Anbarasi MP, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 12, Issue 4, (Series-II) April 2022, pp. 11-19

RESEARCH ARTICLE OPEN ACCESS

Multiple Object Detection and Tracking

Anbarasi MP
Robotics and Automation Engineering
PSG College of Technology, Coimbatore

Aarthi A
Robotics and Automation Engineering
PSG College of Technology, Coimbatore

Manju M
Robotics and Automation Engineering
PSG College of Technology, Coimbatore

ABSTRACT
Object detection and tracking is one of the most common and demanding tasks that surveillance systems need to perform in order to detect meaningful events and suspicious activity and automatically comment and retrieve video content. The reason object detection and tracking are grouped is that object detection can be considered the basis of object tracking, and everyone needs to choose the right features and train for effective classification. Object detection and tracking is one of the key areas of research due to routine changes in object movement, scene size changes, occlusions, appearance changes, and lighting changes. This is relevant for many real-time applications such as vehicle perception and video surveillance. Tracking is performed in terms of object movement and appearance to overcome cognitive problems. Object recognition is one of the most basic and central tasks of computer vision. Its task is to find all the objects of interest in the image and determine the categories and positions of the objects. Object recognition is widely used, has high practical value, and has high research prospects. Object Tracking is a deep learning application that takes an initial set of object detections, creates a unique identifier for each initial detection, and tracks the objects detected as they move within a frame in the video. Algorithms used for implementation of above concepts are CNN, RCNN, Yolo and deep sort methods.

Keywords—CNN, RCNN, Yolo Architecture, Deep sort methods

Date of Submission: 04-04-2022 Date of Acceptance: 19-04-2022

I. INTRODUCTION

This paper provides an overview of object detection and tracking. Machinevision technique that identifies and locates objects in an image file. Object detection can be used to count objects in a spot and determine and track them using this type of recognition and locating.If only one object was present in the video series, tracking an object may be as simple as detecting it in each frame and computing its movement. Convolutional neural network, Region-based convolutional neural network, and Yolo are the three techniques used to train the models [11-15]. Convolutional neural networks (CNNs) are deep neural networks that are used for image analysis, language processing, and other tasks, while an RCNN is utilized for computer vision tasks, namely multiple object detection. YOLO is a real-time object recognition technique that employs neural networks. The goal of thisprojectis todetectandtrack multipleobjectsataninstanceinanimageoravideo and put bounding boxes around them with their names displayed above it.

II. LITERATURE SURVEY

Deep Learning for Multiple Object Tracking, summarizes and analyzes deep learning-based multi-object tracking methods. Methods fall into three categories: extended description with deep capabilities, embedding deep networks, and building end-to-end deep networks. It then examines the deep network structure in these ways and describes the use and training of these networks for multi-object tracking problems [14].A performance improvement in the detection and tracking of multiple objects in real-time and online applications, and is concerned

with surveillance and video traffic systems. A significant improvement in terms of tracking performance could be achieved with an accuracy of approximately 65% [16-20].

Discovery-based Multi-Object Tracking Method, describes how to track multiple objects. The total number of objects are unidentified and will modify during tracking. The multi-object tracking method maintains a graph structure that upholds theories based on experimental results of object detection for every image that may be absent and/or prone to misleading detection [1-4].The survey, Deep Learning Architecture for Face Recognition in Surveillance Videos, mentions face recognition (FR) systems for video surveillance versus applications that attempt to reliably confirm the location of particular people along a distributed network [9].

For many computer vision applications, multiple object tracking is a necessary step [8]. Detecting and tracking objects in a complicated background, on the other hand, remains a difficult issue. The work provides a method for recognizing and tracking numerous targets that combines an improved Gaussian mixture modelling (GMM) with multiple particle filters (MPFs). GMM during the model updating phase by employing the expectation maximization approach and M recent frames with Gaussian distribution weight parametersget improved [7].In computer vision, multiple object tracking (MOT) is a difficult job. For video surveillance, team-sport analysis, or human–computer interaction, many algorithms have been developed to follow multiple targets. Multiple object monitoring has already been shown to provide useful information in team sports analysis in recent studies. As a result, in this work, we look into object tracking systems for goalball, a paralympic team sport. Various tracking systems have been implemented and evaluated, with the goal of evaluating prediction accuracy and performance speed in players as well as ball tracking [6].

The authors in [7]looked at two lines of research that use deep learning to improve video comprehension, video classification and video captioning. Since video classification focuses on automatically classifying video clips based on their conceptual contents, such as human behaviour or complicated events, clip encoding tries to create a complete and natural sentence, supplementing video classification's single label with the most informative dynamics in videos. Today's digital content includes text, music, photos, clip, and other forms of multimedia. Video, in specific, has emerged as a new method of interaction among Web users with the rise of sensor-rich smart phones. As a result of the massive increase in Internet data transfer and extra storage, video data has been generated, published, and spread explosively, becoming an essential component of today's massive data. As a result, advanced techniques for a wide range of video comprehension applications, such as online advertising, video retrieval, video surveillance, and so on, have been developed.

The aims to provide a comprehensive overview of MOT's evolution over the last few decades, as well as an examination of current MOT breakthroughs and some prospective future research avenues is provided in [8]. Following are the discussions of the article published.

1. Clear explanation of the MOT's main solutions to problems
2. Classification of prior MOT methodologies into 12 perspectives and conversations of the main methods at each
3. A review of test functions and quality assessment evaluating the MOT
4. A debate of numerous MOT issues and challenges through the analysis of related references
5. An overview of the most recent MOT technologies.

An efficient visual detection and tracking framework for object counting and monitoring is developed in [2], that fits consumer electronics criteria such as off-the-shelf equipment, quick installation and configuration, and unattended working situations. This is accomplished through the use of an unique Bayesian tracking model that can handle multimodal distributions without having to explicitly compute the link between tracked objects and detections. It's also resistant to false, distorted, and missing detections.

The work published by Patrick Emami et al. uses on learning algorithms primarily for the multi-object tracking assignment phase is providing an idea to unify the diverse methods by highlighting their ties to linear assignment and the MDAP. The work has begun by discussing probabilistic and end-to-end optimization techniques to data association, then exposed to methods for learning association affinities from data. The performance of the methodologies described in the survey is then compared, and future research options are elaborated.The benchmark's two previous releases offers an in-depth examination of around 50 cutting-edge trackers that were put to the test on over 11000 frames are compiled in [6]. Current trends and flaws in a variety of people monitoring technologies, as well as recommendations for where academics should concentrate their efforts in order to advance the area are provided.

Following are the information with regard to the patents received in the areas closer to the current work.

Objecttracking(11122248)by

ZuoguanWang,JizhangShandescribesamethodofdepth estimationutilizingheterogeneouscameras,comprising, homogenizingafirstcameraimageand a second camera image based on a first camera calibration dataset and a secondcameracalibrationdatasetrespectively,whereint hefirstcameraimageandsecondcamera image are distortion corrected and are zoom compensated, determining an initial image pairrectificationtransformationmatrixofthehomogeniz edfirstcameraimageandthehomogenizedsecondcamera image,determiningadeltaimagepairrectificationtransfo rmation matrix based on the initial image pair rectification transformation matrix,determining a final image pair rectification transformation matrix based on the initial imagepair rectification transformation matrix and the delta image pair rectification transformationmatrix resulting in a final rectified image pair and disparity mapping the final rectified imagepairbasedonadepthnetregression

Multiple objects tracking method, device and storage medium (EP2299406A2)by WoonTack The current paper describes a technique, a device, and a storage medium for tracking multiple objects. Quite specifically, the current discovery relates correlate to a technique and device for monitoring objects that accomplishes object detection of one subcategory per input data by operating only objection detection of one sample for each image taken irrespective of the number N of objects to be tracked and keeps track all objects among image data whilst objects are detected to detect multiple objects in the scene, as well as a storage device. The exemplary embodiment's method for tracking multiple objects involves

(a)at a specified time, conducting object detection with regard to only each sample of multiple objects in an input image.;and

(b)While step (a) is being executed, all objects among image data from a period prior to the predefined period are being monitored with regard to all objects in the input data.

III. PROPOSED SYSTEM

Convolutional Neural Network, Regional convolutional Neural Network, You Only Look once and Deep Sort Methods were implemented for single object and multiple object detection.

A. *Convolutional Neural Network*

Convolutional neural networks (CNNs) are neural networks with one or more convolutional layers that are commonly used for image processing, classification, segmentation, and other co-related data. A convolution is simply a filter that is dragged over the input. Convolutional neural networks are utilized in this proposed study to create a model with numerous layers that can categorize objects into any of the defined classes.

B. Regional convolutional Neural Network

The R-CNN series revolves around the concept of region proposals. To locate items inside an image, region proposals are employed.

C. You Only Look Once

YOLO is a neural network-based real-time object detection technology. This approach is quite effective along with its quickness. In a numerousapplications, it has been used to detect traffic lights, people, parking meters, and animals.

D. Deep sort method

Deep sort is a tracking-by-detection method that generates both the detection outcomes bounding box metrics and data about the presence of the monitored objects to correlate detections in a new frame with earlier tracked objects. It is a tracking algorithm used online.

IV. METHODOLOGY

Based on the literature survey conducted in the preceding chapter, this chapter outlines the stages involved in creating the approach for this project. The flowchart created depicts the step-by-step procedure for obtaining the final goal. The literature survey revealed a variety of methodologies that were studied and inferred.

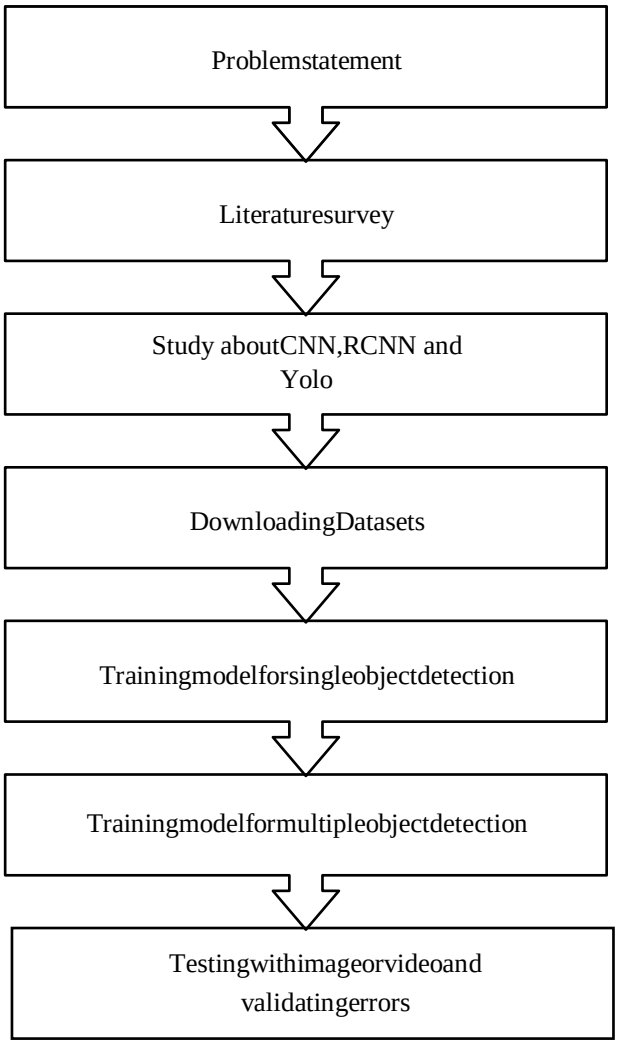


Fig.1.Methodology

A. Problem Statement

Thereareseveralmodelsbasedonneuralnetworkswhered etectingmultipleobjectswithhigh speed and accuracy is not up to what expected. The reasons are camera stability,viewpoint variation, deformation and occlusion.

Camera Stability - If the camera is unstable, then it would be blur and it can't differentiate between objects.

Viewpoint variation - This happens when an object is viewed from different angle the system will find it difficult to detect it. An object can appear radically different depending on the angle from which it is viewed. Using a basic glass as an e.g. (Fig. 2 and Fig. 3) first picture, which shows a spectacular view of a glass filled with black caffeine, appears very unique from the 2nd picture, that reveals a side profile of a cup filled with caffeine.


Fig.2.Side viewFig.3.Top view

Deformation - It happens only when model is trained only with particular images, it is unable to distinguish between distorted things. Many interesting objects are not rigid bodies and can be distorted dramatically. Taking a look at the photographs below of yogi as shown in Fig.4 in a different pose as an example.If the object analyzer has only been trained to identify people seated, walking, or moving, it may be unable to detect people in this picture because the features in such image data do not correlate to those ’ve learnt about people while training.


Fig.4.Yogi

Occlusion - This happens when part of object is hidden as shown in Fig. 5. When only a little fraction of an image, as few frames, is recognizable, the objects of interest can be obscured.


Fig.5.Hidden object

So, the above factors are considered whiletrainingmodeltogetefficientoutput.

B. Literature Survey

Aftergoingthroughmanyjournalsrelatedtothisproject,it helpedustoknowaboutdifferentconcepts andalgorithmsusedinobjectdetectionandtracking. Algorithms selected are CNN, RCNN and Yolo.

Convolutional neural networks (CNNs) are deep neural networks that are used for image analysis, language processing, and other tasks, while an RCNN is utilized for computer vision tasks, namely multiple object detection. YOLO is a real-time object recognition technique that employs neural networks.

C. Objective

After downloading datasets for training model, images are tested with the trainedmodeltolocateobjectsina sceneanddrawrectangularboundingboxesaroundthemf ollowed bydeterminingthenameof each objects discovered with their names displayed.

D. Study on CNN, RCNN & Yolo

After going through journal papers, it is known that how convolutional neuralnetwork and region-based convolutional neural networks are implemented in object detectionand tracking applications. There are categories such as neural networkbased summary advancement utilizing convolutions, deep learning incorporating, and edge deep learning building, as well as reviews of deep network architectures in such techniques, and depth the the use and preparation of these networks for multi-object locating issues, which help us in understanding the potency of neural architectures for monitoring work in different ways, and correlate the benefits of such networks.

E. Downloading datasets

For the training model, roughly 6000 images of six different classes such as cats, dogs, persons, cars, motorbike, bicycleas shown in Fig. 6have been imported in the library.

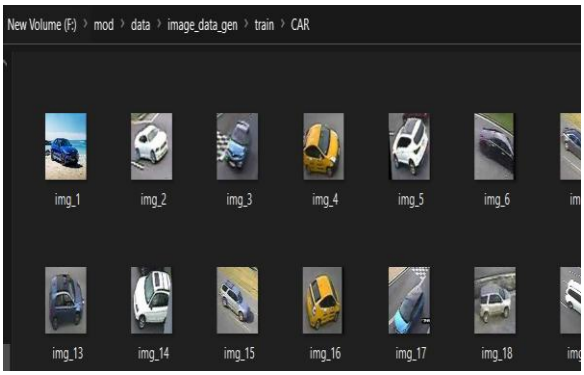


Fig.6.Datasets

F. Training model to detect single objects

This is achieved by using CNN classifier which is trained by downloading datasets containing around 6000 images of 6 different classes. This model has reached expected accuracy. Platform used for the implementation is Jupyter Notebook. The Jupyter Notebook is a free, easily accessible web application that allows data researchers to develop and share information that include live programming, algorithms, computing results, visualisations, and other multimodal features, as well as textual descriptions.

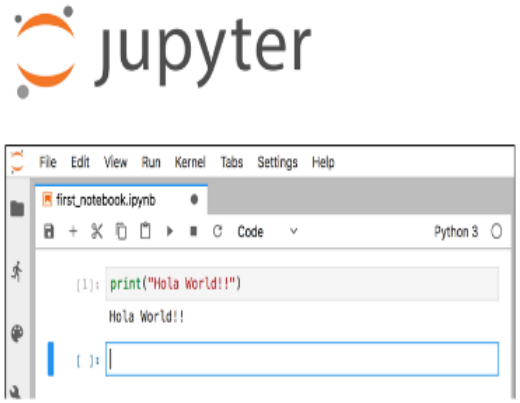


Fig.7.Jupyter Notebook

G. Training Model to detect multiple objects

This is achieved by RCNN, Yolo architecture and deep sort method. Here trained CNN classifier has been used for implementation. Platform used for code execution is Visual Studio Code. Visual Studio Code is a lightweight scripting language with features for error handling, implementation, and versioning. Its purpose is to provide programmers with only the resources they required to finish a quick software cycle.

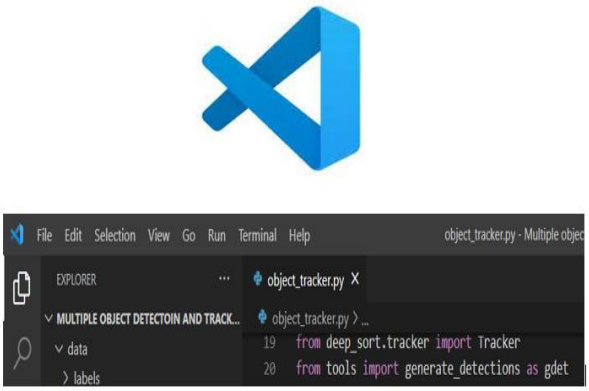


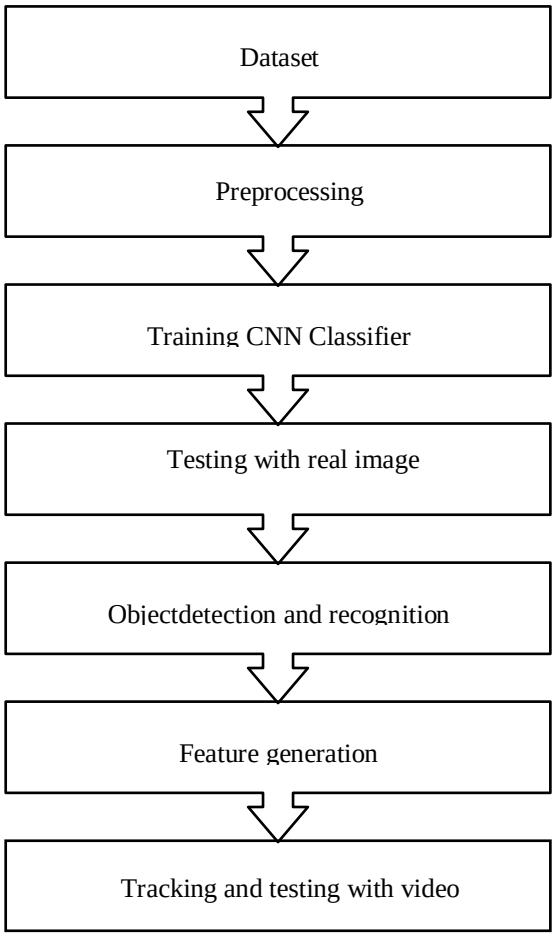
Fig.8.Visual Studio Code

V SOFTWARE ARCHITECTURE

Software architecture refers to the fundamental frameworks of a software platform as well as the restraint of establishing such mechanisms.Each of the structuresbelow contains software elements, their relationships, and the

properties of both elements and relationships. The flowchart in Fig. 9 involves steps, processes and methods carried out in this project that are downloading datasets, preprocessing those images, training CNN classifier, feature generation, tracking followed by testing with images and videos.

Fig.9.Software Architecture



A. Dataset

6000 images of 6 different classes like cats, dogs, pedestrians, cars, bikes and bicycles for training model are utilized. The classes in the library resemble as that of the one as shown in Fig. 10.

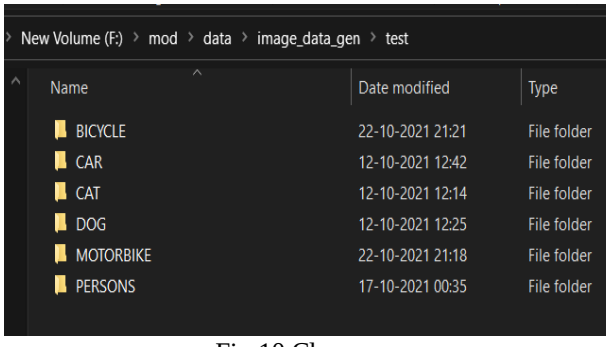


Fig.10.Classes

B.Preprocessing

To prepare picture data for model input, preprocessing is required. Fully linked layers in convolutional neural networks, for example, demanded that all images be of the same size arrays. Process of image can also cut model implementation time and quicken image evaluation. If variance is an informative quantity, these strategies reduce rather than increase object contextual information. Pre-processing is used to optimise picture data and remove undesired deformities or increasing particular spatial features that are relevant for further analysis and testing. The basic goal of machine learning is to accurately train the machine and predict the results. The image must be improved

C. Training CNN classifier

It includes four layers such as convolution, Max pooling, flatten and Dense.

a)Convolution - Convolution is the simple method of assigning a filter to an input in order to generate an induction. Since the same kernel is applied to an input loads of times, a feature map is generated, which displays the locations and attributes of acknowledged attributes in a feedback, i.e. photo. It's being used to make images narrower.

b)Max pooling -Max pooling selects the brightest frames in the picure. It's useful whenever the background of the picture is dull and we're involved in the picture's milder image. It's commonly used after a convolutional layer. It is employed in order to decrease the size of the pictures.

c)Flatten -The process of converting data as a 1D array to be used in the upcoming layer is called as flattening. People straighten the output of the convolutional layers to create a single long feature vector.

d)Dense - Dense Layers are used to detect pictures based on the output of convolutional layers. The dense layer, which is the most popular and commonly utilized layer, is the standard deeply connected neural network layer.

D.Testing with real image

Real images which are captured from a camera have been used for testing and desired output is attained

E. Object detection and recognition

This is the first step of Multiple object detection. Object detection implies it detects whether or not there are any objects in an image, but it doesn't tell you what they are. Object recognition refers to determining the type of detected objects, with the recognition output being a class label.

F. Feature Generation

This is the second step of Multiple object detection. We need to be able to predict motion and generate visual characteristics. Before the data associations, an estimating model is built that takes into account the status of each track, such as bounding box centers box heights, and box width. To model these states and anticipate motion, the Kalman filter is used. A CNN is used to construct feature prediction or bounding box descriptors.

G. Tracking

This is the final step of Multiple object detection. Image tracking is a machine learning implementation that takes a sequence of basic image detection methods and assigns each one an unique identification, then monitors the discovered images as they travel throughout pixels in a clip.Object tracking is utilized in a wide range of applications using various forms of input video. The methods used to create object tracking apps are influenced by whether the anticipated input is an image or a video, and whether it is a real-time video or a preset video. Given the predictive status Kalman filtering, associations are created for the current detections with all object monitors in the prior frame using past information and the newly detected box in the current frames. For the matching update, the cosine feature distance, IoU distance, and Kalman state distance are calculated.

*a)*IOU distance – The intersection over union (IoU) operation is performed when determining an outline. It is a number between 0 and 1 that suggests how the predicted and observed bounding boxes intersect.

*b)*Cosine feature distance – The cosine of the angle between two vectors projected in a multi-dimensional space is measured via cosine similarity.

H. Testing with real video

The laptop webcam is used for real-time testing. The representation of that is shown in testing and validation

VITESTING AND VALIDATION

A. Single object detection

This is achieved by using CNN classifier which is trained by downloading datasets containing around 6000 images of 6 different classes. This model has reached expected accuracy. Below in Fig.11 is the image representing an undesired output i.e., 'Cloth' as 'Persons'. It is because of overfitting which raised while training the CNN classifier. Overfitting occurs when an algorithm learns the data and disturbance in the training examples to the extent where it impairs the performance of the model on updated information. The problem is that these rules don't apply to new information, reducing the model's generalizability. Overfitting is sorted out by retraining the entire CNN classifier and output is predicted as indicated in Fig. 12.and Fig. 13.

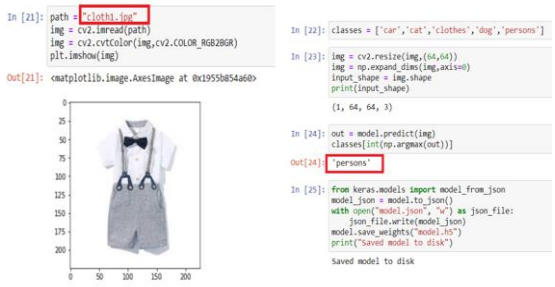


Fig.11.Overfitting

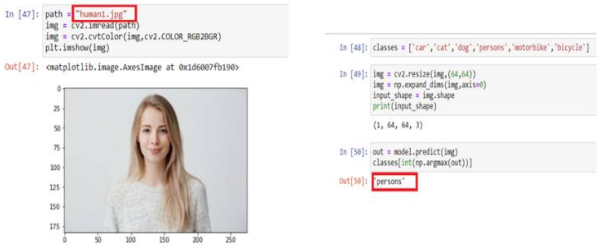


Fig.12.Detecting Person

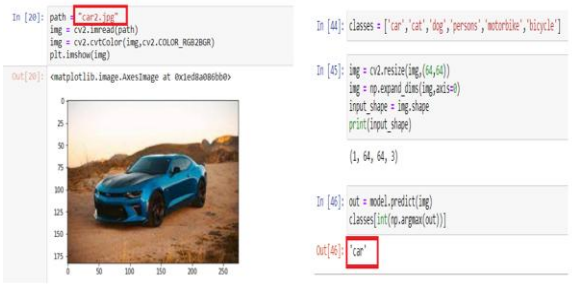


Fig.13.Detecting cars

B.Multiple Object Detection

This is achieved byRCNN, Yolo architecture and deep sort method. Here, CNNclassifier which was trained before has been used for implementation. Platform used for code execution is Visual Studio Code.The snapshot from the uploaded video which depicts the classification of various objects is shown in Fig. 14 and Fig.15 shows how a person is predicted using webcam.

Fig.14.Detecting multiple objects

Fig.15.Real time Detection using webcam

V. CONCLUSION

Deep learning-based object detection has been a major research area in recent years because of its strong learning ability and benefits in having dealt with occlusion and size variation. This paper provides a comprehensive review of deep learning-based object detection architectures that handle various sub-problems, such as occlusion, viewpoint variation and deformation. Several promising methods and algorithms such as CNN, RCNN, Yolo and deep sort method were implied for detecting single and multiple objects. Though limitations such as overfitting,time taken for training were faced,it was overcome by retraining whole CNN classifier. Also, speed and accuracy of our model is as expected up to a percentage of 80. This means that 80 percent of datasets were able to be detected by the trained CNN and for detecting multiple objects, desired outputs were attained. This review is also appropriate for improvements in neural networks and related learning systems, as it offers insightful information and scope for future advances.

REFERENCES

[1]. H. Nuha, N. Abdulghafoor Hadeel, Abdullah, "Enhancement performance of Multiple object detection and tracking for real time and online application", International journal of intelligent Engineering and systems, December 2020

[2]. Saman Bashbaghi, Eric Granger, Robert Sabourni, Mostafa parchami, "Deep learning Architecture for Face Recognition in Video surveillance", December 2019

[3]. Mei Han, Amit sethi, Yihong Gong, "A Detection based Multiple object tracking method", October 2020

[4]. Shengyong Chen, Yingkun Xu, Xiaolong Zhou, Fenfen Li, "Deep learning for Multiple object tracking", IET Computer vision, January 2019

[5]. Yingdong Ma,Yuchen Liu,Shuai Liu,Zhibin Zhang,"Multiple Object Detection and Tracking in Complex Background", International Journal of Pattern Recognition and Artificial Intelligence, 2017

[6]. Julius Gudauskas, Zygimantas Matusevicius,"Multiple objects tracking for video-based sports analysis", 2019

[7]. Zuxuan Wu, Ting Yao, Yanwei Fu, Yu – gang Jiang, "Deep learning for video classification and captioning", 22 February 2018

[8]. Yesul park, L .Minh Dang, Sujin Lee, Dongil Han, Hyeonjoon Moon, "Multiple object tracking in Deep learning approaches", October 2021

[9]. F. Joy ir V . V. Kumar, "A review on multiple object detection and tracking in smart cityvideo analytics," Research gate, 2018, January.

[10]. H. Tao, H.S. Sawhney, and R. Kumar, "A sampling algorithm for tracking multiple objects," in Vision Algorithms 99, 1999

[11]. Alex Bewley, Zongyuan Ge,Lionel Ott, Fabio Ramos,
Ben Upcroft, "SIMPLE ONLINE AND REALTIME TRACKING",July 2017

[12]. Patrick Emami, Panos M. Pardalos, Lily Elefteriadou, Sanjay Ranka, "Machine Learning Methods For Data Association In Multi-Object Tracking", August 2020

[13]. Zhongdao Wang, Liang Zheng, Yixuan Liu, Shengjin Wang, "Towards Real-Time Multi-Object Tracking", September 2019

[14]. Paul Voigtlaender, Michael Krause, Aljosa Osep,Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, Bastian Leibe, "Multi-Object Tracking and Segmentation",April 2019

[15]. Peize Sun,Jinkun Cao,Yi Jiang,Rufeng Zhang,Enze Xie,Zehuan Yuan,Changhu Wang,Ping Luo, "Multiple Object Tracking with Transformer",May 2021

[16]. Laura Leal-Taixe, Anton Milan, Konrad Schindler, Daniel Cremers, "Object Tracking", April 2017

[17]. Carlos Roberto Del-Blanco, Fernando Jaureguizar," An Efficient
[18]. Multiple Object Detection and Tracking Framework for Automatic Counting and Video Surveillance Applications", August 2012
[19]. D. S. Bolme, J. R. Beveridge, B. A. Draper ir M. L. Yui, „Visual object tracking using adaptive correlation filters,“ įtraukta Computer Vision and Pattern Recognition (CVPR) 2010
[20]. SERBY, D., KOLLER-MEIER, S., AND GOOL, L. V,”Probabilistic object tracking using multiple features”. In IEEE International Conference of Pattern Recognition (ICPR)2004
[21]. Jiao L, Zhang F, Liu F, Yang S, Li L, Feng Z, Qu R,”A survey of deep learning-based object detection”,IEEE 2009

Anbarasi MP, et. al. “Multiple Object Detection and Tracking.” *International Journal of Engineering Research and Applications (IJERA)*, vol.12 (04), 2022, pp 11-19.