ISSN No:-2456-2165

Live Object Recognition using YOLO

[1]Prathamesh Sonawane
Department of Information Technology, VPPCOE & VA,
University of Mumbai

[2]Rupa Gudur
Department of Information Technology, VPPCOE & VA,
University of Mumbai

[3]Vedant Gaikwad
Department of Information Technology, VPPCOE & VA,
University of Mumbai

[4]Harshad Jadhav
Department of Information Technology, VPPCOE & VA,
University of Mumbai

Abstract:- Live object recognition refers to the real-time process of identifying and categorizing objects within a given visual input, such as images. This technology utilizes computer vision techniques and advanced algorithms to detect objects, determine their dimension, area and weight and often classify them into predefined categories. Our system proposes R-CNN and YOLO to determine the dimensions of the objects in real time. YOLO takes a different approach by treating object detection as a single regression problem. A single neural network is trained to directly predict bounding boxes and class probabilities for multiple objects in an image. The input image is divided into a grid, and each grid cell is responsible for predicting the objects whose center fall within that cell. Live object recognition finds applications in various fields, including autonomous vehicles, surveillance systems, robotics, augmented reality, and more. By providing instantaneous and accurate insights into the surrounding environment, this technology contributes to enhanced decision-making, interaction, and automation across numerous domains. The objects which can be recognized are solid objects which we use daily such as electronic items, stationery items, culinary items and many more. Our system "Live Object Recognition using YOLO" is aimed to detect and determine the objects and dimensions in real time.

Keywords:- *Live Object Recognition, YOLO, Accuracy, Dimension Measure, CNN, Deep Learning, Convolution Neural Network, Object Detection, Machine Learning*

I. INTRODUCTION

The capacity for comprehension and interaction with the visual world has become increasingly important in the current era of rapid technological innovation. "Live Object Dimension Recognition" is one of this domain's outstanding accomplishments. Object recognition is the process of identifying items in pictures and videos. The autonomous vehicles can recognize and classify items in real time thanks to this computer vision approach. An autonomous vehicle is a car that can sense and respond to its surroundings in order to navigate on its own without assistance from a human. Since object detection and recognition enable the vehicle to detect impediments and determine its future trajectory, they are regarded as among the most crucial duties. This breakthrough is extremely significant since it has applications in many other fields, including manufacturing, logistics, augmented reality, and more. Through the use of cutting-edge object recognition algorithms such as YOLO (You Only Look Once) or related technologies, this combination aims to give machines the ability to recognize, classify, and comprehend the spatial properties of things in real time. But this is not an easy task; in order to deal with perspective distortions and other real-world difficulties, sophisticated computer vision algorithms must be incorporated. Furthermore, this endeavor's development of user-friendly interfaces, calibration capabilities, and smooth interaction with current systems are crucial components that, in the end, promise to completely transform how we communicate with and comprehend our physical environment.

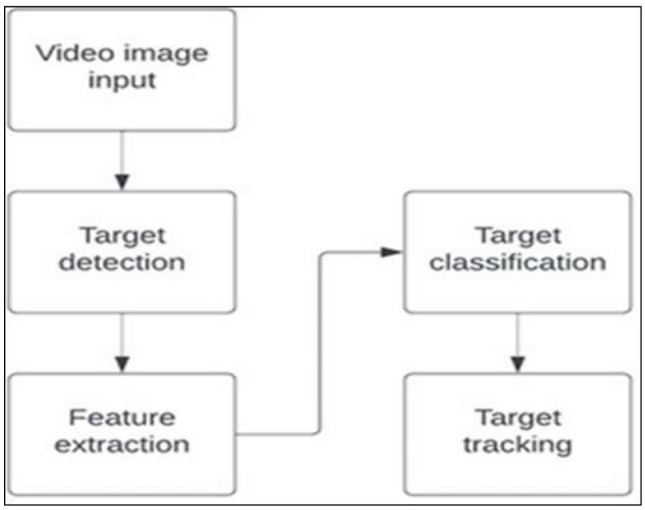

Fig 1 A General Object Detection System [1]

II. LITERATURE REVIEW

- Tracking objects based on morphology is a more sophisticated and intelligent method. Considering their morphology, it tracks the item. Three fundamental procedures are used in this technique: initial background estimation 2) Distinctive geometry 3) Registration of objects. Following a binary threshold calculation, the two subsequent frames are subtracted from one another. From frame to estimated background, this produces the moving items. Then, morphological traits such as width, area, height, and histogram are used to register these items. The backdrop estimation, frame differencing, and

object registration processes are repeated for the upcoming frames. Next, optical flow is computed by comparing the recently registered objects with the previously registered objects using the cost function. Although it mostly depends on the quantity of contours or moving objects in the video sequence, this method's computational efficiency seems to be significantly higher than that of the other approaches. Its usual generation of many contours and exponential growth in operations makes it computationally expensive. This has a significant impact on its computational efficiency. Yet compared to the Census transform and absolute difference approaches, this one is unquestionably more intelligent. The primary benefit of this system is its ability to monitor many objects in a video sequence and its efficient resolution of the object merging issue during object tracking. Complicated and repetitive, with constant processing overhead, are its downsides.

- FPN: To increase scale invariance, feature pyramids built upon image pyramids, or featurized image pyramids, have been widely used in a variety of object identification systems. On the other hand, training time and memory usage rise quickly. In order to achieve this, some methods describe high-level semantics using only one input scale, increasing their resilience to scale changes. However, when testing, image pyramids are constructed, leading to inconsistent inferences across train and test times. A deep ConvNet's in- network feature hierarchy creates feature maps with varying spatial resolutions but also introduces significant semantic gaps due to varying depths. Pioneer works typically create the pyramid starting from intermediate levels or simply sum transformed feature responses, missing the higher-resolution maps of the feature hierarchy, in order to avoid employing low-level features. In contrast to these methods, FPN employs an architecture that combines high-resolution and semantically weak characteristics with low-resolution and semantically strong features through a top-down pathway, a bottom-up pathway, and several lateral connections. The fundamental forward backbone of ConvNet, the bottom-up pathway, creates a feature hierarchy by downsampling the corresponding feature maps with a stride of 2. The layers that possess identical output map sizes are categorized into identical network stages, and the final layer's output from each stage is selected as the reference set of feature maps needed to construct the subsequent top-down pathway. Feature maps of higher network stages are first upsampled and then improved with those of the same spatial scale from the bottom-up pathway via lateral connections, forming the top-down pathway. To minimize channel dimensions, a 1 x 1 convolution layer is added to the upsampled map, and element-wise addition is used to accomplish the merging. The final feature map is constructed and each merged map additionally has a 3×3 convolution attached to it to lessen the aliasing effect of upsampling. The best resolution map is produced by repeating this technique. State-of-the-art representation can be obtained without losing speed or memory since feature pyramid can extract rich semantics from all levels and be trained end-to-end with all scales. While this is going on, FPN is not dependent on the core CNN architectures and may be used for a variety of computer vision tasks, such as instance segmentation, and other phases of object recognition, such as region proposal creation.

- First, we must eliminate the superfluous noise, which is the undesired information that is unrelated to the data that we required for training, in order to recognize the object from the input, which can be an image, video, or any real-time footage. Similar to an image, a video clip requires the same procedure. Simultaneously, the video input and real-time footage input are equally complex. In real-time film input, all detection processes must be completed immediately; in video input, however, the footage is pre-recorded. This is the main distinction between the two types of input. Convolutional neural networks make up the fundamental building blocks of this object detection model. CNN, or neural network, is a widely used acronym. Though it may seem simple, every significant component of this model is handled by CNN, and it is incredibly large and intricate. The procedure seems straightforward, which is first collect input, extract the video, and then use that data to train the model to determine the things, then refine the model to yield more precise forecasts, and finally use this model to forecasts or real-world item recognition. One of the greatest machine learning approaches, CNN offers researchers greater simplicity of use and flexibility in their work. My colleagues and I have developed numerous other object-detecting models around the globe, but all of them—even the most recent ones—are only compatible with large computers and Macs. With this project, we aim to address that problem by developing an object-detecting model that is also compatible with smartphones and other small devices. Because of its large user base, we primarily concentrated on the Android system. It is hoped that the recently developed Object Detecting Artificial Intelligence (ODAI) would also lead to changes in this domain.

- Before processing an image for analysis of categorization, it is necessary to eliminate the image's input noise. It is easy to extract the required item because the same technique is used in video processing, where object detection is dependent on video resolution. The tracking procedure is then simpler to apply in real time from the frame sequence and circumstance. Classification and are both involved in the more complex object detection problem translation. The system will get an image as input in this instance, and a bounding box will be the result along with the type of object in each box, that relate to every thing in the picture. On the CNN principle, the basic framework is created. Finding patterns in photos to identify items, faces, and scenes is a particularly helpful use for CNNs. They derive their knowledge directly from the picture data, classifying images with the aid of patterns rather than manually extracting features. It is reliably detected by the object detection module. CNN algorithm is utilized to track the

ISSN No:-2456-2165

identified item. System utilization proposal Caffe (Convolutional Architecture for Fast Feature Embedding) is an object detection framework. Developed initially at UC Berkeley, it is a deep learning framework. To train, test, fine-tune, and deploy models, Caffe offers a comprehensive toolkit.

- This paper, present a method for identifying objects, measuring their dimensions, and identifying them using an input of a video or image of the surrounding area taken by the computer's webcam or external camera.By using a stand, we may change the camera's height, width, and depth in addition to maintaining it at a predetermined distance. We may detect numerous things at once using this approach, as well as get their dimensional dimensions and other details like the object's region of occupancy. A Python application that uses the queue, math, numpy, and computer vision libraries (opencv-cv2) is used to develop the system. Three modules (a, b, and c) make up this system, and they each perform different system-related functions. Module A: This module sets the camera's width, height, and frame rate in addition to configuring our environment. It is the initial module used to get the input of video frames. Module B: The second module uses the boundaries of the items it detects to identify them. This module does several different tasks. At first, the input frame is converted to grayscale in order to improve comprehension of the details. Next, we use a technique called Gaussian blurring, which replaces a box filter with a Gaussian kernel. The task of the cv.GaussianBlur() method is finished.As a result, the width and height of the kernel should be odd and positive. Furthermore, we need to include the standard deviations in the X and Y, or sigmaX and sigmaY.If just sigmaX is provided, then sigmaY is taken to be equal to sigmaX. Values are selected according to the kernel size if one or both of the inputs are given as zeros. Gaussian blurring works incredibly well to remove Gaussian noise from an image. After that, the picture thresholding process cv.threshold() is used to achieve the thresholding. The first argument is the supplied image, which needs to be in grayscale. The second input is the threshold value, which is utilized to categorize the pixel values. The maximum value that will be applied to pixel values over the threshold is determined by the following input. OpenCV is used to extract the function's final parameter, which allows for various kinds of thresholding. When contours are found, three options are available when employing the cv.findContours() function.source image, contour approximation technique, and contour retrieval mode. Additionally, the outlines and hierarchy are created. Each contour in the picture is listed and labeled as such. For every contour, the border points of an object are organized in (x,y) coordinates. Module C: The third module is utilized to determine the dimensions of the item. It does this by utilizing the contour data to determine the object's various lengths from the math library. The hypot technique, which determines the euclidean distance, enables us to determine the object's length, breadth, and area.

Because computer vision facilitates the processing, analysis, and comprehension of digital videos by computers, it has been used. The suggested work demonstrates computer vision application that demonstrates a notable degree of accuracy for object detection and dimension measurement. Most of the time, the proposed work is utilized to obtain object dimensions with an accuracy of above 95%; these dimensions are then used to calculate the area that the thing occupies.

- We first outline the general architecture of our multi-scaled deformable convolutional object detection network, which is based on YOLO v3, in this section. The deformable convolutional network and the multi-scaled feature fusion via upsampling are next described. Lastly, we present the framework's overall training loss.The YOLO backbone network is used by our picture object detector, together with a new method for convolution operation and feature information fusion. the general structure. The Darknet53 network is the first backbone. It is a hybrid strategy that combines the network utilized in YOLO v2, Darknet-19, and the more recent residual network tactics to achieve feature extraction. With shortcut connections, the larger network consists of successive 3 × 3 and 1 × 1 convolutional layers. Furthermore, we include three deformable convolution layers before the convolutional layers with a size of 52 × 52, 26 × 26 and 13 × 13 to modify the feature extraction. The portion of the detecting network is the second component. Seven by Seven grids are created from the input image by the YOLO detection network. Twenty class probabilities and three bounding boxes with their corresponding confidences are predicted for each grid if the ground truth's center position falls inside it. Additionally, we employ the convolutional set—consisting of 3 × 3 and 1 × 1 convolutional layers—to regulate the output, which comprises the IOU location, three frame positions, and twenty different categories of categorization information. The previously mentioned novel technique pertains to the detection network doing the aforementioned procedures on three distinct feature map scales, namely 13 × 13, 26 × 26, and 52 × 26, in that order. The feature maps at the upper level will be combined and up-sampled with the low-level layer features by the channel.Errors in object recognition resulting from motion or disparate viewing angles might alter an object's shape, size, or stance, making it difficult to identify things in real-world scenarios. In general, there are two approaches that are frequently used to address this query. The first is the data argument, which simulates object deformation by augmenting the diversity of the data and manipulating the object's size, shape, and rotation angle beforehand. Nevertheless, the preparation cost of the data will go up using this strategy, and the data will never cover all actual application scenarios. As a result, the model's capacity for generalization will be somewhat diminished.

ISSN No:-2456-2165

Utilizing a transform invariant feature algorithm, such as SFIT. However, even for known excessively complex transformations, this handcrafted design of invariant features and algorithms may be challenging, if not impossible. In order to improve the modeling capability for the geometric transformation of detected objects, this study suggests applying a deformable convolution network to the one-step object detection network and modifying the fixed geometry of the convolution kernel in the traditional convolutional network in order to address the aforementioned issues. This paper suggests a new multi-scaled deformable convolutional object identification network structure based on the tricks of both the FPN and deformable convolutional networks. Rather than using a standard convolution operation, this network employs a flexible convolution structure to improve the model's learning capacity with regard to object geometric deformation, as well as increasing the accuracy of object detection.In order to extract target object position information, this study also makes use of multi-scaled feature maps, which incorporate low-level characteristics by upsampling. This improves the model's capacity to identify dense and small target objects and substantially mitigates the flaw in missing detections that other object detection models consistently have. The effect on the calculation rate is also sufficiently optimistic, while ensuring accuracy, because the deformable convolution structure and the multi-scale fusion techniques used in this study do not significantly raise processing expenses. Extensive trials demonstrate that the performance (speed-accuracy trade-off) for object detection in images is continuously improved by our multi-scaled deformable convolutional object detection network. When compared to alternative object detection methods, our network's FPS is almost four times higher than R-CNN series. Furthermore, the MAP is roughly 7% and 12% greater than the SSD and YOLO v1 models, respectively. Additionally, compared to the original backbone network without the multi-scaled deformable convolutional operation, the MAP is raised by about 4% under the same backbone. For upcoming object detection challenges, the deformable convolution and multi-scale feature fusion remain novel and viable research approaches. To reduce structural alterations to the feature extraction backbone network even more, we will keep investigating the configuration and application of the deformable convolution structure. By avoiding additional incremental training of the backbone network, we intend to apply the deformable convolution and lessen the overall training job under the pre-training model. Furthermore, we shall investigate the use of multi-scale deformable convolution networks for object detection in videos. Important insights for the real-time detection of distorted objects following motion in the films can be obtained from our method.

III. PROPOSED METHODOLOGY

This technology detects objects, measures their size and area, and frequently classifies them into predefined categories using computer vision techniques and sophisticated algorithms. The approach suggests using an R-CNN and a convolution regional neural network to calculate the object dimensions in real time. By addressing object detection as a single regression problem, YOLOv3 adopts a different strategy. To directly forecast bounding boxes and class probabilities for several objects in an image, a single neural network is trained. A grid is created from the input image, and it is the job of each grid cell to forecast which objects fall inside its borders. "You Only Look Once version 4," or YOLOv4, is a cutting-edge object identification model that marks a substantial improvement over the YOLO series of deep learning models. With a variety of improvements and advances, the YOLOv4 model achieves outstanding results in real-time object detection applications. YOLOv4 has many improvements over its predecessors, including improved model training, more sophisticated data augmentation approaches, and an architecture design that is more efficient. Thanks to these advancements, object recognition in applications like as autonomous vehicles, surveillance, and picture analysis may now be accomplished with greater accuracy and speed.

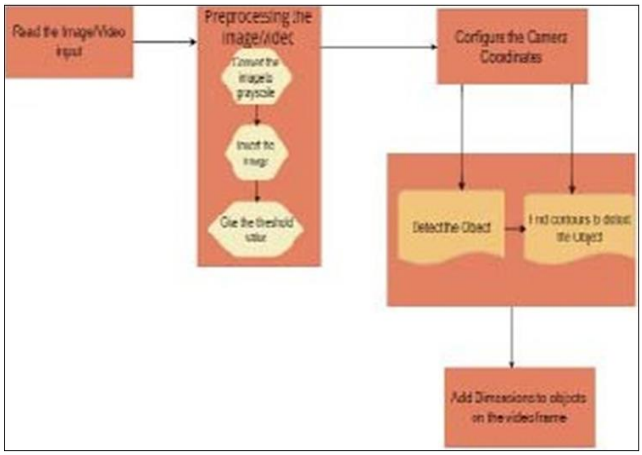

Fig 2 Proposed Architecture

IV. CONCLUSION

In conclusion, a bug localization project that seamlessly integrates deep learning, supervised learning, information retrieval, and crash report correlation could solve modern software development's complex bug identification and resolution problems. This experiment showed the ability to improve problem localization precision and recall, lowering the time and effort needed to find and resolve software defects. Deep learning extracts complex patterns and relationships from code and bug reports, while supervised learning guides localization. Information retrieval methods leverage software repositories' rich history data, making bug localization more context-aware. Crash report correlation gives us a new perspective to prioritize and validate problem reports.

This study shows how data-driven and trans disciplinary bug localization methods improve software maintenance and user experiences. The success of this project highlights the need for continued study and development in this field to improve bug localization for software developers and end-users. These integrated strategies help improve software maintenance agility and reliability as software systems become more sophisticated, ensuring that software issues are quickly recognized and fixed.

ISSN No:-2456-2165

FUTURE WORK

Object recognition will become more accurate and efficient as AI models grow more complex and capable of tackling difficult jobs. Developing object recognition is crucial to the advancement of autonomous vehicles. In order to improve the safety and dependability of autonomous vehicles, object recognition technology will be increasingly integrated in the future. The creation of smart cities will heavily rely on object recognition. It can be applied to trash management, traffic control, security, and other municipal service optimization, all of which will enhance the standard of living in cities.In order to ensure responsible and secure deployment, continued advancements in privacy-preserving methods and ethical considerations will also be crucial in determining the direction of this subject.It is anticipated that object recognition technology will be more deeply integrated into the creation of smart cities in the future. Urban applications for object recognition are numerous and include trash management, traffic control, security monitoring, and service optimization. The potential for this thorough integration of AI-powered item detection might completely transform urban living. In the end, it can improve urban dwellers' overall quality of life by resulting in more efficient city planning, safer, better traffic flow, and lower energy use.

REFERENCES

[1]. Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu, “Object Detection with Deep Learning”, IEEE Transactions on Neural Networks and Learning Systems (Volume: 30, Issue: 11, November 2020), pp. 3212-3232, doi**:** 10.1109/TNNLS.2018.2876865

[2]. Hammad Naeem, Jawad Ahmad_and Muhammad Tayyab, “Real-Time Object Detection and Tracking”, HITEC University Taxila Cantt, Pakistan,Conference: 2020 16th International Multi Topic Conference (INMIC), doi:10.1109/INMIC.2020.6731341

[3]. Anjali Nema, Anshul Khurana "Real Time Object Identification Using Neural Network with Caffe Model." *International Journal of Computer Sciences and Engineering* 7.5 (2021), p.175-182.

[4]. Abhishek Kamble, Abhijit G. Kavathankar, Prathamesh P. Manjrekar, Priyanka Bandagale, “Object Detecting Artificial Intelligence (ODAI)”;International Conference on Smart Data Intelligence, DOI:10.2139/ssrn.3851990

[5]. Madhavi Karanam1, Varun Kumar Kamani1,*, Vikas Kuchana1, Gopal Krishna Reddy Koppula1 , and Gautham Gongada1:Object and it’s dimension detection in real time;E3S Web of Conferences 391, 01016 (2023) ICMED-ICMPC2023;https://doi.org/10.1051/e3sconf/202339101016

[6]. Danyang Cao1,2* , Zhixin Chen1 and Lei Gao1 : An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks Cao et al. Hum. Cent. Comput. Inf. Sci. (2020) 10:14 ;https://doi.org/10.1186/s13673-020-00219-9

[7]. K. Shreyamsh, UAV: Application of object detection and tracking techniques for unmanned aerial vehicles, Texas A & M University, (2015).

[8]. M. Pietikinen, T. Ojala, T. Maenpaa, Multi-resolution gray scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 24, Issue 7, 971 - 987, (2002).

[9]. R. Girshick, Fast R-CNN, In Proceedings of the IEEE international conference on computer vision pp. 1440-1448, (2015).