Journal of Theoretical and Applied Information Technology
31st May 2022. Vol.100. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EMPIRICAL INVESTIGATIONS TO OBJECT DETECTION IN VIDEO USING RESNET-AN IMPLEMENTATION METHOD

[1] Y SUREKHA, [2] DR K KOTESWARA RAO, [3] DR G LALITHA KUMARI, [4] N RAMESH BABU, [5]Y. SAROJA

Assistant Professor, Associate Professor, Sr. Asst. Professor
Department of Computer Science and Engineering,
PVP Siddhartha Institute Of Institute Of Technology, Vijayawada, India
Assistant Professor, RGKUT, Srikakulam, Andhra Pradesh, INDIA
Assistant Professor, Mallareddy College Of Engineering For Women, Hyderabad, India
E-mail: Yalamanchili.surekha@gmail.com, koteswara2003@yahoo.co.in, lalithajoy.nuthakki@gmail.com, ramesh.nuthakki@gmaill.com, sarojaphani@gmail.com

ABSTRACT

Real-time object detection and tracking is a vast, vibrant yet inconclusive and complex area of computer vision. Due to its increased utilization in surveillance, tracking system used in security and many others applications have propelled researchers to continuously devise more efficient and competitive algorithms. However, problems emerge in implementing object detection and tracking in real-time; such as tracking under dynamic environment, expensive computation to fit the real-time performance, or multi-camera multi-objects tracking make this task strenuously difficult. Though, many methods and techniques have been developed, but in this literature review we have discussed some famous and basic methods of object detection and tracking. In the end we have also given their general applications and results.

Keywords: *Image-Processing, Deep Learning, Object Detection, Object Recognition*

1. INTRODUCTION

Object recognition is a computer vision technique for identifying objects in images or videos. Object recognition is a key output of deep learning and machine learning algorithms. When humans look at a photograph or watch a video, we can readily spot people, objects, scenes, and visual details. The goal is to teach a computer to do what comes naturally to humans: to gain a level of understanding of what an image contains. Object recognition is a key technology behind driverless cars, enabling them to recognize a stop sign or to distinguish a pedestrian from a lamppost. It is also useful in a variety of applications such as disease identification in bio imaging, industrial inspection, and robotic vision. Object detection and object recognition are similar techniques for identifying objects, but they vary in their execution. Object detection is the process of finding instances of objects in images. In the case of deep learning, object detection is a subset of object recognition, where the object is not only identified but also located in an image. This allows for multiple objects to be identified and located within the same image. You can use a variety of approaches for object recognition. Recently, techniques in machine learning and deep learning have become popular approaches to object recognition problems. Both techniques learn to identify objects in images, but they differ in their execution.

1.1. Residual Neural Network:

A residual neural network (ResNet) is an artificial neural network (ANN) of a kind that builds on constructs known from pyramidal cells in the cerebral cortex. Residual neural networks do this by utilizing skip connections, or shortcuts to jump over some layers. Typical ResNet models are implemented with double- or triple- layer skips that contain nonlinearities (ReLU) and batch normalization in between. An additional weight matrix may be used to learn the skip weights; these models are known as Highway Nets. Models with several

Journal of Theoretical and Applied Information Technology
31st May 2022. Vol.100. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

parallel skips are referred to as DenseNets. In the context of residual neural networks, a non-residual network may be described as a plain network. The structure of a simple cell is shown Fig. 1.1.

One motivation for skipping over layers is to avoid the problem of vanishing gradients, by reusing activations from a previous layer until the adjacent layer learns its weights. During training, the weights adapt to mute the upstream layer, and amplify the previously-skipped layer. In the simplest case, only the weights for the adjacent layer's connection are adapted, with no explicit weights for the upstream layer. This works best when a single nonlinear layer is stepped over, or when the intermediate layers are all linear. If not, then an explicit weight matrix should be learned for the skipped connection (a *HighwayNet* should be used).Skipping effectively simplifies the network, using fewer layers in the initial training stages. This speeds learning by reducing the impact of vanishing gradients, as there are fewer layers to propagate through. The network then gradually restores the skipped layers as it learns the feature space. Towards the end of training, when all layers are expanded, it stays closer to the manifold and thus learns faster. A neural network without residual parts explores more of the feature space. This makes it more vulnerable to perturbations that cause it to leave the manifold, and necessitates extra training data to recover.

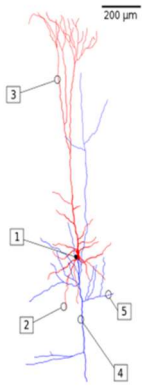

Figure 1.1: A reconstruction of a pyramidal cell. Soma and dendrites are labelled in red, axon arbour in blue.
(1) Soma, (2) Basal dendrite, (3) Apical dendrite, (4) Axon, (5) Collateral axon.

1.2 Instance Segmentation:

There are various techniques that are used in computer vision tasks. Some of them include classification, semantic segmentation, object detection, and instance segmentation. Classification tells us that the image belongs to a particular class. It doesn't consider the detailed pixel level structure of the image. It consists of making a prediction for a whole input. Semantic segmentation makes dense predictions inferring labels for each pixel so that every pixel in the image is labelled with the class of its enclosing object. Object detection provides not only the classes but also indicate the spatial location of those classes. It takes into account the overlapping of objects. Instance segmentation includes identification of boundaries of the objects at the detailed pixel level. For example, in the image below i.e Fig. 1.2, there are 7 balloons at certain locations, and these are the pixels that belong to each one of the balloons.

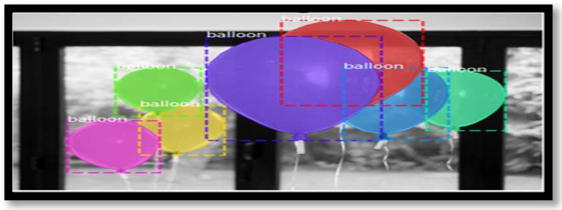

Figure 1.2: Image Segmentation example

1.3 Region Based Convolution Neural Network (R-CNN)

R-CNN is widely used in solving the problem of object detection. It creates a boundary around every object that is present in the given image. It can be done in two steps: region proposal step and the classification step. Classification step consists of extraction of feature vectors and set of linear SVMs.

To solve the problem of selecting a huge number of regions, a selective search is used to extract just 2000 regions from the image and this is known as region proposals. Therefore, instead of trying to classify a large number of regions, we can just work with 2000 regions. The selective search algorithm can be performed in the following steps:

1. Generate initial sub-segmentation (many candidate regions)
2. Use a greedy algorithm to recursively combine similar regions

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. Use generated regions to produce the final region proposals

These proposed regions are then fed into the convolutional neural network and produces a 4096-dimensional feature vector as output. CNN extracts a feature vector for each region which is then used as an input to the set of SVMs that outputs a class label. The algorithm also predicts four offset values to increase the precision of the bounding box.

The main problem with R-CNN is that it still requires a large amount of time to train and thus cannot be implemented for real-time problems.

1.4 Mask R-CNN:

Mask R-CNN is an instance segmentation technique which locates each pixel of every object in the image instead of the bounding boxes. It has two stages: region proposals and then classifying the proposals and generating bounding boxes and masks. It does so by using an additional fully convolutional network on top of a CNN based feature map with input as feature map and gives matrix with 1 on all locations where the pixel belongs to the object and 0 elsewhere as the output.

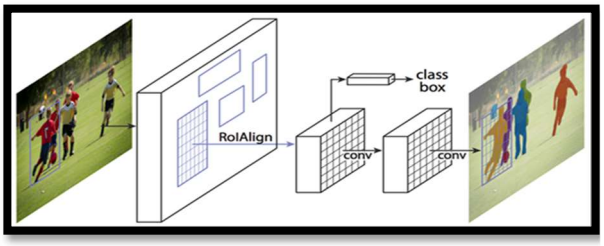

Figure 1.3: Mask R-CNN

It consists of a **backbone network** which is a standard CNN such as ResNet50 or ResNet101. The early layer of network detects low-level features, and later layers detect higher-level features. The image is converted from 1024x1024px x 3 (RGB) to a feature map of shape 32x32x2048. The **Feature Pyramid Network (FPN)** was an extension of the backbone network which can better represent objects at multiple scales. It consists of two pyramids where the second pyramid receives the high-level features from the first pyramid and passes them to the lower layers.

This allows every level to have access to both lower and higher-level-features. It also uses the **Region Proposal Network (RPN)** which scans all FPN top to bottom and proposes regions which may contain objects. It uses anchors which are a set of boxes with predefined locations and scales itself according to the input images. Individual anchors are assigned to the ground-truth classes and bounding boxes. RPN generates two outputs for each anchor — anchor class and bounding box specifications. The anchor class is either foreground class or a background class.

Another module that is different in Mask R-CNN is the **ROI Pooling**. The authors of Mask R-CNN concluded that the regions of the feature map selected by RoIPool were slightly misaligned from the regions of the original image. Since image segmentation requires specificity at the pixel level of the image, this leads to inaccuracies. This problem was solved by using **RoIAlign** in which the feature map is sampled at different points and then a bilinear interpolation is applied to get a precise idea of what would be at pixel 2.93 (which was earlier considered as pixel 2 by the RoIPool).

Then a convolutional network is used which takes the regions selected by the ROI classifier and generates masks for them. The generated masks are of low resolution- 28x28 pixels. During training, the masks are scaled down to 28x28 to compute the loss, and during inferencing, the predicted masks are scaled up to the size of the ROI bounding box. This gives us the final masks for every object.

2. LITERATURE SURVEY:

Cruz, Jerome Paul N., et al. used Artificial Neural Networks in evaluating a frame shot of the target image. The system utilizes three major steps in object recognition, namely image processing, ANN processing and interpretation. With an optimum distance for recognition at 40cm achieved an accuracy of 99.99996072%. [1]

Ren, Xiaofeng, and Chunhui Gu., et al. computed dense optical flow and fit it into multiple affine layers. Then used a max-margin classifier to combine motion with empirical knowledge of object location and background movement as well as temporal cues of support region and color appearance. Evaluated their segmentation algorithm on the large Intel Egocentric Object Recognition dataset with 42 objects and 100K frames. They show that, when combined with temporal integration, figure-ground segmentation improves the accuracy of a SIFT-based recognition system from 33% to 60%, and that of a latent-HOG system from 64% to 86%. This is a serious piece of work that inspires everyone. [2]

Koubaroulis, Dimitri, Jiri Matas, and Josef Kittler., et al. proposed a colour-based object

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

recognition method for video annotation. A colour-based method, the multimodal neighborhood signature (MNS) is used. Despite the poor quality of some of the images and a wide range of appearance variations object recognition and sport classification was achieved for a set of four selected objects/sports [3].

Benxian, Xiao, et al. proposed a method of moving object detection and recognition based-on the frame difference algorithm and moment invariant features. In object recognition algorithm, moment invariant features were extracted from moving object region firstly, and vector standardization was done for these moment invariant features, then wavelet neural network with genetic algorithm was used as pattern recognition and automatic recognition. Results show that the proposed approach is a fast and effective method for moving object detection and recognition. [4].

Foresti, Gian Luca, et al. stated the statistical morphological skeleton, which achieves low computational complexity, accuracy of localization, and noise robustness has been considered for both object recognition and tracking. Recognition is obtained by comparing an analytical approximation of the skeleton function extracted from the analysed image with that obtained from model objects stored into a database. Tracking is performed by applying an extended Kalman filter to a set of observable quantities derived from the detected skeleton and other geometric characteristics of the moving object[5].

Mazumdar, Meghajit, V. Sarasvathi, and Akshay Kumar, et al. . A sequential frame extraction method of videos and also deep learning approach of Convolutional Neural Networks along with Fully Connected Neural Networks is used for this task. The method gives good accuracy of average 77 percent. [6]

Ju, Ting-Fung, et al.. This paper presents a vision-based moving objects detection work which attracts much attention in intelligent automobile applications recently. Vision-based objects detection provides object behaviour information of objects and is an intuitive detection method similar to human visual perception. Accordingly, this paper presents a robustness enhancing method for vision-based moving objects detection. [7]

Rakumthong, Waritchana, et al.. This paper proposes a new design and implementation method in supporting a smart surveillance system that can automatically detect abandoned and stolen objects in public places such as bus stations, train stations or airports. The correctness of object classification is approximately 76%, and the correctness of event classification 83%. [8]

K He, Kaiming, et al. explained Deeper neural networks are more difficult to train. They present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. They explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. They provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers---8x deeper than VGG nets but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. They also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, they obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions, where they also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation. This was our main inspiration to start out framework using ResNet.[[9]

Lowe, David G, et al. Stated an object recognition system has been developed that uses a new class of local image features. These features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. These features share similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. Features are efficiently detected through a staged filtering approach that identifies stable points in scale space. Image keys are created that allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. The keys are used as input to a nearest neighbour indexing method that identifies candidate object matches. Final verification of each match is achieved by finding a low residual least squares solution for the unknown model parameters. Experimental results show that robust

object recognition can be achieved in cluttered partially occluded images with a computation time of under 2 seconds [10].

Raghunandan, Apoorva, Pakala Raghav, and HV Ravish Aradhya, et al. discussed various Object Detection Algorithms such as face detection, skin detection, colour detection, shape detection, target detection is simulated and implemented using MATLAB 2017b to detect various types of objects for video surveillance applications with improved accuracy. [11].

Tripathi, Rajesh Kumar, Anand Singh Jalal, and Charul Bhatnagar, et al elaborated A Strong Background subtraction technique followed by contour evaluation to classify everything into living & non-living things and these non-living objects are tracked. [12]

Experimental results show that proposed system is efficient and effective for real-time video surveillance, which is tested on IEEE Performance Evaluation of Tracking and Surveillance data set (PETS 2006, PETS 2007) and our own dataset.

3. MOTIVATION:

Our approach towards Object Detection & Recognition is quite different as our aim is to create a hybrid or a cross-integrated deep-learning artificial intelligence neural network model which uses pre-trained weights or correlations obtained by training a dataset of images relevant to the application was being used to detect and recognize the objects exist in the image.

While the object detection and recognition are just processing which are involved as a major part in creating a platform that allows us to detect and recognize objects in a video. As there exists a lot more algorithms already in use our aim is to integrate a best object recognition algorithm (i.e ResNet) for video processing applications.

ImageNet classification challenge is a global event that was used to be held every year until 2015. The main purpose of this challenge is to obtain the best efficient algorithm for image classification challenge for real-world applications. For this challenge a sub-set of the ImageNet dataset is made available to the developers who participates in this challenge.

The developers use this dataset to train and test the model they have developed for object recognition. Mostly the participants are MNC's and private developers. The Resnet algorithm is the winner of ImageNet classification challenge 2015 developed by a group of engineers at Microsoft where this model is trained with approximately 32,000 categories of object images and is used for completely research purpose and large-scale applications.

This model requires a relatively more training time depending on the size of dataset with which we are training. It also requires more predicting time as there are up to 32,000 categories of classifiable objects. But this model can be optimized for a regular use small scale application which is what we have done.

4. PROBLEM STATEMENT

The traditional approach for object detection is by using contours or pre-defined or user-defined bounding boxes where these contours which determine the boundaries of the objects in the images using the edges present in the image may be inappropriate when there are objects either adjacent or overlapping to one another. And for the object recognition the most widely used neural network for detecting commonly used objects is YOLO (You Only Look Once) algorithm which uses a neural network model namely Darknet where the images which consists of objects along with bounding boxes around the objects with labels are used to train the Darknet model which gives it higher efficiency in detecting and recognizing objects using a single neural network model. The main drawback of YOLO is it cannot be trained as per the needs of the user but a pre-trained weight must be used to predict the images. The Darknet has a limit of recognizing only 90 categories of objects only which are most commonly found objects in our daily life surroundings. The YOLO is one of the fastest and efficient techniques for object detection which requires relatively less training too.

For this reason, the YOLO is only used for small scale applications and is not appropriate for large scale applications. Even in the small-scale applications it results in less accurate results and cannot classify non-trained categories. Hence as an alternative we developed a small-scale utilizable version of a large-scale algorithm with a bit faster speed. The algorithm we selected for this purpose is Resnet as it was the most efficient algorithm in object detection and recognition up to now.

5. DESIGN AND TECHNICAL DESCRIPTION

Up to now we have only discussed about what are we doing and why are we doing? As we exactly know the methods or

Journal of Theoretical and Applied Information Technology
31st May 2022. Vol.100. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

modules to implement, let us now discuss about the design and implementation of these modules. Our main part here is to integrate several modules having different inputs & outputs which are independent of each other but has to work in a serial manner to get the work done. At first comes the image processing part where we create the platform to open & read images or videos and then process them for object detection & recognition and finally save the output images to output/results.

As per the input & output we have several ways & several libraries to read and save image/video files. Here we used the methods of OpenCV library to read & write images. The methods cv2.imread() & cv2.imwrite() allows us to access the image files as numpy arrays either 3D/4D arrays basing on the file is either an image/video on which we can easy pre-processing using OpenCV library methods and easily perform operations like cropping, filtering & modifying image at greater speeds.

```
img1 = cv2.imread(pic_name)
cap = cv2.VideoCapture(dir_name)
    while(cap.isOpened()):
        ret, frame = cap.read()
        if not ret:
            break
```

Once the image/video frame is loaded into a numpy array we perform object detection by processing the image/frame into an instance segmentation model which uses a pre-trained neural network namely Mask R-CNN to identify the regions of proposals where the probability of existence of an object is relatively high and these object are merged according to similarity to propose final regions which are the passed to Convolution layer which finds the necessary feature maps/vectors by which then the neural network decides the proposed region is valid or not. And those regions that were filtered by the network were returned along with the co-ordinates of the region and the accuracies/confidence values of the object in that region. We are using the sub-module dnn from the module OpenCV to instantiate a neural network precisely the Mask R-CNN model by creating and loading the necessary weights into the function/model.

Those regions that have an object with more than an optimum confidence were cropped out of the original image/frame and these regions were passed to the object recognition where first we have to pre-process the region array/image into the format that was accepted by the ResNet and these regions were processed through the network and returns the prediction labels and confidences of the objects it determines.

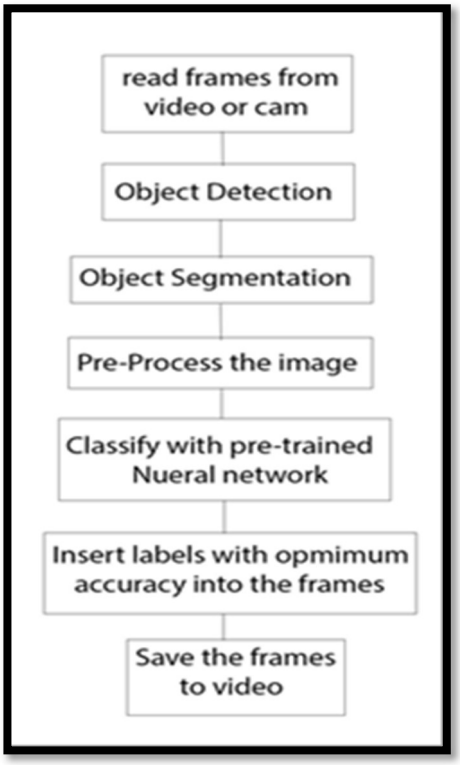

Figure 6.1: Modules & Flow-Chart of framework

For those regions where the object recognition confidence is greater than the optimum value those regions are highlighted by drawing rectangles and by keeping the text of the label name on the rectangle of that specific region. We used the ResNet API developed by the Microsoft to construct and use the ResNet model. All these additional operations were manually created as per the process using OpenCV in-built functions and some by raw code.

In technical terms these rectangles are termed as Bounding Boxes and the region objects are termed as Anchors and these were very essential for the object detection and recognition.

Once the images/frames are drawn with these bounding boxes and labels these were save as new images/videos for result/output. Based on these output accuracies we compute the overall efficiency of the framework or the platform.

6. PROBLEM METHODOLOGY AND SOLUTION

Our approach towards Object Detection & Recognition is quite different as our aim is to create a hybrid or a cross-integrated deep-learning

artificial intelligence neural network model which uses pre-trained weights or correlations obtained by training a dataset of images relevant to the application used here to detect and recognize the objects exist in the image. While the object detection and recognition are just processing which are involved as a major part in creating a platform that allows us to detect and recognize objects in a video. As there exists a lot more algorithms already in use our aim is to create a network which had a unique application.

As an alternative for YOLO which had its drawback, we developed a small-scale utilizable version of a large-scale algorithm with a bit faster speed. The algorithm we selected for this purpose is Resnet as it was the most efficient algorithm in object detection and recognition up to now.

Resnet is custom trainable neural network model which can be trained by any image dataset with necessary format. ImageNet classification challenge is a global event that was used to be held every year until 2015. The main purpose of this challenge is to obtain the best efficient algorithm for image classification challenge for real-world applications. For this challenge a sub-set of the ImageNet dataset is made available to the developer who participates in this challenge.

The developers use this dataset to train and test the model they have developed for object recognition. Mostly the participants are MNC's and private developers. The Resnet algorithm is the winner of ImageNet classification challenge 2015 developed by a group of engineers at Microsoft where this model is trained with approximately 32,000 categories of object images and is used for completely research purpose and large-scale applications.

This model requires a relatively more training time depending on the size of dataset with which we are training. It also requires more predicting time as there are up to 32,000 categories of classifiable objects. But this model can be optimized for a regular use small scale application which is what we have done. The traditional approach for object detection is by using contours or pre-defined or user-defined bounding boxes where these contours which determine the boundaries of the objects in the images using the edges present in the image may be inappropriate when there are objects either adjacent or overlapping to one another. In order to avoid such cases, we have used a different approach called as connected Objects() which can provide bounding boxes for adjacent/overlapping objects in the image up to a good level of accuracy.

But beyond that there exists the revolutionary segmentation techniques (semantic segmentation & Instance segmentation) which are run by again some neural network models which are basically designed for imaging in medical purposes and later on used for detection applications due to its high precision and accuracy in detection.

Hence, we have developed a small-scale version of a large-scale algorithm which can applied for some specific applications.

7. RESULTS & ANALYSIS

7.1 Independent Images:

These are the output resultant images when single images are processed through our network. Fig. 8.1.1, Fig. 8.1.2, Fig 8.1.3, and Fig. 8.1.4 consists of the bounding boxes and labels of the objects identified in the process of object recognition.

1. Here the project manager can login into system, by entering the user name and password.

Figure 8.1.2: A Man wearing Jeans
Figure8.1.2: A wallet

Figure8.1.3: A Mouse
Figure 8.1.4: An Analog Watch

8.2 Images from a Video:

These are the images/frames extracted from the output video file generated by the object detection & recognition framework when a video is processed through it. The images Fig. 8.2.1, Fig. 8.2.2, Fig 8.2.3, Fig 8.2.4 depict the output frames extracted from the output video at random time instances and in each of the image we can observe

the bounding boxes and labels around almost all the object that can be detected and recognized.

Figure 8.2.1: A frame from the video: Time-lapse of Hot Air Balloons by NATGEO

Figure 8.2.2: A frame from the video: Time-lapse of Hot Air Balloons by NATGEO

Figure 8.2.3: A frame from the video: Time-lapse of Hot Air Balloons by NATGEO

Figure 8.2.4: A frame from the video: Time-lapse of Hot Air Balloons by NATGEO

In the table (i.e Table 8.2.1) below these four CNNs are sorted w.r.t their top-5 accuracy on ImageNet data-set. The number of trainable parameters and the Floating-Point Operations (FLOP) required for a forward pass can also be seen.

A number of comparisons can be drawn:

- AlexNet and ResNet-152, both have about 60M parameters but there is about 10% difference in their top-5 accuracy. But training a ResNet-152 requires a lot of computations (about 10 times more than that of AlexNet) which means more training time and energy required.
- VGGNet not only has a higher number of parameters and FLOP as compared to ResNet-152, but also has a decreased accuracy. It takes more time to train a VGGNet with a reduced accuracy.
- Training an AlexNet takes about the same time as training Inception. The memory requirements is 10 times less with an improved accuracy (about 9%)

8. CONCLUSION & FUTURE WORK

We have successfully built and integrated the platform for object detection and recognition using ResNet. It also gave interesting results and a rather efficient accuracy. We have used a video of a Time Lapse of Hot Air Balloons by natgeo as a testing video input and our platform went very well with an average detection rate of 84% and an average recognition rate of 100% accuracy. Hence the large-scale algorithms can be integrated with small-scale applications with cost of some space and time complexities but can make wonders with it. By observing the results above, we can conclude that not only in the programming perspective but in the perspective of human recognition system the framework went very well and given detection of all possible objects in an image. Few objects cannot be detected as they may be either not completely in the image (partial appearance), too small to detect, blurred shape, not focused and many more reasons. By comparing the error rate and detection accuracy we are saying that we have successfully developed a framework of object detection and recognition system using a large scale neural network.

REFERENCES:

[1] Cruz, Jerome Paul N., et al. "Object recognition and detection by shape and color pattern recognition utilizing Artificial Neural Networks." 2013 International Conference of Information and Communication Technology (ICoICT). IEEE, 2013.

[2] Ren, Xiaofeng, and Chunhui Gu. "Figure-ground segmentation improves handled object recognition in egocentric video." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] Koubaroulis, Dimitri, Jiri Matas, and Josef Kittler. "Colour-based object recognition for video annotation." Object recognition supported by user interaction for service robots. Vol. 2. IEEE, 2002.
[4] Benxian, Xiao, et al. "Moving object detection and recognition based on the frame difference algorithm and moment invariant features." 2008 27th Chinese Control Conference. IEEE, 2008.
[5] Foresti, Gian Luca. "Object recognition and tracking for remote video surveillance." IEEE Transactions on circuits and systems for video technology 9.7 (1999): 1045-1062.
[6] Mazumdar, Meghajit, V. Sarasvathi, and Akshay Kumar. "Object recognition in videos by sequential frame extraction using convolutional neural networks and fully connected neural networks." 2017 International conference on energy, communication, data analytics and soft computing (ICECDS). IEEE, 2017.
[7] Ju, Ting-Fung, et al. "Vision-based moving objects detection for intelligent automobiles and a robustness enhancing method." 2014 IEEE International Conference on Consumer Electronics-Taiwan. IEEE, 2014.
[8] Rakumthong, Waritchana, et al. "Unattended and stolen object detection based on relocating of existing object." 2014 Third ICT International Student Project Conference (ICT-ISPC). IEEE, 2014.
[9] K He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[10] Lowe, David G. "Object recognition from local scale-invariant features." Proceedings of the seventh IEEE international conference on computer vision. Vol. 2. IEEE, 1999.
[11] Raghunandan, Apoorva, Pakala Raghav, and HV Ravish Aradhya. "Object detection algorithms for video surveillance applications." 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2018.
[12] Tripathi, Rajesh Kumar, Anand Singh Jalal, and Charul Bhatnagar. "A framework for abandoned object detection from video surveillance." 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). IEEE, 2013