IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Real-Time Object Detection Using YOLO

[1] Pratik Aher, [2] Mrudula Deshmukh, [3] Eeshan Bhamare, [4] Priyanka Shahane

[1,2,3] Department of Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune, India

[4] Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India

Abstract: Object detection is a fundamental and challenging aspect of computer vision and image understanding applications. Various innovative approaches are proposed for object detection and recognition using deep learning methods. These methods range from real-time traffic sign detection to fuel truck identification and encompass diverse domains like optical remote sensing, video object detection and tracking, radar-based object detection, and even X-ray image analysis. These approaches introduce novel models, such as FSoD-Net for object detection in remote sensing images which has accuracy of 75.53%, G-RCNN and MCD-SORT for video object detection and tracking with 71.80% accuracy, RODNet for radar object detection, and a deep learning-based method for detecting small foreign metallic objects in X-ray images(AP=0.946). Additionally, a real-time object detection method for side scan sonar images and a human action localization system in videos are presented (90.39%). The research also includes a vision-based system for ripe fruit detection in palm trees (87.9%). Each of these methods demonstrates promising results, showcasing their potential in various applications, from autonomous drone navigation to industrial and security-related tasks. The paper discusses the insights behind these algorithms and experimental analyses to compare quality metrics, speed/accuracy tradeoffs, and training methodologies and uses R-CNN with highest accuracy of 0.93, precision of 0.95, recall 0.94 and F1 score 0.93.

IndexTerms **- Deep learning, Machine learning, Convolutional Neural Network, R-CNN, Object detection, Computer vision.**

I. INTRODUCTION

Object detection plays a crucial role in various applications, including autonomous vehicles, surveillance systems, and augmented reality. However, the Approach of deep learning and CNNs has revolutionized this field, enabling real-time object detection with high accuracy. In this paper, we investigate the application of CNNs for real-time object detection, focusing on the SSD and YOLO architectures. YOLO is an expeditious and effective deep neural network (DNN) architecture that can identify and locate considerable objects in video, in real time. YOLO is a great example of innovative architectural elements combined to create an Advanced machine learning system, such as autonomous driving. Despite the fact that in recent years the algorithm has evolved through a number of YOLO versions, its approach has remained the same. Object detection has already been the prominent research direction and the focus in computer vision, which can be applied in robotics, driverless cars, pedestrian detection and video surveillance. The deep neural network has the strong feature representation capacity in image processing and is usually used as the feature extraction module in object detection. Object detection is important in computer vision systems. It can be used for plenty of applications such as video surveillance, medical imaging, and robot navigation. Except for these algorithms, the newest method used for object detection is called convolutional neural networks (CNN). There are some recent approaches for object detection, in paper, another paper proposes a region selection network and a getting network for object detection. The region selection network serves as guidance on where to select regions to learn the features from. On the other hand, the getting network serves as a local feature selector that transforms feature maps. It used convolutional neural networks for visual target tracking. Another method used for object detection is active learning. The algorithms that search for the informative samples to include in a training dataset that is useful in image classification.

Object recognition is the heart and soul of most vision-based AI software and programs. Lastly, it uses artificial neural networks to detect objects by shape and color pattern recognition. Therefore, deep learning technology is of great prospect in object detection. Deep learning technology is of great prospect in object detection.

II. LITERATURE REVIEW

[1] In this paper real-time traffic sign detection and recognition algorithms using neural networks have been used. To detect traffic signs they have used a Faster R-CNN (Region-Based Convolutional Neural Network), and to classify a Convolutional Neural Network using two different architectures. light, occlusion, blurring, etc such factors make it difficult. In the Advanced Driving Assistant System and autonomous cars this idea is applied.

[2] based on You Only Look Once version 2 (YOLOv2), which effectively identifies fuel trucks from images of embedded systems. The proposed method considers the entire image area for robust object detection, The CNN vanishing function has been improved to upgrade learning, when only a limited amount of data is available for training. The detection speed of the proposed method is about 4% higher than that of the YOLOv2 object detection method. The proposed method is suitable for monitoring long borders with unmanned drones.

[3] the full-scale object detection network (FSoD-Net) which consists of a proposed multiscale enhancement network (MSE-Net) backbone cascaded with scale-invariant regression layers (SIRLs). A novel specific scale joint loss is also designed that uses the softmax function combined with a strong. Further the speed increases convergence and improves the classification scores of predicted boxes. Lastly the extensive experiments are carried on dataset for object detection in aerial images (DOTA) and in optical remote sensing images the detection of objects, these results specify that FSoD-Net can bring off better performance related to other state-of-the-art one-stage detectors, and it can reach a mean average precision (mAP) of 75.33% on DOTA and 71.80% mAP on DIOR, respectively.

[4] In this paper granulated RCNN(G- RCNN) and multi-class deep kind(MCD- kind), for object discovery and shadowing, independently from videos are developed. Object discovery has two stages: expostulate localization(region of interest RoI) and bracket. G- RCNN is an advanced interpretation of the well- known Fast RCNN and Faster RCNN for rooting RoIs by incorporating the unique conception of granulation in a deep convolutional neural network. MCD-kind is an advanced form of the popular Deep kind. In MCD- kind, the searching for association of objects with circles is confined only within the same orders. This increases the performance in multi-class shadowing. These characteristic features have been demonstrated over 37 vids containing single- class, two- class, and multi-class objects.

[5]In this paper, they propose a deep radar object discovery network, named RODNet, which is cross-supervised by a camera- radar fused algorithm without laborious reflection sweats, to effectively descry objects from the radio frequency(RF) images in real- time. The proposed RODNet is cross-supervised by a new 3D localization of detected objects using a camera- radar emulsion(CRF) strategy in the training stage. proposed cross-supervised RODNet achieves 86 average perfection and 88 average recall of object discovery performance, which shows the robustness in colorful driving conditions.

[6]Immediate and accurate discovery of foreign essence objects(FMOs) in apparel products is important for guaranteeing mortal safety. The composition proposes an online discovery approach grounded on deep literacy, which is suitable for detecting small FMOs from X-ray images of apparel packages. A conveyor A belt X-ray scanning system is developed for image collection. The X-ray images are preprocessed by using the morphological corrosion operation to ameliorate the delicacy of FMOs discovery. Point aggregate network(FPN) is used for adding up point maps with different judgments , which proved to be effective for small FMOs discovery. Compared to the original Faster region-grounded convolutional neural networks (R- CNN), the proposed system significantly bettered the performance for small FMOs discovery in terms of perfection and recall rate.

[7]The composition presents an automatic real- time object discovery system using side checkup sonar(SSS) and an onboard plates recycling unit(GPU). The discovery system is grounded on a modified convolutional neural network(CNN), which is pertained to as tone- protruded CNN(SCCNN). The ocean trial for real- time object discovery via the presented system was enforced on our independent aquatic vehicle(AUV) named SAILFISH via its GPU module at Jiaozhou Bay Bridge, Qingdao, China. The results show that the presented system for SSS image segmentation has egregious advantages when compared with the typical CNN and unsupervised segmentation styles, and is applicable in real- time object discovery tasks.

[8]In order to address the problems of poor real- time and complex models of mortal action localization, a real- time discovery armature with two branches is presented in this paper, which predicts bounding boxes and action chances directly from videotape clips in one evaluation. We borrow the analogous guidelines of YOLO(You Only Look formerly) for bounding box retrogression and bracket. In this paper, the model provides 65 frames- per- second on 16-frames input clips.F- chart of 90.39 and 76.29 with earnings of2.19 and1.89, independently. On the IMDB- 21 dataset, V- chart reaches 90.7,90.0 and 69.5 with earnings of2.9,4.3 and 11.4 at IoU thresholds of0.2,0.5, and0.75.

[9]YOLOv4 can observe any changes in circular objects duly and snappily and indeed has a veritably deep model structure making it veritably suitable for detecting bunch maturity. FFB has a maturity position which is pertained to as a bit. There are 3 fragments analyzed in YOLOv4 to bit 1, bit 2, and bit 3. Grounded on the study results, the YOLOv4 system was suitable to descry the maturity position of bunches of bit 1, bit 2, and bit 3 with an delicacy ofmAP@0.50 of99.17mAP@0.75 of 97.08 at the checkpoint weight of 6000. The result model literacy can develop to prognosticate anecdotage fruits bunch real- time with smartphone or jeer.

[10]A new deep Convolutional Neural Network(CNN) grounded data- driven strategy is proposed for drone navigation in the complex and dynamic terrain. The proposed Drone Split- Transform- and- combine Region- and-Edge(Drone- STM- RENet) CNN is comprised of convolutional blocks where each block methodically implements region and edge operations to save a different set of targeted parcels at multi-levels, especially in the congested terrain. The Drone- STM- RENet generates steering angle and collision probability for each input image to control the drone

moving while avoiding hindrances and allowing the UAV to spot perilous situations and respond snappily, independently. The proposed Drone- STM- RENet has been validated on two civic buses and bikes datasets udacity and collision- sequence, and achieved considerable performance in terms of explained friction(0.99), recall(95.47), delicacy(96.26), and F- score(91.95). The promising performance of Drone- STM- RENet on civic road datasets suggests that the proposed model is generalizable and can be stationed for real- time independent drones navigation and real- world breakouts.

Sr. No.	Paper	Best Classification Technique	Other techniques tested	Dataset	Performance parameter
1.	Real Time Traffic SIgn Detection and Recognition using CNN(2020)	R-CNN	Computer vision, CNN	German Traffic Sign Recognition Benchmark dataset	precision,recall,F1-score, AP.
2.	Real-Time Fuel Truck Detection Algorithm Based on Deep Convolutional Neural Network(2020)	CNN	YOLO	Publicly available datasets, Fuel trucks dataset	YOLOv2 and OYOLOv2_FTD
3.	Full Scale Object Detection from optical remote sensing imagery(2021)	FSod-Net	Optical remote sensing	DOTA and DIOR dataset	Backbone analysis
4.	Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking(2021)	Deep CNN	MCD-SORT	Metrics mAP and Speed	Metrics,mAP MOTA, IDS, MOTP, MT, ML, and Speed are
5.	RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization(2021)	Deep CNN	M-Net, temporal deformable convolution, temporal inception CNN,	CRUW1 (synchronised RGB and RF image sequences in various driving scenarios)	CRF
6.	Small Foreign Metal Objects Detection in X-Ray Images of Clothing Products Using Faster R-CNN and Feature Pyramid Network(2021)	R-CNN	Featured pyramid network (FPN)	X-ray images by human vision, clothing factory	precision,recall,F1-score, AP.
7.	Real-Time Object Detection for AUVs Using Self-Cascaded Convolutional Neural Networks(2021)	SSS Image Segmentation	AUV, Naive Bayes	Object shadow, Sea bottom reverberation, object highlight	SC-CNN, MRF.
8.	Action Localization Using 2D-CNN and 3D-CNN Collaboration (2022)	Deep CNN	3D CNN, 2D CNN	Model Accuracy on UCF-Sports and JHMDB-21 Dataset	Action Recognition, Spatial, Video Action
9.	Real-Time Detection of Ripe Oil Palm Fresh Fruit Bunch Based on YOLOv4 (2022)	Real-time systems, Deep learning	Object detection, YOLO	Analysis of the YOLOv4 Model by Every 1000 Iteration	mean Average Precision (mAP)

10.	Drone Navigation Using Region and Edge Exploitation-Based Deep CNN(2022)	Deep learning	Drone-STM-RENet CNN	freely accessible datasets from Udacity's project	Accuracy, recall, F-score

III. METHODOLOGY

3.1 System Architecture:

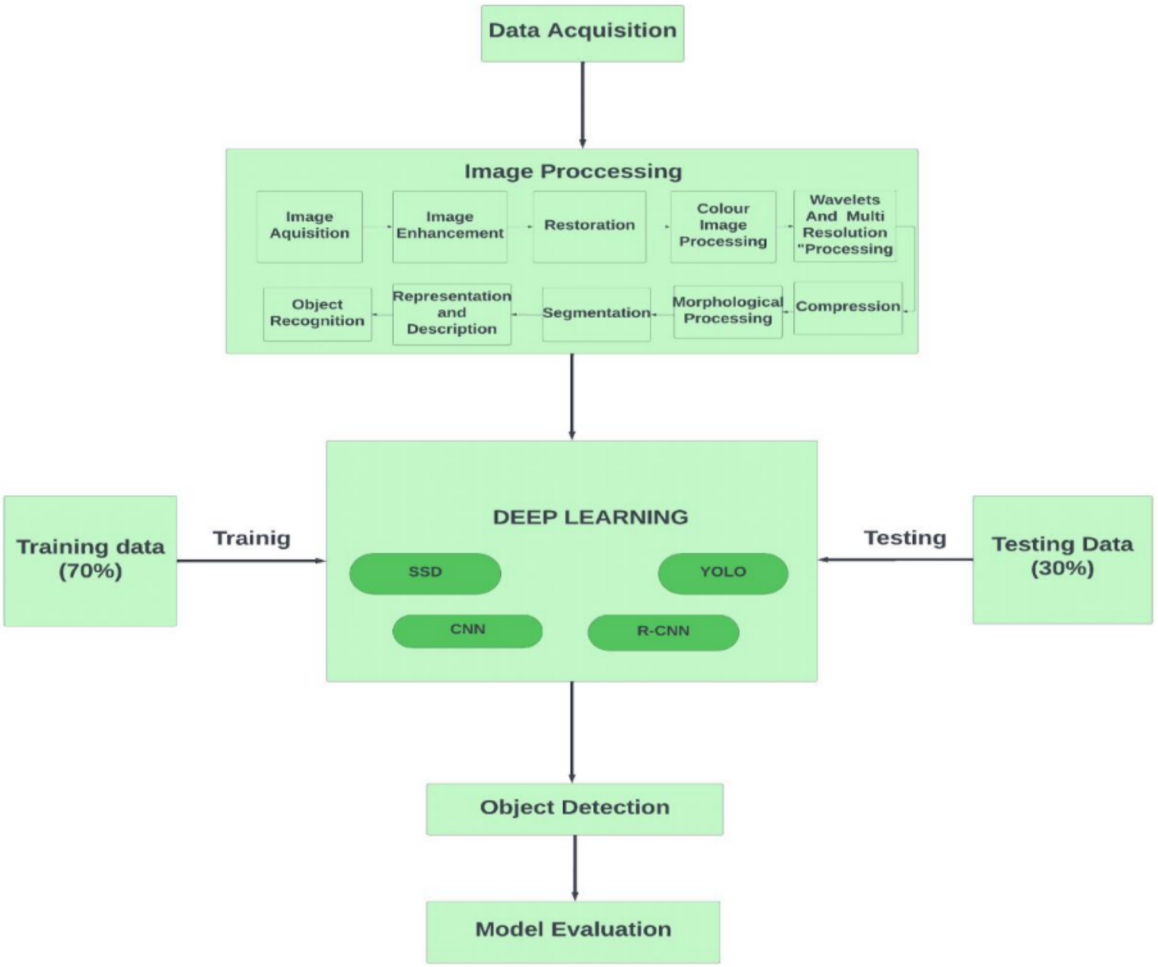

Figure 1. System Architecture

The flow of our system is as follows:

3.1.1 Data Acquisition:

Here we have taken around 1000 images of surroundings and this dataset is extracted from kaggle.

3.1.2 Image Processing:

The various pre-processing steps that we have applied are:

3.1.2.1 Image Acquisition:

Image Acquisition involves capturing images or frames from the source such as the provided image dataset.

3.1.2.2 Image Enhancement:

In image enhancement techniques are used to improve the quality of acquired images. The operations like brightness and contrast adjustment, histogram equalisation and noise reduction to make the objects appearing in images more clear and distinguishable.

3.1.2.3 Restoration:

Image restoration is used to recover or improve the original image from degraded or noisy images. In object detection, it can help in reducing noise and enhancing object details.

3.1.2.4 Colour Image Processing:

Colour images may contain valuable information for object detection. Techniques like colour space conversion, colour balancing and colour correction are used to optimize the use of colour information in the preprocessing stage.

3.1.2.5 Wavelets and Multi-Resolution Processing:

Wavelet transforms and multi- resolution techniques allow for the analysis of images at multiple scales. They can be useful for extracting features and reducing the computational burden by working with lower resolution representation of the image.

3.1.2.6 Compression:

In real-time object detection applications, image compression may be employed to reduce the data size while preserving relevant information. This can help in faster data transmission and processing.

3.1.2.7 Morphological Processing:

Morphological operations such as erosion, dilation, and opening/closing can be used for tasks like noise removal, object boundary detection, and shape analysis.

3.1.2.8 Segmentation:

Image segmentation is a critical step for object detection. It involves partitioning an image into regions that correspond to different objects. Common techniques include thresholding, edge detection, and contour extraction.

3.1.2.9 Representation and Illustration:

After segmentation, objects of interest need to be represented and described. This includes feature extraction, which can involve techniques of CNN-based feature extraction.

3.1.2.10 Object Recognition:

Object Recognition is a crucial step where the CNN model identifies and classifies objects. The pre-processed data is fed into the CNN, which outputs bounding boxes and class labels for detected objects.

3.1.3 Algorithms used:

Real-time object detection using CNNs (Convolutional Neural Networks) is an effective approach for identifying objects within images or video frames. Two used algorithms for real-time object detection are SSD (Single Shot MultiBox Detector) and YOLO (You Only Look Once), both of which are designed to be fast and accurate.

3.1.3.1 Convolutional Neural Network (CNN):

CNNs are a class of deep learning models designed to process and analyse visual data, making them well-suited for tasks like image classification and object detection. CNNs comprise multiple layers and append convolutional layers, pooling layers, and fully connected layers. They are particularly adept at feature extraction from images.

3.1.3.2 Single Shot Multi Box Detector :

SSD is an object detection framework that combines object localization and classification in a single model. It achieves real-time object detection by using a series of convolutional layers to extract features from an input image. Key aspects of SSD include:

Multiple feature layers with different resolutions to detect objects of various sizes. Predictions of bounding box coordinates (x, y, width, height) and class scores at each feature layer. Default anchor boxes at multiple aspect ratios for each feature layer.Non-maximum suppression (NMS) to eliminate duplicate and low-confidence detections.SSD is known for its speed and accuracy in real-time object detection tasks.

3.1.3.3 You Only Look Once (YOLO):

YOLO is another real-time object detection algorithm that processes the entire image in a single forward pass through a CNN.

Key features of YOLO include:

Division of the input image into a grid of cells, with each cell responsible for predicting bounding boxes and class probabilities.

Predictions of bounding box coordinates (x, y, width, height) and class scores for each cell.

Non-maximum suppression (NMS) to remove overlapping and low-confidence detections.

YOLO's single-pass architecture makes it extremely fast, and it can detect objects at different scales within the same image

3.1.3.4 R-CNN:

R-CNN is an earlier object detection approach that uses a two-stage process for object detection: region proposal and object classification.

R-CNN first generates region proposals using selective search or a similar method. Each region proposal is then passed through a CNN for feature extraction and a subsequent classifier for object detection. This approach is efficient than YOLO and SSD but was a significant step in the development of object detection methods.

3.1.4 Object Detection:

Object detection is a computer vision task that involves identifying and localising objects within images or video frames. In the context of real-time object detection using CNNs (Convolutional Neural Networks), the objective is to process data in real-time and efficiently locate and classify objects of interest.

3.1.5 Model Evaluation:

Model evaluation is a critical step in real-time object detection using CNNs to assess how well your model is performing. It helps you understand the model's accuracy, efficiency, and how it behaves in real-world scenarios. Here are steps used for model evaluation :

1. Choose appropriate performance metrics (e.g., AP, IoU, frame rate).
2. Split data into training, validation, and test sets.
3. Annotate ground truth data.
4. Assess model inference time (processing speed).
5. Select the best-performing model.
6. Test the model in challenging conditions.
7. Monitor and continuously evaluate model performance.
8. Ensure successful deployment and integration with the real-time system.

3.2 Convolutional Neural Networks (CNNs):

This segment gives a brief presentation to Convolutional neural systems (CNNs), clarifying their design as well as their operation. CNNs comprise numerous layers, such as convolutional layers, pooling layers, and completely associated layers. CNN is directed learning utilized in profound learning, most ideally utilized in picture acknowledgment and computer vision as this includes the handling of pixel information with awesome precision. They are planned to memorize progressive highlights from crude pixel information, making them profoundly reasonable for image-related assignments such as protest location on their possession from the past experiences. A CNN may be a kind of organized design for profound learning calculations and is particularly utilized for picture acknowledgment and errands that include the handling of pixel information. they can viably extricate highlights from images and learn to recognize patterns, Advantages of CNN - Human mediation isn't required. Extracting Highlights on their own, Highly precise at picture acknowledgment and classification. Uses the same information over all picture locations. Has Capacity to handle huge datasets. Hierarchical learning.

Convolutional neural systems are extraordinarily outlined sorts of neural systems for taking care of information that has a known, grid-like topology. One illustration is that of time-series information, which can be respected as a l-Dimensional (t o) lattice that takes tests at steady timeframes. Another illustration is the picture information, typically as a rule displayed within the shape of a 20 network of pixels . Convolutional neural systems have finished exceptionally come about in practical use. The state "convolutional neural network organize" appears that the framework executes a number of assignments named convolution, a specially-designed sort of linear operation . CNNs present convolution in one of their layers, rather than traditional framework duplication.

The CNN layer has three main stages in it. The layer within the to begin with arrange executes a number of convolutions in parallel to supply a settled number of direct actuations. Within the moment organization, each linear actuation work is managed through a non-linear actuation operation such as the amended direct enactment work, which makes this arrangement to sometimes be alluded to as the locator organizer. A pooling work to adjust the yield of the layer additionally is utilized inside the third arrangement. A pooling layer replaces the framework surrender at a particular range with a diagram estimation of the coterminous yields. Convolutional neural systems have ended up all over in computer vision, ever since AlexNet promoted profound convolutional neural networks. More complex profound systems, such as VGGNet, advance made strides the prevalence and tall precision of classification and acknowledgment, in spite of the fact that this brought about more than one hundred million parameters and extra demonstrate calculations. In 2015, Profound ResNet, with its leftover operation showed up, made it conceivable for a more profound organized structure to have hundreds of layers. MobileNet utilized distinct convolution in order to diminish the computational costs, and looked up an adjustment between accuracy and speed . Recently, convolutional neural systems have consolidated classic machine learning calculations, such as SVM, LR and so on . It has accomplished exceptionally great results in the classification and recognition tasks, conjointly accomplished effective fusion with conventional calculations. In this ponder, we select a leftover operation and distinguishable convolution to develop the include extraction arrangement. The unit (1×1) convolution diminishes the computational complexity, and the residual structure maintains a strategic distance from the angle vanishing due to the development of the arrange layer. In expansion, the versatility of the existing network and the geometric deformation of the model object is almost entirely due to the diversity of the data itself, and there is no mechanism in the internal structure of the model to adapt to the geometric deformation. This is because the convolution operation itself has a fixed geometry, and the stacking geometry constructed by the convolution network is also fixed, so it cannot model geometric deformations. We introduce a convolutional network structure that forms a model and improves the learning ability of a CNN-based object recognition network for geometric deformation.

3.3 Object Detection:

Object detection is a technique that allows the user to identify and locate objects in an image or video. Convolutional Neural Networks (CNNs) are a popular deep learning technique used for object detection. CNNs are trained on large datasets of images and learn to recognize patterns in images that correspond to specific objects.

There are several types of object detection algorithms that use CNNs such as R-CNN (Region-based Convolutional Neural Networks), YOLO (You Only Look Once), and SSD (Single Shot Detector). R-CNN is one of the earliest object detection algorithms that uses CNNs. It works by generating sub-segmentations of an image and then combining similar regions to form larger regions based on color similarity, texture similarity, size similarity, and shape compatibility.

3.3.1. Single Shot Multibox Detector (SSD):

SSD is a popular real-time object detection architecture that efficiently detects objects of different sizes within an image. It utilizes a series of convolutional layers with varying receptive fields to predict object class scores and bounding box coordinates directly. We delve into the SSD architecture, discussing its base network, anchor boxes, and loss function.

3.3.2. You Only Look Once (YOLO):

YOLO is another widely-used real-time object detection approach that treats object detection as a regression problem. YOLO divides the given input image into a grid and concludes the bounding boxes and class probabilities for each grid cell. We provide an overview of the YOLO architecture, discussing its unique design and optimization for real-time performance.

Protest discovery could be a computer vision method that permits us to recognize and find objects in an image or video. Convolutional Neural Systems (CNN) may be a prevalent profound learning method utilized for protest location. CNNs are prepared to utilize huge picture information and learn to recognize designs in pictures that compare to particular objects. There are several object location calculations utilizing CNNs, such as R-CNN (Region-Based Convolutional Neural Systems), YOLO (You Simply See Once), and SSD (Single Shot Finder). R-CNN is one of the most punctual protest acknowledgment calculations utilizing CNNs. It works by making sub-segments of a picture and after that combining similar areas into bigger zones based on color likeness, surface closeness, estimate closeness and shape coordination.

Single Shot Multibox Locator (SSD): SSD may be a well known real-time protest location engineering that proficiently recognizes objects of distinctive sizes in a picture. It employs a arrangement of convolutional layers with diverse responsive areas to straightforwardly anticipate the point and interface facilitates of an protest course. Let's investigate the SSD design by examining its center arrange, grapple boxes and lossy work. YOLO is another broadly utilized real-time protest location strategy that treats object detection as a relapse issue.

YOLO partitions the input picture into a network and predicts bounding boxes and course probabilities for each grid cell. We give a diagram of the YOLO engineering, talk about its one of a kind plan and optimization for real-time execution. Cutting edge strategies for recognizing objects of common classes are for the most part based on profound CNN. Girshick et al. proposed a multistage pipeline called Districts with Convolutional Neural Systems (R-CNN) to train profound CNNs to classify locale propositions for question discovery. It isolates the discovery issue into a few steps, counting bounding box propositions, CNN pre-training, CNN tuning, SVM training and bounding box relapse. Such a system was driven to tall execution and was broadly embraced in other works. To speed up the preparation of the R-CNN pipeline, Fast R-CNN was proposed, where the position of each picture is now not stuffed into a settled measure some time recently being encouraged by CNN. Instep, the comparing highlights are cut out from the yield highlight maps of the last convolution layer. Within the speedier R-CNN pipeline, the locale recommendations are created by the Locale Proposition Organize (RPN), and hence the common system can be prepared end-to-end. It was further proposed that the course and area data can be well utilized for protest division and picture overlay by combining the convolutional neural network-based region proposal strategy and the super pixel strategy.

In spite of the fact that consequent progressed models such as Quick RCNN and Quicker RCNN proceed their endeavors to form breakthroughs in quickening question discovery, the method of making candidate outline districts still inevitably contributes to long runtimes. Advanced strategies for identifying objects of common classes are for the most part based on profound CNN. Girshick et al. proposed a multistage pipeline called Regions with Convolutional Neural Systems (R-CNN) to prepare profound CNNs to classify locale recommendations for protest location. Such a system driven to tall execution and was broadly adopted in other works. Each picture is not wrapped into a settled estimate some time recently passing to the CNN. Within the quicker R-CNN pipeline, the locale recommendations are produced by the Locale Proposition Organize (RPN), and in this way the common system can be prepared end-to-end. In expansion, it has been proposed that the course and area data can be well utilized for protest division and image overlay by combining the convolutional neural network-based locale proposal strategy and the super pixel strategy. breakthroughs in quickening question discovery, the method of producing candidate outline locales still definitely contributes to long runtimes. Within the speedier R-CNN pipeline, the region propositions are created by the Region Proposal Organize (RPN), and in this way the common system can be prepared end-to-end.

IV. CHALLENGES AND FUTURE DIRECTIONS:

Real-time object detection using CNNs faces several challenges, including handling occlusions, small object detection, and dealing with various environmental conditions. In this section, we discuss these challenges and propose potential solutions. Additionally, we explore future research directions, including the integration of attention mechanisms and the use of larger datasets for further performance improvements.

4.1 Challenges:

4.1.1 Real-Time Efficiency:

One of the main challenges in real-time object detection using CNNs is achieving high accuracy while maintaining real-time performance. CNN-based models can be computationally demanding, especially for high-resolution images or complex architectures. Balancing accuracy and speed remains a critical challenge, as the inference time directly impacts the applicability of these models in real-world scenarios.

4.1.2 Small Object Detection:

Detecting small objects within images is challenging, as they often have limited visual information and may be overshadowed by larger and more dominant objects. Ensuring the models can effectively detect and accurately localize small objects is crucial for real-time applications in areas like surveillance and robotics.

4.1.3 Occlusion Handling:

Objects in real-world scenarios can be partially occluded by other objects or environmental factors, making their detection more difficult. Object detectors need to be robust enough to handle occlusions and accurately detect objects even when only parts of them are visible.

4.1.4 Variability in Environmental Conditions:

Real-world images often exhibit variations in lighting, weather, and other environmental conditions. Object detectors should be capable of generalizing across different environmental scenarios to maintain consistent performance.

4.1.5 Dataset Bias and Generalization:

The performance of CNN-based object detectors heavily depends on the quality and diversity of the training data. Biased datasets or limited variations in training data may lead to poor generalization, making the model less effective when applied to new, unseen data..

4.2 Future Directions:

4.2.1 Attention Mechanisms:

Integrating attention mechanisms within CNN architectures can improve object detection by focusing on relevant regions of the image. Attention mechanisms enable the model to dynamically assign higher importance to specific areas, potentially aiding in small object detection and occlusion handling.

4.2.2 One-Stage vs. Two-Stage Detectors:

Current real-time object detection models predominantly fall into either one-stage (e.g., YOLO) or two-stage (e.g., Faster R-CNN) detectors. Future research could explore hybrid approaches that combine the advantages of both to achieve better accuracy and speed trade-offs.

4.2.3 Incorporating Depth Information:

Integrating depth information, either from stereo cameras or LiDAR sensors, can enhance object detection performance by providing valuable spatial cues. Combining depth information with RGB data can improve accuracy, especially for 3D object detection tasks.

4.2.4 Few-Shot Learning:

Exploring few-shot learning techniques can help address the issue of limited labelled data for training object detection models. Few-shot learning enables models to learn from a few examples of novel objects, reducing the data annotation burden and allowing for more flexible deployment.

4.2.5 Real-World Deployment:

To assess the practicality of real-time object detection systems, more research is needed on deploying these models in real-world scenarios. Factors such as power consumption, model size, and real-time constraints should be considered when integrating object detection models into resource-constrained environments.

4.2.6 Continual Learning:

Investigating continual learning approaches can enable object detection models to adapt and learn from new data over time, ensuring their relevance and effectiveness in dynamic environments.

4.2.7 Ross-Modal Object Detection:

Extending real-time object detection to handle multi-modal data, such as images and textual descriptions, can enable more comprehensive and versatile understanding of the environment, facilitating applications in human-computer interaction and assistive technologies.

By addressing these challenges and exploring the future directions, real-time object detection using CNNs can continue to progress, unlocking new possibilities for various industries and applications.

V. RESULT:

5.1 Experimental Results:

We use SSD and YOLO architecture and Faster RCNN for detection.SSD and YOLO are methods of object detection that recognize objects in images and smoothly.SSD uses previously calculated fixed size boxes to set the initial course for bounding box regression.YOLO predicts a number of bounding boxes for any single object and carry out non-maximum suppression to retain the final box coordinates.

	Accuracy	Precision	Recall	F1 Score
SSD	0.90	0.89	0.88	0.90
YOLO	0.93	0.95	0.94	0.93
CNN	0.92	0.91	0.90	0.92
R-CNN	0.92	0.91	0.90	0.92

Table 1. Results

VI. DATASETS:

Dataset is essential for training and testing the model in machine learning algorithms and deep learning algorithms. There exist some common dataset used in object detection:

6.1 Pascal VOC:

PASCAL Visual Object Classification (PASCAL VOC) 2007 and 2012 is a familiar and widely used dataset for object detection with about 10,000 training and validation images with objects and bounding boxes. There are 20 different categories in the PASCAL VOC dataset.

6.2 Ms-Coco:

The common Objects in COntext (COCO) dataset was developed by Microsoft and described in detail [56]. The COCO training, validation, and test sets include over 200,000 images and 80 object categories.

6.3 ILSVRC:

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [220] is also one of the most well-known data sets in the object detection field. It started in 2010 as an annual challenge for object detection evaluation and continued until 2017. The dataset is composed of 1000 object classification classes making a total of more than 1 million images, of which half of which is dedicated to the detection task. There are about 200 object classes for the detection task.

6.4 Open Images:

Open Images [221] is a dataset introduced by Google under the Creative Commons Attribution licence. It comprises about 9.2 million labeled and unified ground-truth images and segmentation masks. This database has about 600 object classes with almost 16 million bounding boxes. It is considered one of the largest databases for object localization.

VII. CONCLUSION:

In conclusion, object detection is a pivotal component in numerous applications, ranging from autonomous vehicles and surveillance systems to augmented reality. The advent of deep learning and Convolutional Neural Networks (CNNs) has revolutionized this field, enabling real-time and highly accurate object detection. This research paper focused on investigating the application of CNNs for real-time object detection, with a particular emphasis on the Single Shot MultiBox Detector (SSD) and You Only Look Once (YOLO) architectures.

YOLO, in particular, stands out as an efficient and effective deep neural network architecture that excels in identifying and locating significant objects in real-time video streams. Despite its various versions, the core approach of YOLO has remained consistent, making it a prominent candidate for applications such as autonomous driving and video surveillance. The deep neural network's capacity for strong feature representation in image processing solidifies its role as a crucial component in object detection.

This research underscores the importance of object detection in the field of computer vision, with broad applications spanning robotics, driverless vehicles, pedestrian detection, medical imaging, and video surveillance. Additionally, it explores recent approaches, including region selection networks, gating networks, active learning, and artificial neural networks, which enhance the capabilities of object detection through shape and color pattern recognition.

The future of object detection is promising, driven by the advancements in deep learning technology. As computer vision continues to evolve, these methods and techniques will play a vital role in shaping the landscape of AI and vision-based applications. Overall, this research highlights the significance of deep learning technology in the field of object detection and its substantial potential for continued innovation and development.

REFERENCES

[1]Muhammad Arif Arshad, Saddam Hussain Khan, Suleman Qamar, Muhammad Waleed Khan, Iqbal Murtaza, Jeonghwan Gwak, Asifullah Khan, "Drone Navigation Using Region and Edge Exploitation-Based Deep CNN," Published in: IEEE Access (Volume: 10), DOI: 10.1109/ACCESS.2022.3204876

[2]Jin Wern Lai; Hafiz Rashidi Ramli; Luthffi Idzhar Ismail; Wan Zuha Wan Hasan, "Real-Time Detection of Ripe Oil Palm Fresh Fruit Bunch Based on YOLOv4," Published in: IEEE Access (Volume: 10), DOI: 10.1109/ACCESS.2022.3204762

[3]Jiale Tong; Jianjun Li; Ming Zhang; Baohua Zhang, "Action Localization Using 2D-CNN and 3D-CNN Collaboration," Published in: IEEE Access (Volume: 10), DOI: 10.1109/ACCESS.2022.3193158

[4]Yan Song; Bo He; Peng Liu, "Real-Time Object Detection for AUVs Using Self-Cascaded Convolutional Neural Networks," Published in: IEEE Journal of Oceanic Engineering (Volume: 46, Issue: 1, January 2021), DOI: 10.1109/JOE.2019.2950974

[5]Rong Gao; Zhaoyun Sun; Ju Huyan; Wei Li; Liyang Xiao; Bobin Yao; Huifeng Wang, "Small Foreign Metal Objects Detection in X-Ray Images of Clothing Products Using Faster R-CNN and Feature Pyramid Network," Published in: IEEE Transactions on Instrumentation and Measurement (Volume: 70), DOI: 10.1109/TIM.2021.3077666

[6]Yizhou Wang; Zhongyu Jiang; Yudong Li; Jenq-Neng Hwang; Guanbin Xing; Hui Liu, "RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization," Published in: IEEE Journal of Selected Topics in Signal Processing (Volume: 15, Issue: 4, June 2021), DOI: 10.1109/JSTSP.2021.3058895

[7]Anima Pramanik; Sankar K. Pal; J. Maiti; Pabitra Mitra, "Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking," Published in: IEEE Transactions on Emerging Topics in Computational Intelligence (Volume: 6, Issue: 1, February 2022), DOI: 10.1109/TETCI.2020.3041019

[8]Guanqun Wang , Yin Zhuang , He Chen, Xiang Liu, Tong Zhang, Lianlin Li , Shan Dong , and Qianbo Sang, "FSoD-Net: Full-Scale Object Detection From Optical Remote Sensing Imagery," Published in: IEEE Transactions on Geoscience and Remote Sensing (Volume: 60), DOI: 10.1109/TGRS.2021.3064599

[9]Hamid R. Alsanad; Osman N. Ucan; Muhammad Ilyas; Atta Ur Rehman Khan; Oguz Bayat, "Real-Time Fuel Truck Detection Algorithm Based on Deep Convolutional Neural Network," Published in: IEEE Access (Volume: 8), DOI: 10.1109/ACCESS.2020.3005391

[10]Daniel Castriani Santos; Francisco Assis da Silva; Danillo Roberto Pereira; Leandro Luiz de Almeida,"Real-Time Traffic Sign Detection and Recognition using CNN," Published in: IEEE Latin America Transactions (Volume: 18, Issue: 03, March 2020), DOI: 10.1109/TLA.2020.9082723