As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(05):385-391, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0805057
Available online at: http://www.ijmtst.com/vol8issue05.html

YOLOv3 Approach for Object Detection and Tracking

Bandapally Saikiran[1], Kairamkonda Devendar[1], Malothu Balaram[1], Dr. A. Prashanth Rao[2]

[1]Department of Information Technology, Anurag Group of Institutions.
[2]Professor, Department of Information Technology, Anurag Group of Institutions.
Corresponding Author Email ID: 18h61a1269@cvsr.ac.in, 18h61a1284@cvsr.ac.in, 18h61a1268@cvsr.ac.in, prasanthraoit@cvsr.ac.in

To Cite this Article
Bandapally Saikiran, Kairamkonda Devendar, Malothu Balaram and Dr. A. Prashanth Rao. YOLOv3 Approach for Object Detection and Tracking. International Journal for Modern Trends in Science and Technology 2022, 8(05), pp. 385-391. https://doi.org/10.46501/IJMTST0805057

Article Info
Received: 16 April 2022; Accepted: 15 May 2022; Published: 19 May 2022.

ABSTRACT

In the field of object detection, recently, tremendous success is achieved, but still, it is a very challenging task to detect and identify objects accurately with fast speed. Human beings can detect and recognize multiple objects in images or videos with ease regardless of the object's appearance, but for computers it is challenging to identify and distinguish between things. Object detection and object tracking is a pivotal ability required by most computer vision systems. The latest research in this field has been making tremendous development in many areas. Object detection and tracking have a variety of uses, our project presents a general trainable framework for object detection in images and videos including live video. The detection technique we are using is based on YOLO (You Look only once). The ability to identify and classify objects, either in a single scene or in more than one frame, has gained huge importance in a variety of ways. YOLO is a powerful technique as it achieves high precision whilst being able to manage in real time.

1. INTRODUCTION

During the last years, there has been a rapid and thriving expansion of computer vision research. Parts of this success have come from adopting and adapting machine learning methods, while others from the event of the newest representations and models for specific computer vision problems or from the event of efficient solutions. One field that has accomplished exceptional progress is object detection. Object detection is a technology affiliated with computer vision and image processing, this field deals with recognizing and identifying instances of specific objects of a chosen class (cars, humans, laptops, human faces, etc.) in digital images and videos. Object localization refers to identifying characteristics of one or more objects in an image or a video and drawing a bounding box around their extent. Object detection does the work of blending these two tasks and localizes and classifies one or more objects during a picture. When a user or practitioner refers to the term "object recognition", they often mean "object detection" As we move towards more complete image understanding, having a more precise and detailed beholding becomes crucial. During this context, one cares not only about classifying images, but also about precisely estimating the category and site of objects contained within the photographs, a haul mentioned as object detection.

Object detection aims to detect all instances of objects from a known class, like people, cars, or faces during an image or a video. Generally, only a small number of instances of the object are present within the image, but there is a sizable number of possible locations

and scales at which they're going to occur which require to somehow be explored. Each detection of the image is reported with the name of the object that's being detected, this is often as simple due to the position of the object, location, and scale or the extent of the thing defined in terms of a bounding box. In different circumstances, the pose data is more detailed and holds the parameters of a linear or non-linear transformation. As an example, a face detector during face detection may compute the locations of the eyes, nose, and mouth, additionally to the bounding box of the face.

1.2 Overview:

The work presented in this chapter aims to improve the accuracy of object tracking algorithms by employing HDR imaging and a method for scene illumination normalization. The algorithm takes as input an HDR video stream where the illumination can have potentially significant variation in intensity. Additionally, a set of HDR fisheye images is also used that capture the scene illumination (incoming light). By analyzing the scene illumination, the method is able to reverse the illumination effects that appear on the input images and produce an illumination- neutral image stream. Tracking objects using the illumination-neutral images is shown to be much more robust.

As the HDR video stream of the tracked area is captured, a separate low dynamic range (LDR) camera, mounted with a fisheye lens, takes snapshots of the scene's incoming illumination and assembles them in the form of an HDR environment map. This is done at regular time intervals (typically every few seconds). Given an estimate of the scene geometry visible through each pixel, the incoming illumination from the current HDR environment map is used to compute in real-time the scene lighting for each point. Using this, lighting disparities, such as shading and shadows are removed, producing an illumination-neutral video stream that can be used as input by any tracking algorithm.

1.3 Objective:

The main objective of our project is to recognize the object and describe the locations of each detected object in the image using a bounding box with good accuracy. 1.4 Purpose: As there is tremendous growth in the field of object detection and tracking, we are fascinated to include ourselves in this kind of project Recently in US, supermarkets introduced object detection for purchasing of items in which it automatically calculates bill by detecting the object which we've put in the trolley from above cameras and sensors.

This kind of project is useful in vehicle detection where if the vehicle is violating the rules it is easy identify the vehicle by tracking speed.

It is also useful in Security surveillance, Crowd counting, Face detection and recognition, online examinations, etc.

2. PROBLEM STATEMENT

Quick, exact calculations for object detection would permit computer to drive vehicles without particular sensors, empower assistive gadgets to pass on constant scene data to human clients, and open the potential for universally useful, responsive automated frameworks.

[1]. Object discovery includes identifying locale of interest of object from given class of picture

[2]. There are basically two algorithms for object discovery and they can be arranged into two kinds:

1. Classification-dependent algorithms are performed in two steps. First, they define and select areas of significance for an image. Second, these regions are organized into convolutional neural networks. The above-mentioned arrangement is mild, since it is required to make estimates for all chosen regions. A commonly recognized case of this type of algorithm is the Regional Convolutional Neural Network (RCNN) and Medium RCNN, Faster RCNN, and the most recent: Mask RCNN[2].

2. Algorithms based on regression – rather than selecting a field of interest for an image, they estimate groups and bounding boxes for the whole picture in one run of the algorithm. The two most common models in this set are the YOLO family algorithms which provides maximum speed and precision for multiple object detection in a single frame [3] and the SSD this algorithms that are typically used to track objects in real-time.

Convolutional neural network

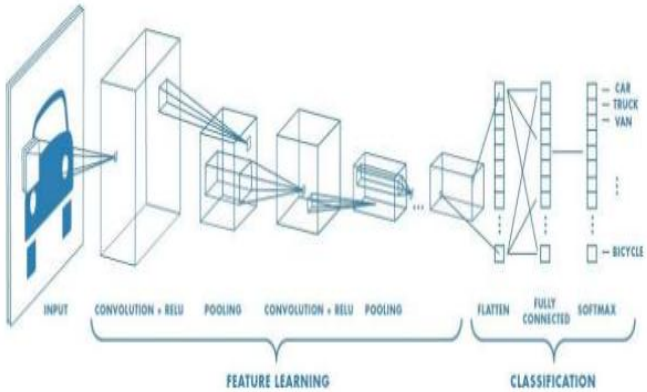
In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of artificial neural network (ANN), most commonly applied to analyze visualimagery. CNNs are also known as Shift Invariant or Space Invariant Artificial Neural Networks (SIANN), based on the shared-weight architecture of the convolution kernels or filters that slide along input features and provide translation- equivariant responses known as feature maps. Counter-intuitively, most convolutional neural

networks are only equivariant, as opposed to invariant, to translation. They have applications in image and video recognition, recommender systems, image classification, image segmentation, medical image analysis, natural language processing, brain–computer interfaces, and financial time series.

CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "full connectivity" of these networks make them prone to overfitting data. Typical ways of regularization, or preventing overfitting, include: penalizing parameters during training (such as weight decay) or trimming connectivity (skipped connections, dropout, etc.) CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.

Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.

CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns to optimize the filters (or kernels) through automated learning, whereas in traditional algorithms these filters are hand- engineered. This independence from prior knowledge and human intervention in feature extraction is a major advantage.



Convolutional Neural Network

Digital image processing:

Digital image processing is the use of a digital computer to process digital images through an algorithm. As a subcategory or field of digital signal processing, digital image processing has many advantages over analog image processing. It allows a much wider range of algorithms to be applied to the input data and can avoid problems such as the buildup of noise and distortion during processing. Since images are defined over two dimensions (perhaps more) digital image processing may be modeled in the form of multidimensional systems. The generation and development of digital image processing are mainly affected by three factors: first, the development of computers; second, the development of mathematics (especially the creation and improvement of discrete mathematics theory); third, the demand for a wide range of applications in environment, agriculture, military, industry and medical science has increased.

Many of the techniques of digital image processing, or digital picture processing as it often was called, were developed in the 1960s, at Bell Laboratories, the Jet Propulsion Laboratory, Massachusetts Institute of Technology, University of Maryland, and a few other research facilities, with application to satellite imagery, wire-photo standards conversion, medical imaging, videophone, character recognition, and photograph enhancement. The purpose of early image processing was to improve the quality of the image. It was aimed for human beings to improve the visual effect of people. In image processing, the input is a low-quality image, and the output is an image with improved quality. Common image processing includes image enhancement, restoration, encoding, and compression. The first successful application was the American Jet Propulsion Laboratory (JPL). They used image processing techniques such as geometric correction, gradation transformation, noise removal, etc. on the thousands of lunar photos sent back by the Space Detector Ranger 7 in 1964, taking into account the position of the sun and the environment of the moon. The impact of the successful mapping of the moon's surface map by the computer has been a huge success. Later, more complex image processing was performed on the nearly 100,000 photos sent back by the spacecraft, so that the topographic map, color and panoramic mosaic of the moon were obtained, which

achieved extraordinary results and laid a solid foundation for human landing on the moon.

The cost of processing was fairly high, however, with the computing equipment of that era. That changed in the 1970s, when digital image processing proliferated as cheaper computers and dedicated hardware became available. This led to images being processed in real-time, for some dedicated problems such as television standards conversion. As general-purpose computers became faster, they started to take over the role of dedicated hardware for all but the most specialized and computer-intensive operations. With the fast computers and signal processors available in the 2000s, digital image processing has become the most common form of image processing, and is generally used because it is not only the most versatile method, but also the cheapest.

Signal processing is a discipline in electrical engineering and in mathematics that deals with analysis and processing of analog and digital signals, and deals with storing, filtering, and other operations on signals. These signals include transmission signals, sound or voice signals, image signals, and other signals etc.

Out of all these signals, the field that deals with the type of signals for which the input is an image and the output is also an image is done in image processing. As it name suggests, it deals with the processing on images.

Digital Image Processing

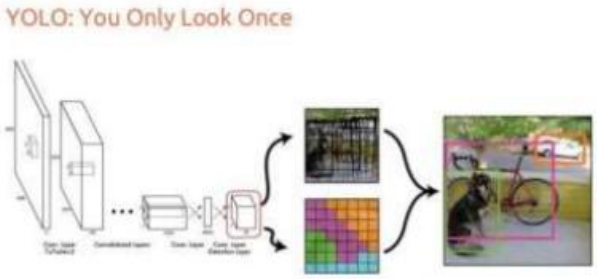
3. PROPOSED SYSTEM

Yolov3:

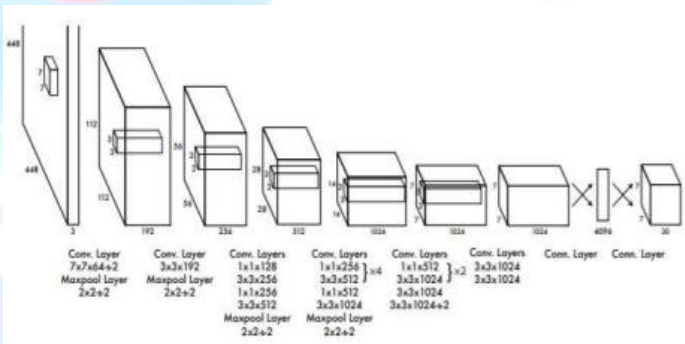
In 2015, Redmon J et al. Proposed the YOLO network, which is characterized by combining the candidate box generation and classification regression into a single step. When predicting, the feature map is divided into 7x7 cells, and each cell is predicted, which greatly reduces the calculation complexity. Accelerate the speed of target detection, frame rate up to 45 fps!.

After a lapse of one year, Redmon J once again proposed YOLOv2. Compared with the previous generation, the mAP on the VOC2007 test set increased from 67.4% to 78.6%. However, because a cell is only responsible for predicting a single object, facing the goal of overlap, the recognition was not good enough.

Finally, in April 2018, the author released the third version of YOLOv3.The mAP-50 on the COCO dataset was increased from 44.0% of YOLOv2 to 57.9%. Compared with RetinaNet with 61.1% mAP, RetinaNet has an input size of 500. In the case of × 500, the detection speed is about 98 ms/frame, while YOLOv3 has a detection speed of 29 ms/frame when the input size is 416 × 416. Further, It is popular because it has a very high accuracy while also being able to run in real-time or used for real-time applications. The YOLO algorithm "only looks once" at the input image that is it needs only one forward propagation pass through the network to make the predictions.



4. YOLO Architecture:



5. sYOLO Architecture

YOLO is a convolution neural network. It consists of a total of 24 convolutional layers and followed by 2 fully connected layers. Each layer has its own importance and the layers are separated by a functionality.

— The First 20 convolutional layers followed by an average pooling layer and a fully connected layer is pre-trained on the ImageNet dataset which is a 1000-class classification dataset.

— The pretraining for classification is performed on the dataset with the image resolution of 224 x 224×3.

— The layers comprise 3×3 convolutional layers and1x1 reduction layers.

— For object detection, in the end, the last 4 convolutional layers followed by 2 fully connected layers are added to train the network.

— Object detection requires more precise detail hence the resolution of the dataset is increased to 448 x 448

— Then the final layer predicts the class probabilities and bounding boxes.

— All the other convolutional layers use leaky ReLU activation whereas the final layer uses a linear activation.

— The input is of 448 x 448 image and the output is the class prediction of the detected object enclosed in the bounding box.

6. RESULT & APPLICATIONS

Result:

By applying Yolo to our images we are the below output for our input images. The objects are detected with an accuracy of greater than 85%.

Output for Object Tracking:

a)

b)

7. APPLICATIONS:

Security Surveillance:

Surveillance is a fundamental element of security and safeguarding. Recent advances in laptop vision technology have junction rectifiers to the event of assorted automatic police work systems, but their effectiveness is adversely stricken by several factors, and that they aren't utterly reliable. Several police work cameras are put in however can't be closely monitored throughout the day.

Since events are additionally doubtless to occur whereas the operator isn't looking, several vital events go undetected, even after they are recorded. Users cannot be expected to trace through hours of video footage, particularly if they are not positive or sure about what they are searching for.

Crowd Counting:

Crowd tally is another valuable application of object detection. For densely inhabited areas like theme parks, malls, city squares, analyzing store performance or crowd statistics throughout festivals. These tend to be harder as folks move out of the frame quickly (also as a result of folk's area unit non-rigid objects). Object detection will facilitate businesses and municipalities a lot of effectively live completely different sorts of traffic—whether on foot, in vehicles, or otherwise.

Crowd Counting

Image Fire Detection:

Real-time vision identification of fireplace has now been enabled for monitoring equipment, guaranteeing a first-class trend in the field of fireplace security.

Image fire detection

Anomaly Detection:

Anomaly detection is applicable in an exceeding form of domains, like intrusion detection, fraud detection, fault

detection, system health monitoring, event detection in sensor networks, detecting ecosystem disturbances, and defect detection in images using machine vision. As a result, manufacturing costs are reduced thanks to the avoidance of manufacturing and marketing defective products. Anomaly detection, in factories, could be a useful gizmo for internal control systems due to its features.

Optical Character Recognition:

The mechanical or electronic translation of printed, hand-written or published text images, into machine-coded textual material, regardless of whether the scanned report is, the photograph of a report, or a real time state, is the optical recognition of character regularly abbreviated by the term OCR.

Optical character recognition

8. CONCLUSION

Object detection is done on videos and images by training detector for a custom dataset consisting of 10000 images for 12 specified classes. The object detection is done using YOLO. Accuracy and precision can be controlled by training the system for more iterations and fine-tuning the training dataset. For Future work, the system can be trained for more classes or more types of objects as it can be used for different domains of videos and different objects can be detected. Our detection system includes Book, Bottle, Car, Computer mouse, Human face, Laptop, Mobile phone, Pen, Person, Picture frame, Weapon, as a class/objects, this can be expanded to more multiple objects or can be dedicated for a specific object with a varying number of datasets.

By using YOLOv3 algorithm the Object tracking and detection process that benefits in tracking and detecting the objects in an image or a video and removes the noise from an image or signal. YOLO is one of the best-known, most powerful object detection models, dubbed "You Only Look Once." YOLO is the first option for every real-time identification of objects. Both input images are divided into the SXS grid structure by YOLO algorithms. For object detection, any grid is responsible. Now these grid cells forecast the observed object boundary boxes. We have five principal attributes for each box, including x and y for coordinates, w and h for object width and height and an insight into the possibility that the box holds the object. In recent years deep learning-based object identification has become a hot spot for analysis due to its powerful study skills and scale transition. This paper suggests a series of YOLO rules to classify objects using a single neural network for the purpose of detection. The rules are easy to create and can be instantly comprehensively photographed. Limit the classifier to a particular area through position concept techniques. In the prediction of limits, YOLO accesses the whole photograph. Moreover, in history regions it expects fewer false positives. This algorithm "only looks once" as in it requires only one forward propagation to cross through the network to make estimations.

9. Future Scope

The object detection technology is practical for future uses such as self-driving cars, vehicle's plate recognition. It will also open up new avenues of research and operations that will reap additional benefits in the future.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Redmon J, Divvala S, Girshick R, & Farhadi, "You Only Look Once: Unified, Real-Time Object Detection." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2016.91

[2] Computing Applications (ICIRCA 2018) IEEE Xplore Compliant Part Number: CFP18N67- ART; ISBN:978-1-5386-2456-2

[3] hethan Kumar B, Punitha R, and Mohana, "YOLOv3 and YOLOv4: Multiple Object Detection for Surveillance Applications" Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020) IEEE Xplore Part Number: CFP20P17- ART; ISBN: 978- 1-7281-5821-1

[4] Hassan, N. I., Tahir, N. M., Zaman, F. H. K., & Hashim, H, "People Detection System Using YOLOv3 Algorithm" 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE). doi:10.1109/iccsce50387.2020.9204925

[5] Pulkit Sharma, "A Practical Guide to Object Detection using the Popular YOLO Framework – Part III" DECEMBER 6, 2018.
[6] Nikhil Yadav, Utkarsh , "Comparative Study of Object Detection Algorithms", IRJET, 2017.
[7] Viraf, "Master the COCO Dataset for Semantic Image Segmentation", May 2020.
[8] Joseph Redmon, Ali Farhadi, "YOLOv3: An Incremental Improvement", University of Washington.
[9] Karlijn Alderliesten, "YOLOv3 — Real-time object detection", May 28 2020.
[10] Arka Prava Jana, Abhiraj Biswas, Mohana, "YOLO based Detection and Classification of Objects in video records" 2018 IEEE International Conference On Recent Trends In Electronics Information Communication Technology,(RTEICT) 2018, India.
[11] Akshay Mangawati, Mohana, Mohammed Leesan, H. V. Ravish Aradhya, "Object Tracking Algorithms for video surveillance applications" International conference on communication and signal processing (ICCSP), India, 2018, pp. 0676-0680.
[12] S. Geethapriya, N. Duraimurugan, and S. P. Chokkalingam, "Real time object detection with yolo," Int. J. Eng. Adv. Technol., vol. 8, no. 3 Special Issue, pp. 578– 581, 2019.