iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6010

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Real Time Object Detection and Tracking using Deep Learning

Sri Laxmi Kuna[1], Malavika Mancharla[2], Anusha Mahakala[3], Anush Kumar Manda[4], Saikiran Nayini[5]
[1]*Assistant Professor, IT DEPT, Sreenidhi Institute of Science and Technology*
[2, 3, 4, 5]*B.Tech Student, Sreenidhi Institute of Science and Technology*

Abstract: ***Efficient Object Recognition and Tracking are main challenging assignments in computer vision techniques. A very big challenge in many object detection techniques using deep learning may lead to slow and non-accurate performance. This Project Aims to detect and tracking of objects efficiently and accurately in real time .Detecting any object is important in understanding object activities. Here we completely used deep learning networking techniques .The network is trained on most used and challenging dataset COCO. The result is very fast and accurate where object recognition is required.***
Keywords: Object Recognition, Tracking, Detection, COCO

I. INTRODUCTION

Many Object Detection Research topics uses computer vision which helpful in detect ,recognise and track objects in images and over a sequence of images.

Object detection involves locating the object in a image , real time and in a video . Object detection is the process of identifying single object and multiple objects in a image or frame of images in video sequence. The more complicated problem is object recognition involves both image classification and localization. Here the input is webcam or static image or video, the output will be object identified along with a bounding box around it in static image or webcam identification or video and corresponding class of object in each bounding box.

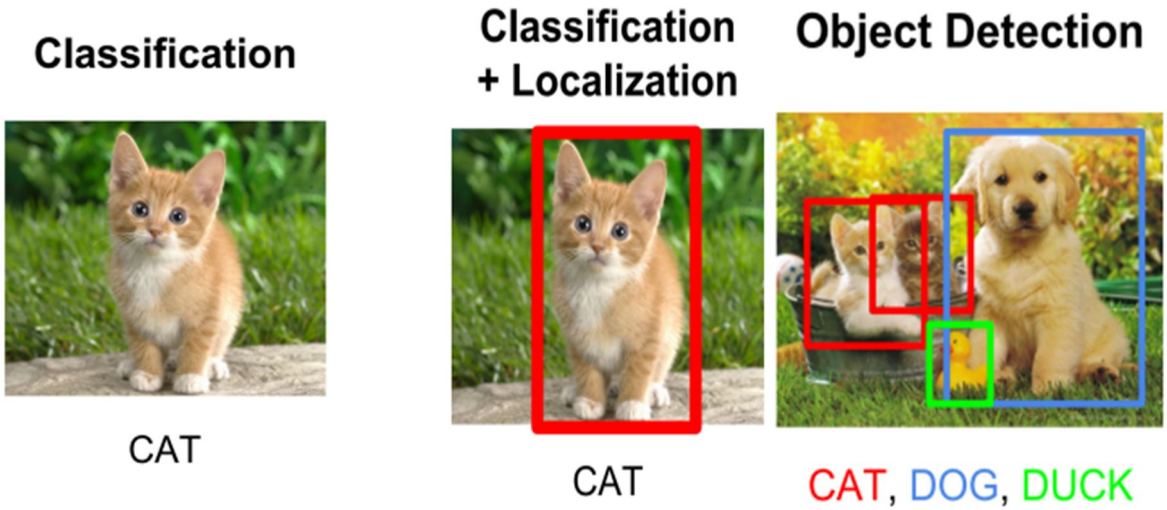


Figure-1: Tasks of Computer Vision

For security purpose many algorithms are explained Detection and tracking of objects based on extracting the features of image and video. Image features are extracted using deep learning and CNN algorithms. Image classification can be done by the use of classifiers .To obtain better results for classification and feature exteaction ,SSD with VGG model along with the concepts of deep learning are used.

A. Dataset

For deep learning ,the dataset importance is crucial for detection of objects . Moreover the dataset for autonomous driving, video surviellances primarly uses COCO dataset. Our SSD model has a large parameters in training .

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

In [1], Qiang Ling et.al, developed a object detection algorithm for feedback. It mainly works based on dual layer updating model to provide updates to the background and segment the foreground with an adaptive threshold method and object tracking is treated as a object matching algorithm. [Background Model Based Detection And Tracking]

A notable feature based algorithm for various object tracking following within the sight of occlusion of object has been proposed. In this method, we consider the features from targeted object from the image and then use a particle filter based approach over there to track the particular featured points in sequence of images from a frame based on different attributes such as place(location), velocity and other. They used rectangular bounding box for representing objects. But this algorithm may not successfully track feature points when it comes to different velocities. Hence this algorithm needs more flexible object representation and also they used static camera for capturing the video.

A unified framework can be considered for both single and cross camera tracking with affinity constraints using graph matching was proposed . In this method, they mainly dealt with the problem of existence occlusion in single camera scenario & the transition in cross

camera scenario and also they consider the data association method in handling occlusion.

Leibe et al. [2] in 2007, presented a novel based on detecting and localizing an object in the form of a visual category compared to the real world scenes.

Their detecting approach consists of an image categorization and figure-ground comparision of two compared processes that closely relate to a common goal and The coupling between the above shown processes allow the detecting process for better actions from each other and improves the overall performance. As they have mentioned, that the resulting method can detect the particular objects and images can persorm segmentation from the obtained results. This received segmentation is turned to again increase detection and recognisation by allowing the system to focus on its efforts. Through this evaluation on the huge consumption of data in data sets shows that the further system was to the huge range of different object or image categories.Further, its flexible feature of representation allowed it to competitive object detecting performance already from training data sets .

Ramya and Rajeswari[3] (2016) proposed an altered technique with consideration of difference in frame which utilizes the correlation method between blocks of the present image in the frame and the background image in the frame to categorize the pixels for both . The blocks in the current picture which are profoundly cor-related with the picture which was in the backgrounf are considered as back-ground. For the another block, the pixel-wise examination is made to

categorize it as anything from to like foreground or background. The experiments directed of the object identification through a frame demonstrated that this methodology may improves the frame distinction method by identifying an object through dividing a frame into blocks particularly as finding accuracy along with the speed. Notwithstanding, this study of frame division should be more concentration towards other data, for example, shape and edge can be utilized to improve the accuracy of detection.

Gang et al. (2010) in [4] for improving accuracy description of objects they developed a kernel locality preserving projections (KLPP). The reproductive results obtained showed that this perspective is appropriate for object recognition which are in space mainly while we consider changes of viewpoints. However, study like this needed concentration on improvement of accuracy level in recognizing which has reference of limited trained objects. Also, effectiveness should be tested by considering other models.

B.M,Nair et al. (2011)[5] developed technique in combining face recognition, tracking and detection to identify individual faces. The reproductive result shows this method can increase accuracy in recognizing and tracking faces which in turn can be applied on real-time applications. However, it fails in different conditions and it may also include lesser background for recognition of face which reduces the performance of system.

Jiang and Zhang (2014)[6] suggested a kernel technique which is based on for identifying multi view objects. The reproductive obtained results are useful in improvement of recognition, while comparing and analysing based on state-of-arts. Afterwards, they authenticated robustness of this approach. However, object space recognition was more focused for resolving this issues .

Foytik et al. (2011)[7] proposed Kalman filter to differentiate between various objects by low-level recognition. Then face recognition at different subspaces formed were analysed in the way of Adaptive Modular Principal Component Analysis. The productive results has more accuracy and can be consider average time for recognizing faces in this approach .

Viola and Jones(2001) [8] proposed in a conference which was held on pattern recognition showed machine learning approaches to detect objects very fastly for processing images and in achieving higher detection rate. Their work was categorised by three concepts. The first was to introduce image representation which was called "integral image". And Second one was an algorithm which was based on Ada Boost, which used for selection of larger set from small number of visual features and in return to get efficient classifiers.

©IJRASET: All Rights are Reserved

Finally Third contribution has been method to combine complex classifiers into "cascade" which allowed background portions of image to be removed quickly. Testing was done on face detection. While using it in real-time applications, per second the detector recognizes 15 frames. Also concentrates on Sliding window approach which detects rigid objects , object can be identified by small region around that object.

In 2008, Koller and Heitz [9] from Stanford university ,presented a project at conference, which improves detection by recognition of two types of objects into system. They regions of each image were separated considering their ability to provide a context for object detection.Detects images based on appearance and relationship group regions, not only providing training set with labels . For individual dataset ,Learning of active set of relationships was presented by them.

[10] The main objective explained in the paper was to detect the target object on roads for driverless systems. The authors proposed an algorithm for object detection using deep neural network concepts. The approach of the algorithm is to provide input images in which the target objects need to be detected. From the input images the rear end or front end or side part information are extracted. The framework of the algorithm is firstly to collect input data. Then machine learning concepts require training data set, so the author has used CIFAR data set for training and loaded to system. Data set had 50,000 images which were in CNN tuning. Thus for verification of algorithm road images were provided and has achieved 100% accuracy for these images. Thus author has proposed a reformative CNN algorithm which also reduced the cost incurred during usage of CNN algorithms. Also by increasing the recognition rate the accuracy of the system was also greatly increased.

Guo et al. in (2012) for detecting and tracking of objects in a video they suggested one way. The simulation result obsolutely indicating that this technique was more efficient and accurate, and also for generic object classes it provides good performance and robustness. In Future it needs to be more focused to increase the classification accurateness for detecting and recognizing real-time objects[14].

III. OBJECT DETECTION AND TRACKING ALGORITHMS

This section explains implementation of the algo-rithms used.

A. Single Shot Detector Algorithm

Many Researches has done in region and most hearing detecting algorithm is YOLO, which was the first research towards real time detection .Main problem with this version of YOLO is low performance. Then Here comes SSD.SSD is also a popular object detection algorithm, Which gives accuracy reasonably in a single pass compared to two stage algorithms.

The main idea is using SSD in default boxes(anchors).Default boxes are the selected bounding boxes on the image based on their positions, aspect ratios and size .

Architecture of SSD contains 3 parts

1) Base network
2) Extra Feature layers
3) Prediction layers

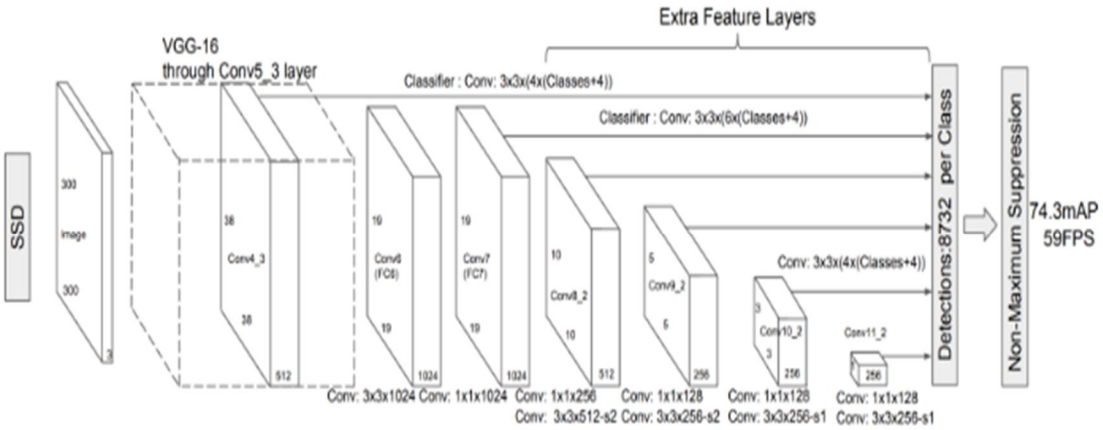


Figure-2: Architecture of SSD

©IJRASET: All Rights are Reserved

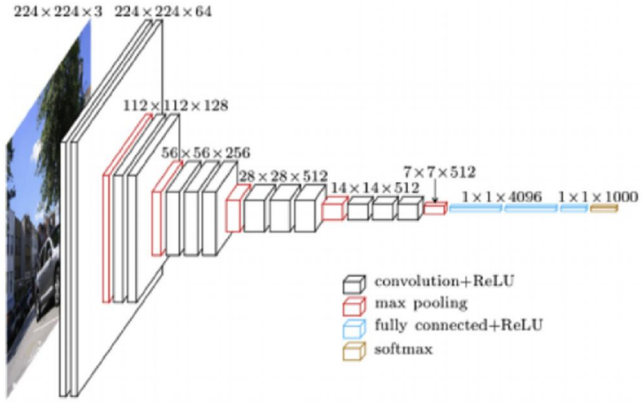


Figure-3: Visualization of the VGG-16 Architecture

The base network is VGG-16 which contains initial layers of any standard image classification, these layers are implemented as convolutional. This is VGG-16 used because the base network, thanks to its robust nature in efficiency of image classification and quality of problems, here transfer learning derives results. Figure-3 visualization of VGG. The two important components of SSD:

a) It mainly used to decrease the size of volume in deep layer, as it would stand with only standard CNN.

b) Every convolutional layer connects to the final layer.

i) *Step1.* From the pre-trained images in a set select one image.

ii) *Step2.* For every image, the small blocks are identified , overlap ratio of image with object is 0.1, 0.3, 0.5, 0.7 and the aspect ratio of each block is set to[1/2,2].

iii) *Step3.* The central portion of the bounding box overlapped with sample block , that portion is retrained.

iv) *Step4.* The output network is a feature map with sizes 19 x 19 x 1024. Four additional layers are added In top of base network which gives feature map with 1 x 1x 256 size. In a convolutional manner default boxes are specifically made to pass by many feature maps to detect objects .If the image bounded boxes are being adjusted to match localization boxes, for that box confidence levels are predicted. Loss function is computed during training which involves confidence loss and localization loss. If a default box not get matched with any truth box is considered as confidence loss, localization is the loss in smooth L1 loss between actual offsets of ground truth to predicted offsets.

B. Mobile nets Algorithm

Mobile nets is the effective convolutional neural network for embedded vision based applications, maximum use of the mobile nets in researches because of its light weight in architecture. Mobilenets only concentrates on speed and low latency. Mobile nets uses depth wise convolutional networks, it applies a single filter to each input rather than more filters at a time. The point wise convolution then applies a 1 x 1 convolution to add the outputs to depth wise convolution. Computation quality
is reduced when these two operations are completed. In a single step standard convolutional filters combines both inputs are converted into a new set of outputs. Depth wise convolutional filter then divide into two layers, one layer for combining and another layer for filtering .

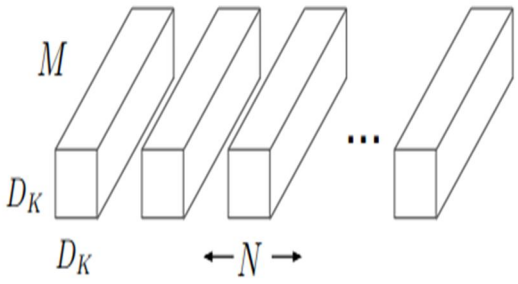


Figure-4 : Normal convolution

©IJRASET: All Rights are Reserved

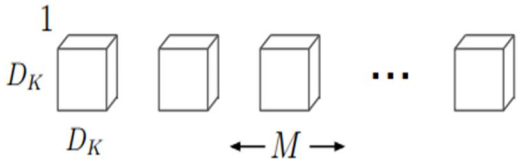


Figure-5 : Depth wise convolutional filters

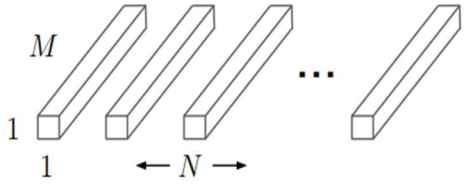


Figure-6 : 1 x 1 convolutional filters are point-wise convolution in the context of depth-wise separable convolutions.

Figure -4: represents the conventional convolutional filter, which is identified for four dimensional. Figure-5: shows depth wise convolutional filter for 3 dimensional and Figure-6: shows purpose wise convolutional filter that may be a vector operation. The combination of 4, 5 and 6 steps take less complexity while computation done with step 4 alone

The complete architecture of Mobile nets is like

1) Convolutional layer
2) Depth wise convolutional layer
3) Point wise layer which can be used to doubles the count of channels
4) Depth wise layer
5) Point wise layer which can be used to doubles the count of channels.

IV. METHODS OF IMPLEMEMTATION

Most efficient language Python is employed to implement these algorithms.

A. Object detection

Region proposal

Region proposal is alternatively called region proposal network which could find some regions for automatically, then we just apply on those regions for object detection.

Frame differencing

Frame differencing is the simple technique to know the which part of the video are moving.

Figure-7: Object which is after applying Frame Difference

In Figure-7 only the moving part of video get highlighted.

©IJRASET: All Rights are Reserved

B. Object Tracking

Object tracking involves three steps are important those are 0bject detection, object recognition and tracking. Some challenges in video-processing are Analyzing the video, Segmentation of video, Compressing the video, video-indexing. For video analysis there are three steps mainly involved: detection of moving object which is intresting, frame to frame tracking of these objects and analysis of tracking those objects to acknowledge their behavior. Next it comes video segmentation it means separation of objects from the back ground. It also consists of three important steps: 0bject detection, object tracking and object- recognition. Here it is given a lot of focus towards the investigation, video analysis and video segmentation.

Figure-8: Tracking of a car

Figure-8: shows the tracking of a car. From the traffic CCTV video, the above image is captured. Suppose anybody wants to track an object which was in motion, he need to take images or frames of video at different intervals .Through these images or frames one can identify the object motion.

V. RESULTS AND ANALYSIS

Here object detection is implemented using openCV, python programming . We have used COCO model which contains 90 object classes. The following results identified after successful object detection of webcam, static images and video sequence.

Figure-9: Detecting bottle with confidence level 95.96%

Figure-10: Detecting train with confidence level 99%

The COCO model was trained to detect different objects such as chair, person, bottle ,bus, car, train, dog, cat, tv moniter, remote, cell phone, apple ,banana, kite, air plane, sofa, bicycle etc. With the confidence level accurately 99%.

©IJRASET: All Rights are Reserved

A. Comparing the Performance of Object Detection

SSD along with MobileNet has the highest mAP among the models implemented for real-time processing. SSD architecture was introduced to improve the memory consumption and speed of the model without sacrificing on accuracy. There are more small objects than large objects in COCO where contains approximately 41% small objects , 34% medium objects , and 24% large objects .According to previous study the objects were detected using SSD with an accuracy of 84%,87%, with our approach objects are detected with an accuracy of 95%,99% etc.

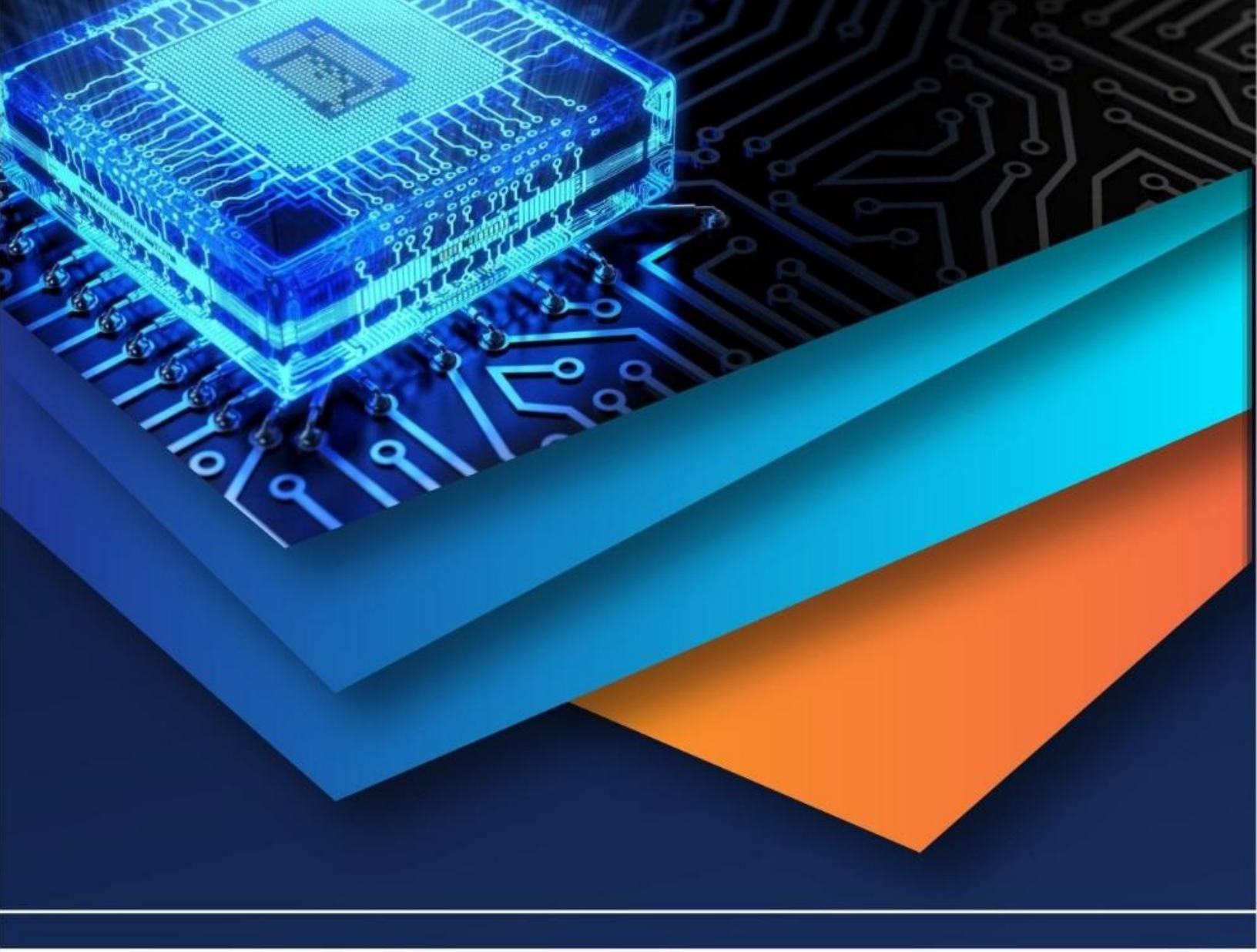
VI. CONCLUSION

Object Detection is a main key for any computer and robot based systems. Although many Achievements are observed in the last years, and some techniques are now part of some face detection tools and automated driving cars and etc. But we are still far away from achieving open world learning. Also we should note that object detection is not used efficiently where it can provide a great help.

REFERENCES

[1] Qiang Ling n, JinfengYan , FengLi and Yicheng Zhang,‖ A background modeling and foreground segmentation approach based on the feedback of moving objects in traffic surveillance systems‖, Journal of Neuro Computing, Elsevier, 2014.
[2] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," Int. J. Comput. Vis., vol. 77, nos. 1-3, pp. 259-289, May 2008.
[3] Ramya, P. & Rajeswari, R., 2016. A Modified Frame Difference Method Using Correlation Coefficient for Background Subtraction. Procedia Comput. Sci. 93, 478–485. doi:10.1016/j.procs.2016.07.236.
[4] Gang, M., Zhiguo, J., Zhengyi, L., Haopeng, Z. & Danpei, Z., 2010. Fullviewpoint 3D space object recognition based on kernel locality preserving projections. Chinese J. Aeronaut. 23, 563–572. doi:10.1016/S1000- 9361(09)60255-7.
[5] Nair, B.M., Foytik, J., Tompkins, R., Diskin, Y., Aspiras, T. & Asari, V., 2011. Multi-pose faces recognition and tracking system. Procedia Comput. Sci. 6, 381–386. doi:10.1016/j.procs.2011.08.070.
[6] Zhang, H. & Jiang, Z., 2014. Multi-view space object recognition and pose estimation based on kernel regression. Chinese J. Aeronaut. 27, 1233–1241. doi:10.1016/j.cja.2014.03.021.
[7] Foytik, J., Sankaran, P. & Asari, V., 2011. Tracking and recognizing multiple faces using Kalman filter and ModularPCA. Procedia Comput. Sci. 6, 256– 261. doi:10.1016/j.procs.2011.08.047
[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1. Dec. 2001, pp. 511-518.
[9] G. Heitz and D. Koller, ``Learning spatial context: Using stuff to find things," in Proc. 10th Eur. Conf. Comput. Vis. (ECCV), 2008, pp. 30-43.
[10] Soin, Akhil, and Manisha Chahande. "Moving vehicle detection using deep neural network." In Emerging Trends in Computing and Communication Technologies (ICETCCT), International Conference, IEEE (2017), pp. 1-5.
[11] Torralba, K. P. Murphy, and W. T. Freeman, ``Using the forest to see the trees: exploiting context for visual object detection and localization," Commun. ACM, vol. 53, no. 3, pp. 107-114, Mar. 2010.
[12] L. Wang, Y. Wu, T. Lu, and K. Chen, ``Multiclass object detection by combining local appearances and context," in Proc. 19th ACM Int. Conf. Multimedia (MM), 2011, pp. 1161-1164.
[13] S. Kumar and M. Hebert, ``A hierarchical eld framework for unified context-based classi cation," in Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV), vol. 2. Oct. 2005, pp. 1284-1291. [11] A Ferri - 2016" Object Tracking in Video with TensorFlow".
[14] https://www.ripublication.com/ijcir17/ijcirv13n5_07.pdf.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)