are often. For that, model will have some limitations on however shut objects are often. For each grid cell, it predicts B boundary boxes and every box has one box confidence score, then it detects one object solely despite the amount of boxes B, so it predicts C conditional category chances (one per category for the likelihood of the thing class).

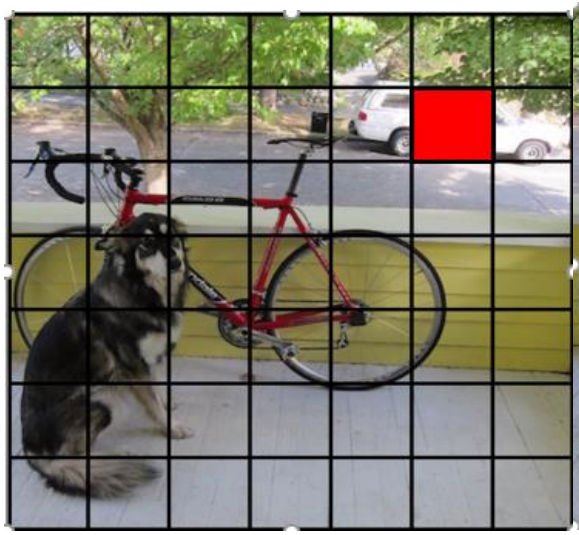


Fig. 10. Each cell predicts boxes and confidence

Fig. 11. Predicted boxes and confidence

Let's get into a lot of details. Every boundary box contains five elements: (x, y, w, h) and a box confidence score. the arrogance score reflects however possible the box contains an object and the way correct is that the boundary box. we tend to normalize the bounding box dimension w and height h by the image dimension and height. x and y square measure offsets to the corresponding cell. Hence, x, y, w and h square measure all between zero and one. every cell has twenty conditional category chances. The conditional category likelihood is that the likelihood that the detected object belongs to a selected category (one likelihood per class for every cell). So, the model's prediction encompasses a form of (S, S, B×5 + C) = (7, 7, 2×5 + 20) = (7, 7, 30). The class confidence score for every prediction box is measured as the confidence on each the classification and the localization (where Associate in Nursing object is located). The rating and chance terms can simply be errored.

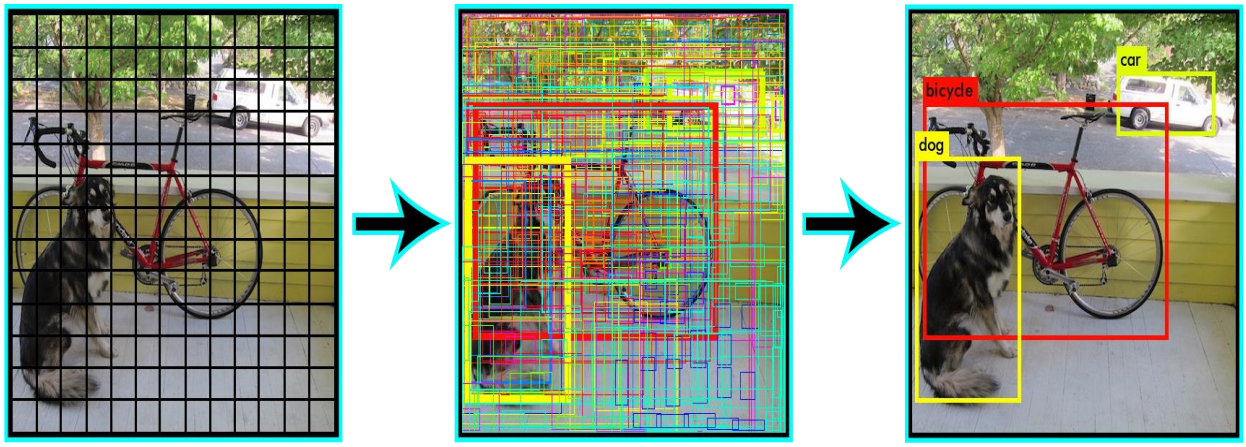


Fig. 12. Stages of Detection

The above stages of detection show that first the grid is split into S x S, the algorithm repeats the process and draws the bounding boxes around the object, when the boxes become darker it is decided to be an object identified at that position. Once the feature extraction is done, then the object is moved to the classifier layer for detection purpose. In the detection layer, it finally matches with the available classes and finds the respective class.

The model can make duplicate detections for the same object. To overcome this issue, non-maximal suppression with lower confidence can be added. Non-maximal suppression can increase the mean average precision by two to three percentage.

The bounding box may be a rectangular box which will be determined by the x and y axis coordinates within the upper-left corner and the x and y axis coordinates within the lower-right corner of the rectangle. we are going to outline the bounding boxes of the dog and the cat within the image supported the coordinate info within the higher than image. The origin of the coordinates within the higher than image is that the higher left corner of the image, and to the proper and down square measure the positive directions of the x axis and the y axis, severally.

5.2 Output of Different version

The model has been built with both frame differencing as well as without frame differencing. On comparing the two version which ran with CPU, it can be easily understood the reason behind the selection of frame differencing algorithm. The process of frame differencing will save a lot of time and amount in terms of deployment.

5.2.1 No Frame Differencing

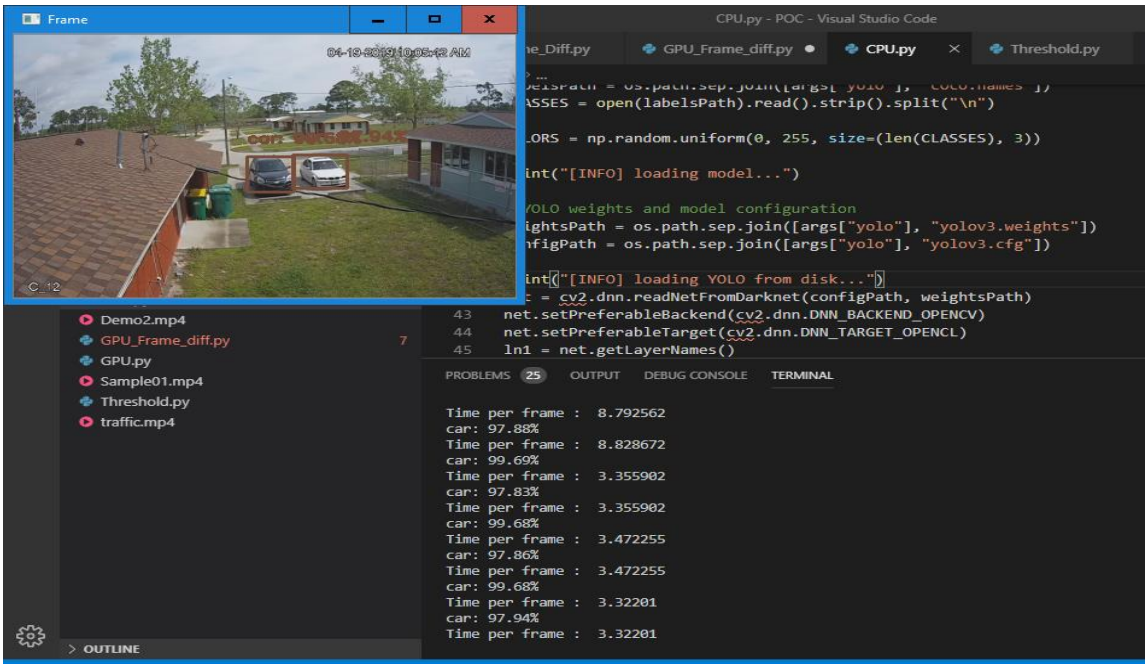


Fig. 13 CPU Version with No Frame Differencing

In fig.13, The code for no frame differencing using CPU model is run and the model has loaded and the detection window appears on top left of the screen, where in terminal window the time taken per frame and the object detected with its confidence score is printed. Since this is a no frame differencing version the time taken per frame is almost three to four microseconds. The difference can now be understood from next screenshot of frame differencing.

5.2.2 Frame differencing

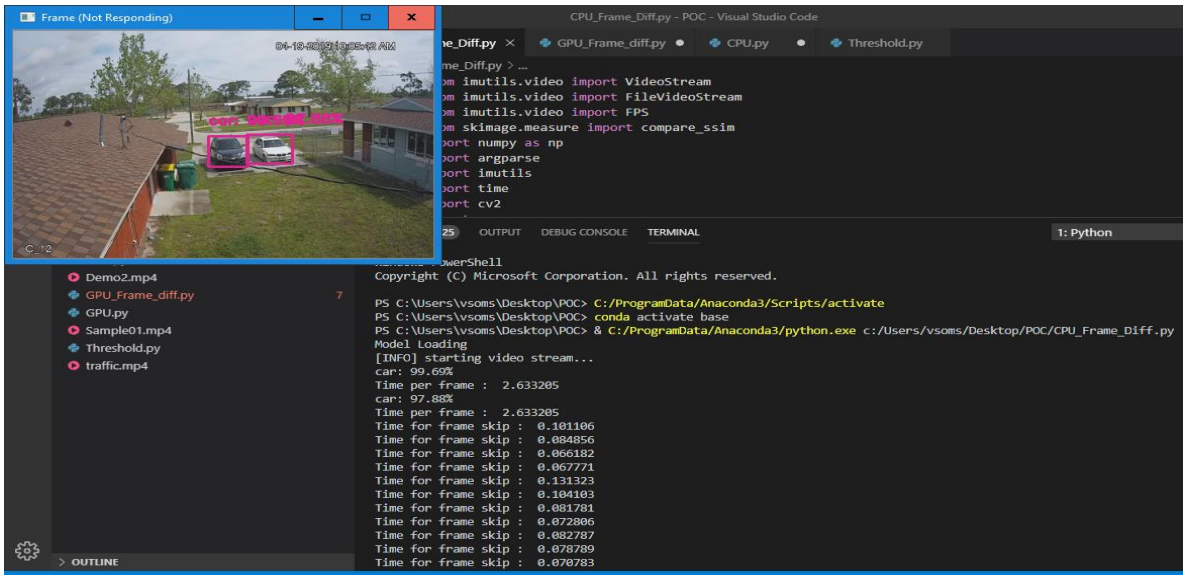


Fig. 14 CPU Version with Frame Differencing – Non-Congested

In Fig 14, It is evident that once the object is detected once and since the video is a non-congested video, there is no motion in the video. This means there will be no pixel change between the frame, this make the model to skip the frame by saving 80 percent of time per frame. This can be seen in the terminal as immediately after detecting two cars, frames are skipped continuously as there is no motion in the video.

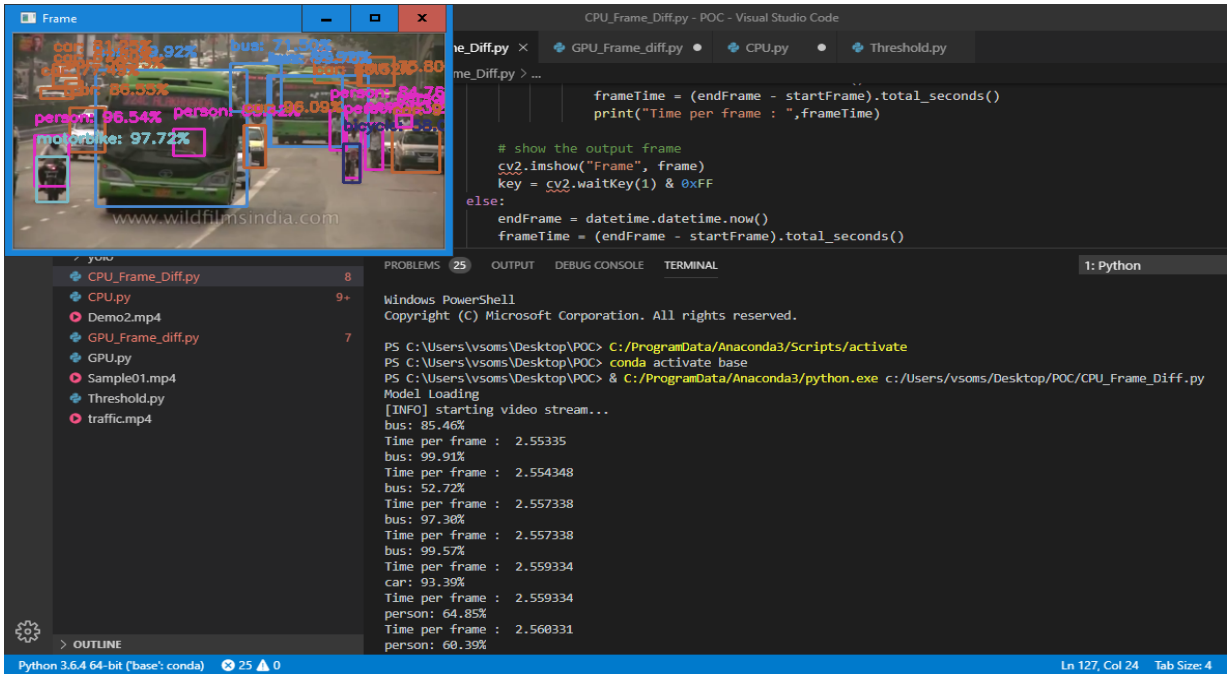


Fig. 15 CPU Version with Frame Differencing - Congested

In Fig 15, It is evident that once the object is detected and since the video is a congested video, there will be a motion or pixel difference between the frames. As expected, the objects are detected consecutively without skipping the frames.

In Fig 16, It is evident that the algorithm works on grey scale as well. The below screenshot shows that the frames are skipped once the objects are detected in the frame, as there is no motion.

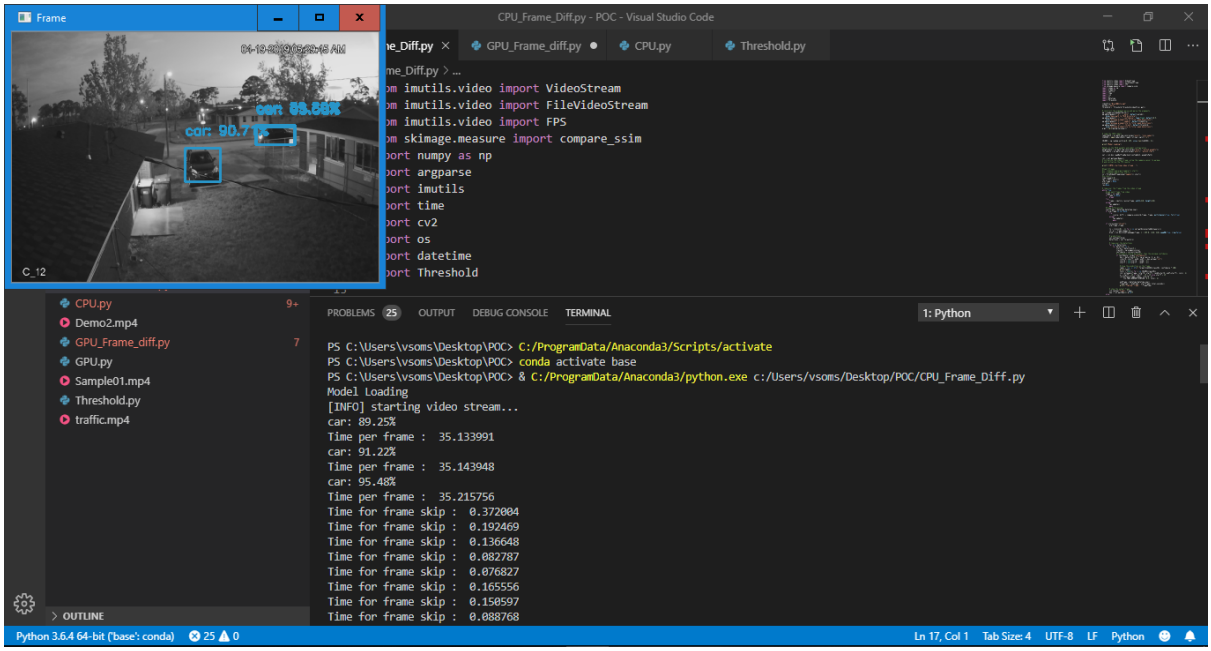


Fig. 16 CPU Version with Frame Differencing – Black and White

5.3 Discussion and Improvisation

To decrease the time even more bounding boxes can be removed, as it will save the processing time as much as it can, even on working with GPU this shows a huge variation.

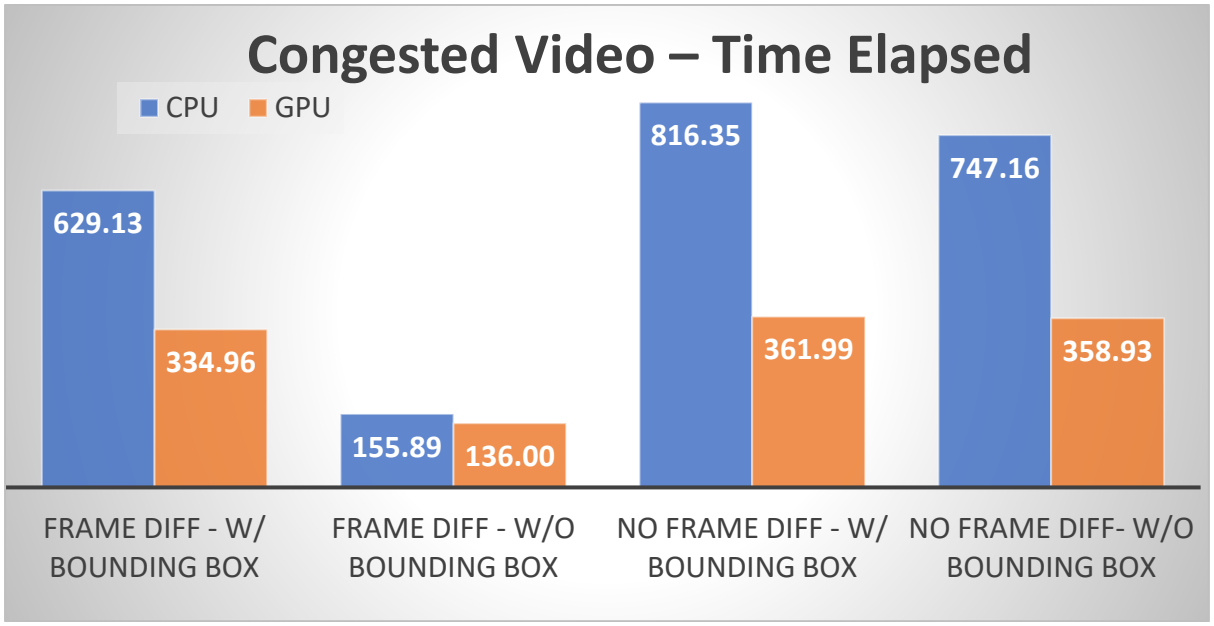


Fig. 17. Time elapsed of Congested video

Frame differencing is the difference between the current frame and the previous frame. Even when there is a slight frame difference in the motion the frame differencing is made. If there is no movement detected, then the current frame is skipped and moves to the next frame. The

analysis is made on traffic video that has lot of objects in a single frame and thus there is a lot of congestion and the performance of the algorithm is tested here.

The above graph distinguishes between the four different methods of time elapse. The Y-Axis indicated the time in millisecond and X-Axis is the different video elapse time.

The different methods are Frame differencing with bonding box, frame differencing without bounding box, No -frame with bounding box and No-Frame without bounding box.

The different bar graphs show the CPU (in blue) and GPU (in orange) performance during the time object detection.

When Frame differencing is taken, The CPU and GPU takes 629.13 milliseconds to form a bounding box, that is to detect the object. The GPU takes 334.96 milliseconds for the same. However, without bounding box the same CPU and GPU takes less time in framing the object. CPU takes around 156 milliseconds and GPU takes around 136 milliseconds to process the data. The performance of CPU and GPU in detecting and object is faster when there is no bounding box.

The next in the graph is about no framing. No framing has two methods, one is no framing without bounding box and No framing with bounding box.

No Frame differencing with bounding box: The CPU takes about 900 milliseconds to process the frame differencing and GPU takes 400milliseconds to process the frame differencing. When No frame differencing without bounding box is taken, The CPU takes 750 milliseconds and GPU takes about 360 milliseconds. The traffic video has many objects in it within a single frame.

On the whole, after analysing the CPU and GPU process time for the four methods to detect the object in a single frame it is easily seen that Fame differencing without bounding box takes very less time and performs the best in detecting an object movement in the frame. However, no frame differencing method in both the cases takes too much time to process and the performance is also not that efficient.

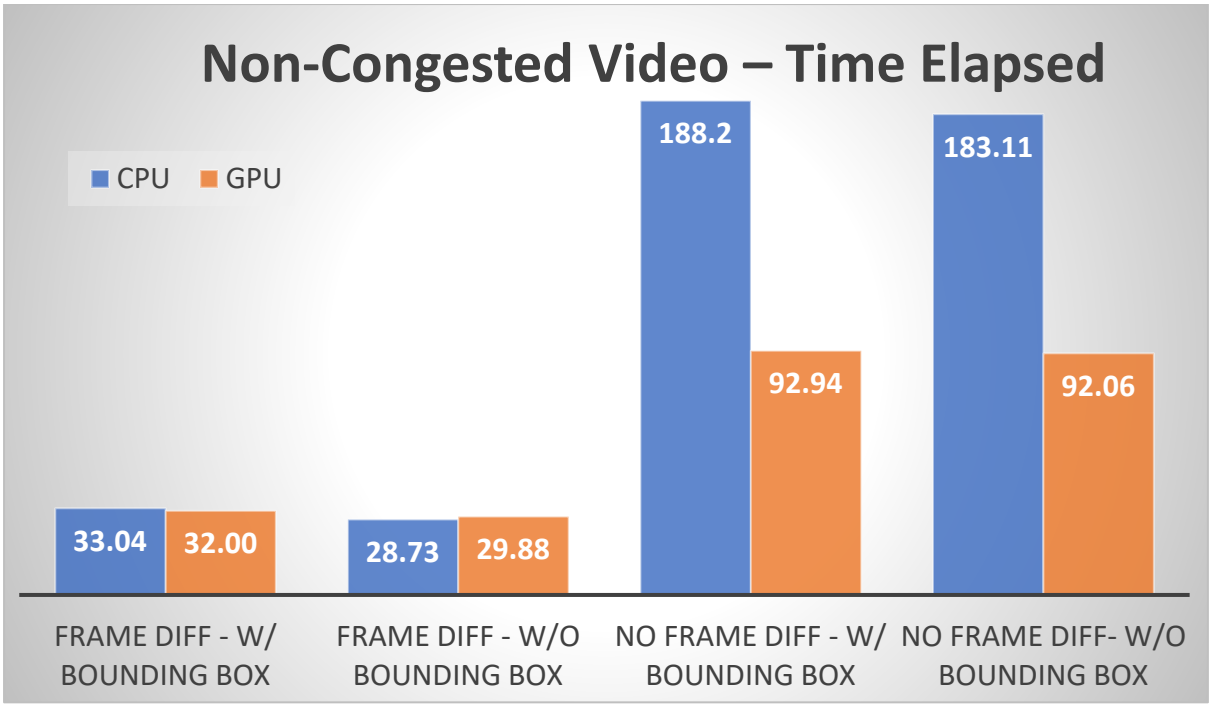


Fig. 18. Time elapsed of Non-Congested video

In the previous graph, the analysis was made on a traffic video that seems to be highly congested and the algorithms performance was tested for the same and analysed frame differencing without bounding box perform well in that case. This is an analysis made on a Demo video that has less objects to be detected. Frame within differencing with bounding box has CPU that detects the object in 33.04 millisecond and GPU takes 32 milliseconds. There is not much difference seen in bars using frame differencing concept, but when there is no frame differencing it shows a huge difference in the bars, which shows that processing the video without bounding boxes will be easier and time efficient, compared to the video with bounding boxes.

5.4 Challenges to be considered

5.4.1 Analysis on Background Displacement

Three image simulated sequences that depict the background relocation scenario, namely Parking, Sofa, and Winter Drive Away. Background displacement strongly depends on the scene under observation. It is quite impractical to set a predefined threshold of time beyond which all stationary foregrounds can be absorbed into the background. On the contrary, it is very tough to strict an absence duration threshold beyond which an existing background class will be removed from the developed model. These two parameters must be varied with respect

to the underlying environment. The scene knowledge along with the information of possible stationary objects to be learned to reduce such false alarms.

5.4.2 Analysis on Bootstrapping

Both Bootstrapping and Wondering Students videos well reflect the bootstrapping scenario. Bootstrapping can be considered as a special case of background relocation wherein the knowledge of possible objects, their size, average halt duration etc, have to be learned over the initialization sequence to remove the faulty background classes from the developed model.

5.4.3 Analysis on Camera Shake

Camera oscillation can be observed in several areas. The oscillation periodicity owing to camera-shake needs to be learned with enough initialization frames.

5.4.4 Analysis on Camouflage

The attire similarity between the foreground and background can be observed in the Camouflage and Curtain sequence. Complementary cues, texture features along with colour cues need to be incorporated to tackle this disguise issue. In addition, the morphological processing and other low pass filtering can be applied as a post improvisation module to minimize the camouflage gap.

5.4.5 Analysis on Gradual Illumination Variation

The varying sunlight illumination, over the time, can be seen in the time of day sequence. Multilayer is the only method that produces acceptable results. Recursive models often fail to tackle such eventual variations because their underlying model parameters are skewed towards the long past data. On the other hand, non-recursive methods efficiently handle the problem at the cost of high memory overhead in terms of a finite buffer at each pixel location.

5.4.6 Analysis on Sudden Illumination Variation

The rapid variation in illumination can be observed in the Lobby and Light Switch sequence. To the best of our knowledge, the literature still lags any immediate fool proof solution to tackle such rapid variation. Such rapid variation completely alters the colour and intensity characteristics of the underlying scene. One time-consuming yet reliable solution is to re-initialize the model as soon as such rapid variation is observed. The usual

background update strategy also adapts the changed pixel values in the model with few successive frames.

5.4.7 Analysis on Shadow

Shadow effect can be visualized in the tall building sequences. Shadow is the scaled down value of illumination. Methods based on RGB or grey colour space miss-classify shadow as foreground. Gradient or texture features along with invariant colour models are suitable candidates to counter this phenomenon.

5.4.8 Analysis on Uninteresting Background Oscillation

Unimodal methods fail to incorporate dynamic background in the model. Multi-modal systems usually assign equal number of classes, and therefore fail in situations, where the waving periodicity differs across the scene. The obvious strategy is to learn sample variation of pixel sequence at each location to determine the oscillation periodicity. Then, a suitable clustering method can distribute the input sequence into the required number of classes.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion:

Object detection and recognition is one in all the difficult and challenging analysis tasks of computer vision aimed toward identification of moving objects from video sequence. Then followed by predicting the trail of moving object for the length of its presence in video frame sequences. This study has provided a comprehensive review of the progressive methods on object detection and following with specialise in soft computing primarily based approaches. The proposed technique of object detection based on frame differencing has been tested on various levels of video frames comprised of complex visuals. The model provides a better output compared to other models as well. The final output is interactive, engaging and robust.

6.2 Future Works:

Object Detection finds scope in various areas such as defence and border security, medical image processing, video surveillance, astronomy and other security related applications. The various object detection algorithms such as skin detection, colour detection, face detection and target detection can be invoked into the model and simulated for detection purposes. Further a single algorithm maybe designed by considering various detection parameters such as Colour, Face, Skin and Target of interest to meet video surveillance applications. With the model, it can be further developed in a way that license plate can be detected using ANPR algorithm. Occlusions result in partial detection in videos because of high density of objects and low angle of camera for observance video sequence. within the experiments, it's been found that once occlusion happens, object following suffers considerably. It shows that planned approaches are often improved victimisation a lot of subtle and strong object following technique. an efficient occlusion handling approach are often targeted for any improvement of results. The reliable detection of shadow could be a difficult task as shadows have same magnitude and movement pattern just like foreground objects. The existence of shadows causes distortion in form of the item, object loss, merging of the object. The projected approach must be improved during this context. Effective handling of problems like dynamic background and surprising object motion can even become subject of future work. Solutions supported deep leaning techniques may be a possible candidate for integration within the planned technique.

REFERENCES

1. El Baf, F., Bouwmans, T. and Vachon, B. (2008) 'Foreground Detection using the Choquet integral', in *WIAMIS 2008 - Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 187–190. doi: 10.1109/WIAMIS.2008.9.
2. Bouwmans, T. (2011) *Recent Advanced Statistical Background Modeling for Foreground Detection-A Systematic Survey*, *Recent Patents on Computer Science*. Bentham Science Publishers. Available at: https://hal.archives-ouvertes.fr/hal-00644746 (Accessed: 25 August 2019).
3. Brutzer, S., Höferlin, B. and Heidemann, G. (2011) 'Evaluation of background subtraction techniques for video surveillance', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, pp. 1937–1944. doi: 10.1109/CVPR.2011.5995508.
4. Calderara, S. *et al.* (2006) 'Reliable background suppression for complex scenes', in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 211–214. doi: 10.1145/1178782.1178814.
5. Cires, D. C. *et al.* (2003) 'Flexible , High Performance Convolutional Neural Networks for Image Classification', pp. 1237–1242.
6. Courbariauxécole, M. and Bengio, Y. (no date) *BinaryConnect: Training Deep Neural Networks with binary weights during propagations*. Available at: https://github.com/MatthieuCourbariaux/BinaryConnect (Accessed: 25 August 2019).
7. Cucchiara, R. *et al.* (2003) 'Detecting moving objects, ghosts, and shadows in video streams', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), pp. 1337–1342. doi: 10.1109/TPAMI.2003.1233909.
8. Dauphin, Y. N. and Bengio, Y. (2013) 'Big Neural Networks Waste Capacity'. Available at: http://arxiv.org/abs/1301.3583 (Accessed: 25 August 2019).
9. Girshick, R. (no date) *Fast R-CNN*. Available at: https://github.com/rbgirshick/ (Accessed: 25 August 2019).
10. Hati, K. K., Sa, P. K. and Majhi, B. (2013) 'Intensity range based background subtraction for effective object detection', *IEEE Signal Processing Letters*, 20(8), pp. 759–762. doi: 10.1109/LSP.2013.2263800.
11. Horé, A. and Ziou, D. (2010) 'Image quality metrics: PSNR vs. SSIM', in *Proceedings - International Conference on Pattern Recognition*, pp. 2366–2369. doi:

10.1109/ICPR.2010.579.

12. Huerta, I. *et al.* (2009) 'Detection and removal of chromatic moving shadows in surveillance scenarios', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1499–1506. doi: 10.1109/ICCV.2009.5459280.
13. *IEEE Xplore: IEEE Transactions on Pattern Analysis and Machine Intelligence* (no date). Available at: https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=34 (Accessed: 25 August 2019).
14. *INTERSPEECH 2011 Abstract: Seide et al.* (2011). Available at: https://www.isca-speech.org/archive/interspeech_2011/i11_0437.html (Accessed: 25 August 2019).
15. Ioffe, S. and Christian Szegedy (2015) 'Batch Normalization: Accelerating Deep Network Training by Reducing', *Journal of Molecular Structure*, 1134, pp. 63–66. doi: 10.1016/j.molstruc.2016.12.061.
16. Jodoin, P. M., Saligrama, V. and Konrad, J. (2009) 'Behavior Subtraction'. Available at: http://arxiv.org/abs/0910.2917 (Accessed: 26 August 2019).
17. Krizhevsky, A., Sutskever, I. and Hinton, G. E. (no date) *ImageNet Classification with Deep Convolutional Neural Networks*. Available at: http://code.google.com/p/cuda-convnet/ (Accessed: 25 August 2019).
18. Lai, A. H. S. and Yung, N. H. C. (no date) 'A fast and accurate scoreboard algorithm for estimating stationary backgrounds in an image sequence', in *ISCAS '98. Proceedings of the 1998 IEEE International Symposium on Circuits and Systems (Cat. No.98CH36187)*. IEEE, pp. 241–244. doi: 10.1109/ISCAS.1998.698804.
19. Lecun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, pp. 436–444. doi: 10.1038/nature14539.
20. Lo, B. P. L. and Velastin, S. A. (no date) 'Automatic congestion detection system for underground platforms', in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*. IEEE, pp. 158–161. doi: 10.1109/ISIMP.2001.925356.
21. Nowozin, S. and Lampert, C. H. (2010) 'Structured learning and prediction in computer vision', *Foundations and Trends in Computer Graphics and Vision*, 6(3–4), pp. 185–365. doi: 10.1561/0600000033.
22. Redmon, J. and Angelova, A. (2014) 'Real-Time Grasp Detection Using Convolutional Neural Networks'. Available at: http://arxiv.org/abs/1412.3128 (Accessed: 25 August 2019).
23. Santosh, D. H. and Mohan, P. G. K. (2015) 'Multiple objects tracking using Extended

Kalman Filter, GMM and Mean Shift Algorithm-A comparative study', in *Proceedings of 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT 2014*. Institute of Electrical and Electronics Engineers Inc., pp. 1484–1488. doi: 10.1109/ICACCCT.2014.7019350.

24. Saxena, A., Driemeyer, J. and Ng, A. Y. (2008) 'Robotic grasping of novel objects using vision', in *International Journal of Robotics Research*, pp. 157–173. doi: 10.1177/0278364907087172.
25. Schmidhuber, J. (2015) 'Deep Learning in neural networks: An overview', *Neural Networks*. Elsevier Ltd, pp. 85–117. doi: 10.1016/j.neunet.2014.09.003.
26. Sermanet, P. *et al.* (2013) 'OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks'. Available at: http://arxiv.org/abs/1312.6229 (Accessed: 25 August 2019).
27. Smail Avcıbas¸uludag, I. ˙ and Sayood, K. (2002) 'Statistical evaluation of image quality measures'. doi: 10.1117/1.1455011.
28. Stauffer, C. and Grimson, W. E. L. (2000) 'Learning patterns of activity using real-time tracking', *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 22(8), pp. 747–757. doi: 10.1109/34.868677.
29. Stauffer, C. and Grimson, W. E. L. (2016) '2016-06-24 Stipulation and Order (TO LENGTHEN THE PAGE LIMIT OF THE OPPOSITION AND REPLY OF THE MOTION TO STRIKE PURSUANT TO C.C.P. SEC. 425.16, BY FIVE PAGES; ORDER)', *Bc606667*, pp. 246–252. doi: 10.1109/CVPR.1999.784637.
30. Wang, Z. *et al.* (2004) *Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE TRANSACTIONS ON IMAGE PROCESSING*. Available at: http://www.cns.nyu.edu/~lcv/ssim/. (Accessed: 26 August 2019).
31. Wright, J. *et al.* (2010) 'Sparse representation for computer vision and pattern recognition', *Proceedings of the IEEE*, 98(6), pp. 1031–1044. doi: 10.1109/JPROC.2010.2044470.